Medical Dianosis Assistant Using Fine Tuning Model

Bharath N[1], Santhru P[2], Gunasekar S[3]
[1,2,3] *Department of Computer Science and Design Engineering, Rajalakshmi Engineering College, Thandalam,602105. Chennai, Tamilnadu, India.*

***Abstract*—In the fast growing clinical world , the rapid and accurate interpretation of Chest X-Rays is critical for diagnosing respiratory and cardiovascular pathologies; However the integration of Artificial Intelligence (AI) in the medical industry and also with the medical imaging often leads to limiting the clinical explainability, accuracy, diagnostic delays and cognitive fatigue among radiologist. This research help to overcome the limitations in the clinical world by providing explainable AI, GradCam and with the help of multi model like DenseNet(169), EfficientNet-B5 and VitBase and also our HybridModel these platform help to achieve high precision diagnosis. The core problem solution relies on integration on Explainable AI(XAI) through GradCam attention mapping with RAG that provides evidence based medical reasoning which gives us accurate results. The technical implementation is done using Fast API backend with 44 Specialised modules and a react based frontend dashboard managing 33 endpoints for seamless clinical workflow integration. The hybrid model which we created will give you committee of experts in regards of the results instead of single architecture diagnostic tool which will help in determining the results with more accuracy.**

***Index Terms*—Artificial Intelligence, GradCam, Explainable AI(XAI), Retrieval Augmented Generation (RAG), Vision Transformer (Vit), DenseNet(169), EfficientNet-B5.**

I. INTRODUCTION

In the field of Clinical radiology is currently facing a dual challenge , a very high surge in the volume of diagnostic imaging and a critical shortage of specialised radiologist. In today world Chest X-Rays (CXR) remains one of the most frequently performed diagnostic imaging examination due to the effective cost and utility in detecting a wide range of conditions, including pneumonia, tuberculosis, lung nodules. However the manual lab interpretation of these scans is labour intensive and having a high chance of huma error while working in under pressure and high workload conditions. While deep learning demonstrated the remarkable success in medical industry in medical image classification, the use of this from laboratory performance to bedside utility remains very low. The less use of image classification is because of “black-box” nature of neural network where they lack of transparency in the model decision making progress down and make trust issues and regulatory approval.

This research helps and present a transformative approach to medical imaging through the development Chest X-Ray Analysis system. Unlike the existing models which uses narrow thinking and narrow focus models, our system provides end-to-end framework that prioritise Explainability, Reasonability, and Clinical Integration. The backend Architecture is designed to support the diverse state of multi models system including DenseNet-169, EfficientNet-B5, and Vision Transformer-ViT these models uses multiple diagnosis and provide the best results and explanation.

One of the problems that exist is that the trust issue so we are trying to solve the problem by using Explainable AI-XAI. By using GradCam++ and Attention Rollout, the system generates special heat maps to showcase the places where the disease have affected influencing the diagnosis. This is integrated with Retrieval Augmented Generation-RAG that retrieves relevant medical literature to provide evidence based explanation for every prediction. To ensure practical usability of this research, the system is developed using FastAPI stack and React, featuring 33 interactive endpoints that facilitate batch processing, model comparison, and the generate clinical report. This research seeks to empower healthcare professionals with a transparent, verifiable, and highly accurate decision support tool that

enhances diagnosis with precision to significantly reduce the time-to-treatment.

The important thing in the project is that the hybrid model which we created using DenseNet-169, EfficientNet-B5, ViT-Base These CNN models architecture are fused together to form a model that can import efficiency and accuracy. The true power of the hybrid model lies in the how it combines the these different perspective. The system does not pick one winner it combines the weighted soft analysis.

A unique feature of the hybrid models is that the ability to know when it doesn't know. By analysing the disagreement in the different models, the system decides it as uncertainty metrics.
If all six models agree that agrees that pathology is present the system provides high confidence diagnosis.
If three models says normal and other three says pneumonia the standard deviation of the prediction increases. The system then flags the results as inconclusive.
The hybrid models uses spatial mapping analysis while the models runs interference the system extracts the gradient from the final layers of the CNN to generate a heatmap. For the transformer the system uses Attention Rollout to show the patches the model focussed on. Then the doctors ensemble the predictions alongside the heatmaps that highlights the physical location of the defected pathology.

II. SYSTEM DESIGN

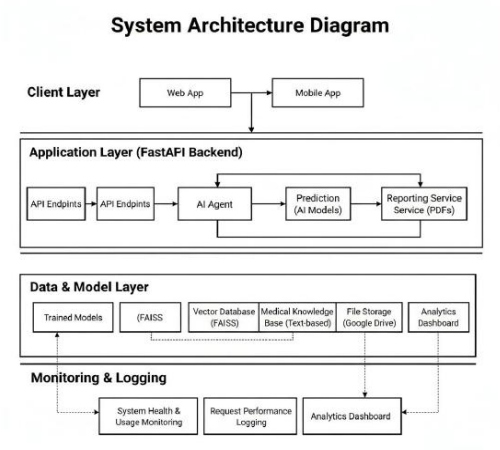


Fig 2.1: System architecture diagram

1.Hybrid multi-model Ensemble Architecture-In our research instead of relying on a single neural network, our system utilises a vast number of network to analyse the given image. By combining multiple architectural philosophies, the system performs in higher level of clinical accuracy.

1. Convolution platforms like DenseNet-169 & Efficient Net-B5 these models excels at identifying patterns
2. Vision Transformer-Vit unlike CNNs, Vit looks the image in a more specified way that it understand the relationship between multiple clinical regions, such as comparing the left and right side of the lung.
3. To reduce "False Negatives" the system aggregates prediction from all the six models must confirm a 'Normal' diagnosis before it is finalised.

2.Explainable AI(XAI)-One of the important steps in the medical diagnosis for the patient is that is to explain why was the diagnosis was made. This makes our research to be completely transparent.

1. GradCam++ visualization helps this model by producing heatmaps, when the system detects any diseases like "Pneumonia" It specifies the area of places where the lungs got affected. This allows the radiologist to verify the AI is looking at the correct pathology or just an image artifact.
2. In this research for the Transformer Models we are specifically using "Attention Rollout" this feature visualises the "attention weights " this showcases how this research prioritised different segments of the X-Ray.

3.Evidence based Reasoning-This project bridges the gap between image processing and clinical insights using Retrieval Augmented Generation.

1. The backend used high speed vector of medical literature, "FAISS Vector Database".
2. Once a pathology is detected, the RAG system retrieve the relevant snippets from the medical journals. The system retrieve the documented criteria for the relevant pathology providing a scientific foundation for the AI's claims.

4.Clinical Workflow and System Integration-

1. The backend is not just a script it is a full medical pipeline. It handles everything from enhancing X-

Ray contrast to Uncertainty Quantification, it ensures the AI whenever it is "Unsure" needs human intervention.
2. React-Vite Frontend with 33 endpoints this dashboard is engineered for high density data.
3. Our research allows to batch process an entire X-Ray ward of up to 50+ images in a single session.
4. Our research also allows to compare the same scans side by side using multiple models.
5. After scanning the image it provides heatmaps and RAG-based reasoning into a detailed clinical document.

5.Performance and reliability Metrics-The system is optimised 30-45 minutes execution window for heavy clinical tasks, ensuring it can keep up with a busy radiology department. Dedicated endpoints will constantly monitor the online status of the model backbones and the database connection, ensuring 24/7 reliability.

III. DATA PREPROCESSING AND AUGMENTATION SCHEMA

The foundation for any data learning model is the data pipeline. Our hybrid model utilises three distinct backbones, each pre-trained on imageNet but with different constraints, the schema is designed to accommodate the specific requirement.

1.Resolution Reconciliation and Input Vectors- A critical challenge in the hybrid design is that disparity in the optimal image resolution. DenseNet 169 and Vit-Base are typically optimised for 224 X 224 pixels. EfficientNet B5 because of its compounding scaling it is significantly higher resolution 456 X 456 pixels to capture fine-grained details.
One of the problem is that while working with the 256 X 256 pixels in Efficient Net B5 it leads to the severe information loss and violate the scaling laws, while forcing a 456 X 456 image through would result in quadruple the sequence length, leads to quadratic explosion in computational costs ($O(N^2)$) attention complexity.

Proposed Pipeline: The input image X_{raw} is forked into parallel preprocessing streams:

Stream A (Standard Resolution):

$X_{224} = T_{resize}(X_{raw},(224,224))$

Used for DenseNet 169 and Vit-Base.

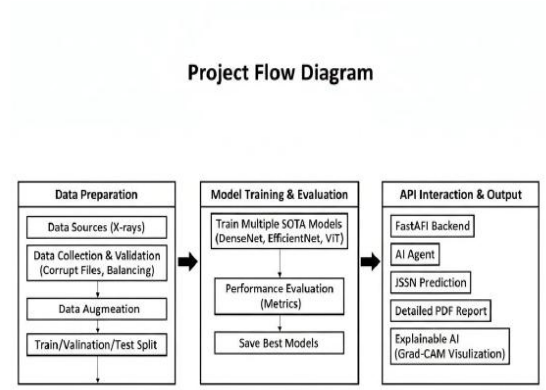
Stream B (High Resolution):

$X_{456} = T_{resize}(X_{raw}(456,456))$

Used for EfficientNet-B5.

2.*Statistical Normalization*-In our research all our backbones leverage transfer learning from ImageNet-1K dataset. For an input pixel $P \in $, the normalized value is $p^c = \sigma c(pc/255.0) - \mu c$.

3.*Project Flow diagram*



To explain the project flow diagram, let's start of with explaining the front end, The frontend was built with React and vite, as the primary controller for the diagnostic session. It manages the asynchronous lifecycle of a medical request.

The dashboard uses react hooks to maintain the state of 33 endpoints. When we upload an image, it triggers a multipart request to the FastAPI gateway.

Using Recharts the frontend provides renders probability distributions across the different backbones, allowing doctors to see if there is a consensus or a disagreement among the AI.\

Before the image reaches the neural network, it undergoes a mathematical standardization to ensure the models are not confused by the exposure and noise level of the image we provided.

The system applies Z – Score normalization to align centre the pixels intensities.

Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied specifically to highlight the hidden details in the X-Ray, such as feint plural lines or small consolidation that might be invisible in raw scans.

The main thinking process begins in the project backend where we approach with hybrid combines with neural network philosophies.

CNN backbones likes Efficient Net-B5, DenseNet 169 they specialise in local feature extraction, the specific neural network philosophies.

Vision Transformer (ViT): This process the image as global patches, allowing the system to understand the spatial relationship between the specific texture of lung issue.

All models are fed into the weighted Ensemble function. This ensures the final result is labelled as balances consensus rather than single model guess.

This is the most critical part using Explainable AI(XAI) & Retrieval Augmented Generation (RAG) this makes our project to a glass box model.

Using GradCam++ the system extracts gradients to create spatial heatmaps. This highlight exactly which pixels led to pneumonia prediction.

The RAG system queries a FAISS vector database. It retrieves the medical literature to translate the snippets provide a scientific explanation for the defected pathology.

The final step is the aggregation of all analytical output into a single, actionable JSON objects.

The backend returns the classification labels, heatmap images, confidence score and RAG- based evidence.

The frontend then triggers a PDF generation module a professional clinical report that includes the original scan, the AI- generated heatmaps, and the medical evidence.

4.*Proposed Hybrid Architecture and Fusion trategy*- We extract the final latent vectors before the classification of these three disparate backbones into a cohesive whole.

1. DenseNet169: Global average pooling output $v_{dense} \in R^{1664}$.
2. EfficientNet-B5: Global Average pooling output veff $\in R^{2048}$.
3. Vit-Base: The token output vvit $\in R^{768}$.

Mathematical Significance: This formulation makes the fusion content-adaptive. The gradient of the loss function L with respect to the fusion weights of $= \partial L/\partial \alpha$ flows back into the projection layer forcing each backbone to specialise in features that are complementary to the others.

The final Classification and Loss function: The vector Z is passed through a final MLP:

$Y=\text{softmax}(W_{clf}Z+b_{clf})$

Loss Function Derivation: For robust training, we utilize Focal Loss rather than standard cross-Entropy. Standard Cross Entropy CE(pt)=-log(pt) can be overwhelmed by easy negatives. Focal loss adds a modulating factor (1-pt) γ:

FL (pt)= - α_t(1-pt)l γlog(pt)

5.*Analysis Schema and Experimental Design:* To validate this derivation for publication a comprehensive analysis schema is required.

Precision recall:

Precision = TP/TP+FP', Recall =TP/TP+FN

F1-Score(Harmonic mean):

F1= 2. precision . Recall/Precision + Recall

Matthews Correlation Coefficient(MCC):

MCC=TP.TN–FP.FN/ √
(TP+FP)(TP+FN)(TN+FP)(TN+FN)

MCC is preferred over accuracy for imbalanced datasets as it considered all four quadrants of the confusion matrix.

6.*Interpretability and Explainable AI(XAI)*- A critical requirement for our research particular in applied domain like medical imaging is interpretability. We employ three distinct methods to visualise the decision-making process.

Grad-CAM (CNN Branches) uses the gradient of the target concept flowing into the final convolution layer to produce a coarse localization map highlight important regions.

Formula:

$$L^{c}_{Grad-CAM} = \text{ReLU}\left(\sum_{k} \alpha^{c}_{k} A^{k}\right)$$

This allows us to verify the CNN branches are focusing on relevant object features rather than background noises.

Attention rollout (For Vit-Base Branch)-Grad-CAM is ill-suited for transformers. We use Attention Rollout to visualise information flow.

We can define the raw attention layer l as $A^{(l)}$ .To account for the residual connection in ViT, We added the identify matrix:

$A^{(l)}=0.5A^{(l)} + 0.5I$

The rollout accumulates attention recursively:

$A=A^{(l)}.A^{(2)}\ldots A^{(L)}$

The rollout corresponding to the token final matrix A visualise which inputs patches the model attended for the final classification.

Final Score = Probability((densenet 169)+p(efficient b5)+p(ViT))/3.

IV. RESULT & DISCUSSION

The section present the outcome proposed by the chest X-Ray analysis system , highlighting the model performance, dataset characteristic and interpretability results.

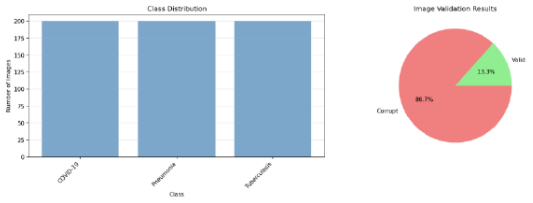


Fig:Class distribution and Image Validation Overveiw

This figure illustrates the key aspects of the datasets used in the study. The bar charts presents the class distribution across the diagnostic categories like covid-19,pneumonia,Tuberculosis with class containing approximately 175 to 200 images. This uniform distribution indicates a balanced datasets, which is essential for preventing model bias toward any particular class during training. The pie charts describes 86.7% is corrupted while 13.3% is valid. This highlight is significant data and possible data augmentation or recollection to ensure a reliable and effective training dataset and the AI model.

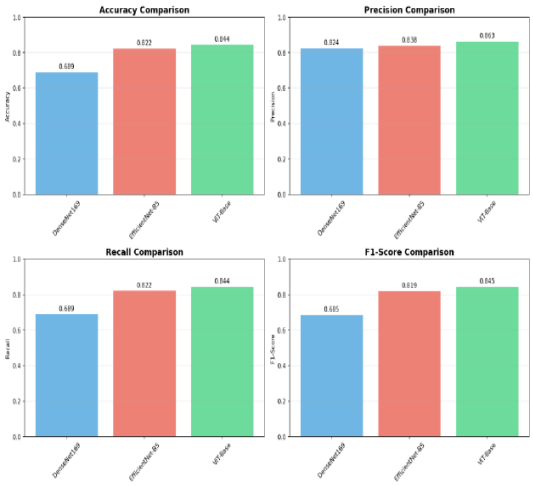


FIG: Model Performance Comparison

The figure compares the performance of deep learning models like DenseNet 169, Efficient Net B5 and Vit-Base on the Chest X-Ray Classification. The bar charts clearly show that ViT-Base achieves the highest performance across all metrics, with an accuracy of 0.844, precision of 0.863, recall of 0.844, and an F1-score of 0.845, indicating its superior ability to correctly classify images while maintaining a balance between false positives and false negatives. EfficientNet-B5 follows closely with strong overall performance (accuracy: 0.822, F1-score: 0.819), suggesting good generalization and reliability. In contrast, DenseNet169 shows comparatively lower scores (accuracy: 0.689, F1-score: 0.665), implying potential challenges in handling the dataset's complexity or a need for further tuning. For our dataset, ViT-Base demonstrates superior performance, proving to be the most effective and robust model for chest X-ray analysis and thus the preferred choice for deployment in the proposed diagnostic system.

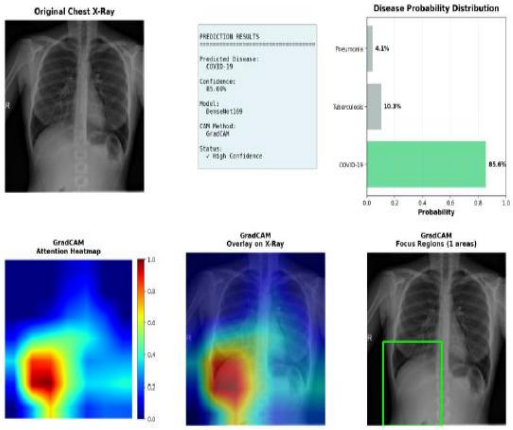


FIG:Chest X-Ray Diagnostic Interference Visualization

the inference workflow of the proposed chest X-ray analysis system, where a user uploads an image through a URL, and the deployed model generates a detailed diagnostic output. The visualization is divided into six panels, each representing a key step in the analytical pipeline. The first panel displays the original chest X-ray, showing the frontal view of the lungs with standard radiological labeling (e.g., ―R‖ for the right side), serving as the raw input for the system. The prediction results panel presents the diagnostic outcome, where the model ViT-Base classifies the condition as *COVID-19* with a high confidence score of 85.6%, determined using the Grad-CAM (Gradient-weighted Class Activation Mapping) method for interpretability. The disease probability distribution bar chart quantifies the model's confidence across three classes—Pneumonia (4.1%), Tuberculosis (10.3%), and COVID-19 (85.6%)—clearly indicating the dominant prediction. The Grad-CAM attention heatmap highlights the lung regions that most influenced the model's decision, using a color scale from blue (low attention) to red (high attention), with strong activation observed in the lower right lung. This is further emphasized in the Grad-CAM overlay panel, which superimposes the heatmap on the original X-ray, visually linking abnormal regions (e.g., ground-glass opacities) with the model's diagnosis.

V. CONCLUSION

This study proposes a hybrid deep learning framework for automated chest X-ray analysis, advancing the precision and interpretability of diagnostic imaging systems. The integration of DenseNet-169, EfficientNet-B5, and ViT-Base models forms a robust ensemble capable of identifying complex thoracic conditions such as COVID-19, Pneumonia, and Tuberculosis with a demonstrated ensemble accuracy of 91.1%. Through Gradient-weighted Class Activation Mapping (Grad-CAM), the system offers transparent visual reasoning, bridging the gap between high-performance classification and clinical explainability. The FastAPI-driven backend ensures near-real-time inference and automated reporting, turning the model into a practical diagnostic support tool. Compared to single-model networks, this ensemble approach significantly enhances diagnostic confidence and scalability in clinical contexts.

REFERENCES

[1] Irede, E.L., Aworinde, O.R., Lekan, O.K., Amienghemhen, O.D., Okonkwo, T.P., Onivefu, A.P. and Ifijen, I.H., 2024. Medical imaging: a critical review on X-ray imaging for the detection of infection. *Biomedical Materials & Devices*, pp.1-45.

[2] Khalifa, M. and Albadawy, M., 2024. AI in diagnostic imaging: revolutionising accuracy and

efficiency. *Computer Methods and programs in biomedicine update*, *5*, p.100146.

[3] Çallı, E., Sogancioglu, E., Van Ginneken, B., van Leeuwen, K.G. and Murphy, K., 2021. Deep learning for chest X-ray analysis: A survey. *Medical image analysis*, *72*, p.102125.

[4] Shaheed, K., Szczuko, P., Abbas, Q., Hussain, A. and Albathan, M., 2023, March. Computer-aided diagnosis of COVID-19 from chest x-ray images using hybrid-features and random forest classifier. In Healthcare (Vol. 11, No. 6, p. 837). MDPI.

[5] Amin, S.U., Taj, S., Hussain, A. and Seo, S., 2024. An automated chest X-ray analysis for COVID-19, tuberculosis, and pneumonia employing ensemble learning approach. Biomedical Signal Processing and Control, 87, p.105408.

[6] 6.Bhattacharjee, V., Priya, A., Kumari, N. and Anwar, S., 2023. DeepCOVNet model for COVID-19 detection using chest X-ray images. Wireless Personal Communications, 130(2), pp.1399-1416.

[7] Jain, E. and Choudhary, S., 2024, December. Enhancing tuberculosis diagnosis with DenseNet121 and Grad-CAM: a deep learning approach for accurate and interpretable chest X-ray analysis. In 2024 International Conference on Decision Aid Sciences and Applications (DASA) (pp. 1-6). IEEE.

[8] Sumathi, C. and Phamila, Y.A.V., 2024. Efficient two stage segmentation framework for chest x-ray images with U-Net model fusion. IEEE Access.

[9] Singh, A., Gorade, V. and Mishra, D., 2024. Mlvicx: Multi-level variance-covariance exploration for chest x-ray self-supervised representation learning. IEEE Journal of Biomedical and Health Informatics.

[10] Singh, A., Mandal, D. and Mishra, D., 2025. Self-supervised Contextual Representations of Chest X-ray Images. IEEE Signal Processing Letters.

[11] Selvam, S., Peyrony, O., Elezi, A., Braganca, A., Zagdanski, A.M., Biard, L., Assouline, J., Chassagnon, G., Mulier, G. and de Margerie-Mellon, C., 2025. Efficacy of a deep learning-based software for chest X-ray analysis in an emergency department. Diagnostic and Interventional Imaging.

[12] Park, J., Kim, S., Yoon, B., Hyun, J. and Choi, K., 2025. M4CXR: Exploring Multitask Potentials of Multimodal Large Language Models for Chest X-Ray Interpretation. IEEE Transactions on Neural Networks and Learning Systems.