MedScan AI

A Radiologist's companion in navigating thoracic X-Rays

MSDSP462 Capstone Project

Group 4 (Baazigars)

Abhinay Kumar Singh | Anubhav Bhatnagar | Pratik Parag | Vivek Mishra

March 2025 | Northwestern University MSDS

Abstract

The MedScan AI system is a web-based diagnostic aid designed to assist radiologists and lab technicians in identifying lung diseases from X-ray images. Given the shortage of specialist doctors in regions like sub-urban and rural India, this proof-of-concept study explores the feasibility of an AI-powered diagnostic tool which can aid diagnosis in absence of specialist doctors. The Medscan-AI app takes high quality Chest X-Rays as input and based on trained model it identifies the potential diseases and also shows the region of concern in the X-Ray by demarcating it and showing a heatmap along with the confidence level of predictions.

The system employs an EfficientNetV2-S architecture, GradCAM for interpretability, and an AI-driven prescription generation module. Model performance was evaluated using the NIH Chest X-ray dataset, achieving an AUC of 0.880 across disease classes. Fine-tuning EfficientNet layers and incorporating additional convolutional layers improved classification accuracy, particularly for conditions with low natural separation.

Future enhancements include deeper training, ensemble learning, self-attention mechanisms, and hybrid architectures integrating image and feature-based embeddings. Expanding dataset diversity, addressing class imbalances through synthetic data, and validating the model across varied demographic groups will further enhance robustness. Beyond machine learning, successful deployment requires collaboration with medical professionals, regulatory bodies, and public health organizations to ensure real-world usability and compliance. This study lays the groundwork for scalable AI-driven diagnostics in resource-constrained healthcare environments.

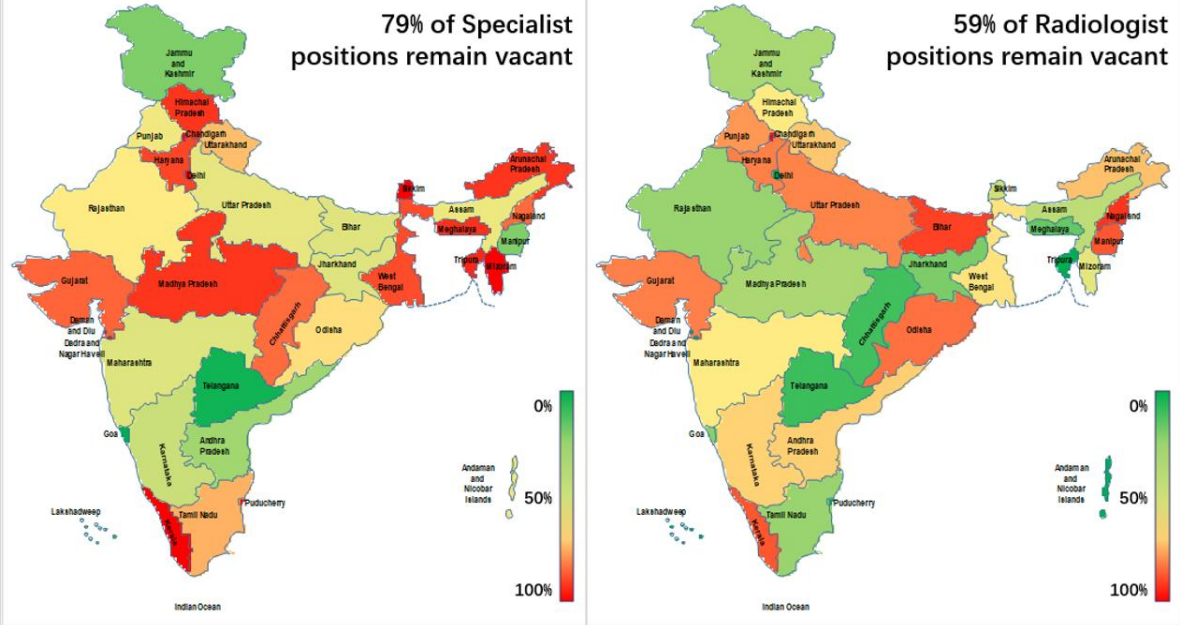
1. Introduction: The Public Health Problem

This paper aims to design a web-based application to assist radiologists and lab technicians in diagnosing common—and often life-threatening—lung diseases from X-ray images. While developing a fully functional, production-ready system would require vast amounts of publicly available and region/hospital-specific data, along with substantial computational resources beyond the scope of this project, it is important to emphasize that such a system would also necessitate a rigorous design-thinking approach. This would involve extensive consultations with patients, radiologists, public health officials, and hospital administrators to define precise user needs.

Instead, this project focuses on developing a Proof of Concept (PoC) application to demonstrate feasibility and generate interest from investors, government agencies, and hospitals. The goal is to lay the groundwork for scaling the initiative to an industrial level.

1.1 Why is a diagnostic aid relevant?

Fig I1 Significant shortages of trained medical personnel cause delayed and potentially poor quality medical care for terminal lung diseases



India faces a critical shortage of specialist doctors capable of diagnosing and managing complex lung diseases. According to the *Rural Health Statistics 2021-22* report by the Ministry of Health and Family Welfare (MoHFW), 59% of radiologist positions and 79% of specialist positions remain unfilled (Rawat et al., 2023). This shortage contributes to several pressing challenges:

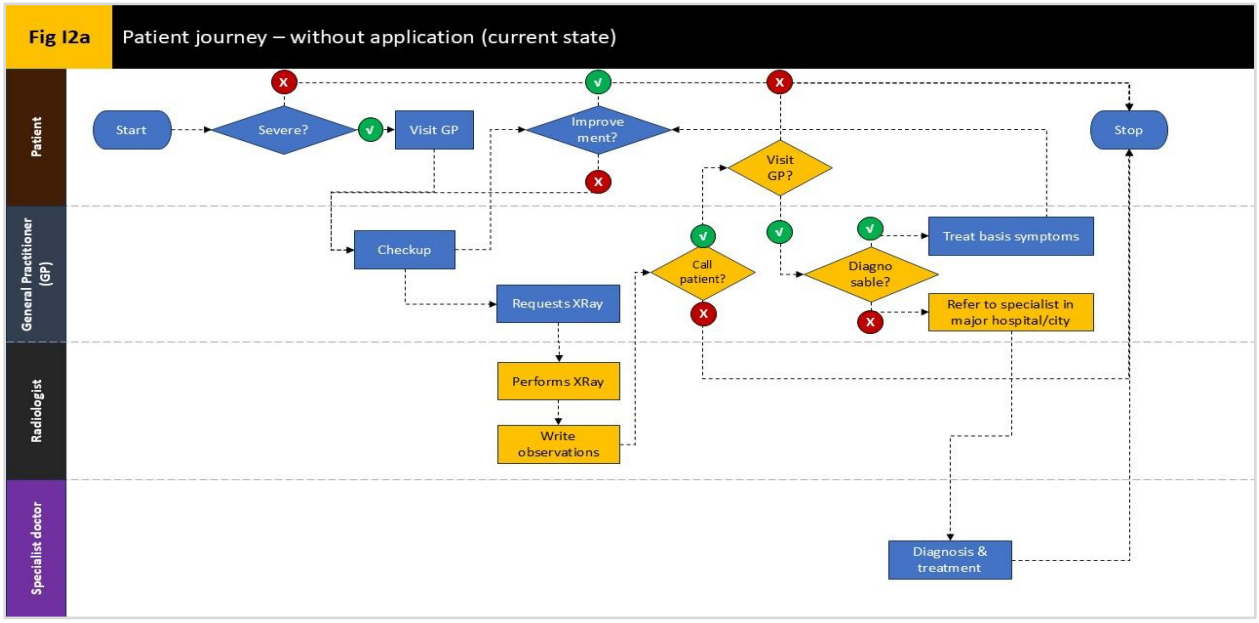
- **Increased burden on less experienced physicians** and a growing reliance on alternative medical sectors to address lung-related ailments.
- **Reduced time per patient for specialists and radiologists**, heightening the risk of diagnostic errors.
- **Longer wait times and frequent doctor visits**, leading many patients (particularly patients belonging to the Economically Weaker Sections of society) to delay seeking medical care until symptoms worsen or reach advanced stages.

These challenges are particularly concerning given India's alarming statistics on lung diseases. According to the *Global Burden of Disease Report 2017* and the International Primary Care Respiratory Group (IPCRG), India has the highest global incidence of Chronic Obstructive Pulmonary Disease (COPD), with 55.23 million cases annually and the second-highest number of COPD-related deaths worldwide (0.85 million deaths per year). Moreover, India accounts for 43% of global asthma-related deaths (IPCRG, 2022).

Given these urgent healthcare gaps, a diagnostic aid leveraging AI-powered image analysis could play a crucial role in enhancing early detection, reducing diagnostic errors, and ultimately improving patient outcomes in India.

1.2 Process related problems:

Figures I2a and I2b show the patient's journey from initial diagnosis through treatment by the appropriate specialist for a specialized lung condition. The blocks in shaded in gold are stages or decisions that would be affected by the presence of an AI powered diagnostic system:



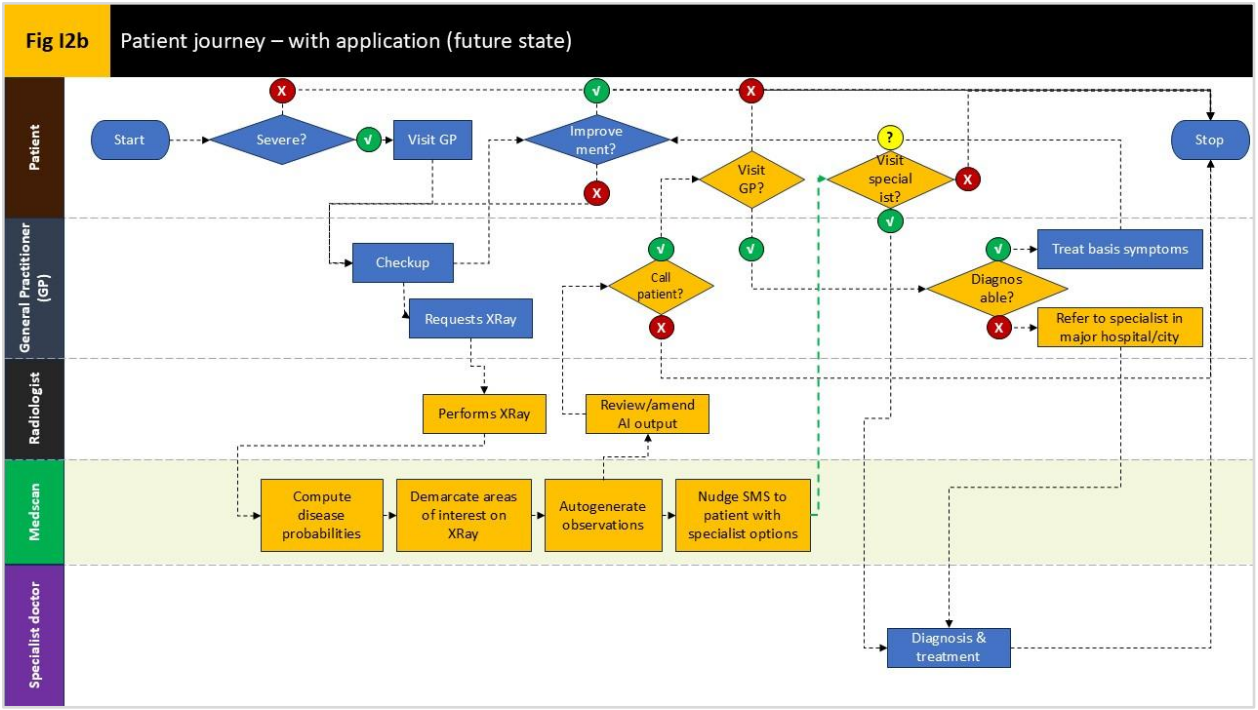


Figure I2a contains stages that present problems that our app may be able to address:

Table I1: Risks present within current patient journey

Person or entity	Stage	Risk presented
Radiologist	Writes observations	India has 20,000 radiologists serving 1.4 billion people. Industry experts estimate that there is one radiologist per 15 clinicians, whereas ideally this ratio should be 1 radiologist per 5 clinicians. Given that there is a simultaneous shortage of clinicians, the workload is likely more than 3x what is recommended, leading to radiologists working long hours and spending very little time per case (News18, 2023). Thus, quality of analysis may not be adequate, leading to late diagnosis of critical cases
General Practitioner (GP)	Call patient?	Fig I1 shows that there is a 79% shortage of doctors in rural areas. This means that doctors have lesser time to diagnose each patient (than industry experts would consider ideal), potentially leading to errors in diagnosis. Additionally, with manual or no patient call back systems, chances of followup are reduced, leading to several critical patients not receiving adequate or timely medical care.
Patient	Visit GP?	While Indian data is unavailable for this step, studies by McNulty (2023), Sandelowsky et al. (2022), Abrahams et al. (2024) among lung disease patients in the US and Sweden suggest that only 22%-49% of patients follow up with a physician post X-Ray within the clinically designated period for their respective diseases. While additional studies would be needed to prove the extent and severity of this problem in India, the US/Swedish data does indicate a potential problem that should be addressed
General Practitioner (GP)	Diagnosable?	Very often, physicians with limited training/experience attempt to treat cases that should ideally be referred to specialists earlier on. In such cases, patients do not have the option or information to consult specialists by themselves, leaving them with late diagnosis, no diagnosis or improper diagnosis.

1.3 Proposed system: lab technician's assistant and patient auto reminder

In order to address the gaps mentioned above, the following features will be factored into the design for the MedScan application:

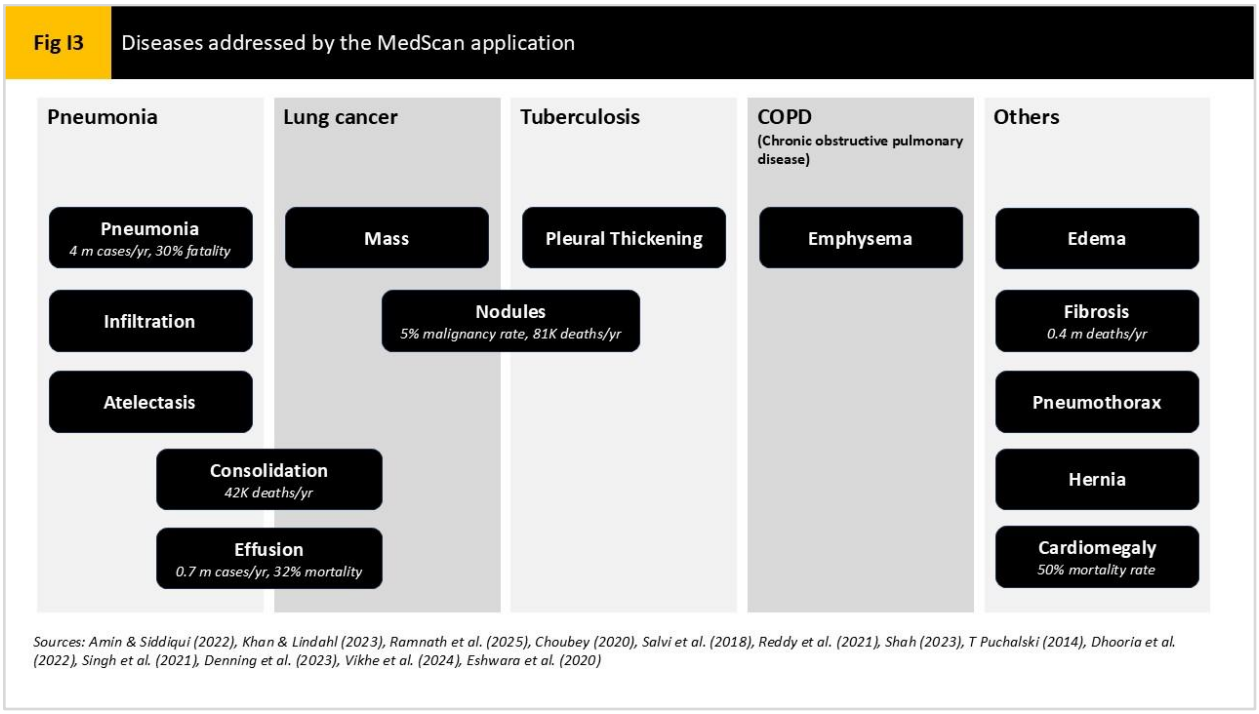
Table I2: Key features of the MedScan app

Feature	Problems addressed	Rationale for including feature
Automated disease detection	● Reduce radiologist workload and burnout ● Improve quality of diagnoses	According to AI assisted radiology startups like RadAI, AI assisted disease detection, heatmap based images and ChatGPT enabled radiologist notes can reduce 60 min per shift (approx 12.5% efficiency) in radiologist effort. This can further reduce radiologists' reported burnout by 84% (RadAI). In other estimates, Babyn & Adams (2025) estimated that radiologist effort reduced from 34.2 seconds/case to 19.8 seconds/case (approx 42% efficiency gains) due to AI assistants with these features.
GradCAM based disease localization	● Improved quality of diagnoses	
Automated radiologist notes generation	● Need for standardized, detailed and exhaustive notes	
Automated SMS reminders to patients	● Large patient dropoff in follow up visits post X-Ray ● Patients' need for access to specialist information to get a second opinion	Per the process flow (Fig I2a), patients only get access to specialist information when they are referred by a doctor. SMS based notifications that contain the contact information of local specialists improve patients' options and access to information - allowing critical patients to seek proper care even if their local doctor is under equipped to diagnose them. This also makes sure that the most critical patients are prompted to pursue follow ups. According to studies by Banerjee et al. (2021), Telukuntla et al. (2021), and Alturbag (2024), SMS based nudge reminders can increase

		patient followup rates for a variety of conditions by 5%-44%.

1.4 Disease scope: Understanding lung diseases:

For the purposes of a Proof of Concept, this study covers X-Ray samples for 14 lung conditions, as well as the possibility of an X-Ray having "No finding". However, since patients often exhibit multiple conditions at once, these conditions can be grouped into 5 disease families, as shown in Figure I3 below. However, depending on the disease load and data of a region where deployment is required, this disease composition can be widened to include other lung conditions or other non lung related diseases.



2. Literature Review

In the interest of building a computationally efficient, accurate model to aid physicians, specialists and lab technicians, we performed a survey of past studies involving the NIH Chest X-Ray dataset as well as similar multilabel classification problems from the medical domain.

Baltruschat et al. (2019) explored deep learning approaches for multi-label chest X-ray classification, evaluating ResNet architectures (ResNet-50, ResNet-38, and ResNet-101) with transfer learning, fine-tuning, and high-resolution imaging. They developed two model families: one incorporating non-image data and another using only image data. For each, they assessed the Area Under the Receiver Operating Characteristic Curve (AUC) across three configurations: (1) off-the-shelf (OTS), (2) fine-tuned (with trainable ResNet layers), and (3) high-resolution OTS. Their findings revealed that fine-tuning improved performance over OTS models for most classes, while high-resolution images yielded results comparable to fine-tuned models. To enhance interpretability, they employed Gradient-weighted Class Activation Mapping (GradCAM) to highlight image features influencing classification decisions.

Chen et al. (2019) introduced a deep Hierarchical Multi-Label Classification (HMLC) approach, demonstrating that modeling conditional probabilities enhances chest X-ray classification. Their model structured diseases into five broad families, subdividing them into individual conditions. Using a DenseNet-121 backbone, they applied a two-stage training strategy: initial training with Hierarchical Label Conditional Probability (HLCP), followed by fine-tuning with Hierarchical Label Unconditional Probability (HLUP). Achieving an AUC of 0.887 across all classes, their approach showed that stagewise classification improves accuracy, particularly for overlapping conditions such as lung cancer, tuberculosis, COPD, and pneumonia.

Holste et al. (2024) conducted the CXR-LT challenge, where nine teams used EfficientNet and DenseNet pretrained feature extractors, ensemble learning, and image augmentation for multi-label classification. Although their dataset differed from the NIH Chest X-Ray dataset used in our study, many conditions overlapped. Their results highlighted the strong performance of DenseNet architectures and the

benefits of class-weighted loss for minority condition classes. Some teams pre-trained models on datasets other than ImageNet, which further improved mean average precision (mAP), underscoring the potential of alternative pre-training strategies.

Huang et al. (2022) investigated deep transfer learning for chest X-ray analysis using a private dataset along with ImageNet, ChestX-ray, and CheXpert. They tested various CNN architectures, including ResNet50 and DenseNet121, across different transfer learning approaches. Their results showed that transfer learning enhanced model performance, particularly when augmenting the NIH Chest X-Ray dataset with additional datasets. The study also highlighted that image augmentation and pretraining on diverse datasets improved minority class performance.

Kufel et al. (2023) demonstrated strong results using EfficientNet for multi-label classification of 14 diseases in the NIH ChestX-ray14 dataset. Their study suggested EfficientNetV2-S as a promising alternative model architecture. To enhance performance, they supplemented EfficientNet with trainable GlobalAveragePooling, Dense layers, and Batch Normalization. Notably, they adjusted train/test splits to improve performance on minority classes. A cross-study comparison showed that EfficientNet-based models performed comparably or slightly better than DenseNet-based models while being computationally efficient.

Wang et al. (2017) explored multi-label disease classification and localization using a Deep Convolutional Neural Network (DCNN). They modified pre-trained models (AlexNet, GoogLeNet, VGGNet-16, and ResNet-50) by adding a transition layer, global pooling, prediction, and loss layers. Experimenting with max, average, and Log-Sum-Exp (LSE) pooling, they found that ResNet-50 performed best, with LSE pooling improving classification accuracy. They also implemented Weighted Cross Entropy Loss (W-CEL) to address class imbalances, significantly enhancing detection of underrepresented conditions. While the model generated bounding boxes for disease localization, resolution limitations led to oversized bounding boxes compared to ground truth annotations.

2.1 Key Insights for Model Design based on Literature Review

Based on these studies, we derive the following considerations for our architecture:

1. **Transfer Learning Model Selection**: Most studies employed EfficientNet, DenseNet, or ResNet architectures. While many used frozen feature extractor layers, Baltruschat et al. (2019) found that fine-tuning the entire model improved performance.
2. **Input Image Resolution**: Higher-resolution images (448×448 pixels by Baltruschat et al. (2019) and 1024×1024 pixels by Kufel et al.) enhanced the detection of subtle abnormalities in lung X-rays.
3. **Pretraining Data & Image Augmentation**: Huang et al. (2022) demonstrated that combining datasets (e.g., NIH Chest X-Ray with CheXpert) improved performance, particularly for minority classes. Several studies incorporated image flipping, brightness adjustments, and saturation changes as augmentation strategies. Kufel et al. (2023) even modified training/test splits to improve class balance.
4. **Tail Architecture**: Studies such as Wang et al. (2017) and Kufel et al. (2023) effectively used GlobalAveragePooling, Dense layers, and regularization techniques in the classification layers after the feature extractor.
5. **Interpretability: GradCAM vs. Bounding Boxes**: Wang et al. (2017) showed that while classifiers could detect conditions, they struggled with precise disease localization. Conversely, Baltruschat et al. (2019) used GradCAM to highlight feature importance, effectively approximating diseased regions. Given the limited availability of bounding box annotations in our dataset, and considering that our model is designed to assist lab technicians rather than provide final diagnoses, GradCAM heatmaps may serve as a viable alternative for interpretability.
6. **Computational efficiency:** Krishna et al. (2022) explores the use of EfficientNet vs DenseNet in a similar problem (among other pretrained models). Their work shows that EfficientNet B0 provides 39.7% time savings over DenseNet201, while providing comparable performance on

AUC across classes. While both architectures merit testing, this may become an important consideration in the choice of final model architecture.

7. **Measuring performance - the broad use and relevance of AUC:** Most studies surveyed used AUC as a performance criterion to measure model effectiveness. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a performance metric that quantifies a classifier's ability to distinguish between positive and negative instances across all possible decision thresholds. AUC measures the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance. In multi-label classification, particularly in imbalanced datasets like medical imaging, AUC is preferred over precision, recall, or F1-score because it is threshold-independent and evaluates the model's ranking ability rather than just its accuracy at a specific threshold. Unlike precision and recall, which can be significantly affected by class imbalance, AUC provides a more stable and holistic measure of classifier performance across all labels. Additionally, it allows per-class assessment and can be averaged to give an overall performance indicator. While precision, recall, and F1-score are useful when false positives and false negatives have unequal costs, AUC remains a more robust and informative metric in scenarios where ranking predictions correctly is more critical than selecting a fixed classification threshold. Therefore, continuing with AUC as a performance metric may be advisable in our case.

These insights guide our architectural decisions, ensuring that our model balances accuracy, interpretability, and computational efficiency.

3. Exploratory Data Analysis

The structure, composition and quality of the underlying training data significantly influences the architectural choices that go into the final model. We approach data exploration in the following manner

3.1 EDA Approach

Exploration of raw data samples the dataset used for predicting lung diseases from chest X-ray images originates from the National Institutes of Health, Bethesda, MD, USA (Wang et al., 2017). It comprises 112,120 lung X-ray images collected from 30,805 unique patients, with each image having a resolution of 1024x1024 pixels. Metadata, including patient ID, age, and sex, is also available. The dataset covers 14 lung conditions, with images potentially belonging to multiple conditions or classified under the "no finding" category. The condition labels were derived using natural language processing (NLP) techniques, achieving an accuracy of approximately 90%. Figure E1 displays sample X-Ray images for 8 major diseases covered (Wang et al., 2017).

Fig E1 Sample X-Ray images from the NIH Chest X-Ray dataset collated by Wang et al. (2017)

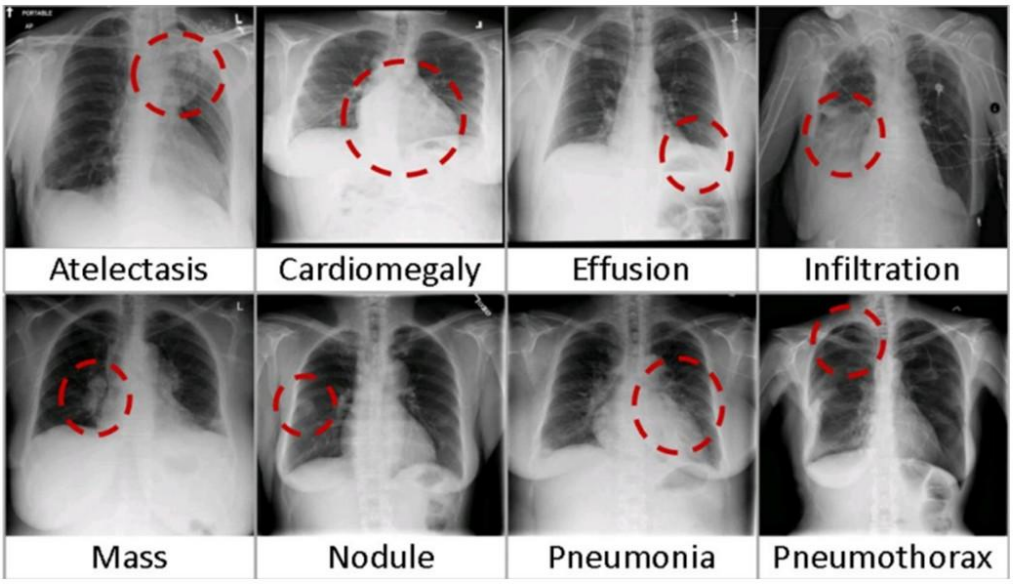


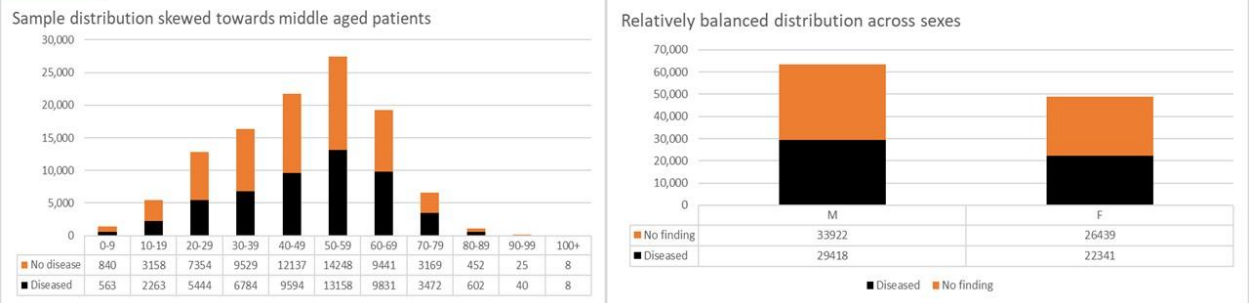
Image source: Wang et al. (2017)

Ensuring demographic balance: Given that this study aims to create a diagnostic aid for an Indian rural population, the dataset used should ideally be composed of Indian patients, evenly distributed

across genders, age bands and condition classes. Such a dataset should ideally reflect the prevalence of smoking, the relatively common consumption of tobacco, paan, gutka and khaini by diseased patients, or the effects of industrial pollution, asbestos exposure or indoor smoke on lung condition. Image resolution and capture should ideally reflect the condition and capabilities of imaging machines across the region of interest/app implementation.

However, this dataset is of North American origin, with limited demographic information about the composition of patients. While we do not possess data regarding comorbidity or the ethnic or physical makeup of patients included in the dataset, we can conclude that the dataset is relatively balanced between genders, and is slightly skewed towards middle aged patients:

Fig E2 NIH Chest X-Ray dataset shows relatively normal distribution across age classes and relative balance between genders



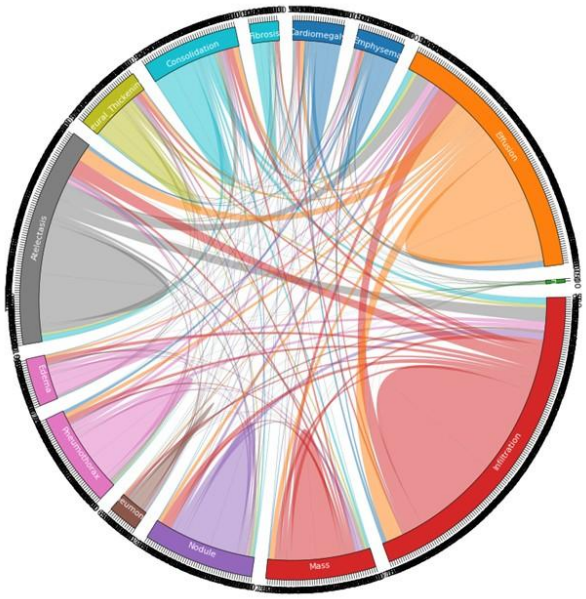
Understanding data limitations: In addition to the demographic concerns articulated above, the fact that the authors used NLP to extract labels with 90% accuracy suggests that there may be a high degree of variation and cross contamination between classes (Wang et al., 2017). For instance, mild cases of cardiomegaly or very small nodules may enter the no finding class. This may cause the classifier to be trained on noisy data, and can hinder performance.

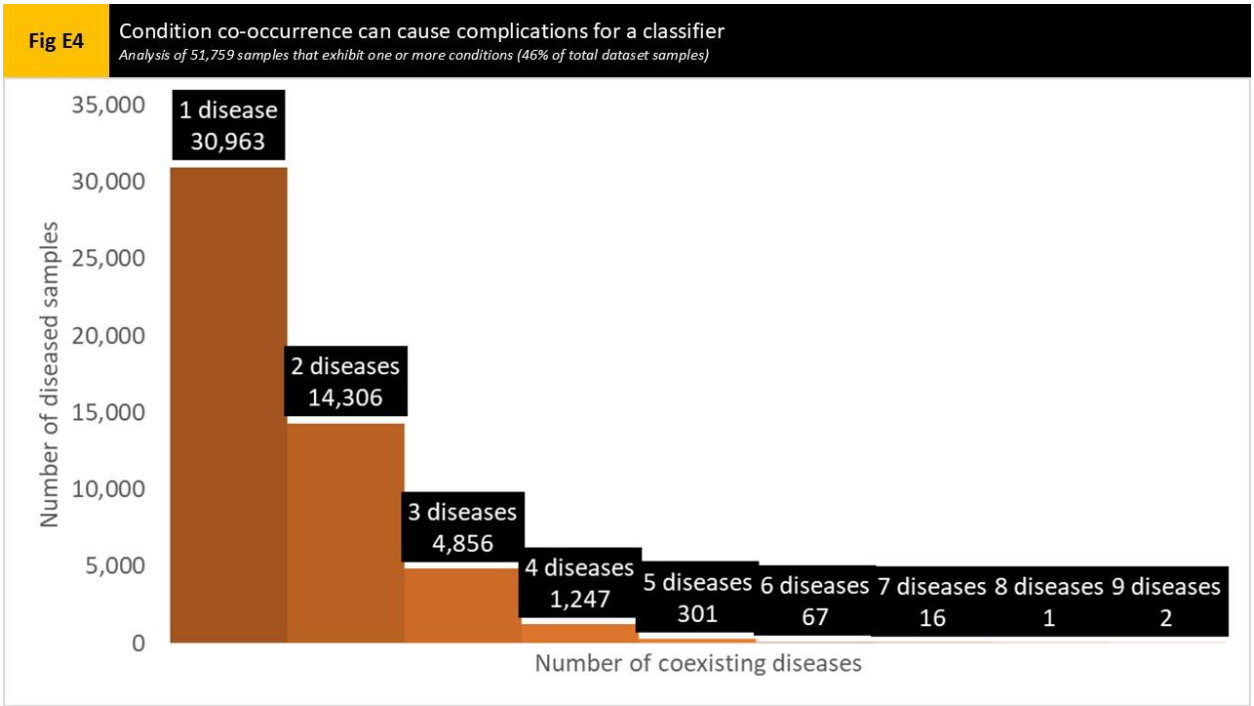
Examining disease co-occurrence: The lung conditions reflected in the dataset may occur in combination, given the disease family that a person is afflicted by. The co-occurrence of diseases can potentially send noisy signals to a classifier, which may end up using the features of one condition to predict the prevalence of an entirely unrelated condition in the dataset. Therefore, it would make sense to train the classifier on a subset of data predominantly composed of single condition samples (to help the

classifier distinguish between pure play instances of each condition for conceptual clarity). Figures E3 and E4 below examine (1) the co-occurrence of disease pairs and (2) the prevalence of single condition samples within the dataset. Thus, we determine that there are approximately 45,000 samples from the "1 disease" and "2 disease" categories in Figure E4 that would be prime examples for training a classifier model. In certain cases, we may choose samples where 3 diseases are present.

Fig E3 Condition co-occurrence is relatively common

Class	No Finding	Cardiomegaly	Emphysema	Effusion	Hernia	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	Edema	Atelectasis	Pleural_Thickening	Consolidation	Fibrosis
No Finding	59946	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cardiomegaly	0	2776	44	1063	7	587	102	108	41	49	127	370	111	169	52
Emphysema	0	44	2516	359	4	449	215	115	23	747	30	424	151	103	36
Effusion	0	1063	359	13317	21	4000	1254	912	269	996	593	3275	849	1287	188
Hernia	0	7	4	21	227	33	25	10	3	9	3	40	8	4	8
Infiltration	0	587	449	4000	33	19894	1159	1546	605	946	981	3264	750	1221	345
Mass	0	102	215	1254	25	1159	5782	906	71	431	129	739	452	610	117
Nodule	0	108	115	912	10	1546	906	6331	70	341	131	590	411	428	166
Pneumonia	0	41	23	269	3	605	71	70	1431	41	340	262	48	123	11
Pneumothorax	0	49	747	996	9	946	431	341	41	5302	33	774	289	223	80
Edema	0	127	30	593	3	981	129	131	340	33	2303	221	64	162	9
Atelectasis	0	370	424	3275	40	3264	739	590	262	774	221	11559	496	1223	220
Pleural_Thickening	0	111	151	849	8	750	452	411	48	289	64	496	3385	251	176
Consolidation	0	169	103	1287	4	1221	610	428	123	223	162	1223	251	4667	79
Fibrosis	0	52	36	188	8	345	117	166	11	80	9	220	176	79	1686





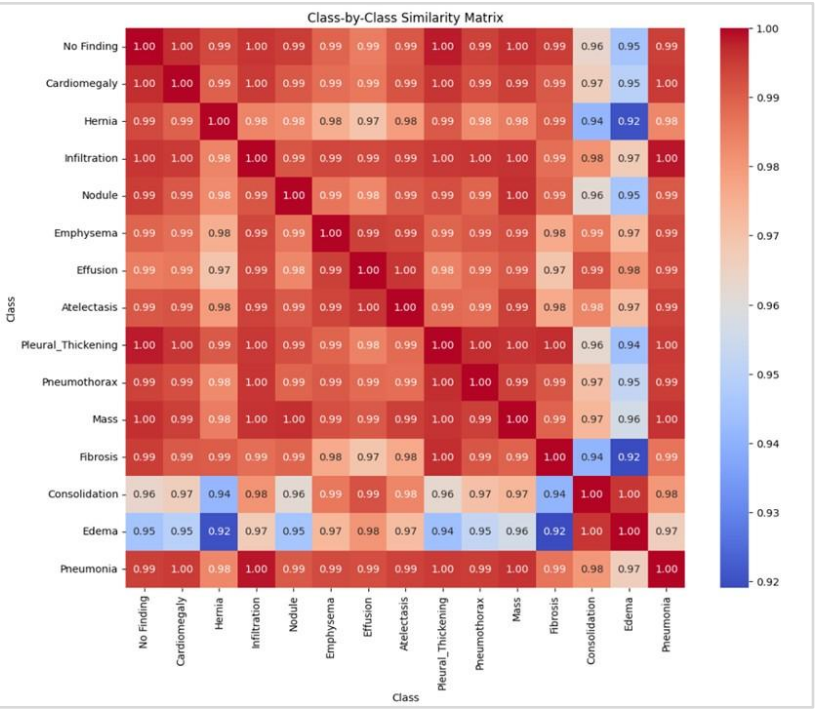
Examining class separation: Insights from clustering: The effectiveness of a classifier largely depends on its ability to distinguish the defining characteristics of each class in a dataset. One key factor influencing this ability is the natural separation between instances of different classes. Clustering methods, such as Agglomerative Clustering and K-Means, provide insight into this separation by grouping data points into k clusters, where k corresponds to the number of classes in the dataset.

One way of looking at class separation is to consider the separation between vector embeddings for single disease/condition images. We select 12,679 such images for analysis to this end. In this analysis, we generate and then cluster standalone EfficientNetV2-S embeddings of each image, which have a vector length of 1280. Ideally, the resulting clusters should align closely with the labeled classes. To quantify this alignment, we use the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) metrics. ARI evaluates how well the predicted clusters match the actual classes while accounting for chance, whereas NMI measures the amount of shared information between clusters and true labels.

Fig E5 Low class separation can hamper classifier discernment

Approach:

- **Select representative sample:**
  - 12,679 single condition images
  - Proportion of each class in dataset maintained
- **Generate image embeddings** for each image with EfficientNetV2-S (vector length:1280)
- **Reduce image embeddings for faster computation**, preserving 95% information using PCA
- **Create the mean embedding vector** for each class
- **Measure cosine distances between mean embedding vectors**
- **Compute similarity scores**



Our results indicate an ARI of 0.0161 and an NMI of 0.0444 (with similar values observed for Agglomerative Clustering). These low scores suggest that the natural structure of the data does not strongly separate classes, meaning that the clustering assignments are nearly random. This highlights the need for a more effective feature extractor that can better capture the distinctive attributes of each class, enabling the classifier to make meaningful distinctions.

3.2 Key Insights for Model Design based on EDA

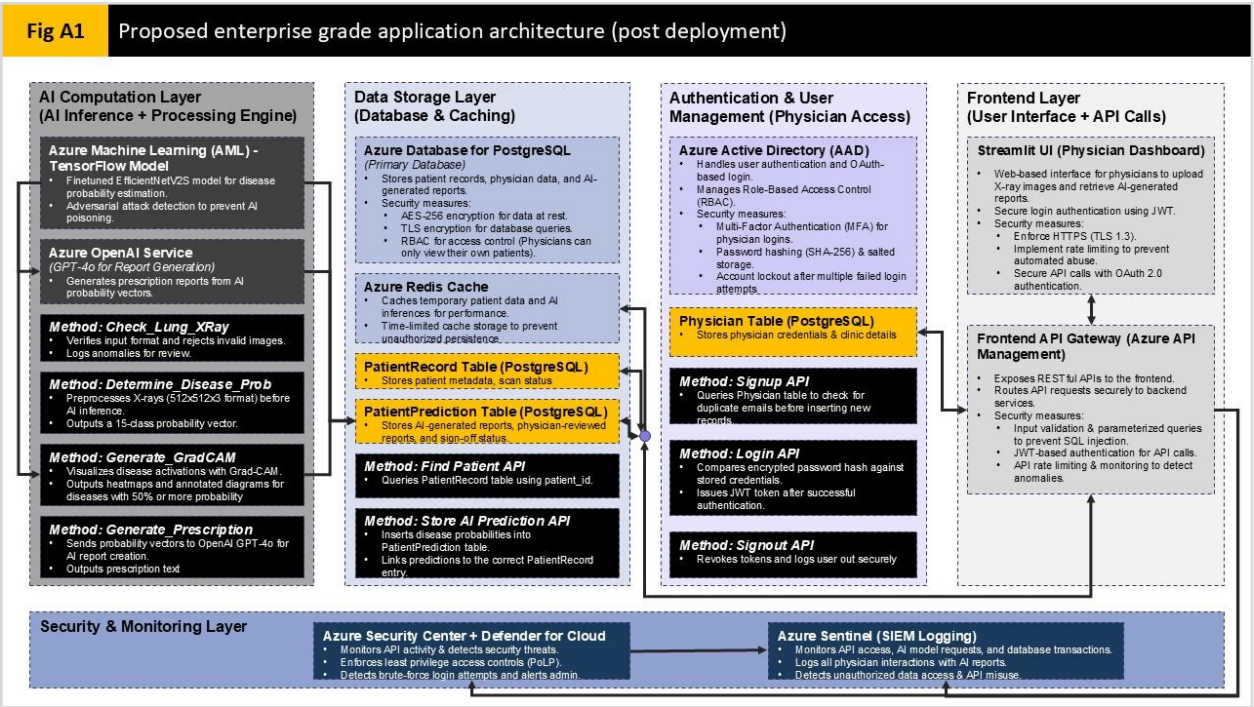
Based on the data reviewed above, the following trends are identified. These insights are likely to find potential application in our model architecture:

Table E1: Key Insights for Model Design from Exploratory Data Analysis (EDA)

Insight	Potential application
Age and sex distribution is relatively balanced	Random selection of samples within each class will not hamper demographic balance
Careful class balancing and sample selection needed	Choosing a subset with 500-1000 samples from each of the 15 classes will balance the training set
Condition co-occurrence can cause complications for a classifier	Limit or remove cases where more than one disease exists from the training set

Low class separation can hamper classifier discernment	Consider pre trained models like DenseNet201 or EfficientNetV2S for robust feature extraction
Disease categories do not naturally form distinct groups	

4. Application architecture



The proposed application architecture for MedScan AI is designed for secure, scalable, and efficient AI-powered medical diagnosis. It consists of five key layers, each serving a distinct role in managing user interactions, AI inference, data processing, security, and monitoring:

4.1 The Frontend Layer (User Interface + API Calls)

- The **Streamlit UI** (created in Python) provides an intuitive interface for physicians to upload X-ray images and view AI-generated reports.

- An **API Gateway (Azure API Management)** is used to manage all API requests, enforce security measures such as JWT authentication, input validation, and rate limiting, and ensure smooth integration with backend services.

4.2 The Authentication & User Management Layer

- **Azure Active Directory (AAD)** handles authentication via OAuth 2.0, enforcing Multi-Factor Authentication (MFA) and Role-Based Access Control (RBAC) to restrict patient data access.
- The **Physician Table (PostgreSQL)** securely stores hashed and salted credentials, preventing unauthorized access.

4.3 The Data Storage Layer (Database & Caching)

- **Azure Database for PostgreSQL** is used for storing structured patient data, AI-generated predictions, and physician annotations. Encryption at AES-256 ensures data security.
- **Azure Redis Cache** enhances performance by caching AI inferences, reducing redundant computation.
- The **PatientRecord Table** stores metadata for medical cases, while the **PatientPrediction Table** tracks AI-generated diagnoses and sign-off status, ensuring full traceability.

4.4 The AI Computation Layer (AI Inference + Processing Engine)

- **Azure Machine Learning (AML) with TensorFlow** is used for disease probability estimation from X-rays, incorporating adversarial attack detection for robustness.
- **Azure OpenAI Service (GPT-4o)** generates structured prescription reports from AI predictions, ensuring that output is validated before being presented to physicians.
- Additional processing techniques like GradCAM (for AI interpretability) ensure that physicians can verify AI-generated predictions visually.

4.4 The Security & Monitoring Layer

- **Azure Security Center + Defender for Cloud** provides real-time monitoring for unauthorized API activity and security threats.
- **Azure Sentinel (SIEM Logging)** logs API requests, physician interactions with AI reports, and potential unauthorized access, ensuring full auditability and regulatory compliance.

4.5 Key Design Considerations:

- **Security-First Approach**: Although not legally binding in India, HIPAA & GDPR compliance is ensured by enforcing end-to-end encryption (TLS 1.3, AES-256), role-based access control (RBAC), and multi-factor authentication (MFA). This is done to mitigate reputational risks and provide best in class security in anticipation of new legislation that is similar to GDPR/HIPAA.
- **Scalability & Performance**: Azure Kubernetes Service (AKS) or containerized deployments can scale AI models and backend services based on demand.
- **Optimized AI Inference**: Using Redis caching and precomputed GradCAM visualizations reduces AI response time.
- **Observability & Compliance**: Security logging and API monitoring provide complete auditability, ensuring physicians can trust AI-generated reports.

The above architecture should make MedScan secure, scalable and efficient.

5. Inference model architecture

Broadly, the inference model is composed of the following distinct modules. Each of these work in sequence, and more than one module depends on the output of the previous module:

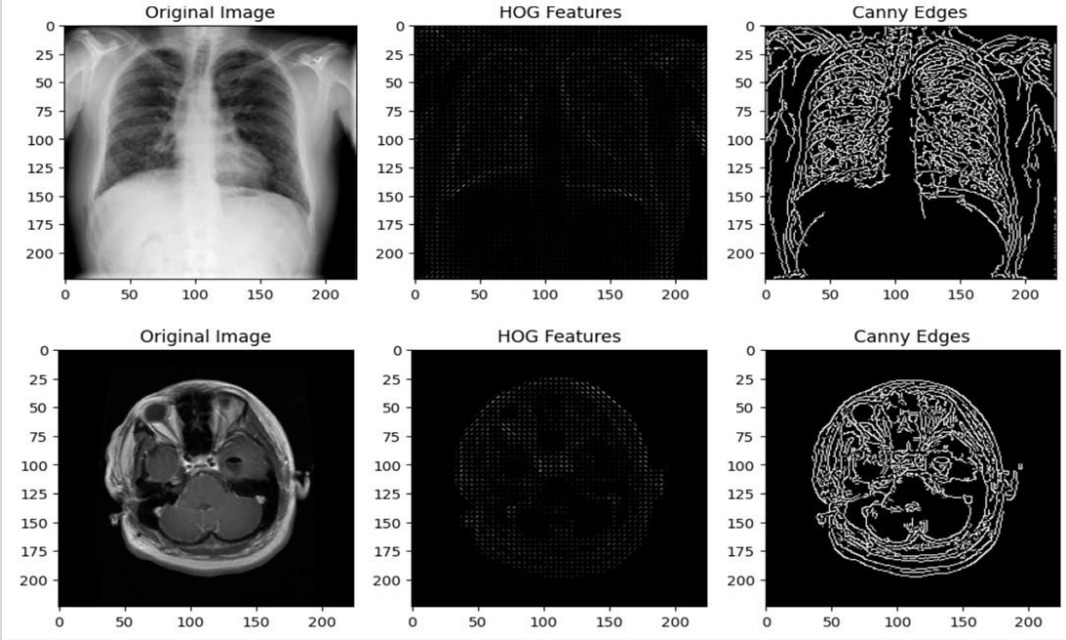
1. Check_Lung_XRay

2. Determine_Disease_Prob
3. Generate_GradCAM
4. Generate_Prescription

We will discuss each module in turn.

5.1 The Check_Lung_XRay Module

Fig IM1 Sample training data for the *Check_Lung_XRay* module



The first step in the inference process is to verify whether the provided image is indeed a thoracic cavity X-ray. To achieve this, we developed a preprocessing model that validates chest X-ray images before they are passed to a disease classification system. The model first converts medical images from DICOM to JPG, applies resizing and normalization, and extracts features using Histogram of Oriented Gradients (HOG) and Canny Edge Detection. HOG was chosen for its ability to capture essential shape and texture details, making it highly effective in distinguishing medical X-rays from non-X-ray images. Canny Edge Detection was incorporated to highlight key structural edges, helping to differentiate medical

scans from regular photographs or unrelated medical images. These extracted features were then used to train a Random Forest (RF) classifier, selected for its efficiency in handling high-dimensional feature spaces and its robustness against overfitting, making it a reliable choice for validating chest X-rays.

Since images in the final application will primarily come from hospital or diagnostic clinic databases, it is reasonable to train this validation model on a balanced dataset comprising 1,000 brain scans and 1,000 lung scans in DICOM file format (split 70%, 20% and 10% between training, validation and test datasets). This ensures that the model is well-exposed to distinguishing features across different types of medical scans. The primary goal of this validation step is to reduce false inputs in disease classification models by filtering out incorrect or irrelevant images. The combination of HOG, Canny Edge Detection, and an RF classifier provides a lightweight yet effective approach that generalizes well across diverse medical imaging scenarios. The model achieved 98.7% classification accuracy on unseen test data, ensuring robust validation. This preprocessing pipeline enhances the reliability of the application by preventing misclassification due to improper inputs, ultimately improving the accuracy of downstream disease detection models.

5.2 The Determine_Disease_Prob Module

From our literature review (LR) and exploratory data analysis (EDA), we have determined that the disease prediction module of the MedScan application should address the following considerations:

Table IM1: Key design considerations for the *Determine_Disease_Prob* module

Group	Source	Insight or design consideration	Actions taken
Data quality	EDA	Age and sex distribution is balanced, so random sampling does not affect demographic balance.	No action needed for the NIH Chest X-Ray dataset, but this may become a question if other datasets are combined to boost the number of minority samples
	EDA	Condition co-occurrence complicates classification; removing such cases may help.	**Train Test split: Preventing data leakage:** We will first segregate patients into train and test sets (80:20), to make sure that before and after scans of the same patient do not end up in both sets.

Table IM1: Key design considerations for the *Determine_Disease_Prob* module

Group	Source	Insight or design consideration	Actions taken
	EDA	Careful class balancing and selection (500-1000 samples per class) is necessary for training.	**Removing noise:** We proceed to select upto 1000 "single condition" samples for each disease class (see figures E4 and IM2). Where we fall short of "single condition" samples, we will pick relevant "2 disease" or "3 disease" samples for minority classes. This should help the classifier get trained on salient features rather than noise from multi disease samples, improving classification quality
	LR	Higher resolution images (448X448 or 1024X1024 pixels) improve detection of abnormalities.	**Input size and kernel size:** To make sure that this advantage is preserved, we only scale images down to 600x600x3, and use 3x3 kernels in convolutional layers to preserve regional detail as far as possible. This should help identify conditions like nodules, which may appear miniscule in a lung X-Ray
Feature extraction	LR	EfficientNet, DenseNet, and ResNet are commonly used, with fine-tuning improving performance.	**Choice of pretrained model:** We will compare the prediction and time performance of DenseNet201 and EfficientNetV2-S to determine an appropriate feature extractor. See figure IM3 for details.
	LR	EfficientNet offers 39.7% time savings over DenseNet201 with similar AUC performance.	
	EDA	Low class separation can impact classifier performance; robust feature extractors are needed.	**Model architecture: Will additional Conv2D layers help?** Training a classifier model to our specific use case in the most computationally efficient manner is required. Therefore, while we will keep the pretrained model as frozen/untrainable, we will add 3 convolutional layers to capture unique features of our X-Ray dataset. See figure IM4 for more details. **Choice of activation function: ReLU** ReLU activation is computationally more efficient and prevents overfitting by allowing for sparsity
	EDA	Disease categories do not form distinct clusters, affecting model separation ability.	
Model design	LR	GlobalAveragePooling, Dense layers, and regularization enhance classification layers.	**Model architecture: Faster, generalized training:** This will be added to the model after the pretrained model block, with moderate dropout layers introduced wherever appropriate. Dropout layers allow generalization before Global Avg Pooling and Dense layers, preventing the model from overreliance on certain feature maps. See figure IM4 for more details.
Robust training	LR	Combining datasets enhances performance, particularly for minority classes.	Due to budgetary and compute related restrictions, we will be training on a subset of the NIH Chest X-Ray dataset itself for the Proof of Concept; however, this will become an important consideration when training a full scale model for an actual government/diagnostic lab/hospital client at a later stage
	LR	Image augmentation techniques (flipping, brightness adjustments, etc.) improve robustness.	**Image augmentation** Lateral flips of X-Ray images, as well as variations of +/-20% on brightness, contrast and saturation should simulate real conditions. However, vertical shifts of images will not be included as machine generated images typically come with a uniform orientation and will be sourced directly from a hospital/diagnostic lab database without human intervention

Table IM1: Key design considerations for the *Determine_Disease_Prob* module

Group	Source	Insight or design consideration	Actions taken
Measurement	LR	AUC-ROC is a preferred performance metric for multi-label medical imaging classification.	**Evaluation metrics** AUC-ROC will continue as the chief evaluation metric. This is discussed in detail in the "Results and Analysis" section below

In addition to the considerations above, the following architectural choices merit more detailed discussion:

5.2.1 Creating a relatively class balanced dataset

Fig IM2 Creating a relatively class balanced training dataset for the *Determine_Disease_Prob* module

14,064

Samples chosen across 15 condition classes

Number of conditions/ sample	Number of samples	Proportion of dataset
Single condition	12,679	90.15%
2 conditions	1,210	8.60%
3 conditions	166	1.18%
4 conditions	5	0.04%
5 conditions	3	0.02%
6 conditions	1	0.01%

As discussed in Table IM1, we bifurcate the dataset into training and test patients sets along a 80:20 split. Within the training set, we choose upto 1,000 samples from each condition/disease class (without replacement) where only a single disease is present. This can be challenging for Hernia, Emphysema, Edema, Fibrosis and Pneumonia, where the total number of single disease samples is limited. In such cases, we resort to choosing samples with 2-6 conditions to represent minority diseases,

causing the counts for majority diseases like Cardiomegaly or Infiltration to cross the 1,000 thresholds in the process.

There is an alternative method to keeping sample size as close to 1,000. In this alternative methodology, we start with minority diseases and populate them with single and multiple disease samples first. With every new disease class taken up (in ascending order of single disease samples present), we end up counting the number of multiple disease instances of that disease already present in the sample, and subtracting that number from 1,000 to get the shortfall of samples to be freshly selected from the overall dataset. However, a key weakness of this methodology is that it prioritises noisy multiple disease samples over single disease samples in order to satisfy the 1,000 samples per class soft constraint. Since this compromises training data quality and therefore classifier performance, we chose to let diseases like Infiltration get more than 1,000 samples in an effort to improve class wise AUC.

Our sampling methodology leaves us with 14,064 training samples which are relatively evenly balanced across our 15 condition classes. In certain cases, like Hernia, limited overall samples can compromise performance on those specific diseases. We take this into consideration in our risk informed approach to the *Generate_Prescription* module below.

5.2.2 Choice of feature extractor: EfficientNetV2-S vs. DenseNet201

Fig IM3 Choosing the "right" pretrained model base for the *Determine_Disease_Prob* module

Reason 1: Better results

EfficientNetV2S outperformed DenseNet201 and other models on the majority of diseases (comparable performance on others)

Condition	AUC scores		Best performance
	DenseNet201	EfficientNetV2S	
Cardiomegaly	0.9219	0.9168	DenseNet
Infiltration	0.7404	0.7397	DenseNet
Effusion	0.8557	0.8551	DenseNet
Pneumothorax	0.8949	0.8940	DenseNet
Fibrosis	0.8444	0.8371	DenseNet
Edema	0.8629	0.8606	DenseNet
Hernia	0.9306	0.9376	EfficientNet
Nodule	0.8130	0.8164	EfficientNet
Emphysema	0.9351	0.9450	EfficientNet
Atelectasis	0.8002	0.8063	EfficientNet
Pleural_Thickening	0.7848	0.8010	EfficientNet
Mass	0.8473	0.8548	EfficientNet
Consolidation	0.7318	0.7394	EfficientNet
Pneumonia	0.6909	0.7291	EfficientNet

Reason 2: Faster processing and less compute intensive

EfficientNetV2S provided 39.7% time savings over DenseNet201 in similar medical image classification tasks (Krishna et al., 2022)

https://www.researchgate.net/figure/Performance-and-efficiency-comparison-of-DenseNet-and-EfficientNet_tbl4_357592591

Choosing the "right" feature extractor involves balancing performance (AUC), prediction speed and computational power. In this case, we created two identical variants of our model - one with DenseNet201 (20 million parameters) and EfficientNetV2-S (5 million parameters). We trained these models on the same training set selected in figure IM2. Each model had an Input layer that took in a 600x600 RGB Chest X-Ray, passed it through a pretrained model block (which was made "not trainable), and subsequently passed the pretrained output through a 16% dropout layer, a GlobalAveragePooling layer, another dropout layer (30%) and a final 15 node dense output layer. The only difference in architecture was the pretrained model used in each variant.

Figure IM3 shows that while EfficientNet variant narrowly outperformed DenseNet variant on the majority of disease classes, the performance difference (in AUC) is minor (ie. +/-2%). Therefore, both can be thought of as equivalent in performance without any particular class bias. The main differentiating factor was training time - EfficientNet took significantly lesser time than DenseNet (39.7% time savings) over the same dataset. This finding confirmed previous work done by Krishna et al. (2022), even though

Krishna et. al had used a more basic version of EfficientNet than V2-S. Thus, we chose to proceed with EfficientNetV2-S as the feature extractor for the eventual model.

5.2.3 OTS or Finetuned? The question of adding more Convolutional layers

Explainability measures like Gradient-weighted Class Activation Mapping (GradCAM) rely on having accessible Convolutional layers which they can use in their analysis. We realised that pretrained models like EfficientNetV2-S were composed of very complex and hierarchical block based architecture, where the last layer was deep within the final block of the pretrained model base (and therefore difficult to access).

Moreover, while making part or all of the EfficientNet model trainable does lead to better AUC outcomes, it would go beyond the budgetary and computational constraints of the present PoC project. We therefore decided to compare two types of models, in an effort to determine whether additional trainable Convolutional layers placed right after the EfficientNetV2-S block would be beneficial to classifier performance. Inspired by Wang et al. (2017) and Kufel et al. (2023), we decided to include GlobalAveragePooling and Dense layers towards the end of each model variant, as well as a certain amount of regularization, as shown in Table IM2 below:

Table IM2: OTS vs finetuned variants of the EfficientNetV2-S classifier
Architectural comparison

"OTS" variant	"Finetuned" variant
Input (600x600x3)	**Input (600x600x3)**
EfficientNetV2S functional (19x19x1280)	**EfficientNetV2S functional (19x19x1280)**
No Convolutional layers added	**Convolutional 2D (19x19x256)**
	Convolutional 2D (19x19x128)
	Convolutional 2D (19x19x64)
Dropout 16% (19x19x64)	**Dropout 16% (19x19x64)**

Table IM2: OTS vs finetuned variants of the EfficientNetV2-S classifier

Architectural comparison

Global Avg Pooling (64)	**Global Avg Pooling (64)**
Dropout 30% (64)	**Dropout 30% (64)**
Dense (15)	**Dense (15)**

We compared the performance of both model variants on the training data. Our findings confirmed what Baltruschat et al. (2019) had discovered in their comparison of OTS and finetuned models - that finetuned models outperformed OTS variants on most classes. Table IM3 shows the comparative performance:

Table IM3: Performance of OTS and finetuned EfficientNetV2S classifiers			
Condition	**AUC scores**		**Best performance**
	Finetuned	**OTS**	
Cardiomegaly	0.9564	0.9168	Finetuned
Hernia	0.8724	0.9376	OTS
Infiltration	0.7733	0.7397	Finetuned
Nodule	0.8536	0.8164	Finetuned
Emphysema	0.9427	0.9450	OTS
Effusion	0.8892	0.8551	Finetuned
Atelectasis	0.8682	0.8063	Finetuned
Pleural_Thickening	0.8720	0.8010	Finetuned
Pneumothorax	0.9232	0.8940	Finetuned
Mass	0.8982	0.8548	Finetuned
Fibrosis	0.8576	0.8371	Finetuned
Consolidation	0.8367	0.7394	Finetuned
Edema	0.9308	0.8606	Finetuned
Pneumonia	0.8394	0.7291	Finetuned

5.2.4 Other architectural choices

In addition to the considerations above, the following regularization and training related decisions helped improve model performance:

- **Binary Cross-entropy** was used as a loss function as each choice is independent (as opposed to **Categorical Cross-entropy** that is used when classes are mutually exclusive). Binary Cross-entropy was also computationally cheaper and more resilient than functions like **Focal Loss**

- **Learning rate was automatically reduced** by 10x (from 10^{-4} up to 10^{-6}) to allow for faster convergence initially but finer weight adjustment when validation loss had not changed for 2 epochs
- **Early Stopping** occurred after 5 epochs of no improvement in validation loss, and prevented overfitting
- **Model checkpointing** allowed for a model with good weights to be saved and restored, even if the model went through overfitting in subsequent epochs

Based on all of the above considerations, we arrived at the model architecture shown in figure IM4 below:

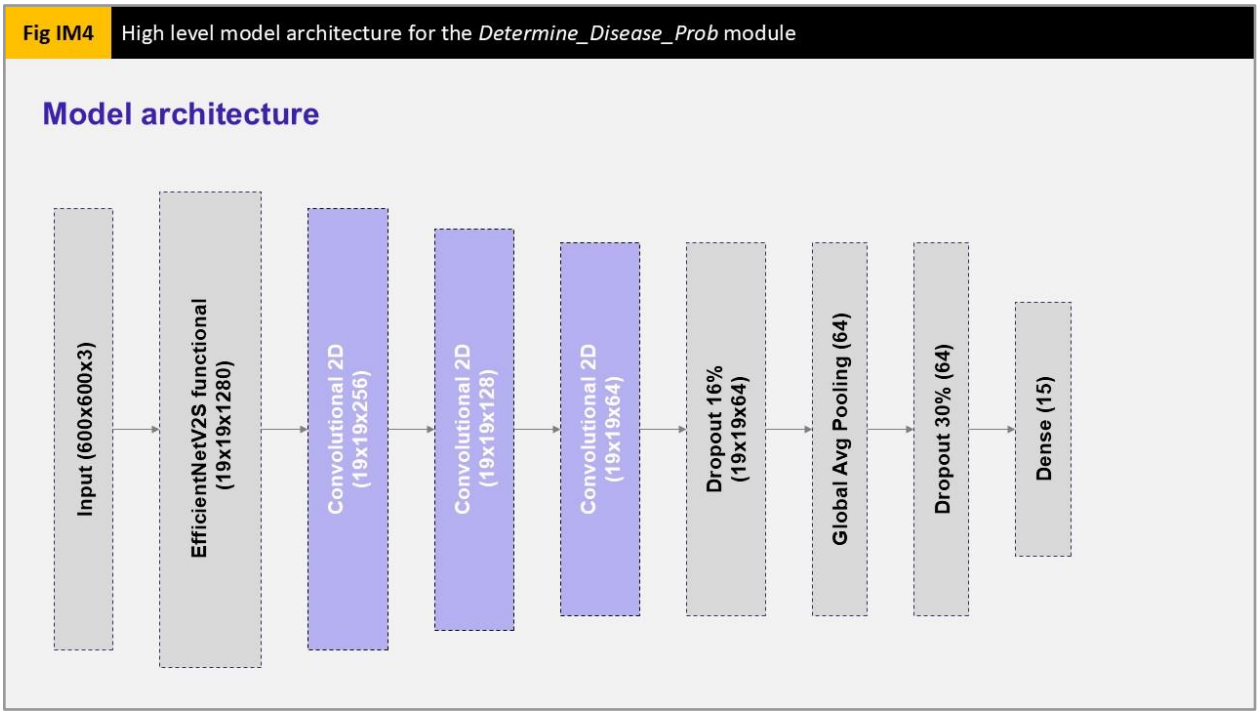
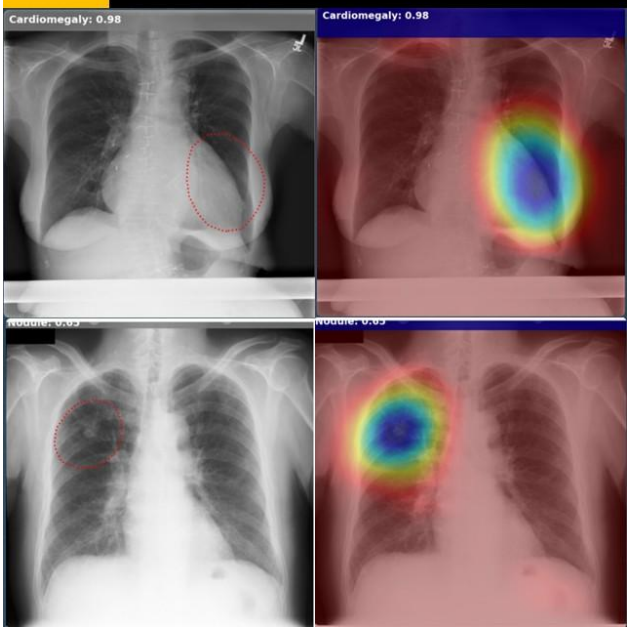


Fig IM4 High level model architecture for the *Determine_Disease_Prob* module

5.3 The Generate_GradCAM Module

We decided to investigate Gradient-weighted Class Activation Mapping (GradCAM) as an alternative to classic bounding box based disease localization in our model architecture. We found that GradCAM is advantageous for localizing diseased regions in chest X-rays because it leverages the model's learned features to highlight relevant areas without requiring explicit annotations. Wang et al. (2017), who worked on similar datasets, found that bounding box data, despite being available, produced suboptimal results due to limited data quality and a scarcity of annotated images. In contrast, GradCAM provides a class-discriminative heatmap, allowing for better visualization of diffuse diseases like pneumonia or COVID-19, where abnormalities may not conform to rigid box boundaries, and where a heatmap may provide a better aid to an overworked lab technician processing cases at 19 seconds/case. Additionally, GradCAM enables model interpretability without requiring additional labeled data, making it more practical for large-scale medical imaging applications. Figure IM5 shows sample outputs from the GradCAM implementation in our application:

Fig IM5 GradCAM contour diagrams and heatmaps for Cardiomegaly (above) and Nodules (below)



5.4 The Generate_Prescription Module

The *Generate_Prescription* module takes in (1) a dictionary of condition classes and their probabilities generated by the previous module, (2) a vector of 15 AUC scores generated during model testing, corresponding to the 15 classes in the module, and (3) a text input taken from the "symptoms" column of the *PatientRecord* table in the PostgreSQL database. The objective of this module is to generate a text prescription of up to 700 words that a lab technician or physician (depending on who is using the system) can copy and/or amend when reviewing MedScan's assessment of a given patient. Figure IM6 provides the sample prompt given to the GPT 4o-mini model.

Fig IM6 Prompt used for prescription generation

```
Patient X-ray Analysis Report
The AI model has analyzed the patient's chest X-ray and categorized the following diseases based on probability and model confidence (AUC scores):

🔹 High Confidence (Strong Model Performance & High Probability)
{high_conf}

🟡 Medium Confidence (Moderate Probability or Model Uncertainty)
{med_conf}

⚠️ Low Confidence (Low Probability or Weak Model Performance)
{low_conf}

🩺 Analysis & Considerations
- High confidence diseases (above 50% with strong AUC) indicate a strong likelihood of presence.
- Medium confidence diseases (30-50%) require additional tests for validation.
- Low confidence diseases may be model misclassifications and should be interpreted cautiously.
- Consider into account of the patient symptoms as well {patient_symptoms} and analyze highlighting this with the scores.

🔬 Next Steps & Recommendations
- 🧪 Recommended Medical Tests(Based on high-confidence diseases)
- 💊 Potential Treatment Options(If applicable)
- 🗓️ Follow-up Consultations(Suggest relevant specialists)
- ⚡ Lifestyle Modifications(If applicable)

Please provide a formal structured prescription report of around 700 words including:
1 Diagnosis Summary
2 Recommended Tests
3 Suggested Medications
4 Consultation Advice (Specialist Referrals)
5 Lifestyle & Health Recommendations

Keep the response clear, concise, and medically relevant.
```

5.5 Key considerations in designing this module

- **Real time processing:** We chose GPT 4o-mini over GPT 4, MedPaLM and GPT 4-Turbo because it is lighter and provides the requisite amount of accuracy. While models like MedPaLM are ideal for complex tasks like diagnosing rare diseases or determining drug interactions, our use case focuses on 14 well defined conditions and provides draft advice to lab technicians, who have the ultimate say on what is eventually prescribed to patients or suggested to physicians/specialists

tending to patients. We also need to account for the fact that lab technicians take approximately 34.2 seconds/case to reach a diagnosis manually (Babyn & Adams, 2025), and that long lead times in generating diagnoses can hamper the value proposition of having a faster and AI driven system. Given these realities, GPT 4o-mini reaches a reasonable balance between processing time, accuracy and usage costs.

- **Risk qualified opinions:** Our prompt is designed to not only provide disease predictions and probabilities, but to also offer an opinion on how valid those predictions may be (basis the *Determine_Disease_Prob* model's test AUC scores). Accordingly, the GPT 4o model uses appropriate language to address disease predictions basis the following confidence and probability tiers for each disease:
    - Confidence tiers: Assessing quality of classification and ranking:
        - High confidence: The model predicts probabilities for a given disease class with an AUC ≥ 0.85
        - Medium confidence: This applies when disease predictions are within the range $0.85 > \text{AUC} \geq 0.75$
        - Low confidence: Disease classes where the AUC falls below 0.75 fall in this category
    - Probability tiers: Determining the likelihood of a sample exhibiting signs of a certain disease:
        - High probability of a certain disease: This applies to diseases where the *Determine_Disease_Prob* model predicts a probability ≥ 0.50 of the disease being present
        - Medium probability of a certain disease: Cases where the *Determine_Disease_Prob* model predicts a probability of the disease being present in the range $0.50 > \text{P(Disease class)} \geq 0.30$ fall in this category.

- Low probability of a certain disease: Diseases where the model generated probability falls below 0.30 are considered “low probability cases.”

The outputs from all four modules above are displayed by the Streamlit UI layer in the form of a patient case that a lab technician can review. The GradCAM image contour diagrams direct the physician/lab technician’s attention to areas of the scan that the model prioritized while generating disease probabilities for a certain class/condition.

6. Results and Analysis

The inference model detailed before was composed of four sections - *Check_Lung_XRay, Determine_Disease_Prob, Generate_GradCAM,* and *Generate_Prescription*. In our analysis, we will focus on the sections pertaining to *Check_Lung_XRay* and *Determine_Disease_Prob*, as both can be objectively and reliably measured and evaluated (as opposed to localization maps or prescriptions which may be more subjective and may vary from physician to physician).

Table RA1 below shows that the *Check_Lung_XRay* module achieved high performance in a balanced manner - the accuracy and F1 scores are both approximately 98.7%, suggesting that the model did not show skewed performance on either the “Lung XRay” or “Non Lung XRay” classes. Therefore, it seems like a reliable component to include into the final application architecture.

Table RA1: Performance metrics for the *Check_Lung_XRay* module

Accuracy	0.9871	**Precision**	1.0000
Recall	0.9748	**F1-score**	0.9872

The remainder of our analysis will focus on the *Determine_Disease_Prob* module’s performance. At this time, it is important to consider our choice of performance metrics carefully.

Choice of measurement metric: Why AUC is more robust than other metrics

We opted for AUC (Area Under the Receiver Operating Characteristic Curve) as our sole performance metric for chest X-ray classification because it provides a more comprehensive evaluation of the model's performance across different threshold settings. Unlike precision and recall, which depend on a specific decision threshold, AUC considers the entire range of classification thresholds, making it threshold-independent and suitable for imbalanced datasets. The NIH Chest X-ray dataset remains somewhat imbalanced despite our best efforts at balanced sampling, and AUC helps measure the model's ability to distinguish between classes regardless of class distribution. Moreover, AUC captures the trade-off between sensitivity (true positive rate) and specificity (false positive rate), which is crucial in medical imaging where both false positives and false negatives have significant implications. While F1-score balances precision and recall, it does not account for how well the model differentiates between normal and abnormal cases at different thresholds, making AUC a more robust choice for evaluating our solution. The fact that most of the papers reviewed used AUC as their primary metric of performance further validates this choice.

Actual AUC performance of the *Determine_Disease_Prob* module

Fig RA1 ROC-AUC curve for the *Determine_Disease_Prob* module

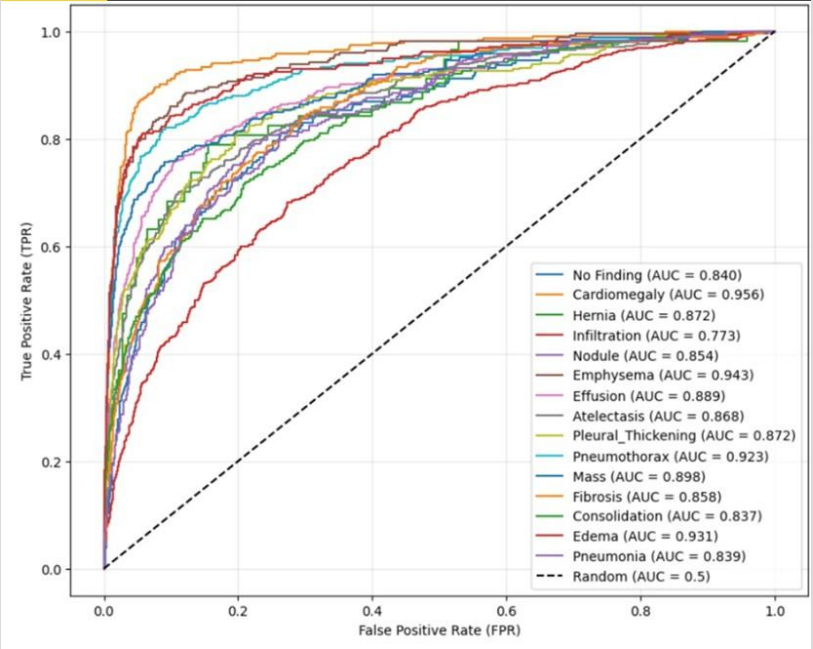


Figure RA1 presents the ROC curves for multiple disease classes in a chest X-ray classification task, with the AUC values indicating the model's ability to distinguish between positive and negative cases for each condition. The model performs best for Cardiomegaly (AUC = 0.956), Emphysema (AUC = 0.943), and Pneumothorax (AUC = 0.923), suggesting strong discriminatory power for these diseases. However, it struggles more with Infiltration (AUC = 0.773), Pneumonia (AUC = 0.839), and Consolidation (AUC = 0.837), indicating that these conditions may be harder to classify due to overlapping features or data limitations.

The performance variation across diseases highlights potential challenges in dataset quality, class imbalance, and the need for improved feature extraction. While the model outperforms a random classifier (AUC = 0.5) for all classes, conditions with lower AUC values may require additional refinement through better data augmentation, class-specific training adjustments, or ensemble learning. The overall results suggest that the model has strong clinical potential for some diseases but needs improvement for others before deployment in a real-world setting.

Performance relative to past research: MedScan vs others:

Fig RA2 ROC-AUC performance across past research

Note: The below compilation of results is adapted from Kufel et al. (2023)

Pathology Label	Ğuendel et al.	Yan et al.	Baltruschat et al.	Kufel et al.	Medscan	Best performance
Official split	Yes	Yes	Yes	No	No	
Atelectasis	0.767	0.792	0.763	0.817	0.868	Medscan
Cardiomegaly	0.883	0.881	0.875	0.911	0.956	Medscan
Effusion	0.828	0.842	0.822	0.879	0.889	Medscan
Infiltration	0.709	0.710	0.694	0.716	0.773	Medscan
Mass	0.821	0.847	0.820	0.853	0.898	Medscan
Nodule	0.758	0.811	0.747	0.771	0.854	Medscan
Pneumonia	0.731	0.740	0.714	0.769	0.839	Medscan
Pneumothorax	0.846	0.876	0.840	0.898	0.923	Medscan
Consolidation	0.745	0.760	0.749	0.815	0.837	Medscan
Edema	0.835	0.848	0.846	0.908	0.931	Medscan
Emphysema	0.895	0.942	0.895	0.935	0.943	Medscan
Fibrosis	0.818	0.833	0.816	0.824	0.858	Medscan
Pleural_Thickening	0.761	0.808	0.763	0.812	0.872	Medscan
Hernia	0.896	0.934	0.937	0.890	0.872	Baltruschat et al.
Average	0.807	0.830	0.727	0.843	0.880	Medscan

Results collated by Kufel et al. (2023) in Figure RA2 show that teams that balanced their own datasets across classes performed marginally better than those that pursued class weighted training strategies. Medscan's average AUC of 0.880 is marginally higher than the next best team, Kufel et al. Since it is close to the other AUC values in the range, it does not suggest that our training or testing practices were significantly different than other teams' training/testing strategies (ie. the chances of our results being anomalous is relatively low). MedScan appears to outperform other teams on all classes except Hernia prediction, where Baltruschat et al (2019) appear to have a superior model. However, given that Hernia is a minority class with 227 overall samples (since it is a medical rarity), this result needs to be interpreted with caution. In general, while the classifier seems to perform well on minority classes, it would be advisable to test it on other datasets (preferably with patients with a different racial or comorbidity makeup). Such a test would provide a more definite idea of model performance, especially when faced with diverse images from varying types of XRay machines in rural India.

7. Conclusion and Further Research

The MedScan AI system marks a significant advancement in applying artificial intelligence to diagnose lung diseases from X-ray images. With an optimized framework featuring EfficientNetV2-S, GradCAM for interpretability, and an AI-powered prescription generation module, the system effectively supports radiologists and lab technicians. Our findings indicate that fine-tuning select layers within the EfficientNet block enhances classification accuracy, particularly for conditions with low natural separation. While the model achieves a strong AUC of 0.880 across disease classes, further refinements are needed before it can be deployed on a large clinical scale.

Despite these achievements, several areas for improvement remain. One key avenue is deepening the training process by making more layers within the EfficientNet block trainable. This adjustment could enhance feature extraction for conditions that currently exhibit lower classification accuracy, such as

Infiltration and Pneumonia. Additionally, incorporating more convolutional layers may refine feature representations and increase the model's sensitivity to subtle disease indicators.

Exploring alternative architectures is another promising direction, particularly integrating recurrent neural networks such as Long Short-Term Memory (LSTM) models, which could capture temporal dependencies in sequential medical imaging. Hybrid models that combine convolutional image embeddings with Histogram of Oriented Gradients (HOG) embeddings may further strengthen classification performance. Another viable approach involves a hierarchical multi-layer classifier that initially categorizes diseases into broader families (e.g., lung cancer, tuberculosis, pneumonia) before refining the classification within each family. The application of self-attention mechanisms could also improve predictions by directing focus toward the most relevant regions in an image.

The potential of ensemble learning remains underutilized. Implementing a soft-voting ensemble of multiple classifiers could improve prediction reliability and reduce individual model biases. Additionally, addressing data limitations through dataset expansion is critical. The current dataset includes 14,064 samples, but sourcing additional images for underrepresented conditions, such as Hernia and Pneumonia, could enhance classification performance. Generating synthetic images using techniques like generative adversarial networks (GANs) offers another approach to addressing class imbalances.

Furthermore, rigorous testing on diverse datasets from different geographic regions is essential to evaluate the model's generalizability. Since the NIH Chest X-ray dataset is primarily North American, assessing model performance on datasets from rural India and other regions with varying imaging conditions will provide a clearer understanding of its applicability in real-world clinical settings. This approach ensures that MedScan AI evolves into a reliable and scalable diagnostic tool for lung diseases across diverse healthcare environments.

In summary, while this study establishes a solid proof of concept, future research should prioritize deeper model training, alternative architectures, data augmentation, and extensive validation to enhance

accuracy and adaptability in global healthcare. If successfully implemented in rural India, this product could be expanded to developing regions in Africa, Asia, and Latin America, where similar resource constraints hinder access to timely and adequate medical care.

Beyond machine learning, developing a fully functional diagnostic aid requires significant investment in usability testing and stakeholder engagement. Conducting design-thinking interviews with lab technicians, general practitioners, and specialists could provide valuable insights to refine the system's user experience. Engaging with government agencies, public health officials, and national regulatory bodies such as the Central Drugs Standard Control Organisation (CDSCO) is crucial to ensuring that the product meets legal and ethical standards. Additionally, pilot studies, alongside educational and awareness campaigns for lab technicians and doctors, will be vital for successful adoption and integration into existing healthcare workflows.

Bibliography

Abrahams, Jacob M., Beth Creekmur, Janet Shin Lee, In-Lu Amy Liu, Mayra Macias, and Michael K. Gould. 2024. “Neighborhood Level Socioeconomic Disadvantage and Adherence to Guidelines for the Evaluation of Patients with Incidentally Detected Pulmonary Nodules.” *CHEST*, December. https://doi.org/10.1016/j.chest.2024.12.011.

Alturbag, Majed. 2024. “Effectiveness of Personalised Phone Calls and Short Message Service Reminders in Improving Patient Attendance at a Radiology Department.” *Cureus*, September. https://doi.org/10.7759/cureus.69568.

Amin, Hina, and Waqas J Siddiqui. 2022. “Cardiomegaly.” StatPearls [Internet]. U.S. National Library of Medicine. November 20, 2022. https://www.ncbi.nlm.nih.gov/books/NBK542296/.

Babyn, Paul S., and Scott J. Adams. 2025. “Harnessing the Power of Generative AI to Enhance Radiologist Efficiency and Accuracy.” *Radiology* 314 (3). https://doi.org/10.1148/radiol.250339.

Baltruschat, Ivo M., Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. 2019. “Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification.” *Scientific Reports* 9 (1). https://doi.org/10.1038/s41598-019-42294-8.

Banerjee, Abhijit, Arun Chandrasekhar, Suresh Dalpath, Esther Duflo, John Floretta, Matthew Jackson, Harini Kannan, et al. 2021. “Selecting the Most Effective Nudge: Evidence from a Large-Scale Experiment on Immunization.” *Preprint-ARXIV*, April. https://doi.org/10.3386/w28726.

Chen, Haomin, Shun Miao, Daguang Xu, Gregory D Hager, and Adam P Harrison. 2019. “Deep Hierarchical Multi-Label Classification of Chest X-Ray Images.” *International Conference on Medical Imaging with Deep Learning* 102: 109–20. https://proceedings.mlr.press/v102/chen19a.html.

Choubey, Ashwini Kumar. 2020. “LOK SABHA UNSTARRED QUESTION NO. 2243: CHRONIC OBSTRUCTIVE PULMONARY DISEASE.” Digital Sansad. Digital Sansad. September 23, 2020. https://sansad.in/getFile/loksabhaquestions/annex/174/AU2243.pdf?source=pqals.

Denning, David W., Donald C. Cole, and Animesh Ray. 2023. “New Estimation of the Prevalence of Chronic Pulmonary Aspergillosis (CPA) Related to Pulmonary TB – a Revised Burden for India.” *IJID Regions* 6 (March): 7–14. https://doi.org/10.1016/j.ijregi.2022.11.005.

Dhooria, Sahajal, Inderpaul Singh Sehgal, Ritesh Agarwal, Valliappan Muthu, Kuruswamy Thurai Prasad, Soundappan Kathirvel, Mandeep Garg, Amanjit Bal, Ashutosh Nath Aggarwal, and Digambar Behera. 2022. “Incidence, Prevalence, and National Burden of Interstitial Lung Diseases in India: Estimates from Two Studies of 3089 Subjects.” *PLOS ONE* 17 (7). https://doi.org/10.1371/journal.pone.0271665.

Eshwara, Vandana Kalwaje, Chiranjay Mukhopadhyay, and Jordi Rello. 2020. “Community-Acquired Bacterial Pneumonia in Adults.” *Indian Journal of Medical Research* 151 (4): 287–302. https://doi.org/10.4103/ijmr.ijmr_1678_19.

Holste, Gregory, Yiliang Zhou, Song Wang, Ajay Jaiswal, Mingquan Lin, Sherry Zhuge, Yuzhe Yang, et al. 2024. “Towards Long-Tailed, Multi-Label Disease Classification from Chest X-Ray: Overview

of the CXR-Lt Challenge." *Medical Image Analysis* 97 (October). https://doi.org/10.1016/j.media.2024.103224.

Huang, Guan-Hua, Qi-Jia Fu, Ming-Zhang Gu, Nan-Han Lu, Kuo-Ying Liu, and Tai-Been Chen. 2022. "Deep Transfer Learning for the MULTILABEL Classification of Chest X-Ray Images." *Diagnostics* 12 (6): 1457. https://doi.org/10.3390/diagnostics12061457.

IPCRG. 2022. "India." International Primary Care Respiratory Group (IPCRG). International Primary Care Respiratory Group (IPCRG). October 28, 2022. https://www.ipcrg.org/india.

Khan, Hina, and Alison Lindahl. 2023. "When to Worry about Lung Nodules." Brown University Health. Brown University. November 7, 2023. https://www.brownhealth.org/be-well/when-worry-about-lung-nodules.

Krishna, Shruthi, Suganthi S S, Shivsubramani Krishnamoorthy, and Arnav Bhavsar. 2022. "Stain Normalized Breast Histopathology Image Recognition Using Convolutional Neural Networks for Cancer Detection." *Preprint - ResearchGate*, January. https://www.researchgate.net/publication/357592591_Stain_Normalized_Breast_Histopathology_Image_Recognition_using_Convolutional_Neural_Networks_for_Cancer_Detection.

Kufel, Jakub, Michał Bielówka, Marcin Rojek, Adam Mitręga, Piotr Lewandowski, Maciej Cebula, Dariusz Krawczyk, et al. 2023. "Multi-Label Classification of Chest X-Ray Abnormalities Using Transfer Learning Techniques." *Journal of Personalized Medicine* 13 (10): 1426. https://doi.org/10.3390/jpm13101426.

McNulty, Rose. 2023. "Annual Lung Cancer Screening Adherence Poor after Initial Screening, Study Finds." AJMC. AJMC. March 3, 2023. https://www.ajmc.com/view/annual-lung-cancer-screening-adherence-poor-after-initial-screening-study-finds.

News18. 2023. "India Has Alarming Ratio of One Radiologist per Lakh People." Medical Buyer. Medical Buyer. October 17, 2023. https://medicalbuyer.co.in/india-has-alarming-ratio-of-one-radiologist-per-lakh-people/.

RadAI. n.d. "Rad Ai: Save Time and Decrease Burnout with Radiology AI Software." Rad AI | Save Time and Decrease Burnout with Radiology AI Software. Rad AI. Accessed March 12, 2025. https://www.radai.com/.

Ramnath, Nithya, Prasanth Ganesan, Prasanth Penumadu, Douglas Arenberg, and Alex Bryant. 2025. "Lung Cancer Screening in India: Preparing for the Future Using Smart Tools & Biomarkers to Identify Highest Risk Individuals." *The Indian Journal of Medical Research* 160 (January): 561–69. https://doi.org/10.25259/ijmr_118_24.

Rawat, Anjali, Anindya Saha, Abhishek Kumar, and Rakesh Sharma. 2023. "RURAL HEALTH STATISTICS 2021-22." Ministry of Health and Family Welfare. Government of India. January 12, 2023. https://mohfw.gov.in/sites/default/files/RHS%202021-22_2.pdf.

Reddy, Surya Pavan, S Srinivas, and CRPS Krishna. 2021. *International Journal of Medical Research Professionals* 7 (6): 20–24. https://doi.org/10.21276/ijmrp.

Salvi, Sundeep, G Anil Kumar, R S Dhaliwal, Katherine Paulson, Anurag Agrawal, Parvaiz A Koul, P A Mahesh, et al. 2018. "The Burden of Chronic Respiratory Diseases and Their Heterogeneity across

the States of India: The Global Burden of Disease Study 1990–2016." *The Lancet Global Health* 6 (12). https://doi.org/10.1016/s2214-109x(18)30409-1.

Sandelowsky, Hanna, Christer Janson, Fredrik Wiklund, Gunilla Telg, Sofie de Fine Licht, and Björn Ställberg. 2022. "Lack of COPD-Related Follow-up Visits and Pharmacological Treatment in Swedish Primary and Secondary Care." *International Journal of Chronic Obstructive Pulmonary Disease* Volume 17 (August): 1769–80. https://doi.org/10.2147/copd.s372266.

Shah, Harsh. 2023. "Burden of TB in India." Burden of TB in India | Knowledge Base. National TB Elimination Program - NTEP. February 17, 2023. https://ntep.in/node/352/CP-burden-tb-india.

Singh, Sheetu, Mohan Bairwa, Bridget F. Collins, Bharat Bhushan Sharma, Jyotsana M Joshi, Deepak Talwar, Nishtha Singh, et al. 2021. "Survival Predictors of Interstitial Lung Disease in India." *Lung India* 38 (1): 5–11. https://doi.org/10.4103/lungindia.lungindia_414_20.

Smithuis, Robin. n.d. "Chest X-Ray - Lung Disease." The Radiology Assistant : Chest X-Ray - Lung Disease. The Radiology Assistant. Accessed March 12, 2025. https://radiologyassistant.nl/chest/chest-x-ray/lung-disease.

T Puchalski, Jonathan. 2014. "Mortality of Hospitalized Patients with Pleural Effusions." *Journal of Pulmonary and Respiratory Medicine* 04 (03). https://doi.org/10.4172/2161-105x.1000184.

Telukuntla, Kartik S., Chetan P. Huded, Mingyuan Shao, Tim Sobol, Mouin Abdallah, Kathleen Kravitz, Michael Hulseman, et al. 2021. "Impact of an Electronic Medical Record-Based Appointment Order on Outpatient Cardiology Follow-up after Hospital Discharge." *Npj Digital Medicine* 4 (1). https://doi.org/10.1038/s41746-021-00443-2.

Vikhe, Vikram B, Ahsan A Faruqi, Rahul Patil, Avani Reddy, and Devansh Khandol. 2024. "A Systematic Review of Community-Acquired Pneumonia in Indian Adults." *Cureus* 16 (7). https://doi.org/10.7759/cureus.63976.

Wang, Xiaosong, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. "Chestx-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases." *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July, 3462–71. https://doi.org/10.1109/cvpr.2017.369.