Gemini Proposal:

A Comprehensive Blueprint for an AI-Powered Lung Nodule Analysis System for Low-Resource Health Ecosystems

Part I: Foundational Strategy - Curating a World-Class Data Corpus

The success of any artificial intelligence system in medicine is predicated on the quality, diversity, and strategic curation of its training data. For a multifaceted system designed for lung nodule segmentation, classification, and dual-format report generation, a simplistic approach to data collection is insufficient. It is imperative to construct a comprehensive data corpus that not only fuels each distinct module but also addresses the inherent challenges of clinical variability, linguistic nuance, and regional specificity. This section outlines a meticulous strategy for acquiring and integrating the necessary datasets to build a robust, generalizable, and clinically valid system.

Section 1.1: Core Datasets for Nodule Segmentation and Classification

The primary computer vision tasks of identifying (detection), delineating (segmentation), and characterizing (classification) lung nodules require large-scale, expertly annotated imaging data. The following datasets form the cornerstone of the vision model's training and validation.

- **LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative):** This dataset is the foundational resource for this project. Comprising 1,010 thoracic Computed Tomography (CT) scans, its principal strength lies in the detailed XML annotations provided by four experienced radiologists for each identified nodule.[1] This multi-reader annotation scheme is not a source of noise but a critical feature. It captures the real-world phenomenon of inter-observer variability in radiology, allowing for the development of models that are robust to differing interpretations of nodule boundaries and characteristics.[3] The dataset's scale and diversity are essential for training a segmentation model that can generalize across a wide range of patient anatomies and nodule morphologies.[3]
- **LUNA16 (Lung Nodule Analysis 2016):** As a well-defined subset of LIDC-IDRI, LUNA16 provides a standardized benchmark for the nodule detection task.[1] It consists of 888 CT scans with a unified set of 1,186 annotated nodules, making it an ideal platform for validating the initial detection model's performance against published,

state-of-the-art results.[3] Utilizing LUNA16 for benchmarking ensures that the model's foundational capabilities are quantitatively comparable to established research.

- **Lung Nodule Dataset with Histopathology-based Cancer Type Annotation:** While LIDC-IDRI is excellent for segmentation, it lacks definitive histopathological ground truth for cancer subtyping. This smaller but clinically vital dataset bridges that gap, providing annotations for 330 nodules across 95 patients, categorized as benign, adenocarcinoma, or squamous cell carcinoma.[1] This dataset is indispensable for training the classification head of the vision model, enabling it to provide more specific, clinically actionable diagnoses beyond a simple benign/malignant dichotomy.
- **Cross Spatio-Temporal Pathology-based Lung Nodule Dataset:** This modern dataset introduces a crucial dimension for elevating the project's novelty and clinical utility: time. Containing 317 CT sequences from 109 patients, it provides longitudinal data that captures nodule progression over multiple scans.[1] Training on this data enables the development of models capable of assessing nodule growth rate, a key indicator of malignancy. This feature is particularly relevant for a scalable telemedicine platform, where a patient's follow-up scans can be compared against their baseline, offering a dynamic risk assessment rather than a static one.

Section 1.2: The Crucial Link - Datasets for LLM-Powered Report Generation

The most innovative component of this system is its ability to generate both structured clinical reports and patient-friendly summaries. This requires a different type of data: CT scans paired with their corresponding free-text radiology reports. Sourcing high-quality datasets of this nature is a common bottleneck in medical AI, and a successful strategy here will be a major differentiator for the project.

- **LNDb (Lung Nodule Database):** As identified in the initial project plan, the LNDb dataset, particularly its later versions (e.g., v4), is of paramount importance. It contains 294 CT scans with associated medical report annotations.[1] This direct link between imaging findings and the language used by radiologists to describe them is the essential fuel for fine-tuning a Large Language Model (LLM) to generate contextually appropriate and clinically accurate reports.
- **RAD-ChestCT Dataset:** This is a large-scale, contemporary dataset that provides an invaluable corpus of clinical language. While not exclusively focused on lung nodules, its initial release contains 3,630 chest CT scans (out of a total of over 35,000) paired with structured labels extracted from free-text reports using the Sentence Analysis for Radiology Label Extraction (SARLE) framework.[5] This dataset can be used to pre-train or fine-tune the LLM on the general structure, style, and terminology of chest CT

reporting. The publicly available SARLE code itself is a valuable asset that can be adapted for pre-processing other report datasets.[5]

- **Duke Lung Cancer Screening (DLCS) Dataset:** This recently compiled dataset is another high-value resource, containing 2,061 low-dose CT scans with 3,187 semi-automatically annotated nodules.[6] Its most significant feature for the LLM module is the inclusion of Lung Imaging Reporting and Data System (Lung-RADS) scores, which were extracted directly from the original radiologist reports. Lung-RADS is a standardized system for classifying malignancy risk. Training on this data will enable the LLM to generate reports that include a standardized risk assessment, a critical component for clinical decision-making and ensuring consistency.

The development of a dual-output reporting system necessitates a sophisticated data strategy that recognizes the distinct requirements of each output. The source datasets provide examples of clinical reports but contain virtually no instances of patient-friendly summaries. This gap reveals the need for a dedicated NLP pre-processing and data synthesis pipeline. Before the LLM is fine-tuned, the free-text reports from LNDb, RAD-ChestCT, and DLCS must be parsed using NLP techniques to extract key information into a structured format (e.g., nodule location, size, morphology, calcification, Lung-RADS score). This structured data then serves as the input for the LLM. To train the patient-friendly generation capability, a synthetic dataset must be created. This can be achieved by using a powerful baseline LLM (e.g., GPT-4) or, ideally, by collaborating with medical professionals to translate a subset of these structured summaries into simple, clear language in target regional languages. This "structured-to-dual-report" fine-tuning task, which involves training the model to generate two distinct outputs from a single structured input, represents a significant and publishable area of novelty.

Section 1.3: Addressing the Regional Data Gap and Ensuring Generalizability

A critical challenge identified in the project's planning phase is the absence of large, publicly available, annotated lung cancer CT datasets specifically from the Indian population.[1] This is a significant concern, as lung cancer presentation, risk factors (e.g., prevalence of bidi smoking, environmental pollution), and genetic markers can vary between populations, potentially limiting the generalizability of models trained on Western datasets.[7] A multi-pronged strategy is required to mitigate this risk.

- **Strategy 1: Domain Adaptation and Transfer Learning:** The primary training of the vision models will be conducted on the large international datasets (LIDC-IDRI, DLCS). This establishes a robust foundational model. The project should actively seek collaborations with Indian medical institutions like AIIMS or Tata Memorial Centre, which are conducting research in this area and may have access to local data.[9] If a smaller, local dataset can be secured, domain adaptation techniques should be employed to

fine-tune the pre-trained model. This approach leverages the knowledge from the large dataset while adapting the model to the specific features of the local population, a far more effective strategy than training from scratch on a small dataset.

- **Strategy 2: Augmenting with Epidemiological Data:** The LLM module offers a unique opportunity to incorporate non-imaging, region-specific data. A patient-specific risk score can be calculated based on demographic and lifestyle information provided at the Health ATM (e.g., age, smoking habits, geographic location). This risk model can be informed by India-specific epidemiological data from sources such as the Indian Council of Medical Research (ICMR) and the National Cancer Registry Programme, which highlight regional variations in lung cancer incidence.[7] This calculated risk score can then be passed to the LLM as part of its input context, allowing it to generate more personalized and contextually relevant risk statements in its reports.
- **Strategy 3: Architecting for Future Data Collection:** The system deployed in Health ATMs must be designed from the outset with a secure, ethical, and consent-based mechanism for data collection. This will enable the creation of a valuable, localized dataset over time, which can be used for periodic retraining and continuous improvement of the AI models. This approach aligns with World Health Organization (WHO) recommendations for developing sustainable AI solutions in low-resource settings.[14]

The following table provides a strategic overview of the curated data corpus, mapping each dataset to its specific role within the project.

Table 1: Comprehensive Dataset Analysis and Strategic Utility

Dataset Name	URL/Reference	Scans/Nodules	Key Annotations	Primary Use	Strategic Notes
LIDC-IDRI	TCIA [1]	1,010 Scans / 244,527 Images	Segmentation masks (XML), Nodule characteristics	Primary Segmentation Training	Cornerstone dataset. Multiple radiologist annotations enable modeling of inter-rater variability.

LUNA16	grand-challenge.org [1]	888 Scans / 1,186 Nodules	Nodule locations (CSV)	Detection Model Benchmarking	Standardized subset of LIDC-IDRI for performance comparison against published results.
Histopathology Dataset	Zenodo [1]	95 Scans / 330 Nodules	Bounding boxes, Cancer types (Adenocarcinoma, etc.)	Multi-Class Classification Training	Provides essential ground truth for classifying nodule subtypes, enabling more specific diagnosis.
Cross Spatio-Temporal	Zenodo [1]	317 Sequences / 2,295 Nodules	Pathological info, Spatio-temporal annotations	Longitudinal Analysis Model	Key for novelty. Enables development of models that track nodule progression over time.

LNDb (v4)	Zenodo [1]	294 Scans	Nodule segmentations, Medical report annotations	Primary LLM Report Generation	Critical resource providing direct linkage between CT images and free-text radiology reports.
RAD-ChestCT	Zenodo [5]	3,630 Scans (initial release)	Report-derived labels (abnormality * location)	LLM Pre-training/ Fine-tuning	Large corpus of clinical language to train the LLM on general chest CT reporting style and structure.
DLCS Dataset	Zenodo [6]	2,061 Scans / 3,187 Nodules	Bounding boxes, Lung-RADS scores from reports	LLM Fine-tuning for Risk Stratification	Contemporary dataset. Lung-RADS scores are crucial for training the LLM to generate standardized reports.

Part II: The Vision Pipeline - Lightweight Segmentation and Classification

The core of the system's diagnostic capability lies in its vision pipeline. The central engineering challenge is to develop a model that achieves high levels of accuracy and precision while adhering to the stringent computational and memory constraints of low-power edge devices like the Raspberry Pi 5 or NVIDIA Jetson Nano. This necessitates a move away from large,

state-of-the-art models towards highly efficient, lightweight architectures specifically designed for such environments.

Section 2.1: State-of-the-Art Lightweight Architectures for Medical Segmentation

A thorough review of recent academic literature reveals a clear trend towards developing efficient yet powerful models for medical image analysis on edge devices.

- **The U-Net Foundation:** The U-Net architecture remains the conceptual benchmark for biomedical image segmentation. Its encoder-decoder structure with skip connections is highly effective at preserving high-resolution spatial information while capturing deep semantic features, a combination crucial for accurately delineating complex structures like lung nodules.[15] However, the standard U-Net, with its large number of parameters, is computationally prohibitive for the target hardware.[17]
- **MobileNet-UNet and Efficient Convolutions:** The project plan correctly identifies MobileNet-UNet as a strong starting point. This architecture replaces the standard convolutional layers in the U-Net's encoder with the depthwise separable convolutions pioneered by MobileNet.[18] This modification dramatically reduces the number of parameters and floating-point operations (FLOPs) with only a minor trade-off in accuracy, making it a well-established choice for edge deployment.[18] Numerous open-source PyTorch implementations of this architecture are available, which can significantly accelerate the initial development phase.[18]
- **The Next Generation of Lightweight Models (2023-2025):** The field is rapidly evolving, and several newer architectures offer even greater efficiency. A world-class product should evaluate these cutting-edge options:
    - **LDMRes-Net:** A recently proposed network specifically tailored for IoT and edge platforms, featuring a remarkably low parameter count of just 0.072 million while using dual multiscale residual blocks to enhance feature extraction.[21]
    - **Lightweight Residual U-Net:** This architecture, proposed in early 2025, integrates a Convolutional Block Attention Module (CBAM) and an Atrous Spatial Pyramid Pooling (ASPP) block into a residual U-Net structure. It achieves state-of-the-art performance on lung segmentation tasks with only 3.24 million parameters, making it an extremely strong candidate.[16]
    - **Hybrid CNN-ViT Models (e.g., LM-Net):** Emerging research explores lightweight hybrid models that combine the local feature extraction strengths of CNNs with the global context modeling of Vision Transformers (ViTs), designed to enhance segmentation accuracy without a heavy computational footprint.[22]

- **Ghost Module-Based Networks:** Architectures that incorporate the "Ghost Module" have demonstrated the ability to maintain high accuracy while significantly reducing parameter counts and computational load. This approach, proven effective in models like YOLOv4-GNet for nodule detection, could be adapted for the segmentation backbone.[23]

Section 2.2: Recommended Architecture and Implementation Strategy

Based on the analysis of current research and the project's specific requirements, a **Hybrid 2.5D Mobile-Residual U-Net** is the recommended architecture. This design makes a deliberate engineering trade-off to maximize performance within the given constraints.

- **Architectural Rationale (The 2.5D Approach):** A full 3D CNN, while theoretically optimal for analyzing volumetric CT data, imposes a computational and memory burden that is simply infeasible for the target edge devices. Conversely, a pure 2D approach, which processes each CT slice independently, is highly efficient but discards valuable contextual information from adjacent slices, potentially leading to errors in nodule segmentation and classification. The 2.5D approach offers an elegant compromise. The model processes a small stack of adjacent CT slices (e.g., 3, 5, or 7) as input channels to a 2D CNN. For example, a 5-slice input would be treated as a 5-channel 2D image. This provides the network with crucial local volumetric context—how a nodule appears across a few millimeters of tissue—without incurring the exponential cost increase of full 3D convolutions.
- **Implementation Plan:**
  1. **Encoder Backbone:** The encoder part of the U-Net will be constructed using a pre-trained, lightweight backbone. The **Lightweight Residual U-Net** architecture[16] is the primary recommendation due to its proven performance and efficiency in lung segmentation. Alternatively, a **MobileNetV3** backbone offers a robust, well-supported option.[24] Using a pre-trained backbone (on a large dataset like ImageNet) allows the model to leverage learned low-level features, accelerating convergence and improving performance.
  2. **Decoder Path:** The decoder will consist of standard U-Net up-sampling blocks (e.g., using transposed convolutions) and will be connected to the corresponding encoder levels via skip connections to restore high-resolution spatial details.
  3. **Multi-Task Learning Heads:** To maximize efficiency, the model will be trained in a multi-task fashion. The final feature map from the decoder will feed into two distinct output heads:
     - **Segmentation Head:** A final 1×1 convolution followed by a sigmoid activation function to produce a pixel-wise probability map, indicating the likelihood that each pixel belongs to a nodule.
     - **Classification Head:** The decoder's output feature map will be passed through a global average pooling layer to create a single feature vector,

which is then fed into a small multi-layer perceptron (MLP) with a softmax output layer. This head will perform the multi-class classification task (e.g., Normal, Benign, Adenocarcinoma, Squamous Cell Carcinoma). This shared-representation approach is highly parameter-efficient.

Section 2.3: Building Clinical Trust - The Explainability Engine

For any AI system to be adopted in a clinical setting, it must be transparent. Clinicians need to understand *why* a model arrives at a particular conclusion. An explainable AI (XAI) module is therefore not an optional add-on but a core requirement for building trust and ensuring safe use.

- **Chosen Technique: Grad-CAM:** Gradient-weighted Class Activation Mapping (Grad-CAM) is the ideal XAI technique for this application. It produces a coarse localization map—a heatmap—that highlights the most important regions in the input image for a specific classification decision.[25] For example, if the model classifies a nodule as "Adenocarcinoma," Grad-CAM will generate a heatmap showing which pixels in the nodule region contributed most strongly to that decision. This provides an intuitive, visual form of evidence for the clinician.
- **Implementation and Integration:** The implementation should leverage the comprehensive pytorch-grad-cam library.[27] This well-maintained package supports a wide variety of models and offers numerous CAM-based methods beyond the original Grad-CAM, allowing for future experimentation. In the system's workflow, once the classification head makes a prediction, the Grad-CAM algorithm will be executed. The resulting heatmap will be overlaid onto the original CT slice and displayed in the user interface directly alongside the segmentation mask and the generated report. This tight integration of prediction, segmentation, and explanation provides the remote doctor with a complete and immediately verifiable diagnostic picture, fostering confidence in the AI's output.

The following table provides a systematic comparison of potential lightweight architectures to justify the final selection based on the project's unique constraints.

Table 2: Comparative Analysis of Lightweight Segmentation Architectures

Model Architecture	Key Innovation	Reported Performance (Dice/IoU)	Parameters (M)	Suitability for Edge Deployment

MobileNetV2-UNet	Depthwise Separable Convolutions	Good (Varies by task)	~9.5 [19]	High
U-Net++	Nested, Dense Skip Connections	85.4% DSC (LUNA16) [17]	~47.2 [17]	Low
LDMRes-Net	Dual Multiscale Residual Blocks	High (Ophthalmology task) [21]	**0.072** [21]	Very High
YOLOv4-GNet (Backbone)	Improved Ghost Module, Attention	81.8% CPM (LUNA16) [23]	~11.4 [23]	High
Lightweight Residual U-Net	Residual Blocks + CBAM + ASPP	**99.1%** DSC (MC Dataset) [16]	**3.24** [16]	Very High

This analysis clearly indicates that newer architectures like the **Lightweight Residual U-Net** offer a superior balance of high performance and extremely low parameter count, making them the optimal choice for this project.

Part III: The Language Pipeline - LLM-Powered Clinical and Patient Reporting

This component represents the most significant opportunity for innovation within the project. A system that can automatically generate both a clinically precise report for doctors and a clear, empathetic summary for patients addresses two of the most pressing challenges in modern healthcare: the administrative burden on clinicians and the pervasive issue of low health literacy among patients.[28]

Section 3.1: The LLM Landscape for Medical Applications

The selection of the base Large Language Model (LLM) is a critical decision that balances performance, cost, privacy, and deployability.

- **Proprietary API-based Models:** Models such as OpenAI's GPT-4 or Anthropic's Claude 3.5 offer state-of-the-art performance.[30] However, they are unsuitable for this project's context for several reasons. First, they rely on sending data to external servers, which introduces significant patient data privacy and security risks, a major concern in healthcare.[28] Second, they operate on a pay-per-use model, which conflicts with the goal of creating a low-cost, scalable public health solution. Finally, network latency in rural areas could make API calls unreliable.
- **Open-Source Models:** This is the ideal category for the project. Models like Meta's Llama 3, Mistral's series, and Google's Gemma are freely available and can be deployed on private infrastructure.[32] Models in the 7 to 8 billion parameter range represent a sweet spot, offering powerful language capabilities while being manageable for fine-tuning and deployment on a local server. The LLM module will not run directly on the resource-constrained Raspberry Pi; instead, it will be hosted on a slightly more powerful local server within the Health ATM's local network or on a secure, government-managed cloud instance accessible via a private API. This architecture maintains data privacy and control.

Section 3.2: Fine-Tuning Strategy for the Dual-Output System

The core task is to adapt a general-purpose open-source LLM to the highly specific domain of lung nodule reporting. This requires a specialized, parameter-efficient fine-tuning strategy.

- **Input Formatting:** The LLM will not process the CT images directly. Instead, it will receive a structured text input (e.g., a JSON object) that programmatically summarizes the findings from the vision pipeline. This approach grounds the LLM's output in the quantitative results from the vision model, fundamentally constraining its ability to hallucinate and ensuring the report is based on visual evidence.
    1. **Example Input:** {"nodule_count": 1, "nodules":}
- **Parameter-Efficient Fine-Tuning (PEFT):** Fully fine-tuning a 7B parameter model requires immense computational resources. PEFT techniques, such as **LoRA (Low-Rank Adaptation)** and its memory-efficient variant **QLoRA (Quantized LoRA)**, are the solution.[32] These methods freeze the vast majority of the pre-trained model's weights and introduce a small number of new, trainable parameters in the form of low-rank matrices. This allows the model to be adapted to the new task with a fraction of the memory and compute, making the process feasible on modest hardware. Numerous open-source projects on GitHub provide practical guides and code for implementing LoRA and QLoRA with popular LLMs.[34]

- **Instruction-Based Fine-Tuning Dataset:** The key to teaching the LLM its dual-report task is the creation of a high-quality instruction dataset. Each entry in this dataset will consist of three parts:
    1. **Instruction:** A clear prompt defining the task. This can be varied to improve model robustness. Example: "From the following lung nodule analysis data, generate a structured clinical radiology report and a simplified summary for the patient in Hindi."
    2. **Input:** The structured JSON object from the vision pipeline.
    3. **Output:** The desired, ideal output, containing both report formats separated by special tokens. This structure teaches the model to generate both components in a single forward pass.
        - Example Output:
          FINDINGS: A single 8 x 7 mm solid, non-calcified nodule is identified in the anterior segment of the right upper lobe. **IMPRESSION:** Suspicious pulmonary nodule, corresponding to Lung-RADS category 4A. The nodule is stable in size compared to the prior examination dated 15-Jan-2024. **RECOMMENDATION:** Follow-up low-dose CT in 3 months is recommended to assess for interval change.नमस्ते, आपके फेफड़ों की सीटी स्कैन जांच में दाहिने फेफड़े के ऊपरी हिस्से में एक छोटी गांठ (8 x 7 मिमी) मिली है। यह गांठ पिछली जांच से बदली नहीं है, जो एक अच्छा संकेत है। हालांकि, पूरी तरह से निश्चित होने के लिए, डॉक्टर आपको 3 महीने बाद एक और सीटी स्कैन कराने की सलाह देंगे ताकि यह देखा जा सके कि इसमें कोई बदलाव तो नहीं हो रहा है। कृपया अपने डॉक्टर से आगे की जानकारी के लिए संपर्क करें।

Section 3.3: Ensuring Clinical Fidelity and Mitigating LLM Risks

The deployment of an LLM in a clinical context carries significant responsibility. A rigorous, multi-faceted validation framework is non-negotiable to ensure the outputs are accurate, reliable, and, above all, safe.

- **The Challenge of Hallucination and Inaccuracy:** LLMs are probabilistic models that can generate plausible-sounding but factually incorrect statements, a phenomenon known as hallucination.[35] In a medical report, even a minor factual error (e.g., wrong location, incorrect size) could have severe consequences. The system's design, which relies on structured input, inherently mitigates this risk by grounding the LLM. However, explicit validation is still essential.
- **Validation Framework for Clinical Reports:**
    1. **Quantitative Metrics:** Automated NLP metrics such as BLEU, ROUGE, and BERTScore can be used to compare the generated clinical reports against the ground-truth reports from the test sets of the LNDb and DLCS datasets. It is crucial to recognize, however, that recent studies have shown these metrics often correlate poorly with clinical correctness, as they prioritize stylistic similarity over factual accuracy.[35]

2. **Qualitative Expert Review:** A panel of certified radiologists must review a statistically significant sample of the generated clinical reports. They will score each report on a Likert scale across several dimensions:
    - **Factual Accuracy:** Does the report correctly state all facts from the structured input (location, size, type, etc.)?
    - **Clinical Completeness:** Are any critical details omitted?
    - **Clarity and Conciseness:** Is the report written in standard radiological language and easy for another clinician to understand?
    - **Absence of Hallucinations:** Does the report contain any information not supported by the input data?

- **Validation Framework for Patient-Facing Summaries:** This is a key area of novelty and requires a unique validation approach. The summaries must be evaluated for both clinical safety and patient comprehension.
    1. **Clinical Safety Review:** The same panel of radiologists must review the patient summaries to ensure that the simplification process has not introduced any dangerous ambiguities or omitted critical information, such as the need for urgent follow-up. The tone must be reassuring but not dismissive of potential risks.
    2. **Patient Comprehension Study:** A user study must be conducted with participants who are representative of the target rural population, considering varying levels of literacy. Participants will be given the AI-generated summaries, and their understanding will be assessed using:
        - **Standard Readability Scores:** Metrics like the Flesch-Kincaid Grade Level can provide an objective measure of text complexity.[29]
        - **Comprehension Questionnaires:** After reading the summary, participants will answer a short questionnaire to test their understanding of key points: What was found? Where was it found? What is the recommended next step? Do they need to see a doctor?.[28] This validation step is essential for ensuring the system empowers patients rather than confusing or alarming them, and it will form a cornerstone of the project's academic contribution.

Part IV: From Lab to Field - Edge Deployment and Optimization

The successful transition of the AI models from a high-performance development environment to a low-cost, power-constrained Health ATM requires a meticulous and hardware-aware optimization strategy. This phase is about ensuring that the system can deliver accurate results in near real-time under the demanding conditions of a rural deployment.

Section 4.1: Selecting the Optimal Edge Platform

The choice of hardware is a fundamental decision that will influence the entire optimization workflow and the final performance of the system.

- **Comparative Analysis of Candidate Devices:**
    - **Raspberry Pi 5:** This platform's primary advantages are its extremely low cost and vast community support. Its powerful quad-core ARM CPU makes it well-suited for general-purpose computing and tasks that are CPU-bound.[36] However, its GPU capabilities are minimal, making it a challenging environment for deploying complex deep learning models without aggressive, CPU-specific optimization.[36]
    - **NVIDIA Jetson Nano:** While more expensive and power-hungry than the Raspberry Pi, the Jetson Nano is purpose-built for edge AI applications. Its key advantage is the integrated 128-core NVIDIA Maxwell GPU, which is designed to accelerate the parallel computations inherent in neural networks.[37] Furthermore, it is supported by NVIDIA's mature and highly optimized software stack, including CUDA and TensorRT, which simplifies the process of achieving high-performance inference.[38]
- **Recommendation:** The **NVIDIA Jetson Nano is the technically superior and recommended platform** for this project. The presence of a dedicated GPU will provide a significant performance advantage for the vision pipeline, reducing inference latency and enabling more complex models if necessary. While the Raspberry Pi 5 remains a viable low-cost alternative, achieving real-time performance would require a more intensive and challenging CPU-specific optimization effort using frameworks like NCNN. Initial development and benchmarking should ideally be conducted on both platforms to provide a clear cost-performance trade-off analysis for stakeholders.

Section 4.2: The End-to-End Optimization Workflow

The trained PyTorch models are not directly deployable on edge devices. They must be converted and optimized through a multi-stage pipeline to create a lean, fast, and efficient inference engine.

- **The Optimization Pipeline:**
    1. **Step 1: Export to ONNX:** The first and most crucial step is to export the trained PyTorch models (both the segmentation/classification model and the Grad-CAM components) to the Open Neural Network Exchange (ONNX) format.[39] ONNX serves as a universal, intermediary representation that decouples the model from its training framework, enabling its use with various inference engines.[36]
    2. **Step 2: Model Pruning and Quantization:** Before final compilation, the model's size and computational complexity must be reduced.

        - **Pruning:** This technique involves identifying and removing redundant or unimportant weights and connections from the neural network. This can lead to a smaller model footprint with minimal impact on accuracy.[41]
        - **Quantization:** This is one of the most effective optimization techniques. It involves converting the model's weights and activations from high-precision 32-bit floating-point (FP32) numbers to low-precision 8-bit integers (INT8).[42] This conversion reduces the model size by a factor of four and can dramatically speed up inference, as integer arithmetic is much faster on most processors, including those on edge devices.
    3. **Step 3: Compilation to a Target Engine:** The optimized ONNX model is then compiled into a final, hardware-specific engine.
        - **For the NVIDIA Jetson Nano:** The **NVIDIA TensorRT** framework is used. TensorRT takes the ONNX model and performs a series of aggressive optimizations, including layer and tensor fusion, kernel auto-tuning, and precision calibration, to generate an engine that is maximally optimized for the underlying NVIDIA GPU architecture.[39]
        - **For the Raspberry Pi 5:** A CPU-centric inference engine is required. **NCNN**, developed by Tencent, is an excellent choice as it is highly optimized for ARM CPUs and has demonstrated superior performance over other frameworks in Raspberry Pi benchmarks.[43]

A critical aspect often overlooked in academic projects is that the optimization process should not be an afterthought. A "hardware-in-the-loop" development methodology should be adopted. This means that early in the architectural design phase, individual candidate layers and small network blocks should be benchmarked for their latency and power consumption on the actual target hardware (Jetson Nano with TensorRT and Raspberry Pi 5 with NCNN). An architectural choice that seems optimal on a development GPU might be highly inefficient on the edge if it relies heavily on operations that are not well-supported by the target's optimization framework. This proactive, hardware-aware design process de-risks the deployment phase and is a hallmark of sophisticated engineering, adding significant novelty to the project's methodology.

Section 4.3: Designing an Intuitive User Interface (UI)

The system will be operated by healthcare technicians in rural Health ATMs, who may have varying levels of technical expertise. Therefore, the user interface must be exceptionally simple, robust, and intuitive.

- **Design Principles and Workflow:**
    - **Simplicity and Accessibility:** The UI should feature large, touch-friendly buttons, clear iconography, and a guided, step-by-step workflow.
    - **Multi-lingual Support:** The interface must be available in English and key regional Indian languages to ensure usability across different states.

- **Core Workflow:**
    1. **Patient Identification:** The operator starts by inputting the patient's ID, ideally by scanning their Ayushman Bharat Health Account (ABHA) QR code.
    2. **Data Ingestion:** The operator loads the patient's chest CT scan from a connected device or local storage.
    3. **One-Click Analysis:** A single, prominent "Analyze Scan" button initiates the entire AI pipeline.
    4. **Results Display:** The results are presented in a clean, tabbed layout:
        - **Scan Viewer:** An interactive viewer displaying the CT slices, with options to overlay the AI-generated segmentation mask and the Grad-CAM explainability heatmap.
        - **Clinical Report:** The structured, technical report generated by the LLM for the remote doctor.
        - **Patient Summary:** The simplified, multi-lingual summary for the patient.
- **Technology Stack:** To ensure the UI is lightweight and responsive on kiosk hardware, a simple web-based stack is recommended. A backend service built with a minimal framework like **Flask** or **FastAPI** will run on the edge device to handle requests and trigger the AI models.[1] The frontend can be built with standard **HTML, CSS, and JavaScript**, ensuring it runs smoothly in any modern browser without requiring heavy dependencies.

The following table provides a data-driven comparison of the two potential hardware platforms to aid in the final deployment decision.

Table 3: Edge Deployment Platform and Optimization Strategy Comparison

Feature	Raspberry Pi 5	NVIDIA Jetson Nano
Key Hardware	Quad-core ARM Cortex-A76 CPU	Quad-core ARM Cortex-A57 CPU, 128-core Maxwell GPU
AI Performance	CPU-dependent, lower	GPU-accelerated, significantly higher (472 GFLOPs)[37]

Cost (Approx. USD)	~$80	~$150
Power Consumption	Low (~5W idle, ~12W peak)	Moderate (5-10W) [37]
Primary Opt. Framework	NCNN [43]	NVIDIA TensorRT [39]
Benchmark Inference Time	Slower (CPU-bound)	Faster (GPU-accelerated)
Pros	Very low cost, large ecosystem, low power draw.	Purpose-built for AI, superior performance, mature software stack.
Cons	Limited AI acceleration, requires more intensive optimization.	Higher cost, slightly higher power consumption.

Part V: The Indian Healthcare Ecosystem - Integration and Scalability

A technically proficient AI system can only achieve its public health goals if it is seamlessly integrated into the existing national healthcare infrastructure. For India, this means aligning with the Ayushman Bharat Digital Mission (ABDM), understanding the domestic landscape of health-tech innovators, and designing a strategy for scalable, phased deployment.

Section 5.1: Navigating the Ayushman Bharat Digital Mission (ABDM)

Compliance with ABDM is not merely a technical requirement; it is the key to unlocking the system's potential for nationwide impact. ABDM aims to create a "seamless online platform" through "open, interoperable, standards-based digital systems".[45]

- **Technical Integration Strategy:**
    - **Core ABDM Components:** The system must interface with three fundamental ABDM registries: the **Ayushman Bharat Health Account (ABHA)** for unique

patient identification, the **Health Facility Registry (HFR)** to identify the Health ATM, and the **Healthcare Professionals Registry (HPR)** to identify the remote consulting doctor.[46]

- **Interoperability Standards:** ABDM mandates the use of international standards to ensure data can be exchanged between different systems. The primary standard for clinical data is **FHIR (Fast Healthcare Interoperability Resources)**.[48] The final reports generated by the system must be packaged into FHIR-compliant bundles before being transmitted. The National Resource Centre for EHR standards (NRCeS) has been established to promote the adoption of these standards.[49]
- **Onboarding via the ABDM Sandbox:** The official pathway for integration is through the **ABDM Sandbox**.[50] This is a testing environment where developers can register their applications, integrate with ABDM APIs, and undergo functional and security testing (such as Web Application Security Audits) before being approved for live deployment.[49]
- **API-Driven Workflow:** All interactions with the ABDM ecosystem are managed via the ABDM Gateway APIs. The system's workflow must be designed to handle these API calls:
    1. **Patient Authentication:** At the Health ATM, the patient's ABHA is verified, typically by scanning a QR code on their ABHA card or mobile app.[53]
    2. **Consent Management:** The system must use the ABDM framework to generate a digital consent request, which the patient approves on their mobile device. This consent artifact authorizes the system to link the newly generated lung screening report to their ABHA. This consent-based sharing is a foundational principle of ABDM.[45]
    3. **Health Record Linking:** Once consent is granted, the system pushes the FHIR-compliant report to the patient's longitudinal health record, making it accessible to them and any other healthcare provider they authorize in the future.

- To simplify this complex integration, the project can leverage SDKs and API suites from ABDM-integrated partners like Eka.care, which provide a higher-level abstraction over the core ABDM APIs.[54]

Section 5.2: Competitive and Collaborative Landscape in India

While the project is a public health initiative, it is crucial to understand the landscape of commercial entities operating in the AI-based lung cancer screening space in India.

- **Key Players:**
    1. **Qure.ai:** A global leader in medical imaging AI with a strong presence in India. Their qXR solution for chest X-rays has been deployed at a massive scale for screening tuberculosis and lung abnormalities, often in partnership with governments and international organizations like AstraZeneca.[56] Their focus is broad, covering multiple diseases and modalities.
    2. **NURA.ai:** This company operates high-end, AI-powered health screening centers in major Indian cities. They use advanced imaging technology from Fujifilm and target the urban, private healthcare market with a focus on a premium patient experience and comprehensive screening packages that include low-dose CT for lung cancer.[59]
    3. **Eka.care:** A health-tech startup focused on building tools for the ABDM ecosystem. Their primary products include an EMR for doctors and an AI-powered scribe that transcribes doctor-patient conversations into clinical notes, demonstrating a strong focus on clinical workflow automation.[54]
- **The Project's Unique Value Proposition:** This project does not directly compete with these players but rather carves out a unique and vital niche. Its key differentiators are:
    1. **Public Health Mission:** The primary goal is to improve healthcare access and equity in underserved rural areas, not commercial profit.
    2. **Rural and Edge-First Design:** The entire system is architected from the ground up for the specific technical and operational constraints of rural Health ATMs.
    3. **Open Framework Potential:** As a government-backed academic project, there is potential for the models and source code to be made open, fostering a wider ecosystem of innovation.
    4. **Native ABDM Integration:** The system is designed not just to be compliant with ABDM but to leverage its full potential for creating a connected, longitudinal health record for rural citizens.

Section 5.3: A Phased Rollout and Scalability Strategy

A successful nationwide deployment requires a careful, evidence-based, phased approach.

- **Phase 1: Pilot Deployment and Technical Validation:** The system should first be deployed in a small, controlled set of 5-10 Health ATMs in a single district. The objectives of this phase are to test the end-to-end technical stability, network reliability, workflow integration with on-ground operators, and the tele-consultation loop with remote doctors.
- **Phase 2: Expanded Clinical and User Validation:** Based on the learnings from the pilot, the system will be refined and deployed to a larger cohort of 50-100 Health ATMs. This phase will be dedicated to conducting the rigorous clinical and patient validation studies outlined in Part III. This includes gathering data on diagnostic accuracy compared to standard-of-care and assessing patient comprehension of the simplified reports.

- **Phase 3: State-Level and National Rollout:** Upon successful validation, the project can present a comprehensive report of its technical efficacy and clinical impact to state and national health authorities. The system's modular design and deep integration with ABDM's standardized infrastructure are the key enablers for a scalable rollout across the entire Health ATM network.

The integration with ABDM unlocks a capability that transforms the project's clinical impact. A single screening is a static snapshot. True clinical value in nodule management arises from tracking changes over time. By leveraging the patient's ABHA, the system can retrieve previous scans and reports when a patient returns for a follow-up, regardless of which Health ATM they visit. The AI model, trained on the spatio-temporal dataset, can then perform a direct comparison, quantifying changes in nodule size and density. This elevates the system from a simple screening tool to a powerful **longitudinal monitoring platform**, a paradigm shift in its public health utility and a profoundly novel research contribution that is uniquely enabled by India's digital health infrastructure.

Part VI: Achieving Novelty - A Roadmap to Journal Publication

The ultimate goal of this project extends beyond building a functional product; it is to make a significant, publishable contribution to the scientific community. This requires a clear strategy for addressing existing research gaps and rigorously validating the system's innovative components. The project is well-positioned to produce a high-impact paper that can influence both the technical and public health domains.

Section 6.1: Identifying the Core Research Gaps

A critical analysis of the current literature reveals several key areas where this project can make a novel contribution:

1. **Validated End-to-End Systems for Low-Resource Edge Environments:** While many papers describe lightweight models, there is a scarcity of research that presents a complete, end-to-end medical AI system (from data ingestion to XAI-assisted reporting) that has been specifically designed, optimized, and benchmarked for deployment on ultra-low-cost, resource-constrained edge hardware like that found in rural Health ATMs.[61]
2. **Patient-Centric AI-Generated Reporting:** The use of LLMs to generate clinical reports is an active area of research.[30] However, the concept of a dual-output system that generates both a clinical report and a simplified, multi-lingual patient summary is highly innovative. Crucially, there is a significant gap in the literature regarding the validation methodologies required to ensure such patient-facing reports are both clinically safe and genuinely comprehensible to a lay audience with varying levels of health literacy.[28]

3. **Integration with National Digital Health Infrastructure:** Most academic AI projects are developed in a silo, using isolated datasets. There is very little research demonstrating how a medical AI system can be deeply integrated with, and leverage the capabilities of, a national-scale digital health platform like ABDM to enable advanced clinical functions such as longitudinal patient monitoring across a distributed network.[45]

Section 6.2: Proposed Avenues for High-Impact Contribution

The project should be framed around a central thesis that directly addresses these gaps: **"A validated, end-to-end, hardware-aware AI framework for democratizing lung cancer screening in low-resource settings through explainable edge computing, patient-centric LLM-powered reporting, and integration with a national digital health ecosystem."** The manuscript's results should be structured around three distinct but interconnected validation studies, each representing a core contribution.

- **Contribution 1: The Dual-Report LLM - A Validated Framework for Patient-Centered AI Communication.**
    - **Novelty:** This is the first system of its kind to generate and, more importantly, rigorously validate a dual-output report for lung nodule analysis.
    - **Methodology to Highlight:** The paper will detail the "structured-to-dual-report" fine-tuning methodology using PEFT. The results will feature a two-part validation: (1) The expert radiologist review assessing the clinical fidelity of the structured reports, and (2) The user study with a representative rural population assessing the comprehension and clarity of the patient-friendly summaries.
    - **Impact:** This contribution addresses the globally relevant challenges of clinician burnout and patient health literacy, providing a novel, validated technical solution.
- **Contribution 2: The Holistic Edge AI System - A Comprehensive Benchmark for Real-World Deployment.**
    - **Novelty:** This work will provide one of the first comprehensive performance benchmarks for a complete medical AI pipeline (preprocessing, 2.5D segmentation/classification, Grad-CAM explanation) on widely accessible, low-cost edge devices.
    - **Methodology to Highlight:** The paper will present detailed performance metrics, including diagnostic accuracy (Dice, IoU, Precision, Recall), system latency (milliseconds per scan), power consumption (Watts), and memory footprint. It will showcase the performance gains achieved through the full PyTorch -> ONNX -> TensorRT/NCNN optimization workflow and discuss the findings from the "hardware-in-the-loop" design process.
    - **Impact:** This provides a practical, reproducible blueprint and sets realistic performance expectations for researchers and public health organizations seeking to deploy advanced AI in resource-constrained environments, a topic of immense interest to the global digital health community.[14]

- **Contribution 3 (Advanced/Future Work): Longitudinal Nodule Analysis Leveraging National Health Infrastructure.**
    - **Novelty:** This demonstrates a paradigm shift from static screening to dynamic, AI-assisted monitoring, enabled by the integration of an edge AI system with a national health data backbone.
    - **Methodology to Highlight:** Using the Cross Spatio-Temporal dataset as a proxy, the paper will demonstrate the model's ability to quantify nodule growth between two time points. It will then show how the LLM can integrate this temporal information into its report (e.g., "The nodule has grown by 2mm (25% in volume) since the prior examination 6 months ago, increasing its risk profile.").
    - **Impact:** This showcases a powerful new model for public health screening programs, where AI can provide continuous, data-driven risk assessment within a connected healthcare ecosystem.

Section 6.3: Structuring the Manuscript and Target Journals

- **Manuscript Structure:** The paper should follow the standard IMRaD (Introduction, Methods, Results, and Discussion) format. The "Methods" section will be particularly detailed, meticulously describing the data curation strategy, the vision model architecture, the LLM fine-tuning process, the edge optimization pipeline, and the multi-part validation protocols. The "Results" section will be clearly organized around the findings of the three primary contributions.
- **Target Journals:** Given the project's interdisciplinary nature and high potential for impact, the team should aim for top-tier journals. Suitable venues include:
    - **Specialized Medical AI and Imaging Journals:** *IEEE Transactions on Medical Imaging*, *Medical Image Analysis*, *Radiology: Artificial Intelligence*. These journals would be interested in the technical novelty of the lightweight architecture and the LLM reporting framework.
    - **High-Impact Digital and Global Health Journals:** *The Lancet Digital Health*, *Nature Digital Medicine*, *JAMIA (Journal of the American Medical Informatics Association)*. These journals would be interested in the project's broader implications for healthcare access, patient engagement, and the practical deployment of AI in low-resource settings.

By pursuing this structured approach, the project can transcend the scope of a typical undergraduate endeavor and produce a world-class system with a corresponding high-impact scientific publication.

Of course. Here is a comprehensive analysis of the existing body of work, detailing the contributions and limitations of key papers and outlining how your project can build upon them to achieve its novel objectives.

Analysis of Prior Art and Strategic Opportunities

This report provides a detailed review of approximately 40 relevant academic papers, theses, and technical articles to situate your project within the current landscape of medical AI. For each piece of work, we analyze its contributions, identify its limitations, and articulate how your project is uniquely positioned to overcome these gaps.

Part 1: Lightweight Architectures for Vision Tasks (Segmentation & Classification)

The core of your system's diagnostic power lies in its ability to run accurately on resource-constrained hardware. This requires moving beyond standard, computationally heavy models.

1. Paper: "Lung Segmentation with Lightweight Convolutional Attention Residual U-Net" (2025) [1]

- **What Was Done:** This paper proposes a highly efficient "Lightweight Residual U-Net" that integrates a Convolutional Block Attention Module (CBAM) and Atrous Spatial Pyramid Pooling (ASPP). It achieves a state-of-the-art Dice score of 99.08% on the MC dataset with only 3.24 million parameters.
- **Limitations:** The model was developed for lung *segmentation* (identifying the lung field), not the more granular task of lung *nodule* segmentation and classification. While highly efficient, its direct applicability to nodule detection is not proven.
- **How Our Project Overcomes This:** We can adopt this highly efficient architecture as the foundational encoder-decoder for our vision pipeline. By adding a multi-task head for both nodule segmentation and classification, we leverage their lightweight design but adapt it to our specific, more complex clinical task. This represents a novel application of their architecture.

2. Paper: "An automated end-to-end deep learning-based framework for the early detection and classification of lung nodules, specifically for low-resource settings" (2023) [2]

- **What Was Done:** This study proposes a complete, three-stage framework using a 3D Res-U-Net for lung segmentation, YOLOv5 for nodule detection, and a Vision Transformer (ViT) for classification. It specifically targets low-resource settings and achieves a high lung segmentation Dice score (98.82%).

- **Limitations:** The framework is sequential, running three separate, relatively heavy models (a 3D CNN, YOLO, and a ViT). This multi-model pipeline is computationally inefficient and likely too slow for real-time inference on a Health ATM's edge device.
- **How Our Project Overcomes This:** Our project's core novelty is a single, multi-task lightweight model. Instead of three separate models, our unified architecture will perform segmentation and classification simultaneously from a shared set of features. This is a far more efficient and practical approach for edge deployment.

3. Paper: "Half-UNet: A Simplified U-Net for Medical Image Segmentation" (2022) [3]

- **What Was Done:** The authors challenge the necessity of the complex decoder in U-Net architectures. They propose "Half-UNet," which drastically simplifies the decoder path and incorporates efficient Ghost modules. They demonstrate similar accuracy to standard U-Net but with 98.6% fewer parameters and 81.8% fewer FLOPs.
- **Limitations:** The paper's primary contribution is architectural simplification, but it doesn't fully explore optimization for specific edge hardware. The performance is benchmarked in terms of parameters and FLOPs, not real-world latency on a device like a Raspberry Pi or Jetson Nano.
- **How Our Project Overcomes This:** We can incorporate the core principles of Half-UNet—a simplified decoder and Ghost modules—into our design. However, we will take the crucial next step of performing hardware-in-the-loop optimization, benchmarking the model's actual inference speed and power consumption on the target hardware using frameworks like TensorRT and NCNN. This moves from theoretical efficiency to proven real-world performance.

4. Paper: "A lightweight neural network for lung nodule detection based on improved ghost module" (2023) [4]

- **What Was Done:** This research presents YOLOv4-GNet, a lightweight model for lung nodule *detection*. It improves upon GhostNet by incorporating a spatial-temporal attention mechanism into the backbone, significantly reducing parameters (from ~41M to ~11.4M) and improving detection accuracy (mAP from 34.5 to 54.56).
- **Limitations:** The model is designed for detection (drawing a bounding box) and not the more precise task of pixel-level segmentation. Furthermore, it does not perform classification of the nodule type (e.g., benign, adenocarcinoma).
- **How Our Project Overcomes This:** We can integrate their improved Ghost Module (G-Bneck) with its attention mechanism into the encoder of our U-Net-based architecture. This allows us to benefit from their efficiency gains while applying it to the more advanced tasks of segmentation and multi-class classification, creating a more comprehensive diagnostic tool.

5. Paper: "CPLOYO: A Lightweight and High-Precision Detection Method for Small Pulmonary Nodules" (2025) [5]

- **What Was Done:** This paper introduces CPLOYO, a YOLOv8-based model specifically designed for detecting very small lung nodules. It uses a lightweight RepViT module combined with a Contextual Attention with Multi-scale Feature Fusion (CAMF) to improve small object detection.
- **Limitations:** This is purely a detection model. It is highly specialized for finding small nodules but does not provide segmentation masks or classify their malignancy, which are essential for a full clinical workflow.
- **How Our Project Overcomes This:** The C2f_RepViTCAMF module is a powerful innovation for feature extraction. We can adapt this module and integrate it into the deeper layers of our segmentation model's encoder. This would enhance our model's ability to accurately segment and classify even the smallest, most difficult-to-detect nodules, a key requirement for early-stage cancer screening.

6. Paper: "U-Net and its variants for medical image segmentation: theory and applications" (Review) [6]

- **What Was Done:** This paper provides a comprehensive theoretical overview of the U-Net architecture and its many variants, including 3D U-Net, U-Net++, and Adversarial U-Nets. It explains why the core encoder-decoder structure with skip connections is so effective for medical imaging.
- **Limitations:** As a theoretical review, it does not provide a practical implementation or benchmark for low-resource settings. It discusses the architectures conceptually without considering the computational trade-offs for edge deployment.
- **How Our Project Overcomes This:** This paper validates our foundational architectural choice (U-Net). Our novelty comes from taking this powerful theoretical base and systematically adapting it for the real world by creating a lightweight, multi-task variant and proving its efficacy on constrained hardware.

7. Paper: "AWEU-Net: A Novel Attention-Aware Weight Excitation U-Net" (2021) [7]

- **What Was Done:** This work proposes AWEU-Net, a two-stage model that first uses a Faster R-CNN for nodule detection and then a U-Net with novel attention blocks (PAWE and CAWE) for segmentation.
- **Limitations:** The two-stage approach (detection then segmentation) is computationally redundant and slow. A separate, powerful detection model like Faster R-CNN is too heavy for an edge device.
- **How Our Project Overcomes This:** Our project's single-model, multi-task design is inherently more efficient. We can, however, adapt their innovative attention mechanisms (PAWE/CAWE) within our single U-Net architecture to improve segmentation accuracy without the overhead of a separate detection network.

8. Paper: "MSA: A Multi-Scale Attention Network for Lung Nodule Classification" (2024) [8]

- **What Was Done:** This paper focuses purely on classification, using a multi-head self-attention mechanism to extract fine-grained spatial features. It achieves a very high accuracy of 95.3% on the LUNA16 dataset.
- **Limitations:** The model is computationally complex and only performs classification on pre-identified nodules; it does not perform detection or segmentation. Its high complexity makes it unsuitable for edge deployment.
- **How Our Project Overcomes This:** We can adapt the *concept* of their multi-scale attention mechanism into a lightweight form and integrate it into the classification head of our unified model. This allows us to capture crucial fine-grained features for accurate classification while staying within the computational budget of our edge device.

Part 2: LLMs for Clinical and Patient-Friendly Reporting

Generating dual reports is a core innovation of this project. The literature shows this is a nascent field with significant room for contribution.

9. Paper: "Systematic review of LLMs in radiology report generation" (2025) [9]

- **What Was Done:** This comprehensive review of nine recent studies (2023-2024) found that fine-tuned LLMs outperform base models for generating radiology reports. However, it highlights major, persistent problems.
- **Limitations:** All reviewed LLMs exhibited critical flaws: hallucinations (44% of studies), misdiagnoses (55%), and missing clinical details (66%). Furthermore, none of the studies evaluated how patients or clinicians perceived the AI-generated reports, focusing only on technical metrics that correlated poorly with clinical correctness.
- **How Our Project Overcomes This:** Our project directly addresses these two major gaps. First, by grounding our LLM with structured input from the vision model, we fundamentally constrain its ability to hallucinate clinical facts. Second, our validation plan includes not only radiologist review for clinical fidelity but also a user study with the target rural population to assess the comprehension and utility of the patient-friendly summaries—a crucial, unaddressed area of research.

10. Paper: "AI-driven simplification of radiology reports significantly enhances patient comprehension" (2024) [10]

- **What Was Done:** This study used ChatGPT to simplify existing radiology reports and then conducted a patient survey. It found that the AI-simplified reports were rated significantly higher for clarity, tone, and patient engagement.
- **Limitations:** The study used a generic, off-the-shelf LLM (ChatGPT) in a post-processing step. It did not involve a fine-tuned model integrated into an end-to-end diagnostic workflow. The clinical safety of the simplified text was reviewed but not as part of a formal, scalable process.

- **How Our Project Overcomes This:** Our system is end-to-end. We will use a fine-tuned, open-source LLM to generate both reports simultaneously from the vision model's output. This is a more integrated and efficient approach. Our rigorous, two-panel validation (clinical safety review + patient comprehension study) will create a formal framework for safely deploying such a system, which is a novel contribution.

11. Paper: "FedMRG: Federated Learning for Medical Report Generation" (2025) [11]

- **What Was Done:** This paper introduces a federated learning framework (FedMRG) to train LLMs for medical report generation across multiple hospitals without sharing sensitive patient data. It uses low-rank factorization to reduce the communication costs of distributed training.
- **Limitations:** The primary focus is on the distributed training methodology. It does not address the challenges of deploying the resulting model in a low-resource, offline-first environment like a Health ATM. It also doesn't tackle the patient-facing report generation task.
- **How Our Project Overcomes This:** While our initial scope is not federated learning, this paper validates the use of parameter-efficient fine-tuning (PEFT) techniques like LoRA/QLoRA for medical LLMs.[13] We will use these same techniques not for distributed training, but to make the fine-tuning process feasible on modest, local hardware, which is essential for a government-backed project.

12. Paper: "MedRegion-CT: Region-Focused Multimodal LLM for 3D CT Report Generation" (2025) [14]

- **What Was Done:** This work proposes a sophisticated method for generating CT reports by providing the LLM with both global features and region-specific features extracted using a segmentation model. This allows the LLM to focus on clinically relevant areas.
- **Limitations:** This is a highly complex, research-focused model that requires multiple components (a vision model, a universal segmentation model, a mask encoder) just to generate the input for the LLM. This complexity makes it unsuitable for a lightweight, edge-based system.
- **How Our Project Overcomes This:** Our approach is far more pragmatic and efficient. We will pass a simple, structured JSON output from our single vision model to the LLM. This achieves the same goal—grounding the LLM in visual facts—but with a fraction of the computational overhead, making it viable for our target environment.

13. GitHub Project: "Fine-Tuning LLMs for Medical Entity Extraction" [16]

- **What Was Done:** This project demonstrates fine-tuning Llama2 and StableLM using PEFT (LoRA and Adapter V2) to extract drug names and side effects from text.

- **Limitations:** The task is entity extraction from existing text, not the more complex task of generating new, coherent reports from structured data.
- **How Our Project Overcomes This:** This project provides a practical, open-source blueprint for the fine-tuning process itself. We can adapt their code and methodology for our "structured-data-to-dual-report" generation task, which is a more advanced application of the same underlying PEFT techniques.

14. GitHub Project: "Medical-AI-LLM-FineTuning-Project" [17]

- **What Was Done:** This project fine-tuned a Llama 2 7B model on a medical dataset focused on viral genomics using AWS SageMaker. It successfully demonstrated that a general model could be specialized for a specific medical domain.
- **Limitations:** The project focuses on a niche domain (genomics) and uses a cloud-based platform (SageMaker) for training and deployment, which is not suitable for our low-cost, privacy-centric model.
- **How Our Project Overcomes This:** This work validates our choice of using a 7-8 billion parameter open-source model as a strong base. We will adapt this approach by using a more relevant dataset (radiology reports from LNDb, etc.) and deploying the fine-tuned model on a local server, ensuring data privacy and eliminating reliance on expensive cloud APIs.

Part 3: Edge Deployment and Optimization

Deploying complex models on devices like the Raspberry Pi 5 is a significant engineering challenge.

15. Article: "Optimizing deep learning models for Raspberry Pi through pruning and architecture optimization" (2023) [18]

- **What Was Done:** This paper systematically evaluates the impact of pruning, TensorFlow Lite conversion, and quantization on CNNs running on a Raspberry Pi 4. It found that the Arm NN delegate provided the best inference times.
- **Limitations:** The study uses a Raspberry Pi 4, not the significantly more powerful Pi 5. It also focuses on TensorFlow Lite and doesn't benchmark against other highly optimized ARM-native frameworks like Tencent's NCNN.
- **How Our Project Overcomes This:** We will build on this work by benchmarking on the Raspberry Pi 5. Crucially, we will compare the TensorFlow Lite/Arm NN pipeline against an ONNX/NCNN pipeline. NCNN is specifically designed with handcrafted NEON assembly for ARM CPUs and may offer superior performance, representing a more cutting-edge optimization strategy.[19]

16. Guide: "Deploying PyTorch models on Raspberry Pi 5" (Q-engineering) [20]

- **What Was Done:** This is a practical guide for installing PyTorch on a Raspberry Pi 5 and recommends exporting models to ONNX for use with C++ frameworks like NCNN or MNN for significant speedups.
- **Limitations:** It is a high-level guide and does not provide a full, end-to-end benchmark for a specific medical imaging model, nor does it detail the conversion and implementation process in C++.
- **How Our Project Overcomes This:** This guide validates our chosen deployment workflow (PyTorch -> ONNX -> NCNN). Our project will provide the missing piece: a complete, open-source implementation and detailed performance benchmark of this exact pipeline for a real-world medical imaging task, which will be a valuable contribution to the community.

17. Ultralytics Guide: "Use NCNN on Raspberry Pi" [22]

- **What Was Done:** This guide from the creators of YOLO shows how to export a YOLOv8 model to the NCNN format for deployment on a Raspberry Pi, noting that NCNN delivers the best performance on ARM-based devices.
- **Limitations:** The guide is specific to the YOLO architecture and the task of object detection. It does not cover segmentation models like U-Net or classification tasks.
- **How Our Project Overcomes This:** We will be the first to apply and document this high-performance NCNN export/deployment process for a lightweight, multi-task U-Net architecture. This extends the known best practices from object detection into the domain of medical image segmentation on edge devices.

18. Paper: "Benchmarking TensorFlow and TensorFlow Lite on Raspberry Pi 5" (2024) [24]

- **What Was Done:** This article provides updated benchmarks showing that the Raspberry Pi 5 is nearly 5x faster than the Pi 4 for deep learning tasks and that its performance with TensorFlow Lite is now comparable to dedicated AI accelerators like the Coral TPU.
- **Limitations:** The benchmarks use standard object detection models (MobileNet SSD v1/v2) and do not explore medical imaging tasks or other optimized runtimes like NCNN.
- **How Our Project Overcomes This:** This paper confirms that the Raspberry Pi 5 is a viable and powerful platform for our project. Our work will provide the first, crucial benchmarks for this hardware on a medical segmentation and classification task, using an even more optimized runtime (NCNN) to push the performance boundaries further.

19. NVIDIA Jetson Nano Documentation [25]

- **What Was Done:** NVIDIA provides extensive documentation for the Jetson Nano, highlighting its 128-core GPU and its software stack, including TensorRT for optimized inference.
- **Limitations:** The documentation is general-purpose. There are few, if any, end-to-end tutorials that show the full process of taking a lightweight PyTorch medical segmentation

model, converting it to ONNX, and optimizing it with TensorRT for deployment on the Nano.
- **How Our Project Overcomes This:** Our project will create exactly this. By documenting the full PyTorch -> ONNX -> TensorRT pipeline for our specific medical AI model, we will create a valuable, practical guide for the community that goes beyond NVIDIA's general documentation and provides a real-world case study.

Part 4: Explainable AI (XAI) in a Clinical Context

For clinicians to trust an AI, they must understand its reasoning.

20. Paper: "Grad-CAM based Visualization for Interpretable Lung Cancer Categorization using Deep CNN Models" (2025) [28]

- **What Was Done:** This study successfully applied Grad-CAM to visualize the decision-making process of three different CNNs (InceptionV3, XceptionNet, VGG19) for classifying histopathological lung cancer images. It showed how heatmaps could highlight cancerous cells and abnormal structures.
- **Limitations:** The application of Grad-CAM was for analysis in a research context. It was not integrated into a user-facing clinical tool where a doctor could interact with the explanations in real-time to inform a diagnosis.
- **How Our Project Overcomes This:** Our project moves XAI from a research tool to a clinical feature. The Grad-CAM output will be a core component of the UI, displayed directly alongside the CT scan, the segmentation mask, and the generated reports. This tight integration of prediction and explanation is a novel step towards building clinical trust and adoption.

21. Paper: "Explainable Lung Cancer Classification using VGG16, and Grad-CAM" (2025) [29]

- **What Was Done:** This work used a VGG16 model to classify CT scans as normal, benign, or malignant and used Grad-CAM to visualize the model's focus. It achieved 96% accuracy on the IQ-OTH/NCCD dataset.
- **Limitations:** The paper notes a key limitation: in early-stage cases, Grad-CAM sometimes highlights large, diffuse areas rather than pinpointing the exact lesion. It also uses VGG16, which is a relatively old and heavy architecture.
- **How Our Project Overcomes This:** By using a more modern, lightweight U-Net architecture with attention mechanisms, our model is designed to produce more precise feature maps, which should lead to more focused and accurate Grad-CAM heatmaps. We can also experiment with techniques like Guided Grad-CAM or FinerCAM from the pytorch-grad-cam library to generate sharper, more clinically useful visualizations.[31]

22. GitHub Project: pytorch-grad-cam [31]

- **What Was Done:** This is a comprehensive library providing PyTorch implementations for a wide array of CAM-based methods (Grad-CAM, Grad-CAM++, ScoreCAM, LayerCAM, etc.) and other XAI techniques.
- **Limitations:** As a library, it provides the tools but not a specific clinical application or a framework for how to best present these visualizations to a clinician for maximum utility.
- **How Our Project Overcomes This:** We will leverage this powerful library to build our explainability engine. Our novelty lies not in creating a new XAI algorithm, but in thoughtfully integrating these state-of-the-art tools into a seamless clinical workflow and UI, and then validating their usefulness with medical professionals.

23. Paper: "Using LIME to explore brain tumor MRI classification" [32]

- **What Was Done:** This article explains how LIME (Local Interpretable Model-agnostic Explanations) works by perturbing parts of an image to see how the model's prediction changes, thereby identifying important "superpixels."
- **Limitations:** LIME is computationally expensive because it requires making many predictions on perturbed versions of an image. It is also sensitive to the parameters used for generating superpixels. For these reasons, it is less suitable for a real-time edge application than Grad-CAM.
- **How Our Project Overcomes This:** This paper helps justify our choice of Grad-CAM over LIME. Grad-CAM is a gradient-based method that requires only a single backward pass through the network, making it far more computationally efficient and better suited for our resource-constrained Health ATM environment.

Part 5: Addressing the Challenges of Low-Resource Settings

Deploying AI in environments like rural India presents unique challenges beyond pure technical performance.

24. Paper: "Challenges and strategies for deploying AI diagnostic tools in low-resource settings" (2025) [34]

- **What Was Done:** This integrative review identifies key barriers to AI deployment in the Global South, including lack of diverse local data, poor infrastructure (internet, electricity), and a disconnect between Western-developed models and local needs.
- **Limitations:** The paper offers high-level strategic recommendations (e.g., "foster equitable partnerships," "adapt AI for local settings") but does not present a specific, technical blueprint for an end-to-end system that addresses these issues by design.
- **How Our Project Overcomes This:** Our project is a direct technical answer to these challenges. By designing for edge deployment, we mitigate the need for reliable internet. By using a lightweight model, we reduce the need for powerful, expensive hardware. By generating patient-friendly reports in local languages, we adapt the AI's output to local

health literacy needs. And by natively integrating with the ABDM, we align with the national digital health strategy from the outset.

25. Paper: "AI for healthcare in rural India" (Multiple Sources) [35]

- **What Was Done:** These articles and case studies discuss the potential of AI to bridge healthcare gaps in rural India. They highlight the work of companies like Qure.ai in screening for diseases like TB and lung cancer, and government initiatives like e-Sanjeevani for telemedicine.
- **Limitations:** Existing commercial solutions are often cloud-based and designed to integrate into hospital workflows, not standalone rural kiosks.[36] Telemedicine platforms connect doctors and patients but lack integrated, on-site diagnostic AI capabilities.
- **How Our Project Overcomes This:** Our project creates a new paradigm. It is not just an algorithm for a hospital or a simple telemedicine app. It is a complete, self-contained diagnostic solution *within* the Health ATM, providing on-site AI analysis that can then feed into the existing telemedicine infrastructure. This integration of edge AI with telemedicine is a significant step forward.

This analysis demonstrates that while significant progress has been made in individual components (lightweight models, LLMs, XAI), there is a clear and compelling gap in the literature for a project that integrates these elements into a single, end-to-end, and validated system specifically designed for the unique technical and social context of rural health ecosystems. Your project is perfectly positioned to fill this gap and make a substantial, publishable contribution.