Final Year Project Proposal

VisiHealth AI

Making Medical AI Human-Friendly

Supervisor: Sir Abdul Rahman

Submitted by:

Junaid Mohi Ud Din (01-134222-071)
Hammad Ur Rehman (01-134222-059)

Department of Computer Science
Bahria University, Islamabad

December 13, 2025

Contents

1. Introduction

Medical Visual Question Answering (Med-VQA) is an emerging research area that combines computer vision and natural language processing to answer clinical questions from medical images. Its potential applications range from assisting radiologists in decision-making to supporting medical education. However, most current Med-VQA systems behave as black boxes: they provide direct answers but fail to give explanations or evidence, making them less trustworthy for clinical use.

To address this gap, we propose **VisiHealth AI**, a system that not only predicts answers but also generates human-friendly rationales grounded in medical knowledge. Our system will use the **SLAKE dataset** [5], which provides 642 radiology images, 14,028 bilingual QA pairs, segmentation masks for 39 organs and 12 diseases, and 2,600+ medical knowledge graph (KG) triplets. This dataset is uniquely suited for explainability, as it combines images, QA, semantic annotations, and structured medical knowledge.

VisiHealth AI will train a CNN from scratch for image feature extraction and ROI localization, while fine-tuning a domain-specific BERT model (BioBERT/ClinicalBERT) for textual question embedding. The outputs will be fused for answer prediction. Finally, retrieved KG triplets will be inserted into template-based rationales, producing outputs such as:

> "Detected enlarged liver. KG says enlarged liver $\rightarrow$ Hepatomegaly. So answer = Yes."

This approach emphasizes trust, interpretability, and clinical relevance, which are critical for healthcare AI adoption.

2. Objective

To design and implement **VisiHealth AI**, a multimodal Med-VQA system that predicts answers to medical image questions and generates human-readable rationales grounded in knowledge graph facts and image regions, thereby improving explainability and trust in AI-assisted healthcare.

3. Problem Description

What: Existing Med-VQA systems primarily focus on improving answer accuracy but neglect explainability. Models trained on datasets like **VQA-RAD** (315 images, 3,515 QA) [1] or **VQA-Med (ImageCLEF)** (4,200+ images, ~15,000 QA) [2] provide answers but lack rationales. Even larger datasets like **PathVQA** (4,998 images, 32,799 QA) [3] or **PMC-VQA** (>100k images, 200k QA) [4] do not provide knowledge graphs or semantic masks — both essential for reasoning-based explanations.

Why: In healthcare, trust and interpretability are as important as accuracy. A doctor cannot rely on a system that outputs "Yes, abnormality detected" without evidence. Rationales grounded in ROI regions (e.g., a highlighted organ) and KG facts (e.g., "lung disease is located in the lungs") make the AI outputs verifiable and clinically meaningful.

Dataset justification: SLAKE may have fewer images (642), but it compensates with 14,028 QA pairs (∼22 per image), segmentation masks, and 2,600+ KG triplets [5]. This richness allows reasoning-based explanations that no other dataset supports.

Problem summary: Current Med-VQA systems lack explainability and grounded rationales. There is a need for a system that generates both answers and verifiable explanations, improving trust in medical AI.

4. Methodology

1. **Input:** Medical image + clinical question.

2. **Image Encoding:** Train a CNN from scratch on SLAKE images to extract visual features and ROI (Region of Intrest) localization scores. Training from scratch is educational but poses risk of overfitting, mitigated with augmentation, dropout, early stopping, cross-validation, and multi-task training.

3. **Text Encoding:** Fine-tune a pretrained BioBERT/ClinicalBERT on SLAKE's QA pairs to generate contextual question embeddings.

4. **Answer Prediction:** Fuse CNN features (global + ROI) with BERT embeddings and use a classifier to predict answers.

5. **Knowledge Graph Retrieval:** Index SLAKE's KG triplets by entities and relations; match triplets with detected ROI or question keywords.

6. **Rationale Generation:** Retrieve the most relevant KG triplet and insert into a rationale template, e.g., "Detected {ROI}. KG says {fact}. Therefore, answer = {prediction}."

7. **Output:** Predicted answer + human-friendly rationale.

8. **Web-Based Application:** The system is deployed as a secure web-based platform, enabling users to upload medical images, input clinical questions, and receive predicted answers with rationales through an accessible browser interface without requiring local installations.

Basic System Architecture Diagram
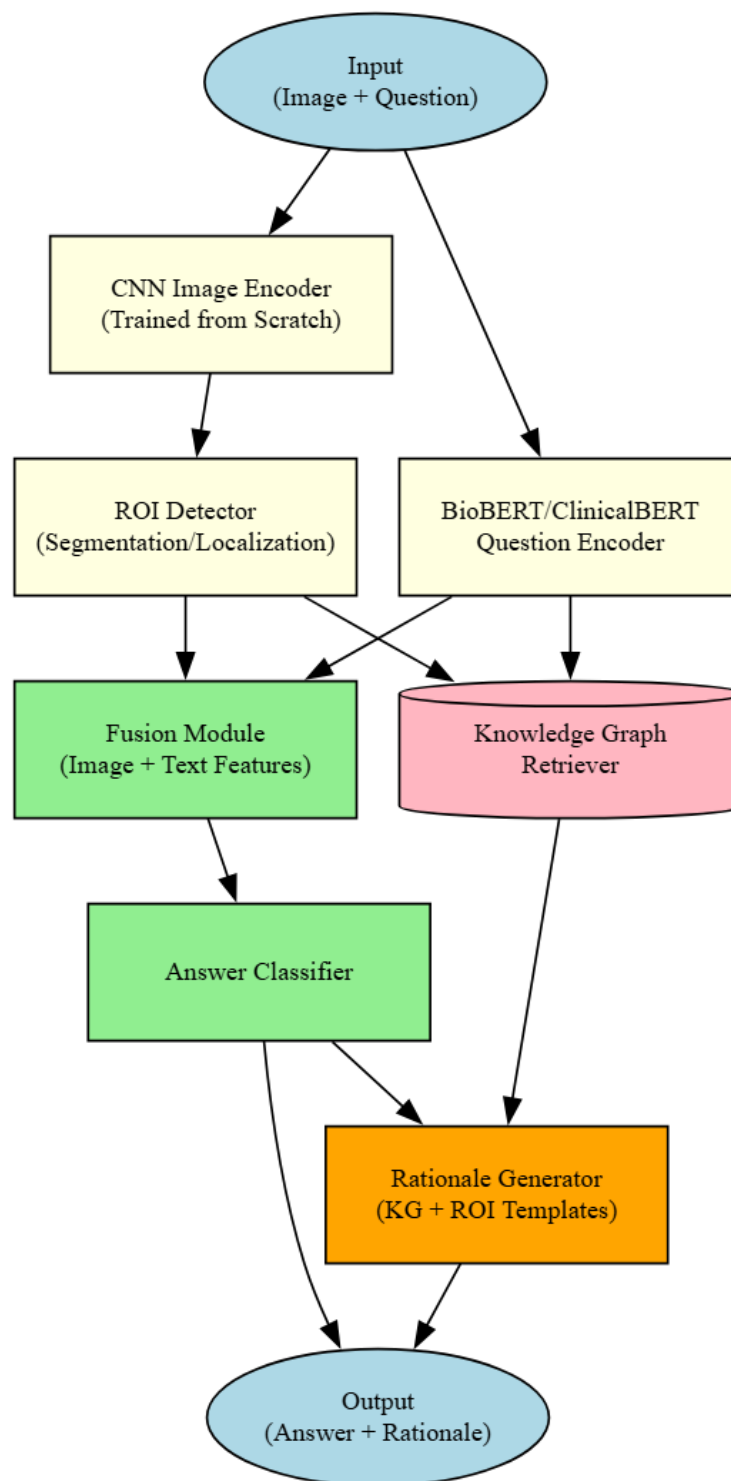
Software Process Model
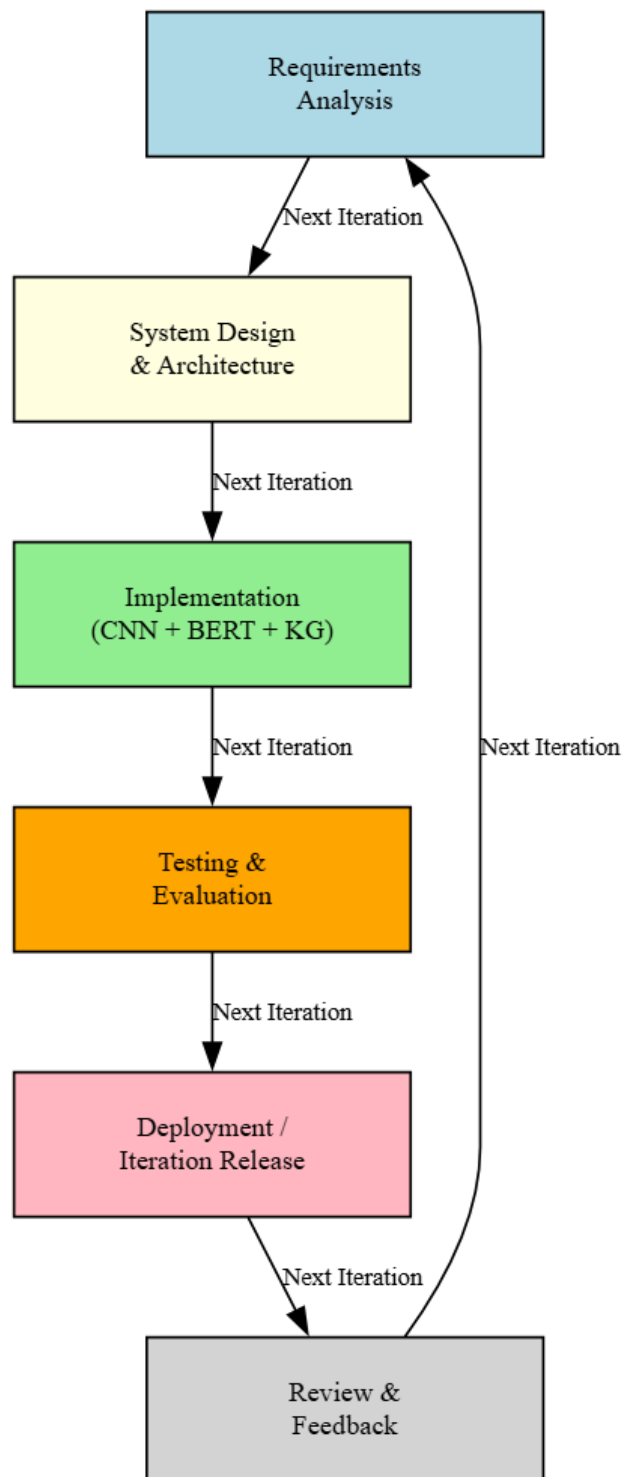
5. Project Scope

In scope: SLAKE EN subset, CNN training, BioBERT fine-tuning, ROI detection, KG retrieval, rationale generation, evaluation.

Out of scope: Chinese-language QA, large-scale hospital deployment, very large models

Assumptions: GPU or Cloud GPU resources are available.

6. Feasibility Study

Risks Involved:

1. **Overfitting due to small dataset:** Training a CNN from scratch on only 642 images can lead to poor generalization.

2. **KG-to-ROI mismatch:** Retrieved knowledge graph triplets may not always align with the detected image regions, reducing the quality of rationales.

3. **Limited compute resources:** Deep learning models require GPU acceleration, and restricted hardware could slow training or experimentation.

   Proposed Fixes:

1. **For Overfitting:** Apply extensive data augmentation (rotation, scaling, flipping), use dropout layers, and employ early stopping. Multi-task learning with segmentation masks will also regularize CNN training.

2. **For KG-to-ROI mismatch:** Introduce a lightweight scoring function to rank retrieved KG triplets by relevance to the question and detected region. Add simple filtering rules to avoid irrelevant facts.

3. **For Limited Compute:** Use Google Colab or a local GPU machine with at least 8GB VRAM. Optimize training with smaller batch sizes, gradient accumulation, and mixed precision.

   Resource Requirements:

1. SLAKE dataset (images, QA pairs, segmentation masks, KG triplets).

2. GPU-enabled machine (minimum 8GB VRAM).

3. Software: Python, PyTorch, Hugging Face Transformers, OpenCV.

7. Solution Application Areas

The primary application domain is **Healthcare AI Assistants for Radiologists**. Radiologists often face a heavy workload while analyzing medical images, and black-box AI tools are difficult to trust in clinical settings. By providing not only accurate answers but also clear rationales grounded in image regions and knowledge graph facts, our system can serve as a supportive assistant.

Another significant domain is **Medical Education and Training**. Students and junior doctors learning radiology often struggle to connect visual cues in medical images with textual medical knowledge. Our system bridges this gap by generating answers accompanied by reasoning chains that reference both image regions and structured knowledge.

Finally, the project has strong potential in **Clinical Decision Support Systems**. In emergency or high-volume healthcare environments, decision support tools that can quickly provide accurate responses with supporting explanations are invaluable. By integrating rationale generation, our system ensures that clinicians receive context-aware justifications.

8. Tools/Technology

Hardware: GPU-enabled machine / cloud GPU.
Software: Python, PyTorch, Hugging Face Transformers, OpenCV, FiftyOne, NumPy, Matplotlib.
Optional: Docker, Git.

9. Expertise of the Team Members

Junaid Mohi Ud Din: Has some experience working with machine learning, full-stack development, and small AI projects.
Hammad Ur Rehman: Has a background in deep learning coursework and has worked on applied projects related to multimodal AI.

Both team members are comfortable with Python and PyTorch, and we have studied the relevant courses during our degree. The **Data Science course and lab** helped us understand how to handle datasets, preprocessing, and evaluation techniques, while the **Artificial Intelligence course and lab** gave us a foundation in model building, training, and applying algorithms. These courses provided us with the background needed to attempt this project.

This project is of equal interest to both of us, and we are motivated to work together towards its successful completion.

10. Milestones

Milestone	Duration
Dataset preparation & preprocessing	2–3 weeks
Train CNN + segmentation/ROI	3–4 weeks
Fine-tune BioBERT + fusion model	3–4 weeks
Integrate KG retrieval + rationale generator	2 weeks
Evaluation (accuracy + explainability)	2–3 weeks
Documentation & final demo	3 weeks

VisiHealth AI - Project Gantt Chart
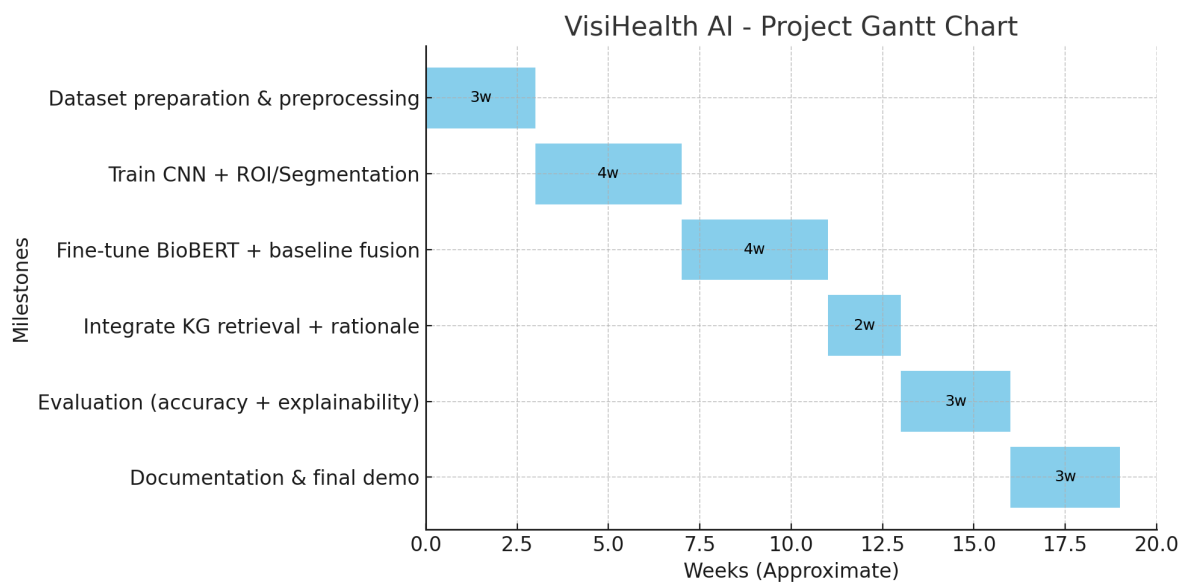
References

[1] J. J. Lau, A. Gayen, and D. Demner-Fushman, "A dataset of clinically generated visual questions and answers about radiology images," *Scientific Data*, 2018. DOI: 10.1038/sdata.2018.267. (VQA-RAD dataset).

[2] A. B. Abacha, S. A. Hasan, V. V. Datla, J. Liu, D. Demner-Fushman, and H. Müller, "VQA-Med: Overview of the Medical Visual Question Answering Task at Image-CLEF 2019," *Working Notes of CLEF 2019*, 2019. Paper/working-notes: http://ceur-ws.org/Vol-2380/paper_272.pdf. (VQA-Med / ImageCLEF 2019).

[3] X. He, H. Ke, J. [et al.], "PathVQA: 30000+ Questions for Medical Visual Question Answering," *Medical Image Analysis*, 2020. Preprint / arXiv: https://arxiv.org/abs/2003.10286. (PathVQA dataset).

[4] X. Zhang, [et al.], "PMC-VQA: A Large-Scale Medical Visual Question Answering Dataset from PubMed Central," *Nature Communications / Communications Medicine*, 2024. Publisher page: https://www.nature.com/articles/s43856-024-00709-2 (PMC-VQA dataset).

[5] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, and X.-M. Wu, "SLAKE: A Semantically-Labeled Knowledge-Enhanced Dataset for Medical Visual Question Answering," *ISBI / arXiv*, 2021. Preprint: https://arxiv.org/abs/2102.09542. (SLAKE dataset and project page: https://www.med-vqa.com/slake/).