AI-Powered Anomaly Detection in Image-Guided Surgeries

1. Problem Statement

Image-guided surgeries rely heavily on human expertise to detect anomalies such as tumors, misalignments, or distortions. Current approaches are prone to fatigue-induced errors, lack real-time feedback, and offer limited explainability—posing risks to patient safety.

---

2. Clinical Needs

Surgeons require more than simple anomaly detection. Key requirements include:

- **Localization** of anomalies within organ structures
- **Visual evidence** for immediate validation
- **Confidence scoring** to gauge reliability
- **Concise, automated reports** for decision support

---

3. Proposed Solution

A real-time, explainable AI pipeline integrating:

- **CNN-based segmentation** for organ delineation (liver, initially)
- **Diffusion models (DDPM + DDIM)** for reconstructing healthy anatomy
- **Residual heatmaps** to highlight deviations
- **Traffic-light scoring system** for intuitive alerts
- **Automated clinical reporting** to summarize findings

---

4. Pipeline Workflow

Step	Description
1	Preprocess input CT slice
2	Segment liver using CNN
3	Reconstruct healthy slice with DDIM
4	Generate residual heatmap (input vs reconstruction)
5	Score anomalies with traffic-light system
6	Auto-generate structured report
7	Display original slice, reconstruction, heatmap, and report

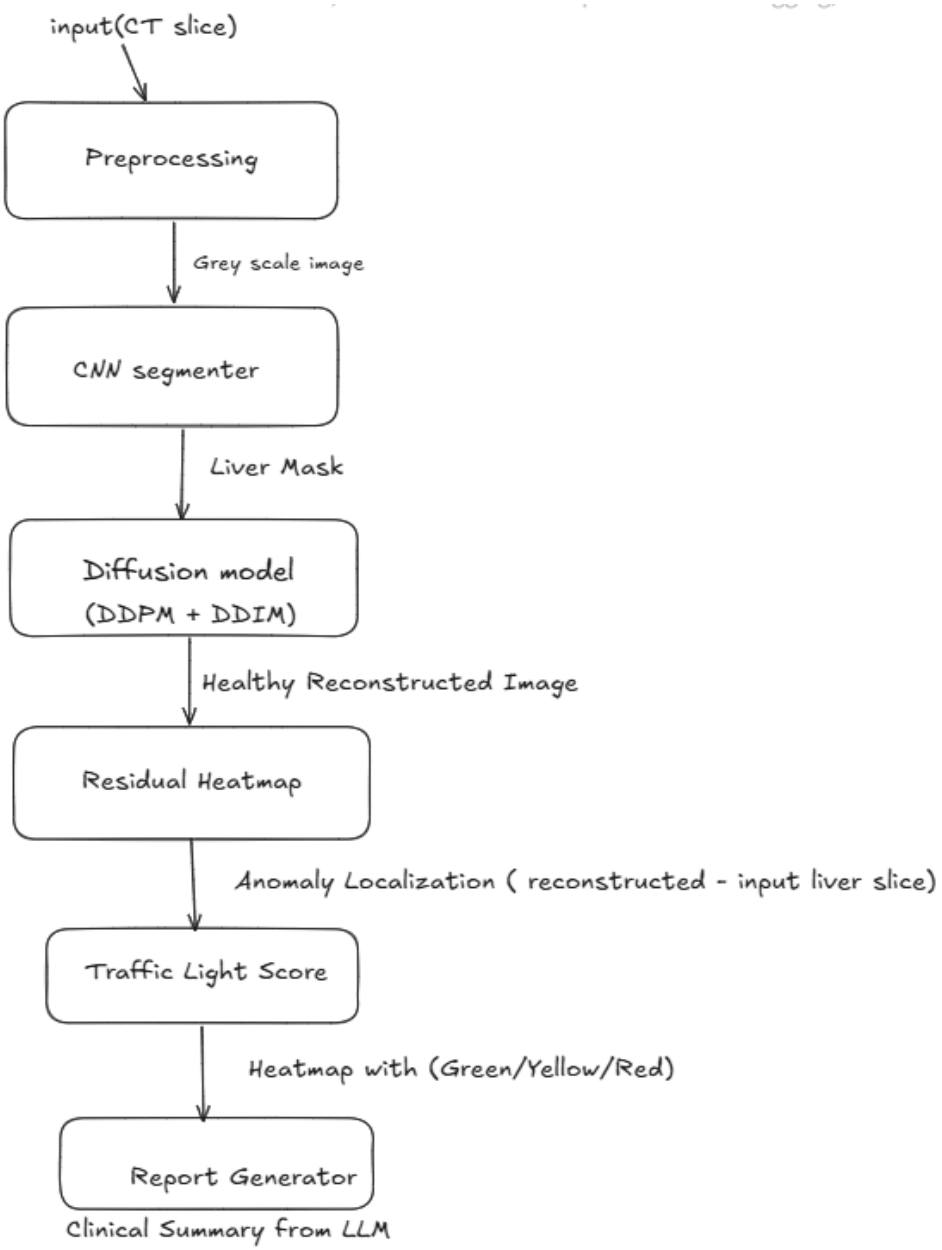
---

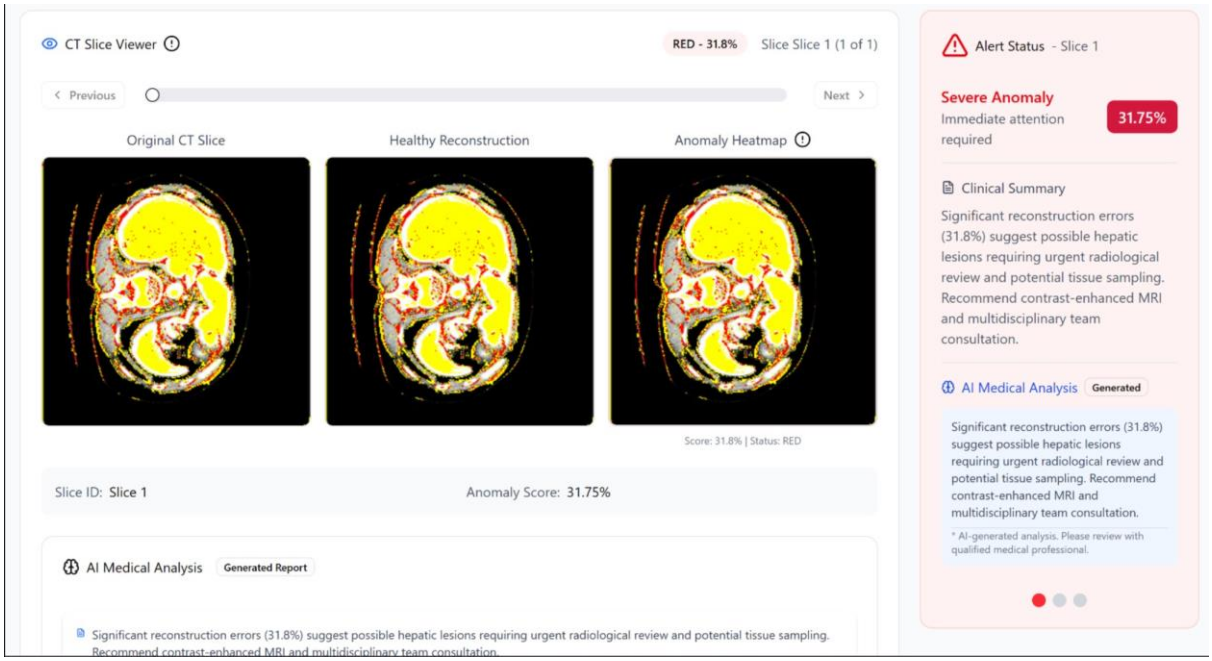
5. Model Architecture

Task	Model	Notes
Liver Segmentation	CNN (U-Net)	Lightweight, trained on chaos
Healthy Reconstruction	DDPM + DDIM	Unsupervised diffusion model for anomaly detection
Anomaly Localization	Residual Heatmap	Pixel-wise difference highlights abnormal regions
Anomaly Scoring	Rule-based	Green/Yellow/Red based on heatmap intensity
Reporting	Agentic AI	Structured clinical summaries generated automatically

---

6. System Block Diagram





---

7. Advantages

- **Label-efficient**: Unsupervised learning reduces dependency on annotated datasets
- **Explainable**: Heatmaps and traffic-light scoring make anomalies interpretable
- **Actionable**: Structured reports support faster clinical decision-making
- **Real-time capable**: Lightweight models allow near-instant feedback

---

8. Feasibility

- Public datasets (CHAOS) available for training
- Modular architecture supports seamless hospital integration
- Lightweight design enables real-time inference

---

9. Future Enhancements

- Expand to **multi-organ detection** (lungs, brain, spine)
- **Local Fine-Tuned LLM for Clinical Report Generation**

  Currently, the system utilizes cloud-based LLM APIs (e.g., Gemini) for automated report generation. For production deployment, transitioning to a locally-hosted fine-tuned language model using frameworks like Ollama offers significant advantages:

  Real-Time Latency: Local inference achieves 50-200ms response times compared to 2-5 seconds for API calls, enabling true quick real-time feedback during surgical procedures.

Cost Elimination: Zero per-inference fees versus recurring API costs exceeding $10,000 over five years for high-volume clinical use, with only one-time hardware investment.

Data Privacy & Regulatory Compliance: Patient CT data never leaves the device, eliminating transmission risks and streamlining approval pathways for on-device Software as a Medical Device (SaMD) classification.

Offline Operation: Ensures 100% system availability in restricted hospital networks, operating rooms with limited connectivity, or air-gapped military/government facilities where internet-dependent solutions are infeasible.

Domain Customization: Fine-tuning on liver-specific radiology reports (MIMIC-CXR, LiTS datasets) enables specialized medical terminology, and evidence-based follow-up recommendations that generic LLMs lack.

Efficient Implementation: Leveraging LoRA (Low-Rank Adaptation) for rapid fine-tuning (2-4 hours on single GPU) and 4-bit quantization compresses models like llama to be deployed on the system.