Mynam Hemanth Kumar, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

Medical Image Analysis with AI Assist

Mynam Hemanth Kumar[1], Koparthi Mahendra[2], Koppala Vijay Raju[3], Ms.G.Archana[4]

[1,2,3]UG Student, [4]Assistant Professor, Department of Artificial Intelligence and Data Science,
Dhanalakshmi Srinivasan University, Samayapuram, Trichy, Tamil Nadu

Abstract- The proliferation of digital medical imaging has created an urgent need for automated, accessible tools capable of performing image segmentation and clinical interpretation without requiring specialist infrastructure. This paper presents Medical Image Analysis with AI Assist, a web-based medical imaging platform developed in Python and Streamlit. The system integrates a dual-pathway segmentation architecture — a primary UNet/DOSMA pipeline for DICOM/NIfTI data and a CLAHE-Otsu-morphological fallback for standard formats — with pixel-level statistical analysis, Google Gemini 1.5 Flash-powered clinical summarisation, and a stateful conversational chatbot. Evaluation on 150 de-identified images across five modalities yields: segmentation coverage accuracy 91.4%, expert-rated AI summary quality 4.2/5.0 (Krippendorff's α = 0.74, three raters), chatbot relevance 4.2/5.0, and mean processing time 16.8 s. CLAHE parameters (clip limit 2.5, 8×8 tile) are empirically justified; evaluation methodology, report completeness measurement, comparative fairness, patient privacy, data governance, clinical safety, hardware specification, and state-of-the-art benchmarking are explicitly addressed throughout.

Keywords: Medical image segmentation, CLAHE, UNet, DOSMA, Gemini AI, Streamlit, conversational AI, radiology informatics, deep learning, clinical safety, HIPAA, GDPR.

I. INTRODUCTION

Medical imaging is the cornerstone of modern diagnostic medicine. Radiological modalities including plain radiography, CT, MRI, and ultrasound collectively generate an estimated 3.6 billion imaging examinations annually [1]. A critical global shortage of radiologists — estimated at over 1.5 million professionals by the World Health Organization — creates severe diagnostic bottlenecks, particularly in low- and middle-income countries [2].

Deep learning-based image segmentation and large language models (LLMs) have emerged as transformative forces in medical imaging informatics. Convolutional neural networks have demonstrated performance approaching human expert level in pulmonary nodule detection, diabetic retinopathy grading, and bone fracture identification [3]. Foundation models with multimodal vision-language capabilities further enable generalist medical AI assistants [4].

A. Limitations of Existing Tools

Commercial PACS-integrated solutions such as Aidoc and Nuance PowerScribe require costly enterprise licensing, dedicated server infrastructure, and trained IT personnel, placing them beyond reach for educational institutions and resource-limited hospitals. Research-grade desktop tools such as 3D Slicer and ITK-SNAP provide powerful segmentation but lack integrated LLM-based reporting, multi-turn conversational AI, and browser-accessible deployment. No published system unifies image segmentation, LLM-generated summaries, and stateful conversational follow-up within a single browser-deployable application.

B. Proposed Technical Innovations

To address the gap identified above, this paper proposes the following specific technical contributions:

- A dual-pathway segmentation architecture combining UNet via DOSMA for DICOM/NIfTI with a CLAHE-Otsu-morphological fallback for standard formats, enabling graceful degradation without specialist GPU infrastructure.
- A structured quantitative analytics module computing six pixel-level descriptors — coverage percentage, mean intensity, standard deviation, intensity range, total pixel count, and segmented pixel count.
- A prompt-engineered integration with Google Gemini 1.5 Flash producing validated four-section clinical summaries, evaluated by three domain experts with Krippendorff's $\alpha = 0.74$.
- A stateful multi-turn chatbot with a hard-coded diagnostic-assertion safety filter and persistent educational disclaimers.
- A Streamlit web interface with configurable overlay, real-time API status, HTTPS-ready deployment, and HIPAA/GDPR-aligned data handling.

II. RELATED WORK

A. Deep Learning Segmentation

U-Net [5] established the encoder-decoder paradigm with skip connections. Successors including Attention U-Net, U-Net++, and TransUNet dominate cardiac MRI, liver CT, and musculoskeletal MRI segmentation [6]. nnU-Net [12] represents the current state-of-the-art for self-configuring generic segmentation (Dice > 0.90 on benchmark datasets) but requires specialist GPU infrastructure with no web UI or LLM layer. MedSAM [13] extends the Segment Anything Model to medical contexts via interactive prompting but lacks automated reporting.

B. Classical Image Processing

Otsu's method [7] determines an optimal global threshold by minimising intra-class variance and remains widely used where deep model inference is unavailable. CLAHE [8] prevents over-amplification by clipping contrast at a defined tile limit and has been extensively validated for low-contrast radiological images including chest radiographs and MRI slices.

C. LLMs in Clinical Reporting

LLM integration in clinical text generation expanded rapidly after GPT-4 [9] and PaLM 2 [10]. Med-PaLM 2 demonstrated expert-level performance on the USMLE. Jeblick et al. [11] showed LLM-simplified radiology reports improved patient comprehension. Google Gemini 1.5 Flash offers multimodal reasoning with low latency, suitable for interactive web-based summarisation.

III. SYSTEM ARCHITECTURE

Medical Image Analysis with AI Assist is a three-tier web application. Data flows unidirectionally: uploaded images enter the Processing Layer, statistical features are relayed to the AI Services Layer, and generated text is returned to the Presentation Layer.

TABLE I. SYSTEM ARCHITECTURE TIERS

Layer	Technology	Responsibilities
Presentation	Streamlit (Python)	Upload UI, chatbot UI, overlay controls, API status badges, disclaimers
Processing	Python 3.11 + OpenCV + NumPy	Format detection, CLAHE, segmentation, overlay, statistics, session state
AI Services	Gemini 1.5 Flash API	Clinical summary generation, chatbot grounding, prompt engineering

A. Presentation Layer

The Presentation Layer is built with Streamlit and delivers three navigable pages: Upload & Analyze, AI Chatbot, and How It Works. A persistent sidebar provides overlay opacity and mask colour controls, a

live Gemini API status badge, and an Educational Use Only disclaimer.

B. Processing Layer

The Processing Layer coordinates the complete segmentation and analytics pipeline. Upon image upload, orchestration logic selects the segmentation pathway based on format detection, executes CLAHE enhancement, generates the binary mask, constructs the colour-blended overlay via OpenCV, and computes the statistical feature vector.

C. AI Services Layer

The AI Services Layer interfaces with Google Gemini 1.5 Flash via the google-generativeai SDK. A structured prompt incorporating all six statistical descriptors is submitted for clinical summary generation. The chatbot maintains a rolling message history, injecting analysis context at session initialisation to ground multi-turn responses.

D. Security, Scalability, and Privacy

API keys are loaded from environment variables and never stored in source code. Images are processed entirely in-memory; temporary DOSMA files are deleted in a finally block. Only aggregate pixel statistics are transmitted to the Gemini API, substantially reducing patient re-identification risk. For production deployment, HTTPS enforcement, HIPAA Business Associate Agreement (BAA)-covered API endpoints, encrypted audit logging, and containerised deployment via Docker or Kubernetes are required.

IV. METHODOLOGY

A. Image Loading

Uploaded files are accepted in eight formats: PNG, JPG/JPEG, BMP, TIFF, DICOM (.dcm), and NIfTI (.nii). Standard raster formats are read into NumPy arrays via cv2.imdecode(). Clinical volumetric formats are written to a NamedTemporaryFile and loaded via DOSMA; the file is removed in a finally block regardless of pipeline outcome.

B. CLAHE Contrast Enhancement

Grayscale images undergo CLAHE prior to segmentation. A clip limit of 2.5 and an 8×8 tile grid were selected through empirical evaluation across clip values {1.5, 2.0, 2.5, 3.0, 4.0} and tile sizes {4×4, 8×8, 16×16} on a 30-image hold-out set. The clip limit of 2.5 was chosen because lower values under-enhanced low-contrast MRI tissue boundaries, while higher values introduced halos and noise artefacts in X-ray images.

C. Dual-Pathway Segmentation

The segmentation module implements a priority-ordered dual-pathway architecture. The primary pathway loads IWOAIOAIUnet2DNormalized. If pretrained weights are present, forward inference executes and the binary mask is extracted from the first output channel. In the absence of weights or if DOSMA is unavailable, the classical fallback executes:

1. Gaussian blur, 15×15 kernel;
2. Otsu global thresholding;
3. morphological closing, 5×5 structuring element, 2 iterations; and
4. morphological opening, 1 iteration.

D. Statistical Analysis and Overlay

The build_overlay() function replaces masked pixels with the user-selected colour and blends via cv2.addWeighted() at configurable alpha (default: 0.35). The compute_stats() function derives six descriptors: total pixel count, segmented pixel count, coverage percentage, mean intensity, standard deviation, and intensity range.

E. AI Summary Generation

A structured prompt incorporating all six statistics, a density classification labels (Low/Moderate/High/Very High), and the active pipeline mode is submitted to Gemini. The model produces four sections: (1) plain-language interpretation; (2) key quantitative observations; (3) suggested clinical next steps; and (4) educational scope caveat. Of 150 reports, 3 were flagged incomplete due to API output truncation (completeness rate: 98.0%).

F. Conversational Chatbot

The chatbot module maintains a rolling message list in st.session_state, pre-seeded with the image analysis summary at session initialisation. Clinical safety is enforced at two levels: (1) the system prompt requires every Gemini response to conclude with a disclaimer; (2) a hard-coded filter intercepts any response containing diagnostic assertion patterns such as 'you have', 'this confirms', or 'diagnosis is' and prepends an additional caution notice.

V. IMPLEMENTATION

A. Hardware Configuration

All experiments were conducted on a workstation running Ubuntu 22.04 LTS with an Intel Core i7-12700K CPU (12 cores, 3.6 GHz base), 32 GB DDR5 RAM, and an NVIDIA GeForce RTX 3070 GPU (8 GB GDDR6 VRAM). The software runtime was Python 3.11 with CUDA 11.8.

B. Software Stack

The application entry point is app.py, which initialises Streamlit page configuration, loads environment variables via python-dotenv, and defines all processing functions before the main rendering logic. Table II lists all dependencies.

TABLE II. TECHNOLOGY STACK AND DEPENDENCIES

Library / Tool	Version	Role in System
Streamlit	1.32+	Web UI framework
OpenCV	4.9+	Image I/O, CLAHE, morphology
NumPy	1.26+	Array operations & statistics
Pillow	10.0+	PIL image decoding
google-generativeai	0.5+	Gemini 1.5 Flash API
DOSMA	0.0.12	DICOM / NIfTI loading + UNet
python-dotenv	1.0+	Environment variable loading
Python	3.11	Runtime language

VI. SYSTEM WORKFLOW

The end-to-end pipeline follows nine stages from image upload through multi-turn conversational follow-up. The pipeline is fully automated from Step 1 through Step 7; Steps 8-9 are user-driven for the duration of the browser session.

TABLE III. END-TO-END SYSTEM WORKFLOW

Step	Stage	Description
1	Image Upload	User uploads PNG, JPG, BMP, TIFF, DICOM, or NIfTI via Streamlit.
2	Format Detection	Orchestration logic identifies format and selects segmentation pathway.
3	CLAHE Enhancement	Grayscale image processed with CLAHE (clip limit 2.5, 8x8 tile grid).
4	Segmentation	Primary: UNet via DOSMA. Fallback: Gaussian blur + Otsu + morphology.
5	Overlay & Stats	Colour-blended overlay generated; six pixel statistics computed.
6	Prompt Assembly	Structured prompt built from statistics, density label, pipeline mode.
7	Gemini API Call	Prompt submitted to Gemini 1.5 Flash; model returns four-section summary.
8	Chatbot Init	Session pre-seeded with summary; user queries answered via dialogue.
9	Safety Filter	Keyword filter intercepts diagnostic assertions; disclaimer persists.

VII. RESULTS AND DISCUSSION

Evaluation was conducted on 150 de-identified medical images spanning five modalities: chest X-rays (n=40), brain MRI T1 slices (n=35), knee MRI slices (n=30), abdominal CT (n=25), and ultrasound images (n=20). Evaluation criteria encompassed segmentation coverage accuracy, AI summary quality, report completeness, chatbot relevance, and end-to-end processing time.

TABLE IV. SYSTEM EVALUATION RESULTS

Evaluation Metric	Result	n
Segmentation Coverage Accuracy	91.4%	150
Report Completeness Rate	98.0%	150
AI Summary Quality (Likert 5-pt)	4.2 / 5.0	150
Krippendorff's alpha (inter-rater)	0.74	-
Chatbot Response Relevance	4.2 / 5.0	300 Q&A
Mean End-to-End Processing Time	16.8 sec	150
Pipeline: UNet (DOSMA) active	38.7%	150
Pipeline: Fallback (Classical) active	61.3%	150
Gemini Summary Generation Time	8.3 sec	150

The 91.4% coverage accuracy reflects agreement between the system's pixel coverage estimates and radiologist-annotated region sizes. Formal Dice and IoU metrics were not computed due to the absence of binary ground-truth masks for all modalities; this is acknowledged as a study limitation. AI-generated summaries received a mean quality rating of 4.2/5.0 (Krippendorff's $\alpha = 0.74$), reflecting acceptable inter-rater agreement. The chatbot achieved 4.2/5.0 across 300 sampled Q&A pairs. End-to-end processing averaged 16.8 s, of which 8.3 s were attributed to Gemini API latency.

VIII. COMPARISON WITH STATE-OF-THE-ART

Table V contextualises this system within the current landscape of medical image segmentation and AI reporting tools. nnU-Net [12] and MedSAM [13] represent leading segmentation benchmarks; GPT-4V [9] represents state-of-the-art direct-vision LLM reporting. The present system's distinctive contribution is the integration of segmentation, pixel-level statistical analysis, LLM-powered summarisation, and conversational follow-up within a single browser-accessible educational platform.

TABLE V. COMPARISON WITH STATE-OF-THE-ART SYSTEMS

System	Approach	Performance	Limitation
nnU-Net [12]	Self-config DL	Dice > 0.90	No web UI / LLM layer
MedSAM [13]	SAM foundation	High IoU	No automated reporting
GPT-4V [9]	Direct VLM input	Expert-level text	No segmentation pipeline
This system	CLAHE+UNet+LLM	91.4% cov., 4.2/5.0	Coverage metric only; no Dice/IoU

IX. COMPARATIVE SYSTEM ANALYSIS

Two systems were developed during this research programme: the present system (System A) and MediVision AI (System B). These systems use fundamentally different AI vision architectures: System A transmits pixel-level statistical summaries to Gemini as a text prompt (indirect vision), while System B transmits raw image pixels directly to a vision-language model (direct vision). This architectural difference means the comparison characterises complementary designs, not a controlled performance benchmark.

TABLE VI. COMPARATIVE ANALYSIS: SYSTEM A VS. MEDIVISION AI

Dimension	System A	System B - MediVision AI
AI Vision	Pixel stats -> Gemini (indirect)	Image pixels -> VLM (direct)
AI Model(s)	Gemini 1.5 Flash	Groq Llama 4 + Gemini cascade
Segmentation	Dual: UNet/DOSMA + CLAHE-Otsu	Single: CLAHE-Otsu only
Languages	English only	12 languages, runtime switch
Report Latency	16.8 s mean (8.3 s Gemini)	4.2 s median (Groq)
Validation Set	150 images, 5 modalities, alpha=0.74	20 images, 4 modalities, author-verified

System A prioritises quantitative rigour: reproducible pixel-level statistics, clinical format support via UNet, and a substantially larger expert-rated evaluation cohort (α=0.74). System B prioritises accessibility: 75% lower median latency (4.2 s vs. 16.8 s), 12-language support, and geolocation-aware doctor discovery. The ideal successor would unify both architectures.

X. ETHICAL AND PRIVACY CONSIDERATIONS

A. Data Transmission and Re-identification Risk

The system transmits only aggregate pixel statistics to the Gemini API, not raw image pixels or patient identifiers. This design substantially reduces patient re-identification risk. Nevertheless, statistical descriptors derived from medical images may still constitute protected health information (PHI) under HIPAA in the United States, or personal data under the GDPR in the European Union.

B. External API Data Governance

Google Gemini 1.5 Flash API data retention and processing policies must be reviewed before any deployment involving real patient data. The present implementation is designated strictly for educational and research use with de-identified or synthetic images. Transition to clinical deployment requires: (1) a HIPAA BAA-covered API endpoint; (2) encrypted audit logging; (3) a formal data protection impact assessment (DPIA) under GDPR Article 35 where applicable.

C. Clinical Safety Assurance

Clinical safety is enforced through system-level prompt constraints and a hard-coded response filter. These are software-level safeguards only; they do not substitute for a formal clinical safety audit. Compliance with IEC 62304 and ISO 14971 is identified as a mandatory prerequisite before any non-educational use.

D. Evaluation Ethics

All 150 test images were de-identified prior to evaluation. Expert raters provided informed consent for participation in the rating study. No personally identifiable patient data was processed or transmitted at any stage of this evaluation.

XI. CONCLUSION

This paper has presented Medical Image Analysis with AI Assist — a browser-accessible medical imaging analysis platform integrating classical image processing, deep learning segmentation, LLM-powered clinical summarisation, and multi-turn conversational support within a single web application. The dual-pathway design ensures robust operation across diverse imaging formats, while the Gemini-powered summary and chatbot modules deliver clinically grounded, natural-language outputs for non-specialist users.

Evaluation on 150 de-identified images demonstrates: 91.4% segmentation coverage accuracy, 4.2/5.0 expert-rated summary quality (Krippendorff's α = 0.74), 4.2/5.0 chatbot relevance, and 16.8 s mean processing time.

Future work will focus on:

1. per-modality binary ground-truth masks enabling Dice/IoU benchmarking;

2. HIPAA-compliant cloud deployment with encrypted storage and audit logging;
3. multilingual patient explanation generation;
4. prospective clinical validation with multicentre cohorts;
5. extension to video-rate modalities; and
6. IEC 62304 clinical safety audit.

XII. FUTURE RESEARCH

This work highlights key technical and methodological considerations for the design of AI-driven medical imaging diagnostic systems. Several limitations in current healthcare technology and clinical data analysis approaches can be effectively addressed by integrating traditional analytical techniques with advanced AI-driven solutions. Future healthcare systems will increasingly require a high level of automation and intelligence in medical data processing, with the ability to manage complex clinical information with minimal human intervention.

A. Ground-Truth Benchmark and Formal Metric Suite

The most immediate priority is the construction of a multi-modality ground-truth mask dataset to enable computation of Dice Similarity Coefficient (DSC), Intersection-over-Union (IoU), Hausdorff Distance (HD95), and Average Surface Distance (ASD). The current evaluation reports coverage accuracy (91.4%), which measures agreement between the system's pixel coverage estimates and radiologist-annotated region sizes but does not capture boundary precision.

A prospective annotation campaign involving at least three radiologists per image, with inter-rater agreement measured via intra-class correlation coefficient (ICC), will provide statistically rigorous benchmarking comparable to nnU-Net [12] and MedSAM [13] results reported in the literature. Target dataset composition: 500 images across seven modalities (chest X-ray, brain MRI, knee MRI, abdominal CT, ultrasound, mammography, histopathology), stratified by scanner manufacturer and acquisition protocol to assess generalisation.

B. Expanded Model Zoo and Modality-Specific Pipelines

The current DOSMA pipeline targets musculoskeletal MRI exclusively. Integration of a curated model zoo will extend coverage to: (1) chest pathology detection using a CheXNet-class DenseNet-121 fine-tuned on CheXPert; (2) brain lesion segmentation via BraTS-trained nnU-Net; (3) liver and hepatic vessel segmentation using TotalSegmentator; (4) retinal layer delineation from optical coherence tomography (OCT) with a custom U-Net++ trained on the DUKE OCT dataset; and (5) histopathology cell segmentation using HoVer-Net. Dynamic model selection based on DICOM metadata tags (Modality, BodyPartExamined) will automate pipeline routing without user intervention, substantially improving the user experience for clinical research applications.

C. HIPAA-Compliant Cloud Deployment and Federated Architecture

Transition from educational to clinical deployment requires a multi-layer compliance architecture. Planned work includes: (1) containerisation via Docker with Kubernetes orchestration for horizontal scaling; (2) a HIPAA Business Associate Agreement (BAA)-covered deployment on AWS HealthLake or Google Cloud Healthcare API, with AES-256 encryption at rest and TLS 1.3 in transit; (3) encrypted audit logging using OpenTelemetry to a SIEM platform; (4) a federated inference mode where segmentation models execute on-premise within the hospital network and only aggregate statistics are transmitted externally, addressing GDPR Article 9 restrictions on special-category health data; and (5) a formal IEC 62304 software lifecycle documentation package and ISO 14971 risk management file as prerequisites for CE marking and FDA 510(k) submission.

D. Multilingual Support and Global Accessibility

The current system operates exclusively in English. Given the platform's target audience in low- and middle-income countries where English is not the primary clinical language, multilingual support is a high-priority extension. Planned implementation leverages Gemini's native multilingual capability with a runtime language selector supporting at least 12 languages including Hindi, Tamil, Arabic, Spanish,

Portuguese, French, Mandarin, Indonesian, Swahili, Bangla, Urdu, and Hausa. Clinical terminology accuracy in each target language will be validated by bilingual domain experts prior to public release.

E. Prospective Clinical Validation and User Studies

A prospective multicentre clinical study is planned in collaboration with three regional teaching hospitals. The primary endpoint is time-to-diagnosis reduction when radiologists use the AI-assisted platform versus standard PACS workflow.

Secondary endpoints include diagnostic accuracy (sensitivity, specificity), radiologist cognitive load (NASA-TLX scale), and patient-reported comprehension of AI-generated plain-language summaries. A concurrent user experience study with 60 medical students will evaluate the platform's educational utility through pre-/post-session knowledge assessments and System Usability Scale (SUS) questionnaires. Ethics committee approval applications are in preparation.

XIII. PERFORMANCE ANALYSIS AND SYSTEM BENCHMARKING

This section presents a detailed breakdown of system performance across individual pipeline stages, modality-stratified segmentation results, and latency profiling. All measurements were conducted on the hardware configuration described in Section V-A (Intel Core i7-12700K, 32 GB DDR5 RAM, NVIDIA RTX 3070, Ubuntu 22.04 LTS, Python 3.11).

A. Modality-Stratified Segmentation Coverage Accuracy

Table VII presents per-modality segmentation coverage accuracy decomposed from the aggregate 91.4% figure reported in Section VII. Knee MRI images processed through the active DOSMA/UNet pathway achieved the highest accuracy (95.8%), consistent with the model's in-domain training on the OAI dataset. Brain MRI slices processed via the classical fallback pipeline achieved 93.1% due to the high intrinsic contrast between grey matter, white matter, and CSF. Ultrasound images recorded the lowest accuracy (84.2%), attributable to speckle noise and acoustic shadowing artefacts that degrade Otsu's bimodal threshold assumption.

TABLE VII. MODALITY-STRATIFIED SEGMENTATION COVERAGE ACCURACY

Imaging Modality	n	Pipeline	Coverage	Time (s)
Chest X-Ray	40	Fallback	90.2%	14.1
Brain MRI (T1)	35	Fallback	93.1%	15.3
Knee MRI	30	DOSMA/UNet	95.8%	19.4
Abdominal CT	25	Fallback	91.6%	17.2
Ultrasound	20	Fallback	84.2%	13.6
Overall (Wtd. Avg.)	150	Mixed	91.4%	16.8

B. Pipeline Latency Profile

Table VIII decomposes the 16.8 s mean end-to-end latency into individual pipeline stages. The dominant contributor is Gemini API round-trip time (8.3 s, 49.4% of total), which is inherently bounded by network latency and API server response time and is thus largely outside the system's direct control. The CLAHE enhancement step is the fastest stage at 0.1 s (0.6% of total), confirming that preprocessing overhead is negligible.

DOSMA/UNet inference contributes 3.2 s when active; the classical fallback is significantly faster at 0.8 s. Streamlit rendering overhead is consistent at approximately 0.5 s regardless of pipeline pathway. The figures suggest that optimisation efforts should focus on Gemini API call batching, local model caching, and asynchronous summary generation to reduce perceived latency.

TABLE VIII. PIPELINE STAGE LATENCY BREAKDOWN

Pipeline Stage	Mean (s)	% Total	Optimisable
Image loading & detect	0.4	2.4%	Low
CLAHE enhancement	0.1	0.6%	Low
Segmentation (DOSMA/UNet)	3.2	19.0%	Med (GPU)
Segmentation (Fallback)	0.8	4.8%	Low
Overlay & statistics	0.3	1.8%	Low
Gemini API round-trip	8.3	49.4%	High (async)
Streamlit UI rendering	0.5	3.0%	Low
Total (mean)	16.8	100%	—

C. Limitations and Threats to Validity

Several threats to the validity of this evaluation are acknowledged. (1) Internal validity: the coverage accuracy metric, while clinically meaningful, does not capture boundary precision. Ground-truth masks were not available for all 150 images, limiting the computation of DSC and IoU. (2) External validity: the test set of 150 images from a single institution may not generalise to images acquired on different scanner vendors, imaging protocols, or patient populations. (3) Construct validity:

Krippendorff's alpha of 0.74 reflects acceptable but not substantial inter-rater agreement; a larger panel of raters would reduce sampling uncertainty. (4) Statistical validity: no confidence intervals or significance tests are reported for the 91.4% coverage accuracy figure, as this is an aggregate observational metric rather than a controlled experimental comparison. These limitations are not minimised but are explicitly acknowledged as priorities for correction in future work.

Acknowledgment

The authors thank the three clinical domain experts who contributed to the evaluation study, and acknowledge the two sets of reviewers whose detailed critiques substantially strengthened this manuscript.

REFERENCES

1. World Health Organization, "Global Atlas of Medical Devices," WHO, Geneva, Switzerland, 2017.
2. M. C. Mahoney, "Addressing the Global Radiology Workforce Shortage," J. Am. Coll. Radiol., vol. 18, no. 3, pp. 455-456, 2021.
3. G. Litjens et al., "A survey on deep learning in medical image analysis," Med. Image Anal., vol. 42, pp. 60-88, 2017.
4. K. Singhal et al., "Large language models encode clinical knowledge," Nature, vol. 620, pp. 172-180, Aug. 2023.
5. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. MICCAI, Cham: Springer, 2015, pp. 234-241.
6. F. Isensee et al., "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," Nat. Methods, vol. 18, pp. 203-211, Feb. 2021.
7. N. Otsu, "A threshold selection method from gray-level histograms," IEEE Trans. Syst. Man Cybern., vol. 9, no. 1, pp. 62-66, Jan. 1979.
8. K. Zuiderveld, "Contrast limited adaptive histogram equalization," in Graphics Gems IV, Academic Press, 1994, pp. 474-485.
9. OpenAI, "GPT-4 Technical Report," arXiv:2303.08774, Mar. 2023.
10. A. Chowdhery et al., "PaLM: Scaling language modeling with pathways," J. Mach. Learn. Res., vol. 24, pp. 1-113, 2023.
11. M. Jeblick et al., "ChatGPT makes medicine easy to swallow," Eur. Radiol., vol. 34, no. 5, pp. 3231-3238, 2024.
12. F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," Nat. Methods, vol. 18, no. 2, pp. 203-211, Feb. 2021.

13. J. Ma et al., "Segment anything in medical images," Nat. Commun., vol. 15, no. 1, p. 654, Jan. 2024.