DeepSpine: MCP-Based LLM Agentic Reasoning Chatbot for Automated Lumbar MRI Pathology Analysis and Diagnosis

Mukhlis Raza
Department of Artificial Intelligence, College of AI Convergence, Daeyang AI Center, Sejong University, Seoul 05006, Korea

Saied Salem
Department of Artificial Intelligence, College of AI Convergence, Daeyang AI Center, Sejong University, Seoul 05006, Korea

Afnan Habib
Department of Artificial Intelligence and Data Science, College of AI Convergence, Daeyang AI Center, Sejong University, Seoul 05006, Korea

Hyunwook Kwon
Department of Artificial Intelligence and Data Science, College of AI Convergence, Daeyang AI Center, Sejong University, Seoul 05006, Korea

Ahmet Arif Aydin
Department of Computer Engineering, Faculty of Engineering, Inonu University, Malatya, Turkey

Mugahed A. Al-antari *
Department of Artificial Intelligence and Data Science, College of AI Convergence, Daeyang AI Center, Sejong University, Seoul 05006, Korea

Abstract

Lumbar spine pathology assessment through MRI analysis remains a complex and time-intensive process requiring specialized radiological expertise. Current automated computer-aided diagnosis (CAD) systems lack sophisticated reasoning capabilities and contextual awareness necessary for comprehensive diagnostic support, particularly in handling diverse pathological presentations and clinical decision-making scenarios. The integration of Model Context Protocol (MCP)-based agentic AI frameworks with advanced contextual protocols represents a promising approach to address these limitations in clinical healthcare practice. DeepSpine employs a novel multi-agent conversational framework enhanced with MCP layer for dynamic context management and reasoning coordination. The system comprises specialized agents for segmentation, measurement extraction, pathology classification, and report generation, orchestrated through Reasoning cycles with action-observation-thought sequences. The MCP layer maintains persistent contextual awareness across agent interactions, enabling sophisticated reasoning chains and tool orchestration. The framework integrates ensemble segmentation models for vertebral and disc structure identification, coupled with quantitative measurement algorithms for spinal parameters. Retrieval-Augmented Generation (RAG) capabilities provide evidence-based responses through integrated vector databases and external medical literature sources. System evaluation utilized the DeepEval framework with LLM-as-a-judge methodology across task completion, contextual relevancy, and tool correctness metrics. Performance evaluation demonstrated superior task completion rates with GPT-4o achieving 96.6% compared to Gemma3 (83.3%) and Llama-3.2 (63.3%), while maintaining perfect tool correctness (100.0%) across all models. Contextual relevancy peaked with Gemma3 at 75.0%, indicating effective context understanding and response generation. Segmentation performance achieved optimal results on sagittal views with DSC of 96.43% and accuracy of 99.66%, substantially outperforming axial view analysis (DSC 92.71%, accuracy 99.88%). Report generation quality metrics revealed Gemma3 delivering superior performance across METEOR 20%, ROUGE-L 14%, and BERTScore-F1 84% evaluations, demonstrating enhanced natural language generation capabilities for clinical documentation. The framework's integration of multi-agent coordination, ensemble vision models, and persistent context management delivers accurate pathological assessment with clinically relevant diagnostic insights. The system's conversational interface and evidence-based reasoning support streamlined radiological workflows, offering substantial potential for clinical deployment and improved diagnostic accuracy in lumbar spine pathology evaluation.

Keywords

Agentic AI, Model Context Protocol (MCP), Lumbar MRI Analysis, Multi-agent Systems, Medical Imaging, Clinical Decision Support, Conversational AI, Retrieval-Augmented Generation (RAG).