Real-time Medical Imaging Interpretation System
System Architecture Document

Executive Summary

This document outlines the architecture for an AI-powered medical imaging interpretation system designed to provide real-time analysis of emergency CT and MRI scans. The system integrates Amazon HealthLake Imaging, Amazon OpenSearch, and IBM Granite LLM to deliver immediate, evidence-based clinical insights to emergency medical teams.

Problem Statement

Emergency departments face critical delays in interpreting CT and MRI scans, particularly during off-hours when radiology specialists may not be immediately available. These delays can significantly impact patient outcomes in time-sensitive conditions such as stroke, hemorrhage, or other life-threatening emergencies where every minute counts.

Solution Overview

The proposed system creates an automated pipeline that processes medical images immediately upon upload, analyzes them using pre-trained AI models, retrieves relevant clinical guidelines, and generates actionable summaries for medical professionals. This reduces diagnostic delays from hours to minutes while maintaining clinical accuracy and evidence-based recommendations.

System Architecture

Core Components

Amazon HealthLake Imaging
- Primary storage for DICOM-formatted CT and MRI scans
- FHIR-compliant healthcare data management
- Built-in HIPAA compliance and PHI protection
- Optimized for medical imaging workflows

AWS Lambda Functions
- Event-driven orchestration of the analysis pipeline
- Serverless architecture ensuring scalability and cost-effectiveness
- Automatic triggering upon new scan uploads
- Integration layer between all system components

Amazon SageMaker
- Hosts pre-trained medical imaging models

- Provides real-time inference capabilities
- Scalable model serving with automatic scaling
- Specialized models for different scan types and conditions

Amazon OpenSearch
- Vector database storing medical guidelines and protocols
- Implements Retrieval-Augmented Generation (RAG) capabilities
- Semantic search for relevant clinical information
- Real-time query processing for contextual recommendations

IBM Granite LLM (via watsonx.ai)
- Large language model optimized for medical text generation
- Processes imaging findings and guidelines
- Generates concise, clinically relevant summaries
- Provides treatment recommendations in plain language

Amazon SNS & API Gateway
- Notification services for immediate alert delivery
- API endpoints for hospital system integration
- Multi-channel communication (mobile, dashboard, email)
- Real-time status updates and result delivery

Data Flow Architecture

Stage 1: Image Ingestion
- CT/MRI scans uploaded to Amazon HealthLake Imaging
- DICOM metadata extraction and validation
- Automatic patient data anonymization for AI processing
- Trigger generation for downstream processing

Stage 2: Event Processing
- AWS Lambda detects new image uploads via CloudWatch Events
- Scan metadata validation and quality checks
- Workflow orchestration initiation
- Error handling and retry logic implementation

Stage 3: AI Analysis Pipeline
- Medical imaging model inference via SageMaker endpoints
- Parallel processing for multiple scan sequences
- Confidence scoring and uncertainty quantification
- Critical finding detection and prioritization

Stage 4: Knowledge Retrieval
- Semantic search in OpenSearch using detected findings

- Retrieval of relevant clinical guidelines and protocols
- Context-aware filtering based on patient demographics
- Evidence-based recommendation compilation

Stage 5: Clinical Summarization
- IBM Granite LLM processes findings and guidelines
- Generation of structured clinical reports
- Plain-language summary creation for rapid comprehension
- Treatment recommendation prioritization

Stage 6: Result Delivery
- Multi-channel notification via Amazon SNS
- Real-time dashboard updates through API Gateway
- Mobile application push notifications
- Integration with existing hospital information systems

Technical Specifications

Performance Requirements

Latency Targets
- Image processing initiation: < 30 seconds
- AI analysis completion: < 2 minutes
- Complete report generation: < 3 minutes
- Notification delivery: < 10 seconds

Throughput Specifications
- Concurrent scan processing: Up to 50 simultaneous analyses
- Daily scan capacity: 1,000+ scans per day
- Peak hour handling: 100 scans per hour
- Multi-hospital deployment scalability

Accuracy Standards
- Critical finding detection: > 90% sensitivity
- False positive rate: < 5%
- Guideline retrieval relevance: > 95%
- Clinical summary accuracy: Validated by medical professionals

Security and Compliance

Data Protection
- End-to-end encryption for all data transfers
- HIPAA-compliant data handling throughout pipeline
- Patient data anonymization for AI processing

- Audit logging for all system interactions

Access Control

- Role-based access control (RBAC) implementation
- Multi-factor authentication for system access
- API key management and rotation
- Network-level security with VPC isolation

Compliance Standards

- HIPAA compliance for healthcare data
- DICOM standard adherence for medical imaging
- HL7 FHIR compatibility for healthcare interoperability
- SOC 2 Type II compliance for cloud services

Integration Architecture

Hospital System Integration

Electronic Health Records (EHR)
- HL7 FHIR API integration
- Real-time patient data synchronization
- Clinical decision support system integration
- Automated report insertion into patient records

Picture Archiving and Communication System (PACS)
- DICOM protocol support
- Direct image retrieval and processing
- Automated workflow integration
- Radiologist review queue integration

Clinical Decision Support
- Real-time alert generation
- Prioritized notification delivery
- Clinical workflow integration
- Treatment protocol recommendations

External Service Integration

Third-party Medical Databases
- Integration with medical literature databases
- Real-time guideline updates
- Evidence-based recommendation validation
- Continuous knowledge base enhancement

Telemedicine Platforms
- Remote radiologist notification
- Virtual consultation facilitation
- Mobile application integration
- Multi-location hospital support

Deployment Architecture

Cloud Infrastructure

Multi-Region Deployment
- Primary region for main operations
- Secondary region for disaster recovery
- Cross-region data replication
- Load balancing and failover mechanisms

Auto-Scaling Configuration
- Lambda function concurrent execution limits
- SageMaker endpoint auto-scaling policies
- OpenSearch cluster scaling parameters
- SNS throughput optimization

Monitoring and Observability
- CloudWatch metrics and alarms
- Distributed tracing with AWS X-Ray
- Custom medical workflow monitoring
- Performance optimization recommendations

Development and Testing

CI/CD Pipeline
- Automated testing for medical AI models
- HIPAA-compliant development environments
- Staged deployment with medical validation
- Rollback mechanisms for critical systems

Quality Assurance
- Medical professional validation workflows
- Automated testing with synthetic medical data
- Performance benchmarking against clinical standards
- Regulatory compliance verification

Business Impact

Clinical Benefits

Improved Patient Outcomes
- 75% reduction in time-to-treatment for critical conditions
- 24/7 automated preliminary analysis availability
- Consistent application of evidence-based guidelines
- Enhanced decision-making support for emergency physicians

Operational Efficiency
- Reduced radiologist workload for routine screenings
- Prioritized attention to critical cases
- Streamlined emergency department workflows
- Cost reduction through automated initial analysis

Economic Value

Cost Savings
- Reduced overtime costs for emergency radiology
- Decreased patient length of stay
- Improved hospital throughput and capacity utilization
- Reduced medical liability through evidence-based recommendations

Revenue Enhancement
- Faster patient turnover in emergency departments
- Improved patient satisfaction scores
- Enhanced hospital reputation for emergency care
- Potential for new service line development

Implementation Roadmap

Phase 1: Foundation (Months 1-2)
- AWS infrastructure setup and configuration
- Basic DICOM ingestion and storage implementation
- Initial Lambda function development
- Security and compliance framework establishment

Phase 2: AI Integration (Months 3-4)
- SageMaker model deployment and testing
- OpenSearch setup and medical guideline indexing
- IBM Granite LLM integration and fine-tuning
- Initial end-to-end testing with synthetic data

Phase 3: Clinical Validation (Months 5-6)
- Medical professional validation workflows

- Clinical accuracy testing and optimization
- Integration with hospital test environments
- Regulatory compliance verification

Phase 4: Production Deployment (Months 7-8)
- Pilot deployment with select emergency departments
- Real-world performance monitoring and optimization
- Staff training and workflow integration
- Full-scale production rollout

Phase 5: Enhancement and Scaling (Months 9-12)
- Multi-hospital deployment
- Advanced AI model refinements
- Additional imaging modality support
- Continuous improvement based on clinical feedback

Risk Management

Technical Risks

AI Model Accuracy
- Continuous model validation and retraining
- Multi-model ensemble approaches
- Human oversight and validation workflows
- Performance monitoring and alerting

System Reliability
- Redundant architecture design
- Automated failover mechanisms
- Comprehensive backup and recovery procedures
- Regular disaster recovery testing

Regulatory Risks

FDA Compliance
- Medical device regulation adherence
- Clinical validation documentation
- Post-market surveillance requirements
- Quality management system implementation

Data Privacy
- HIPAA compliance monitoring
- Regular privacy impact assessments
- Incident response procedures

- Staff training and awareness programs

Conclusion

This real-time medical imaging interpretation system represents a significant advancement in emergency medical care technology. By combining cutting-edge AI capabilities with robust cloud infrastructure and evidence-based medical knowledge, the system addresses critical gaps in emergency radiology services.

The architecture provides a scalable, secure, and clinically validated solution that can dramatically improve patient outcomes while optimizing hospital operational efficiency. The phased implementation approach ensures proper validation and integration with existing clinical workflows, minimizing disruption while maximizing clinical benefit.

The system's design prioritizes patient safety, clinical accuracy, and regulatory compliance while delivering the speed and efficiency required in emergency medical situations. This positions it as a transformative solution for modern healthcare delivery in emergency settings.