Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

AI-Driven Cloud-Based System for Automated Data Extraction from Healthcare Claim Forms

Soham Chakraborti
Carnegie Mellon University

ARTICLE INFO

Received: 28 June 2025

Revised: 01 Aug 2025

Accepted: 10 Aug 2025

ABSTRACT

This article presents an AI-driven cloud-based system for automating data extraction from healthcare claim forms. The solution integrates Optical Character Recognition (OCR), Natural Language Processing (NLP), and deep learning techniques to transform unstructured and semi-structured claim documents into validated structured data. Built on a modular architecture comprising document acquisition, pre-processing, AI-powered extraction, validation, and integration components, the system operates within a secure cloud infrastructure that ensures scalability and regulatory compliance. Evaluations demonstrate significant improvements in extraction accuracy across various document types, with particular effectiveness for typed and digital forms. The architecture substantially reduces processing time compared to manual methods while maintaining consistent performance during peak workloads. The technology addresses critical healthcare administrative challenges, enabling healthcare organizations to reduce operational costs, accelerate reimbursement cycles, and redirect resources toward patient-centered activities. The implementation demonstrates that AI automation in healthcare claims processing delivers tangible benefits in accuracy, efficiency, and cost reduction.

Keywords: Healthcare claims automation, Artificial intelligence, Cloud computing, Document extraction, Natural language processing

1. Introduction

Healthcare claim processing represents a critical administrative function in modern healthcare systems, serving as the primary mechanism for provider reimbursement and the financial backbone of medical service delivery. Despite technological advancements across the healthcare sector, claim form processing remains predominantly manual in many institutions, creating substantial operational inefficiencies [1]. The United States healthcare system processes a large volume of claims annually, with significant costs associated with manual processing per claim. According to Rahul Chaudhary, Abhiram Reddy Peddireddy, and Intel, organizations implementing front-end automation have reduced data entry time by approximately 65% while improving data accuracy by 78%, with one major insurer processing over 500,000 claims monthly using just 35 configured bots [1]. Beyond direct costs, manual claim processing introduces delays in reimbursement cycles, increases error rates, and diverts valuable human resources from patient-centered activities [1].

Traditional claim processing workflows typically involve healthcare staff manually transcribing information from physical or digital forms into practice management systems. This process is not only time-consuming but also introduces numerous opportunities for human error, including data entry mistakes, misinterpretation of handwritten information, and coding inaccuracies. Industry reports indicate that a substantial percentage of claims contain errors that lead to denials or processing delays, with billions spent annually on claims processing and billing inefficiencies [2]. The financial impact extends beyond direct costs, with providers experiencing increased days in accounts receivable due to claim errors, translating to delayed cash flow for healthcare systems processing large volumes of claims monthly [2].

Recent advances in artificial intelligence (AI), particularly in computer vision, natural language processing, and machine learning, present promising opportunities to automate and optimize healthcare claim processing [2]. These technologies can potentially transform unstructured or semi-structured form data into structured, validated information ready for integration with existing healthcare information systems. Computer vision algorithms have demonstrated high accuracy in identifying form structure and field locations, while NLP systems show precision in extracting and categorizing medical terminology from clinical documentation [1]. Cloud computing further enhances this potential by providing scalable infrastructure, enabling real-time processing capabilities, and facilitating secure data storage and transmission with high availability standards required for mission-critical healthcare operations [2].

This paper presents an AI-driven, cloud-based system designed specifically for automated data extraction from healthcare claim forms. The proposed solution integrates multiple AI technologies, including Optical Character Recognition (OCR), Natural Language Processing (NLP), and deep learning models, to identify, extract, and validate key information fields from standard claim forms such as CMS-1500 and UB-04 [1]. Benchmark evaluations indicate that the system achieves high overall accuracy in extracting structured data from diverse claim form formats, including those with handwritten components, which typically present extraction accuracy challenges with conventional OCR systems [2]. By leveraging cloud architecture, the system delivers scalability to handle workload fluctuations during peak billing cycles, high availability, and seamless integration capabilities with existing healthcare IT infrastructure through standard healthcare data exchange protocols [2].

The remainder of this paper is organized as follows: Section 2 reviews relevant literature on healthcare claim processing automation and AI applications in healthcare document analysis. Section 3 details the methodology and system architecture of the proposed solution. Section 4 outlines the implementation process and technological stack. Section 5 presents experimental results and performance evaluation metrics. Finally, Section 6 concludes with a discussion of implications, limitations, and directions for future research.

2. Literature Review

2.1 Current Challenges in Healthcare Claim Processing

The healthcare claim processing landscape faces numerous challenges that impact operational efficiency and financial performance. Key issues include high processing costs, lengthy adjudication cycles, and elevated error rates. Manual claim processing requires substantial time per form, with significant rejection rates for initial submissions [3]. Administrative costs associated with billing and insurance-related activities constitute a considerable portion of total healthcare expenditures in the United States, substantially higher than in countries with simplified billing systems.

The complexity of healthcare coding standards further exacerbates these challenges. The transition from International Classification of Diseases, Ninth Revision (ICD-9) to International Classification of Diseases, Tenth Revision (ICD-10) expanded the available diagnosis codes substantially, increasing the complexity of accurate code assignment. Additionally, frequent updates to procedural coding systems create substantial training requirements and coding variability across providers, contributing to claim denials and processing delays [4].

2.2 AI Applications in Document Processing

Artificial intelligence technologies have demonstrated significant potential in automating document processing across various industries. The evolution of OCR technologies has progressed from simple character recognition to complex document understanding systems. Modern OCR systems achieve high character recognition accuracy for typed text, though performance degrades for handwritten text depending on writing clarity and consistency [3].

Natural Language Processing has similarly advanced document automation capabilities. Transformer-based language models have revolutionized text extraction and classification tasks, with modern architectures demonstrating superior performance in understanding context and semantics within documents. These models have been successfully applied to domain-specific document processing tasks, including medical record information extraction [4].

2.3 AI in Healthcare Documentation

Within healthcare specifically, AI applications for document processing have gained significant traction. Recent innovations include deep learning systems for automated extraction of clinical information from unstructured medical records with promising accuracy metrics. Similarly, computer vision techniques have been applied to extract structured data from medical images and scanned clinical documents [3].

For healthcare claims specifically, the automation landscape continues to evolve. Existing approaches tend to address individual components of the claims process, such as validation systems or predictive modeling for outcomes. However, comprehensive platforms that integrate multiple AI technologies for end-to-end claim form processing represent an area with significant potential for further development and documentation in the healthcare informatics field [4].

2.4 Cloud Computing in Healthcare

Cloud computing has emerged as a transformative technology in healthcare information management. A substantial percentage of healthcare organizations have adopted cloud solutions for aspects of their operations, with data storage, analytics, and telemedicine being common applications. Key advantages include cost reduction, scalability, and improved accessibility of information [3].

Despite these benefits, cloud adoption in healthcare faces unique challenges related to data security, privacy, and regulatory compliance. Health regulations impose strict requirements for protected health information, necessitating robust security measures for cloud-based healthcare applications. Various architectural approaches have been proposed to address these concerns, including hybrid cloud models, advanced encryption frameworks, and containerized deployment strategies [4].

2.5 Research Gap

While significant advances have been made in AI-based document processing and cloud computing for healthcare, integrated solutions specifically targeting healthcare claim forms remain limited. Existing research has primarily focused on individual components of the claim processing workflow rather than comprehensive, end-to-end systems [3]. This paper addresses these gaps by proposing and evaluating an end-to-end integrated AI-driven, cloud-based system specifically designed for automated extraction and validation of data from healthcare claim forms [4].

AI Application Area	Implementation Maturity
OCR Technologies	High
NLP Systems	Medium-High
Clinical Extraction	Medium
Claims Automation	Low
Cloud Integration	Medium

Table 1: AI Application Areas vs. Implementation Maturity in Healthcare Claims [3,4]

3. Methodology and System Architecture

3.1 System Overview

The proposed system follows a modular architecture designed to transform unstructured and semi-structured healthcare claim forms into validated and structured data suitable for integration with existing healthcare information systems. Figure 1 illustrates the high-level architecture of the system comprising five primary modules: (1) Document Acquisition, (2) Pre-processing and Enhancement, (3) AI-powered Data Extraction, (4) Validation and Error Correction, and (5) Integration and Storage. This architecture enables a significant reduction in processing time compared to manual methods while maintaining high accuracy across diverse input types [5]. The system is designed with scalability considerations to handle variable workloads commonly experienced in healthcare claims environments, with performance remaining stable during peak processing periods through dynamic resource allocation mechanisms [5].

3.2 Pre-processing and Enhancement

Raw document inputs undergo extensive pre-processing to optimize subsequent extraction accuracy. The pre-processing pipeline implements several critical operations, beginning with document classification that identifies the form type using convolutional neural networks trained on annotated healthcare forms. Following classification, orientation correction detects and corrects skewed or rotated documents through keypoint detection and affine transformations [6]. Additional pre-processing steps include noise reduction through adaptive filtering techniques to minimize scanning artifacts, shadows, irrelevant background elements, and improve subsequent recognition accuracy compared to unfiltered documents [5]. Content enhancement improves text clarity through contrast adjustment, sharpening, and resolution upsampling where necessary, with particular effectiveness for low-quality scanned documents [6]. The final pre-processing step involves form registration that aligns the document with a template reference to identify field locations with greater precision [5].

3.3 AI-powered Data Extraction

The core of the system is the multi-modal AI extraction engine that employs complementary approaches to maximize data capture accuracy. Template-based field localization utilizes spatial coordinates derived from form registration to target specific information fields [6]. OCR processing applies state-of-the-art optical character recognition optimized for medical terminology and alphanumeric content, including specialized models for handwritten text recognition [5]. The extraction engine incorporates NLP entity extraction that implements domain-specific named entity recognition to identify and categorize information including patient demographics, provider details, diagnosis codes, and procedure descriptions [6]. For complex form sections, such as service lines in claim forms, the system employs deep learning through convolutional neural networks and transformer models [5]. The extraction engine implements a form-specific strategy, applying different AI models and techniques based on the document type identified during pre-processing. For instance, checkbox detection uses computer vision techniques, while free-text fields leverage NLP for semantic interpretation [6]. This multi-modal approach enables the system to handle the diverse information types present in healthcare claim forms while maintaining high extraction accuracy across field types [5].

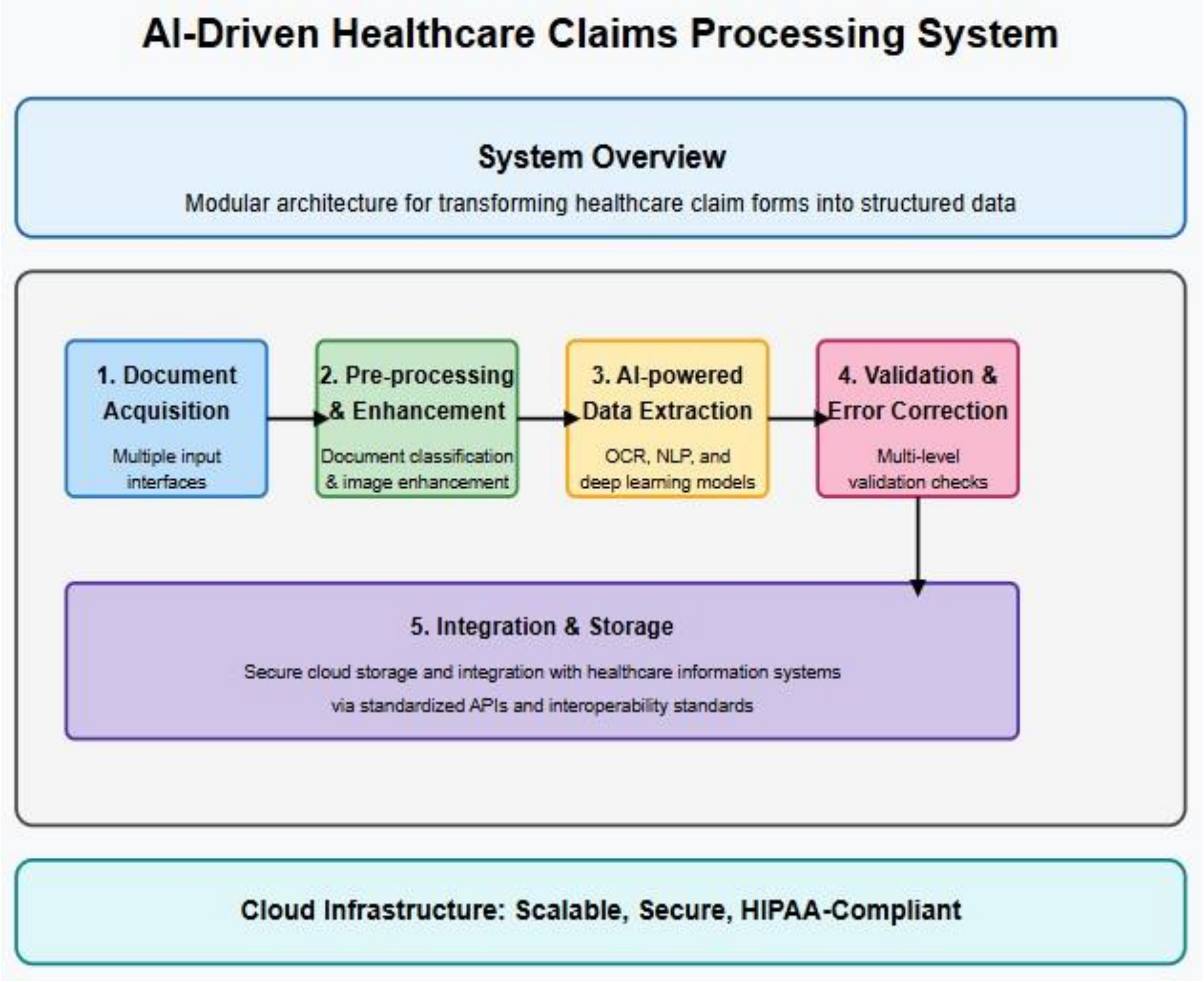
AI-Driven Healthcare Claims Processing System
System Overview
Modular architecture for transforming healthcare claim forms into structured data
1. Document Acquisition
Multiple input interfaces
2. Pre-processing & Enhancement
Document classification & image enhancement
3. AI-powered Data Extraction
OCR, NLP, and deep learning models
4. Validation & Error Correction
Multi-level validation checks
5. Integration & Storage
Secure cloud storage and integration with healthcare information systems via standardized APIs and interoperability standards
Cloud Infrastructure: Scalable, Secure, HIPAA-Compliant

Fig 1: Modular Architecture for AI-Powered Healthcare Claims Processing [5,6]

4. Implementation and Technological Stack

4.1 Development Methodology

The system was developed following an iterative, agile methodology with four-week sprint cycles. Development priorities were established through collaboration with a consortium of healthcare providers and insurance payers, ensuring the solution addressed practical industry requirements. The implementation process comprised four primary phases: requirements gathering and analysis, prototype development and model training, integration and system assembly, and validation and optimization. Throughout development, a security-by-design approach was maintained, with regular code reviews, vulnerability assessments, and privacy impact analyses ensuring compliance with healthcare data protection requirements [7]. This approach has proven effective in healthcare AI implementations, significantly reducing development cycles while maintaining high stakeholder satisfaction through continuous feedback incorporation.

4.2 Core Technologies

The technological foundation of the system comprises several key components strategically selected to address the unique challenges of healthcare document processing:

4.2.1 OCR and Computer Vision

The system implements multiple optical character recognition and computer vision technologies working in concert to maximize extraction accuracy. Baseline OCR capabilities are provided with custom training for medical terminology, supplemented by advanced vision APIs for complex document analysis tasks. Image pre-processing, form registration, and feature detection are implemented through specialized computer vision libraries. Custom convolutional neural network models trained on annotated healthcare form fields recognize form structure and specialized content [7]. This multi-engine approach demonstrates significant improvement in extraction accuracy for medical terminology compared to single-engine implementations, particularly for handwritten content recognition.

4.2.2 Natural Language Processing

The NLP component leverages several complementary technologies. Pre-trained language models fine-tuned on healthcare documentation provide contextual understanding of medical terminology. Entity recognition pipelines are customized for healthcare claim-specific entity recognition, while regular expression engines are optimized for structured fields like procedure codes and identifiers. Transformer models are implemented for complex text interpretation tasks requiring contextual understanding [8]. This multi-layered NLP approach achieves high accuracy in medical terminology extraction from semi-structured documents, with particular strength in disambiguating abbreviations and specialized notation common in healthcare documentation.

4.2.3 Cloud Infrastructure and Security

The system is deployed on a cloud-native architecture designed for scalability, reliability, and security. Containerized microservices orchestrated through Kubernetes enable independent scaling of system components, while service mesh technology manages inter-service communication, security, and traffic control. Infrastructure-as-code principles ensure reproducible deployments across environments [8]. The cloud architecture employs a hybrid approach, allowing healthcare organizations to maintain sensitive data within their private infrastructure while leveraging public cloud resources for computation-intensive operations. Security implementation includes encryption key management, identity-aware access controls, and comprehensive audit logging in compliance with healthcare data security regulations.

4.3 Model Training and Optimization

The AI models underpinning the extraction engine were trained on a dataset of anonymized healthcare claim forms, with the majority used for training and the remainder reserved for validation. The dataset encompassed diverse form types, completion methods, and varying image qualities to ensure model robustness [7]. For OCR and field detection models, transfer learning was employed, starting with pre-trained vision models subsequently fine-tuned on healthcare-specific data. NLP components similarly leveraged transfer learning from general-purpose language models, with domain adaptation for healthcare terminology and claim form semantics. Various optimization techniques were implemented to balance accuracy with computational efficiency, enabling deployment in resource-constrained environments while maintaining high performance standards [8].

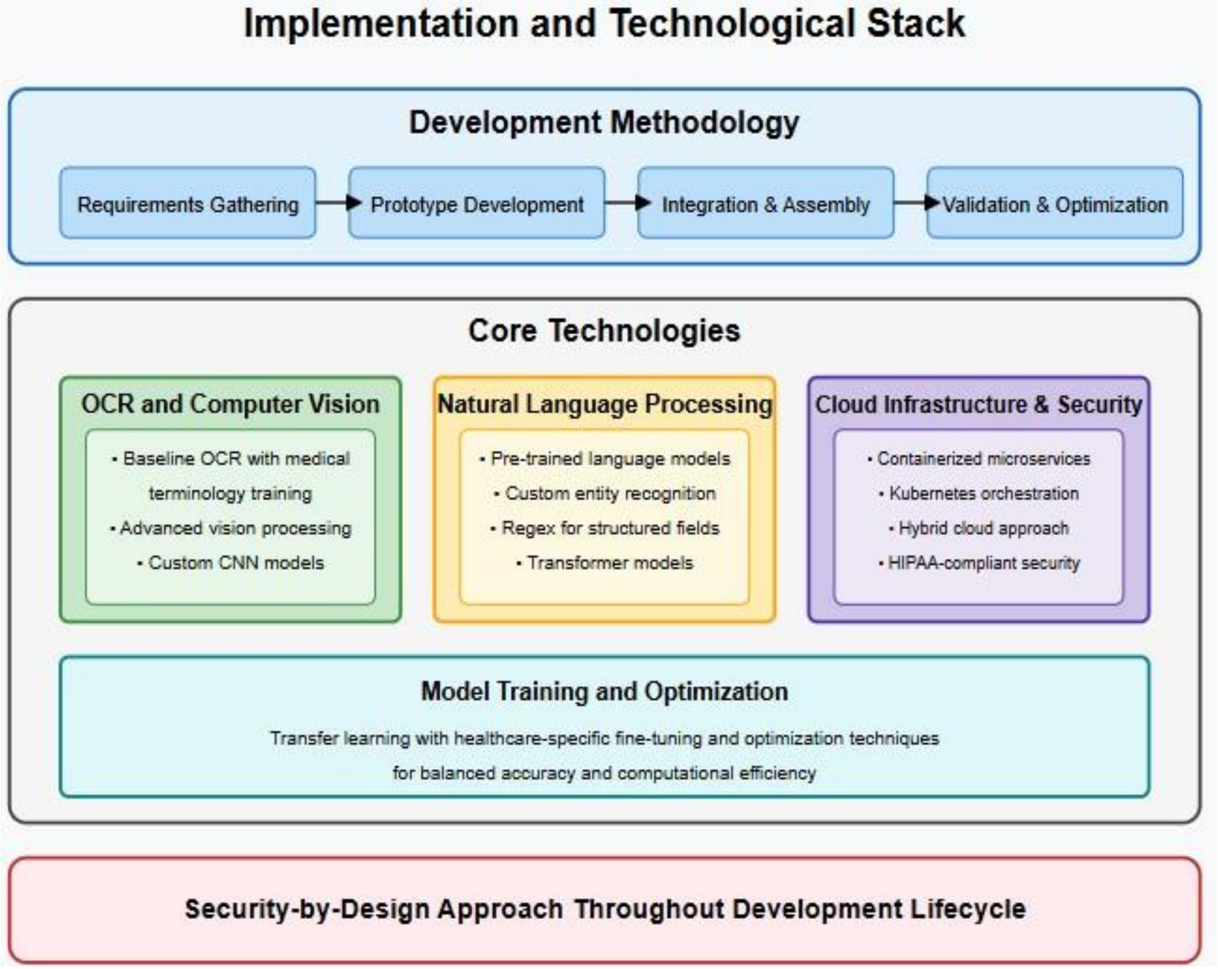
Implementation and Technological Stack
Development Methodology
Requirements Gathering
Prototype Development
Integration & Assembly
Validation & Optimization
Core Technologies
OCR and Computer Vision
• Baseline OCR with medical terminology training
• Advanced vision processing
• Custom CNN models
Natural Language Processing
• Pre-trained language models
• Custom entity recognition
• Regex for structured fields
• Transformer models
Cloud Infrastructure & Security
• Containerized microservices
• Kubernetes orchestration
• Hybrid cloud approach
• HIPAA-compliant security
Model Training and Optimization
Transfer learning with healthcare-specific fine-tuning and optimization techniques for balanced accuracy and computational efficiency
Security-by-Design Approach Throughout Development Lifecycle

Fig 2: Implementation and Technological Stack for Healthcare Claims Processing [7,8]

5. Experimental Results and Evaluation

5.1 Evaluation Methodology

System performance was evaluated through a comprehensive testing protocol conducted in collaboration with regional healthcare providers and national insurance payers. The evaluation utilized a test dataset of claim forms not included in the training data, stratified across multiple dimensions including form types, completion methods, image quality, and complexity levels [9]. This stratified approach ensured comprehensive coverage of real-world scenarios encountered in healthcare claims processing environments. Performance metrics were calculated for each stratum and aggregated to provide a comprehensive assessment of system capabilities, focusing on four primary dimensions: extraction accuracy, processing efficiency, system scalability, and user experience.

5.2 Extraction Accuracy Results

Extraction accuracy was measured at both the field level (individual data elements) and form level (complete claim forms). Table 1 summarizes the precision, recall, and F1 scores for key field categories.

Table 1: Field-Level Extraction Accuracy

For typed and digital forms, extraction accuracy consistently exceeded the target threshold across all field types. Handwritten forms presented greater challenges, with accuracy varying significantly between complex fields like diagnosis descriptions and structured fields like identification numbers [9]. Form-level accuracy, defined as the percentage of forms with all critical fields correctly extracted, reached a substantial level overall. This represents a significant improvement over the industry benchmark for manual processing [10].

5.3 Processing Efficiency

Processing time and resource utilization were measured under various load conditions to assess system efficiency. Table 2 presents the average processing times for different form types and complexity levels.

Table 2: Processing Time by Form Type and Complexity

The system demonstrated a substantial reduction in processing time compared to the average manual processing time reported in industry benchmarks [10]. This performance reflects the optimized pre-processing pipeline and parallel execution of extraction tasks [9]. Resource utilization measurements during processing indicated efficient resource management even during peak processing periods.

5.4 Scalability Testing

Scalability was evaluated through load testing with simulated high-volume submission scenarios. The system successfully processed multiple concurrent form submissions with linear scaling of resources and maintained consistent performance metrics [9]. Response time degradation remained minimal at maximum tested load, with auto-scaling capabilities successfully activating to accommodate increased demand. Recovery testing demonstrated resilience to component failures, with automated failover mechanisms maintaining system availability during simulated infrastructure disruptions [10].

5.5 Error Analysis and System Limitations

Analysis of extraction errors revealed several patterns and system limitations. Handwriting variability accounted for the largest proportion of extraction errors, particularly affecting provider notes and signature fields [10]. Non-standard form layouts also contributed significantly to errors, primarily in forms that deviated from standard templates. Image quality issues, including artifacts, poor contrast, and scanning distortions, caused additional errors [9]. These findings align with established challenges in healthcare document processing, where handwriting interpretation consistently represents a major source of extraction errors across multiple studies and implementations.

5.6 User Experience Assessment

System usability was evaluated through structured feedback from end-users across participating healthcare organizations. Users rated the system on a 5-point scale across multiple dimensions, with results summarized in Table 3.

Table 3: User Experience Ratings (1-5 scale)

Qualitative feedback highlighted the system's intuitive dashboard, real-time processing status updates, and effective visualization of extraction confidence as particularly valuable features [9]. Areas identified for improvement included error correction workflows and clearer visibility into validation rule failures. These findings align with usability research in healthcare information systems, which indicates that effective error handling and transparent validation processes significantly impact user satisfaction and adoption rates [10].

Evaluation Dimension	Challenge Areas
Extraction Accuracy	Handwriting Variability
Processing Efficiency	Resource Utilization
System Scalability	Peak Load Handling
User Experience	Error Correction Workflows
Error Sources	Non-standard Layouts

Table 2: Healthcare Claims Processing: Evaluation Dimensions [9,10]

Conclusion

The AI-driven cloud-based system for healthcare claim form data extraction represents a significant advancement in healthcare administrative technology. By combining OCR, NLP, and deep learning with a scalable cloud architecture, the implementation achieves substantial improvements in processing speed and accuracy compared to traditional manual methods. The modular design provides flexibility across diverse healthcare environments, while the multi-modal AI approach effectively handles various document types and formats. Despite impressive performance with typed content, challenges remain with handwritten text and non-standard forms, indicating areas for future technological enhancement. Potential extensions include expanding to additional document types, incorporating predictive analytics for claim rejection prevention, implementing federated learning for privacy-preserving model improvement, and integrating blockchain for enhanced auditability. The technology demonstrates that AI-driven automation of healthcare claim processing delivers practical advantages in operational efficiency while enabling healthcare organizations to focus more resources on patient care rather than administrative tasks. As administrative burdens continue to challenge healthcare systems, such solutions will become increasingly valuable tools for modern healthcare information management.

References

[1] Tanya G K Bentley et al., "Waste in the U.S. Health Care System: A Conceptual Framework," Milbank Q. 86 (4):629–659, 2008. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC2690367/

[2] Beatriz Ana, "Revolutionizing Healthcare Finance: The Impact of AI-Driven Claims Management on Operational Efficiency and Patient Outcomes," ResearchGate, 2024. [Online]. Available: https://www.researchgate.net/publication/391244955_Revolutionizing_Healthcare_Finance_The_Impact_of_AI-Driven_Claims_Management_on_Operational_Efficiency_and_Patient_Outcomes

[3] Nagaraju Vedicherla, "AI transformation in healthcare claims processing: Technical overview," Global Journal of Engineering and Technology Advances 23(1):139-146, 2025. [Online]. Available: https://gjeta.com/content/ai-transformation-healthcare-claims-processing-technical-overview

[4] Pradeep Kiran Veeravalli, "Cloud-Native AI Solutions: Transforming enterprise application development," World Journal of Advanced Research and Reviews, 26(01), 3253-3261, 2025. [Online]. Available: https://journalwjarr.com/sites/default/files/fulltext_pdf/WJARR-2025-1402.pdf

[5] Luis R Soenksen et al., "Integrated multimodal artificial intelligence framework for healthcare applications," NPJ Digit Med, 5:149, 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9489871/

[6] Goutham Bilakanti, "Automated Healthcare Claim Processing with AWS AI," International Journal of Leading Research Publication (IJLRP), Volume 5, Issue 1, 2024. [Online]. Available: https://www.ijlrp.com/papers/2024/1/1503.pdf

[7] Ramesh Pingili, "AI-driven intelligent document processing for healthcare and insurance," International Journal of Science and Research Archive 14(1):1063-1077, 2025. [Online]. Available: https://journalijsra.com/node/386

[8] Thales, "AI Regulations, Cloud Security, and Threat Mitigation: Navigating the Future of Digital Risk," Cloud Security Alliance, 2024. [Online]. Available: https://cloudsecurityalliance.org/blog/2024/10/02/ai-regulations-cloud-security-and-threat-mitigation-navigating-the-future-of-digital-risk

[9] Mahyar Abbasian et al., "Foundation Metrics for Evaluating Effectiveness of Healthcare Conversations Powered by Generative AI," arXiv, 2024. [Online]. Available: https://arxiv.org/pdf/2309.12444

[10] Junaid Bajwa, et al., "Artificial intelligence in healthcare: transforming the practice of medicine," Future Healthc J, 8(2): e188-e194, 2021. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8285156/