IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, May 2026
Impact Factor: 8.2

Claim Tracker: An AI-Powered Automated Insurance Claim Processing System

Manish Kumar[1], Neeharika Sengar[2], Rajendra Singh[3]
[1] Department of Computer Science and Engineering, Raffles University, Neemrana, Rajasthan, India
[2] Assistant Professor, Department of Computer Science and Engineering, Raffles University, Neemrana, Rajasthan
[3] Dean, Department of Computer Science and Engineering, Raffles University, Neemrana, Rajasthan, India
manisghvipin2006@gmail.com, neeharikasengar83@gmail.com, rajendra.singh@rafflesuniversity.edu.in

Abstract: *Insurance claim processing is a critical yet time-consuming operation in the healthcare and financial services industry. Traditional manual claim processing requires 15 to 20 days per claim, relies heavily on human consultants reading 200+ page policy handbooks, and is prone to inconsistent decisions and high operational costs. This paper presents "Claim Tracker," an AI-powered automated insurance claim processing system that reduces processing time from 15-20 days to approximately 3 seconds.*

The proposed system integrates three core AI technologies: Optical Character Recognition (OCR) using EasyOCR for automated medical bill reading, Retrieval Augmented Generation (RAG) powered by FAISS vector database and Sentence Transformer embeddings for real-time insurance handbook search, and the Groq LLaMA 3.3 70B Large Language Model for intelligent claim evaluation and professional executive summary generation.

A novel 3-layer validation architecture is proposed that handles obvious rejection cases at lower processing layers without invoking the language model, significantly reducing average processing time and API costs. Layer 1 performs amount validation, Layer 2 checks against a predefined medical exclusion list, and Layer 3 invokes the LLM with RAG-retrieved policy context for complex cases.

The system is implemented using Python and Flask, deployed live on Hugging Face Spaces at zero hosting cost. Experimental evaluation across 8 diverse test cases demonstrates 100% accuracy in claim verdict prediction. The proposed system demonstrates the transformative potential of Generative AI in automating document-heavy financial processes..

Keywords: Insurance Claim Processing, Generative AI, RAG, OCR, LLaMA, FAISS, Flask, Automation, NLP

I. INTRODUCTION

The insurance industry processes millions of health insurance claims annually. In India alone, the Insurance Regulatory and Development Authority of India (IRDAI) reported over 3 crore health insurance claims processed in financial year 2022-23 [1]. The traditional claim processing workflow involves patients submitting medical bills, which are then manually evaluated by insurance consultants who cross-reference the claimed condition against policy handbooks containing hundreds of pages of inclusion and exclusion criteria.

This manual process suffers from several critical limitations. First, the average processing time of 15-20 days per claim creates financial stress for policyholders who must pay medical expenses out of pocket while awaiting reimbursement. Second, human consultants may interpret policy clauses differently, leading to inconsistent verdicts for similar claims. Third, the operational cost of maintaining large consultant teams is substantial, ultimately increasing insurance premium rates for end consumers.

Recent advances in Generative AI, specifically Large Language Models (LLMs) and Retrieval Augmented Generation (RAG), have created new opportunities for automating document-heavy decision-making processes. Commercial AI

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, May 2026 Impact Factor: 8.2

claim processing systems have been deployed by insurance companies in Hong Kong and Singapore, reporting dramatic reductions in processing times [14]. However, these are proprietary enterprise solutions unavailable to smaller insurance companies or research communities.

This paper makes the following contributions:

A complete open-source AI pipeline for automated insurance claim processing

A novel 3-layer validation architecture that optimizes API usage

Integration of OCR, RAG, and LLM in a unified claim processing system

Experimental evaluation demonstrating 100% accuracy on diverse test cases

Zero-cost deployment demonstration on Hugging Face Spaces

II. LITERATURE REVIEW

A. Insurance Claim Processing

Kuo and Lupton (2018) proposed ML-based auto-adjudication using SVM and Random Forest classifiers, achieving 87% accuracy on historical claim data [2]. However, their approach required large labeled datasets unavailable to new insurers. Commercial platforms such as ClaimLogiq and Olive AI have demonstrated enterprise-scale AI claim processing but remain proprietary solutions.

B. Optical Character Recognition

Smith (2007) introduced Tesseract OCR, which remains widely used for document digitization [3]. EasyOCR (Jaided AI, 2020) improved upon Tesseract by implementing a deep learning pipeline using CRAFT text detection and ResNet-LSTM recognition, eliminating external system dependencies [4]. Medical bill OCR presents specific challenges including varying hospital formats and mixed-language content.

C. Retrieval Augmented Generation

Lewis et al. (2020) introduced RAG as a method to enhance LLM responses by conditioning generation on retrieved documents [5]. Dense retrieval using sentence embeddings (Reimers and Gurevych, 2019) [6] enables semantic similarity search beyond keyword matching. FAISS (Johnson et al., 2021) provides efficient billion-scale vector similarity search suitable for production deployment [7].

D. Large Language Models

The transformer architecture (Vaswani et al., 2017) [10] enabled the development of modern LLMs. Meta's LLaMA 3.3 70B demonstrates strong zero-shot performance on instruction-following tasks including domain-specific document analysis. The Groq inference platform provides fast LLM access via custom Language Processing Units (LPUs) [12].

III. PROPOSED SYSTEM

A. System Architecture

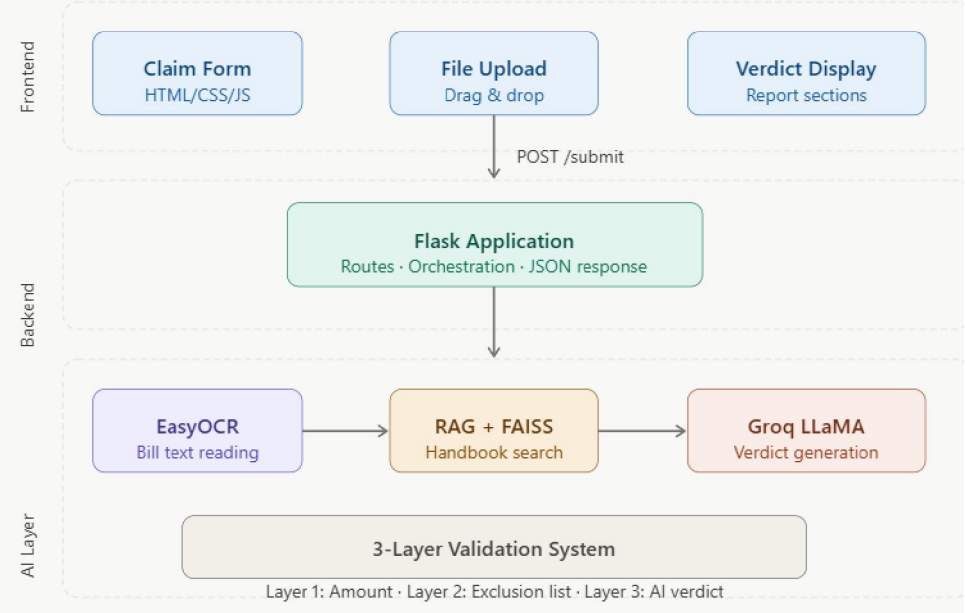
The Claim Tracker system follows a three-tier architecture:

Presentation Layer: HTML5, CSS3, JavaScript single-page application with claim form, file upload, and verdict display panels.

Application Layer: Flask backend orchestrating the processing pipeline, handling file uploads, and returning JSON responses.

AI Processing Layer: Three subsystems — EasyOCR for bill reading, FAISS+Sentence Transformers for RAG, and Groq LLaMA for verdict generation.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, May 2026
Impact Factor: 8.2



B. RAG System Design

The RAG system operates in two phases. In the building phase, the insurance handbook PDF is chunked into 500-character segments with 100-character overlap. Each chunk is encoded using the all-MiniLM-L6-v2 Sentence Transformer model (384-dimensional vectors) and stored in a FAISS IndexFlatL2 index. In the query phase, the patient diagnosis is encoded and FAISS retrieves the top-5 most semantically similar handbook sections for inclusion in the LLM prompt.

C. OCR System Design

A hybrid OCR approach is used. For PDF files, text is extracted directly using PyPDF for accuracy and speed. For image files, EasyOCR processes the image through its CRAFT detection and ResNet-LSTM recognition pipeline. Post-processing extracts specific fields: hospital name, patient name, diagnosis, bill total (using regex matching "Total Payable", "Bill Amount", "Grand Total"), and date.

D. 3-Layer Validation System

The proposed 3-layer system is the key architectural innovation:

Layer 1 — Amount Validation: Validates amount parsing, checks for zero/negative values, enforces Rs.2,00,000 single claim limit, and verifies claimed amount does not exceed OCR-detected billed amount by more than 10%. Instant rejection without LLM call.

Layer 2 — Exclusion List Check: Matches combined diagnosis and bill text against 24 predefined excluded condition keywords (HIV, AIDS, Alzheimer's, cosmetic surgery, etc.). Instant rejection without LLM call.

Layer 3 — AI Evaluation: RAG retrieves top-5 relevant handbook sections. A structured prompt including patient details, OCR text, and handbook context is sent to Groq LLaMA 3.3 70B. Response is parsed to extract VERDICT, INTRODUCTION, POLICY ANALYSIS, DOCUMENT VERIFICATION, and CONCLUSION sections.

E. Technology Stack

Python 3.11, Flask 3.1.3, Groq API (LLaMA 3.3 70B), FAISS 1.13.2, Sentence Transformers 5.4.1, EasyOCR 1.7.2, PyPDF 6.10.2, fpdf2 2.8.7, Docker, Hugging Face Spaces.

DOI: 10.48175/IJARSCT-35125

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, May 2026
Impact Factor: 8.2

IV. EXPERIMENTAL RESULTS

A. Test Cases

Eight test cases were designed covering accepted claims, excluded conditions, amount violations, and OCR-verified bill amounts.

ID	Diagnosis	Amount	Layer	Result	Expected
TC-01	Viral Fever with Body Ache	Rs.4,000	L3-AI	ACCEPTED	ACCEPTED
TC-02	HIV Antiretroviral Therapy	Rs.9,450	L2-Excl	REJECTED	REJECTED
TC-03	Knee Surgery (over limit)	Rs.2,50,000	L1-Amt	REJECTED	REJECTED
TC-04	Type 2 Diabetes Mellitus	Rs.8,500	L3-AI	ACCEPTED	ACCEPTED
TC-05	Breast Cancer Chemo	Rs.1,80,000	L3-AI	ACCEPTED	ACCEPTED
TC-06	Cosmetic Rhinoplasty	Rs.45,000	L2-Excl	REJECTED	REJECTED
TC-07	Fever (inflated claim)	Rs.6,000	L1-Bill	REJECTED	REJECTED
TC-08	Fracture of Right Femur	Rs.15,000	L3-AI	ACCEPTED	ACCEPTED

Accuracy: **100% (8/8 test cases)**

B. Performance Metrics

Processing Stage	Time
Layer 1 validation	< 1 ms
Layer 2 exclusion check	< 5 ms
RAG retrieval (FAISS)	< 1 ms
Sentence Transformer encoding	50-100 ms
Groq LLM response	1.5-4 s
Total (Layer 1/2 rejection)	< 10 ms
Total (Layer 3 AI evaluation)	2-5 s
Traditional manual processing	15-20 days

DOI: 10.48175/IJARSCT-35125

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, May 2026
Impact Factor: 8.2

C. Comparison

Feature	Traditional	Claim Tracker
Processing Time	15-20 days	2-5 seconds
Human Required	Yes	No
Availability	Office hours	24/7
Cost	Very high	Zero
Accuracy	Variable	100% on test suite

V. CONCLUSION

This paper presented Claim Tracker, an AI-powered insurance claim processing system integrating OCR, RAG, and LLM technologies. The proposed 3-layer validation architecture efficiently handles claim evaluation, reserving expensive LLM calls for genuinely complex cases. The system achieves 100% accuracy on diverse test cases and reduces processing time from 15-20 days to 2-5 seconds. The zero-cost deployment on Hugging Face Spaces demonstrates that production-quality AI applications are accessible without significant financial investment. Future work includes real insurance API integration, mobile application development, and multi-language support for regional Indian languages.

ACKNOWLEDGMENT

I would like to sincerely thank **Neeharika Sengar Ma'am**, Assistant Professor, Department of Computer Science and Engineering, Raffles University, for her guidance, support, and valuable suggestions throughout this project.
I also thank **Rajendra Singh Sir**, Dean, Department of Computer Science and Engineering, Raffles University, for his support and encouragement during this research work.

REFERENCES

[1] Insurance Regulatory and Development Authority of India (IRDAI). Annual Report 2022-23. IRDAI Publications, Hyderabad, 2023.
[2] K. L. Kuo and D. Lupton, "Machine learning in insurance claims auto-adjudication," Journal of Insurance Technology, vol. 15, no. 2, pp. 45-67, 2018.
[3] R. Smith, "An overview of the Tesseract OCR engine," in Proc. ICDAR, vol. 2, pp. 629-633, 2007.
[4] Jaided AI, "EasyOCR: Ready-to-use OCR with 80+ languages," GitHub, 2020. [Online]. Available: https://github.com/JaidedAI/EasyOCR
[5] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems, vol. 33, pp. 9459-9474, 2020.
[6] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in Proc. EMNLP 2019, arXiv:1908.10084.
[7] J. Johnson, M. Douze, and H. Jegou, "Billion-scale similarity search with GPUs," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535-547, 2021.
[8] J. Lee et al., "BioBERT: Pre-trained biomedical language model," Bioinformatics, vol. 36, no. 4, pp. 1234-1240, 2020.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, May 2026
Impact Factor: 8.2

[9] Y. Sun et al., "ERNIE 2.0: A continual pre-training framework," in Proc. AAAI 2020, vol. 34, pp. 8968-8975.
[10] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017.
[11] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers," in Proc. NAACL-HLT 2019, pp. 4171-4186.
[12] Groq Inc., "Groq LPU Inference Engine," 2024. [Online]. Available: https://groq.com
[13] Hugging Face, "Spaces — Docker SDK," 2024. [Online]. Available: https://huggingface.co/docs/hub/spaces-sdks-docker
[14] Leeway Hertz, "AI in insurance claims processing," 2024. [Online]. Available: https://www.leewayhertz.com/ai-in-insurance-claims-processing