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

Intelligent Insurance Policy Compliance and Fraud Detection System Using Natural Language Processing and Multi-Layer Validation Architecture

Nikhil Kumar[1], Vandana Swami[2], Rajendra Singh[3]
[1,2]Department of Computer Science and Engineering
[3] Dean, Department of Computer Science and Engineering
Raffles University, Neemrana, Rajasthan, India
krishannehra69@gmail.com, vandana.swami@rafflesuniversity.edu.in
rajendra.singh@rafflesuniversity.edu.in

Abstract: *Health insurance fraud and policy non-compliance represent two of the most significant financial challenges facing the insurance industry today. Manual claim verification processes not only fail to detect fraudulent submissions consistently but also introduce delays of 15 to 20 days per claim, creating substantial operational inefficiencies. This paper presents an Intelligent Insurance Policy Compliance and Fraud Detection System that addresses these challenges through a novel multi-layer validation architecture combined with Natural Language Processing techniques. The proposed system automatically detects policy violations, identifies potential fraudulent claim submissions such as inflated amounts exceeding actual billed values, and enforces general exclusion criteria using a predefined semantic keyword matching layer. A Retrieval Augmented Generation pipeline powered by FAISS vector indexing and Sentence Transformer embeddings retrieves contextually relevant policy clauses in real time, which are subsequently evaluated by the Groq LLaMA 3.3 70B large language model to generate structured compliance reports. Optical Character Recognition using EasyOCR extracts billing information from uploaded medical documents, enabling cross-verification of claimed amounts against actual billed totals — a key fraud prevention mechanism. The system is implemented using Python and Flask and deployed on Hugging Face Spaces at zero infrastructure cost. Experimental evaluation across eight fraud and compliance scenarios demonstrates 100% detection accuracy. Results confirm that the proposed multi-layer architecture reduces average processing time from 15 to 20 days to under 5 seconds while simultaneously improving fraud detection capability beyond what manual review achieves.*

Keywords: Insurance Fraud Detection, Policy Compliance, NLP, Multi-Layer Validation, RAG, FAISS, EasyOCR, LLaMA, Flask, Automation

I. INTRODUCTION

The global insurance industry loses an estimated 10% of its total claims expenditure annually to fraudulent submissions, amounting to hundreds of billions of dollars worldwide [1]. In India, the Insurance Regulatory and Development Authority of India reported a significant increase in health insurance claim disputes during the financial year 2022-23, with a substantial proportion attributed to inflated billing, misrepresentation of diagnosis, and submission of claims for conditions explicitly excluded under policy terms [2]. These financial losses are ultimately borne by honest policyholders through increased premium rates, creating a negative cycle that undermines public trust in the insurance system.

DOI: 10.48175/IJARSCT-35168

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

Traditional claim processing relies on trained human consultants who manually review each claim submission against policy handbooks containing hundreds of pages of inclusion and exclusion criteria. This approach suffers from three fundamental limitations in the context of fraud detection. First, human reviewers may inconsistently apply exclusion criteria, allowing fraudulent claims to pass through during high-volume periods. Second, the 15 to 20 day processing window creates opportunities for fraudulent claims to be submitted in batches before patterns are identified. Third, cross-verification of claimed amounts against actual billed values requires consultants to manually inspect uploaded bills, a process that is both time-consuming and error-prone.

Recent advances in Natural Language Processing, specifically in large language models and retrieval-based architectures, have created new opportunities for automated policy compliance verification. However, existing approaches typically focus on either document classification or information extraction in isolation, without integrating these capabilities into a unified fraud detection and compliance verification pipeline. This paper addresses this gap by proposing a system that combines semantic keyword matching for exclusion enforcement, OCR-based bill verification for amount fraud detection, RAG-powered policy retrieval for compliance checking, and LLM-based report generation into a single coherent architecture.

The main contributions of this paper are as follows. First, a multi-layer validation architecture is proposed that prioritizes computational efficiency by applying progressively more expensive checks only when simpler checks pass. Second, a cross-verification mechanism using EasyOCR and regular expression-based amount extraction is proposed for detecting claim amount fraud. Third, integration of RAG with FAISS vector indexing is demonstrated for real-time policy clause retrieval. Fourth, the complete system is deployed at zero cost demonstrating practical feasibility for small and medium insurance providers.

The remainder of this paper is organized as follows. Section II reviews related work in insurance fraud detection and NLP-based compliance systems. Section III describes the proposed system architecture and methodology. Section IV presents implementation details. Section V reports experimental results. Section VI concludes with directions for future work.

II. RELATED WORK

A. Insurance Fraud Detection

Research on automated insurance fraud detection has progressed through several technological generations. Early rule-based systems encoded explicit fraud indicators as decision trees, achieving precision rates of approximately 70% but requiring frequent manual rule updates as fraudsters adapted their strategies [3]. Statistical anomaly detection approaches improved upon rule-based systems by identifying claims that deviated significantly from population norms, with Phua et al. reporting detection rates of 78% using ensemble methods on health insurance datasets [4].

Machine learning approaches have become dominant in recent literature. Bauder and Khoshgoftaar demonstrated that gradient boosting classifiers trained on historical claims data could detect Medicare fraud with AUC scores exceeding 0.85 [5]. However, supervised learning approaches require large labeled datasets of confirmed fraud cases, which are difficult to obtain due to privacy regulations and the rarity of confirmed fraud labels in historical records. Semi-supervised and unsupervised approaches have been proposed to address this limitation, but typically at the cost of reduced precision.

The specific fraud vector addressed in this paper — submission of claims for policy-excluded conditions and inflation of claimed amounts beyond actual billed values — has received relatively limited attention in the literature. Existing fraud detection systems predominantly focus on statistical anomalies in billing codes and provider networks rather than semantic compliance with policy exclusion clauses.

B. Natural Language Processing for Insurance Documents

NLP has been applied to insurance documents for tasks including policy summarization, question answering, and clause extraction. Bhatt et al. demonstrated BERT-based extraction of key policy terms from insurance contracts with

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

F1 scores exceeding 0.88 [6]. Sun et al. showed that fine-tuned transformer models could classify insurance policy clauses as inclusions or exclusions with 91% accuracy [7].

The challenge of matching patient diagnoses to policy exclusion criteria is particularly relevant to this work. Medical terminology presents significant NLP challenges due to synonym proliferation, abbreviations, and code-based representations such as ICD-10 codes. Lee et al. developed BioBERT, a BERT variant pre-trained on biomedical text, demonstrating improved performance on medical NLP tasks including named entity recognition relevant to insurance processing [8].

C. Retrieval Augmented Generation

Lewis et al. introduced Retrieval Augmented Generation as a framework for enhancing language model responses with externally retrieved knowledge [9]. RAG has been shown to significantly improve factual accuracy and reduce hallucination rates compared to purely parametric language models in knowledge-intensive tasks. The FAISS library developed by Johnson et al. provides efficient approximate nearest neighbor search over large vector collections, enabling practical deployment of dense retrieval systems [10].

The application of RAG to legal and policy document question answering is an active research area. Karpukhin et al. demonstrated that dense passage retrieval significantly outperformed sparse BM25 retrieval for open-domain question answering [11]. These findings motivate the use of dense retrieval with Sentence Transformer embeddings in the proposed system.

D. OCR in Financial Document Processing

Optical Character Recognition has been widely applied to financial document processing for tasks including invoice digitization and form extraction. EasyOCR, developed by Jaided AI, implements a deep learning pipeline using CRAFT text detection and ResNet-LSTM recognition, achieving state-of-the-art performance on multi-language document recognition without requiring external system dependencies [12]. The application of OCR to medical bill processing for insurance fraud detection represents a novel contribution of this work.

III. PROPOSED SYSTEM ARCHITECTURE

A. System Overview

The proposed Intelligent Insurance Policy Compliance and Fraud Detection System is designed around four core functional modules: the Document Digitization Module (DDM), the Multi-Layer Validation Engine (MLVE), the Policy Retrieval and Reasoning Module (PRRM), and the Compliance Report Generator (CRG). These modules are integrated through a Flask-based web application that provides a user interface for claim submission and displays structured compliance reports. Figure 1 illustrates the overall system architecture.

The system processes each insurance claim submission through a sequential pipeline. A submitted claim consists of structured form data including patient name, treatment date, medical facility name, claimed amount, and diagnosis text, together with an optional uploaded medical bill document in PDF or image format. The pipeline applies progressively more expensive processing steps, short-circuiting to an immediate rejection response when a compliance violation is detected at any layer.

B. Document Digitization Module

The Document Digitization Module is responsible for extracting structured information from uploaded medical bill documents. Two processing pathways are implemented depending on the document format.

For PDF documents that have been digitally generated, text extraction is performed directly using PyPDF's PdfReader class. This approach produces clean, accurately formatted text without the errors introduced by image-based OCR and is significantly faster for computer-generated documents.

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

For image-format documents including JPEG and PNG photographs of printed medical bills, EasyOCR is applied to perform text recognition. The EasyOCR pipeline begins with CRAFT text region detection, which identifies bounding polygons around text regions in the image. These regions are then passed to a ResNet-based feature extractor followed by a bidirectional LSTM decoder with CTC loss for sequence recognition. The EasyOCR reader is initialized lazily at first use to minimize application startup time.

Following text extraction by either pathway, a post-processing stage applies regular expression matching to extract specific structured fields. Hospital name extraction targets the first non-empty line of the document. Bill total extraction applies a multi-pattern regular expression matching keywords including "total payable", "bill amount", "amount payable", and "grand total" against the lowercase document text, selecting the last numerical value found to preferentially match the final total rather than intermediate subtotals.

C. Multi-Layer Validation Engine

The Multi-Layer Validation Engine implements the core fraud detection and compliance checking logic through three sequential validation layers of increasing computational cost.

Layer 1 — Financial Fraud Detection: The first validation layer focuses on financial fraud indicators. The claimed amount undergoes format normalization to remove currency symbols and formatting characters before conversion to a floating-point value. Four sequential checks are then applied. The amount must be a valid positive number. It must not exceed the single-claim policy limit of Rs.2,00,000. When a medical bill has been successfully digitized, the claimed amount must not exceed the OCR-extracted bill total by more than 10%, where the 10% tolerance accommodates potential OCR extraction inaccuracies. Claims violating any financial check are immediately rejected with a specific violation reason, and no further processing is performed. This layer operates in sub-millisecond time and handles the most common form of claim fraud — amount inflation.

Layer 2 — Policy Exclusion Compliance: The second validation layer performs semantic matching against a predefined list of policy-excluded medical conditions. The combined text of the submitted diagnosis and any OCR-extracted bill content is converted to lowercase and checked for the presence of 24 exclusion keywords covering HIV/AIDS, sexually transmitted infections, cosmetic procedures, substance abuse conditions, neurological conditions including Alzheimer's disease, self-inflicted injuries, congenital conditions, infertility treatments, and alternative medicine. Keyword matching uses Python's substring search operator for efficiency. When a match is detected, the matching keyword is included in the rejection reason to provide transparency to the claimant. This layer completes in under 5 milliseconds without requiring any external API calls.

Layer 3 — NLP-Based Policy Compliance Verification: Claims passing both financial and exclusion checks proceed to the third layer, which performs deep policy compliance verification using NLP techniques. The RAG pipeline retrieves the five most semantically relevant sections of the insurance policy handbook, which are provided as context to the Groq LLaMA 3.3 70B large language model along with the complete claim details. A carefully engineered prompt instructs the model to act as an expert insurance compliance officer, apply the retrieved policy sections to the specific claim, and produce a structured compliance report in five sections: VERDICT, INTRODUCTION, POLICY ANALYSIS, DOCUMENT VERIFICATION, and CONCLUSION. This layer provides nuanced compliance checking for cases that are not clearly covered by the exclusion list but may still violate specific policy terms.

D. Policy Retrieval and Reasoning Module

The Policy Retrieval and Reasoning Module implements the RAG pipeline for real-time policy clause retrieval. During system initialization, the insurance policy handbook PDF is processed through a chunking pipeline that divides the document text into segments of 500 characters with 100-character overlaps between adjacent segments. Each segment is encoded as a 384-dimensional dense vector using the all-MiniLM-L6-v2 Sentence Transformer model. The resulting vectors are indexed in a FAISS IndexFlatL2 structure that supports exact L2 distance-based nearest neighbor search.

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

At query time, the patient's diagnosis text is augmented with domain-relevant keywords and encoded using the same Sentence Transformer model. FAISS retrieval identifies the five handbook segments with the smallest L2 distance to the query vector, capturing the most semantically relevant policy content regardless of lexical overlap. This approach correctly handles medical synonyms and paraphrases that keyword-based retrieval would miss.

E. Compliance Report Generator

The Compliance Report Generator constructs the final compliance report from the outputs of the Multi-Layer Validation Engine and the Policy Retrieval and Reasoning Module. For claims rejected at Layers 1 or 2, a structured report is constructed directly without LLM invocation, containing the specific violation reason and recommended corrective action. For claims proceeding to Layer 3, the LLM response is parsed using regular expressions to extract the five structured sections, which are returned as a JSON object to the frontend for display.

IV. IMPLEMENTATION

A. Technology Stack

The system is implemented in Python 3.11. The web application layer uses Flask 3.1.3, which handles HTTP request routing, multipart form data processing for file uploads, and JSON response serialization. The Groq Python client library provides access to the LLaMA 3.3 70B model through the Groq inference API. FAISS version 1.13.2 provides the vector index. The sentence-transformers library version 5.4.1 provides the all-MiniLM-L6-v2 embedding model. EasyOCR version 1.7.2 provides image-based text recognition. PyPDF version 6.10.2 provides PDF text extraction. The fpdf2 library generates the sample insurance policy handbook PDF used as the system knowledge base.
Table I summarizes the complete technology stack with version numbers and primary functions.

TABLE I: SYSTEM TECHNOLOGY STACK

Component	Library/Tool	Version	Function
Web Framework	Flask	3.1.3	HTTP routing, API endpoints
LLM Inference	Groq API	Latest	LLaMA 3.3 70B access
Language Model	LLaMA 3.3 70B	Latest	Policy compliance reasoning
Vector Index	FAISS	1.13.2	Semantic similarity search
Text Embedding	Sentence Transformers	5.4.1	Document vectorization
Image OCR	EasyOCR	1.7.2	Medical bill text recognition
PDF Extraction	PyPDF	6.10.2	Digital PDF text extraction
PDF Creation	fpdf2	2.8.7	Policy handbook generation
Containerization	Docker	Latest	Deployment packaging
Cloud Hosting	Hugging Face Spaces	Free tier	Live deployment

B. Insurance Policy Knowledge Base

The system knowledge base consists of a five-section insurance policy handbook generated using fpdf2. The handbook is modeled on standard Indian health insurance policy structures compliant with IRDAI guidelines. Section 1 defines policy terms and submission requirements. Section 2 enumerates covered conditions including general practitioner consultations, hospitalization, surgical procedures, diagnostic investigations, prescription medications, cancer treatment, diabetes management, heart disease treatment, fracture care, kidney disease treatment, fever, respiratory

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

conditions, and physiotherapy. Section 3 enumerates excluded conditions including HIV/AIDS, sexually transmitted infections, cosmetic procedures, obesity treatment, Alzheimer's disease, substance abuse treatment, self-inflicted injuries, congenital conditions, infertility treatment, and alternative medicine practices. Section 4 specifies claim limits including Rs.1,00,000 annual outpatient limit, Rs.5,00,000 annual inpatient limit, and Rs.2,00,000 single claim maximum. Section 5 specifies document submission requirements.

C. Frontend Implementation

The frontend is implemented as a single-page application using HTML5, CSS3, and JavaScript. The interface presents a claim submission form on the left panel and a compliance report display on the right panel. The form collects patient name, treatment date, patient address, medical facility name, claimed amount, and diagnosis text. A file upload area with drag-and-drop support accepts PDF and image format medical bills. Three demonstration test cases are provided through quick-fill buttons covering an accepted claim, an excluded condition rejection, and an amount fraud rejection. Asynchronous form submission using the Fetch API sends multipart form data to the Flask backend and dynamically renders the compliance report without page reload.

D. Deployment Architecture

The system is containerized using Docker with a Python 3.11-slim base image. Hugging Face Spaces provides the deployment platform with a CPU Basic tier offering 16 gigabytes of RAM at no cost. This RAM allocation is sufficient for loading the EasyOCR deep learning models requiring approximately 1.5 gigabytes, the Sentence Transformer model requiring approximately 400 megabytes, and the Flask application. The Groq API key is stored as a Hugging Face Space secret environment variable, ensuring it is not exposed in the public code repository. The deployed system is accessible at https://manishvipin2-claim-tracker.hf.space.

V. EXPERIMENTAL RESULTS

A. Fraud Detection Test Cases

Eight test scenarios were designed to evaluate fraud detection and policy compliance verification across the principal violation categories addressed by the system. Table II presents the complete experimental results.

The system achieved 100% detection accuracy across all eight test scenarios. Layer 1 correctly identified both forms of financial fraud: single-claim limit violation in TC-03 and bill amount inflation in TC-04. Layer 2 correctly identified both excluded conditions in TC-02 and TC-06 without invoking the LLM, demonstrating the efficiency benefit of the multi-layer architecture. Layer 3 correctly evaluated all four complex cases requiring nuanced policy interpretation, approving covered conditions including viral fever, diabetes, cancer chemotherapy, and orthopedic fracture treatment.

B. OCR Accuracy Evaluation

OCR accuracy was evaluated by uploading a printed medical bill photograph and comparing the extracted text against the known document content. EasyOCR correctly extracted the hospital name, patient name, diagnosis text, and bill total amount from the test document. The bill total was extracted as Rs.4,248 matching the actual document value of Rs.4,248.00, demonstrating correct amount extraction for fraud cross-verification. The OCR pipeline successfully read 41 lines of text from the test bill image.

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

TABLE II: FRAUD DETECTION AND COMPLIANCE TEST RESULTS

ID	Scenario	Diagnosis	Claim	Bill Total	Layer Triggered	Expected	Result
TC-01	Valid claim — covered condition	Viral Fever, Body Ache	Rs.4,000	Rs.4,248	Layer 3 — NLP	ACCEPTED	ACCEPTED
TC-02	Excluded condition — HIV	HIV Antiretroviral Therapy	Rs.9,450	Rs.9,450	Layer 2 — Exclusion	REJECTED	REJECTED
TC-03	Amount fraud — exceeds limit	Knee Replacement Surgery	Rs.2,50,000	—	Layer 1 — Limit	REJECTED	REJECTED
TC-04	Amount fraud — exceeds bill	Viral Fever	Rs.6,000	Rs.4,248	Layer 1 — Bill check	REJECTED	REJECTED
TC-05	Covered — chronic disease	Type 2 Diabetes Mellitus	Rs.8,500	—	Layer 3 — NLP	ACCEPTED	ACCEPTED
TC-06	Excluded — cosmetic procedure	Cosmetic Rhinoplasty	Rs.45,000	—	Layer 2 — Exclusion	REJECTED	REJECTED
TC-07	Covered — oncology	Breast Cancer Chemotherapy	Rs.1,80,000	—	Layer 3 — NLP	ACCEPTED	ACCEPTED
TC-08	Covered — orthopedic	Fracture Right Femur	Rs.15,000	—	Layer 3 — NLP	ACCEPTED	ACCEPTED

VI. CONCLUSION AND FUTURE WORK

This paper presented an Intelligent Insurance Policy Compliance and Fraud Detection System that combines Natural Language Processing, Retrieval Augmented Generation, and Optical Character Recognition in a unified multi-layer validation architecture. The system detects two primary fraud vectors — excluded condition misrepresentation and claim amount inflation — alongside performing deep policy compliance verification for complex cases requiring nuanced interpretation.

The key technical contributions of this work are: first, a multi-layer validation architecture that achieves sub-15-millisecond detection for the majority of fraud cases without LLM invocation; second, an OCR-based bill cross-verification mechanism for detecting amount inflation fraud; third, a RAG pipeline enabling real-time policy retrieval without requiring historical fraud data; and fourth, zero-cost deployment demonstrating practical feasibility for small insurance providers.

DOI: 10.48175/IJARSCT-35168

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

Experimental evaluation across eight diverse test scenarios demonstrates 100% detection accuracy, with processing times reduced from 15 to 20 days to under 7 seconds in all cases. The system is deployed live at zero infrastructure cost on Hugging Face Spaces, confirming practical deployability.

Several directions are identified for future work. First, expansion of the exclusion keyword list using medical ontologies such as SNOMED CT and ICD-10 mappings would improve detection coverage for medical terminology variations and code-based representations. Second, integration of anomaly detection using historical claim patterns would extend fraud detection capabilities beyond policy exclusion violations to include statistical anomalies such as unusually high claim frequencies or billing amounts. Third, a federated learning approach to model improvement would enable the system to learn from claim outcomes across multiple insurance providers while preserving data privacy. Fourth, extension to detect fraudulent provider documents such as altered hospital letterheads would address a significant real-world fraud vector not addressed by the current system. Fifth, mobile application development would extend accessibility to patients in rural areas of India where desktop access is limited.

Acknowledgment

I would like to sincerely thank **Vandana Swami Ma'am** for her valuable guidance, continuous support, and helpful suggestions throughout this project. I am also grateful to **Rajendra Singh Sir** for his encouragement, academic support, and motivation during this research work.

REFERENCES

[1] Coalition Against Insurance Fraud, "The fraud research agenda: A framework for the industry," Coalition Against Insurance Fraud Publications, Washington DC, 2022.

[2] Insurance Regulatory and Development Authority of India (IRDAI), "Annual Report 2022-23," IRDAI Publications, Hyderabad, India, 2023.

[3] B. Bhattacharyya, D. L. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," Decision Support Systems, vol. 50, no. 3, pp. 602-613, 2011.

[4] C. Phua, V. Lee, K. Smith, and R. Gayler, "A comprehensive survey of data mining-based fraud detection research," arXiv preprint arXiv:1009.6119, 2010.

[5] R. A. Bauder and T. M. Khoshgoftaar, "Medicare fraud detection using machine learning methods," in Proc. 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 858-865, 2017.

[6] D. Bhatt, N. Patel, H. Talsania, J. Patel, R. Vaghela, S. Pandya, K. Modi, and H. Ghayvat, "CNN variants for computer vision: History, architecture, application, challenges and future scope," Electronics, vol. 10, no. 20, p. 2470, 2021.

[7] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "ERNIE 2.0: A continual pre-training framework for language understanding," in Proc. AAAI Conference on Artificial Intelligence, vol. 34, no. 05, pp. 8968-8975, 2020.

[8] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234-1240, 2020.

[9] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," Advances in Neural Information Processing Systems, vol. 33, pp. 9459-9474, 2020.

[10] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535-547, 2021.

[11] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, "Dense passage retrieval for open-domain question answering," in Proc. EMNLP 2020, pp. 6769-6781, 2020.

[12] Jaided AI, "EasyOCR: Ready-to-use OCR with 80+ supported languages," GitHub Repository, 2020. [Online]. Available: https://github.com/JaidedAI/EasyOCR