Clinical Decision Support System Using GenAI

Sayan Chakraborty, Sayantani Saha, Shreya Banerjee, and Suman Bhunia

Abstract This chapter demonstrates how to use Google's Gemini model to design and implement a Clinical Decision Support System (CDSS) driven by GenAI. The system allows clinical workflows to use multimodal reasoning that considers the context by combining text and image data, such as imaging studies and patient reports. Our framework uses Gemini's advanced language and vision capabilities to understand both structured and unstructured input, route queries through LangGraph logic, and call on domain-specific tools to give medically relevant answers without crossing ethical lines. The case study shows how GenAI improves the accuracy of decisions in areas like symptom triage, patient education, and workflow navigation, all while following strict healthcare safety rules. Implementation of key features include graph-based state transition, system prompt management, and measurement metrics to establish model performance in safety, tone, and medical accuracy. We also define challenges to applying generative models in clinical practice, e.g., handling ambiguity of patient input and aligning AI output with provider accountability. This ride provides grounds for bringing GenAI into intelligent healthcare systems—driving conversational agents from scripted responses to service-based adaptive decision-making.

Keywords CDSS · GenAI · LangGraph · LangChain · Gemini

S. Chakraborty · S. Saha (✉)
Department of Information Technology, Maulana Abul Kalam Azad University of Technology, Kolkata, West Bengal, India
e-mail: sayantani@gmail.com

S. Banerjee
Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India

S. Bhunia
Department of Computer Science and Software Engineering, Miami University, Oxford, OH, USA

S. Banerjee et al. (eds.), *AI in Smart and Secure Healthcare*, Springer Optimization and Its Applications 243, https://doi.org/10.1007/978-3-032-15092-9_17

1 Introduction

Generative Artificial Intelligence (GenAI) [1] in Clinical Decision Support Systems is modern processing and communication of healthcare information. This chapter explores the development and deployment of Clinical Decision Support Systems (CDSS) powered by GenAI from Google's Gemini model [2], with an emphasis on multimodal reasoning over text and vision clinical data. Our case study illustrates how Gemini's sophisticated language comprehension [3] and image analysis can be leveraged to process structured patient records and medical imaging in parallel, enabling real-time decision support. Our framework is primarily based on LangGraph [4], a graph-based platform that manages tool-based query management, state changes, and system prompting. This modular architecture supports accurate triage, patient counsel, and workflow exploration with ethical regulation through restraint of sensitive responses. We show the system's capability for generating medically appropriate intuitions without venturing into clinical boundaries through holistic implementation and evaluation procedures. We also underscored the practical difficulties in applying generative models, e.g., patient language vagueness, adherence to clinical responsibility, and cautionary measures in interaction. This chapter contributes to the existing literature in intelligent health systems by integrating technical innovation with ethical analysis. It reveals the way GenAI can transform CDSS from strict rule-based protocols to service-based, flexible clinical reasoning domains.

2 Overview of Clinical Decision Support Systems (CDSS)

Clinical Decision Support Systems (CDSS) [5] are highly sophisticated computerized systems designed to assist medical professionals in making more informed clinical decisions by utilizing patient information and medical knowledge. CDSS is designed to enhance diagnostic ability, rationalize treatment suggestions, prevent drug side effects, and enhance overall workflow within the clinical setting [6, 7].

Rule-based systems [8] are a standby of classical CDSS models running according to expressed logic statements: if A, then recommend B. Such rules are typically controlled by choosing between clinical guideline, experience, or expert opinion. Determinism and transparency of rule-based systems have made them widely acceptable in regulation settings where explainability is more important.

Nonetheless, even though helpful, rule-based CDSS also possess a few intrinsic drawbacks:

- Inflexibility: Inflexible rules never get a chance to practice dealing with complex or novel situations of patients.
- Bad Generalization: They don't handle incomplete or imprecise information—widespread in real clinical practice.
- Maintenance: Maintenance is required over time as medical guidelines change, with high degree of manual control required.

- Context Sensing: The programs can't enjoy the sense of subtlety of patient discussions or non-linear symptom improvement.
- Binary Logic: Probabilistic argumentation and subtle pattern recognition are typically engaged clinical judgment outside of the scope of formal logic trees.

With health care becoming more patient- and data-specific, such boundaries also point toward the necessity for increasingly adaptive, learning-type systems—like GenAI-based systems—that can offer contextually sensitive, adaptive decision-making far beyond the scope of rule-based systems.

3 Evolution Toward GenAI-Powered Clinical Decision Support Systems

To overcome the limitations of conventional rule-based CDSS, scientists and health-care start-ups are beginning to adopt Generative Artificial Intelligence (GenAI), a family of AI that can generate human-like content by finding patterns in vast datasets. Unlike fixed rule engines, GenAI models provide dynamic reasoning, context understanding, and multimodal data processing, enabling more nuanced interaction with patient data.

One of the most exciting GenAI models in this category is Google's Gemini. It combines state-of-the-art language modeling and computer vision so that the system can understand clinical stories as well as imaging reports such as X-rays or MRIs. This multimodality [9] is essential for real-world clinical deployment where decisions are normally based on both textual documentation as well as visual diagnosis.

Gemini integration within CDSS, particularly when harmonized via frameworks such as LangGraph, enables modular graph-based routing, tool calling, and ethical prompt management. The outcome is a system that:

- Recognizes ambiguous patient words.
- Processes varied data types (text and image).
- Answers medically appropriate, context-sensitive advice.
- Adheres to ethical guidelines and regulatory standards.
- GenAI and Gemini, taken together, are a revolution: shifting CDSS from being static, rule-driven reasoning to being dynamic, responsive, and service-oriented intelligence—facilitating new pathways for scalable, tailored healthcare delivery.

4 Generative AI (GenAI) in Clinical Decision Support

Generative Artificial Intelligence (GenAI) refers to a class of models that can produce original outputs—such as text, images, audio, or code—by learning patterns from extensive datasets. Unlike traditional discriminative models that classify input, GenAI systems generate responses based on context, semantics, and modality awareness. In the healthcare domain, GenAI enables sophisticated reasoning [10] by interpreting unstructured clinical text, diagnostic images, and patient interactions in real time. This adaptive capacity offers transformative potential for Clinical Decision Support Systems (CDSS), particularly when powered by models like Google's Gemini.

4.1 What Are AI Hallucinations?

Hallucinations [11] are when the AI produces information that is factually incorrect, does not exist, or is deceptive but presents it confidently.

Example:

- Citation fabrication.
- Misquoting scientific facts.
- Production of rational rather than accurate medical counsel.
- Reasons.
- Training on noisy or unvalidated data.
- Lack of external or real-time knowledge grounding formation.
- Over-reappearance on probabilistic pattern matching instead of factual corroboration.

4.2 What Are AI Biases?

Biases [12] are systematic AI response mistakes of prejudice or unbalanced representations from training data**.**

Types:

- Gender bias: Stereotypical reinforcement (e.g., woman-nurse connotations).
- Racial bias: Mislabeling or underrepresentation of specific ethnic groups.
- Cultural bias: Priming dominant cultural norms or languages.
- Non-representative or unbalanced training data.
- Algorithm design choices.
- Narrow diverse evaluation benchmarks.

Table 1 contains different strategies to mitigate the effects of biasness produced through the GenAI models.

Table 1 Different strategies to mitigate these issues

Strategy	Purpose
Data pre-processing	Cleans and balances training data to reduce bias
Retrieval-augmented generation (RAG)	Ground outputs in trusted sources to reduce hallucinations
Chain-of-thought prompting	Encourages step-by-step reasoning to expose flaws
Model evaluation	Tests outputs for factual accuracy and fairness
Temperature control	Adjusts randomness in generation to reduce hallucinations

4.3 *Adaptation of AI in CDSS System*

The introduction of AI has significantly changed medical services, from disease prediction to remote patient treatment. AI uses traditional machine learning techniques, deep learning techniques to structure the Electronic Health Records (EHR) data, analyze the insights of the data, augment different alert systems, provide different suggestions. The limitations with EHR data formats and different insights of the data, as shown in Table 2, have been reshaped in GenAI. The large-scale pre-trained model is capable enough to work with any form of healthcare data, with different languages and support fine-tuning to enhance the system's performance. Advances in GenAI support amalgamation of multiple technologies like Natural Language Processing (NLP) [3], Machine Learning (ML), Deep Learning (DL) to transform the healthcare services more patient-centric. Table 3 highlights the glimpses of new functionalities in healthcare using GenAI.

Table 2 Comparison of traditional AI versus generative AI in CDSS

Feature	Traditional AI	GenAI (e.g., Gemini)
Input type	Structured only	Structured + unstructured
Data modalities	Mostly text/numeric	Text, image, code, hybrid
Reasoning ability	Rule-based	Contextual + probabilistic
Adaptability	Static logic	Dynamic contextual learning
Real-time decision support	Limited	Highly adaptive
Patient interaction	Scripted	Natural language dialogue

Table 3 GenAI capabilities applicable to healthcare

GenAI feature	Clinical use case	Benefit
Text generation	Patient education, discharge summaries	Improves clarity and consistency
Image understanding	Radiology and dermatology triage	Accelerates image-based diagnosis
Question answering	Triage and symptom checkers	Provides fast, relevant support
Dialog management	AI-assisted consultations	Enables natural clinical chat
Prompt conditioning	Ethical and regulatory compliance	Prevents unauthorized guidance

5 CDSS Application Development with GenAI

Large language models (LLMs) [13, 14] are the fundamental building block of GenAI. Strong prompts [15] help the model to identify the tasks one by one. The proper instruction to the model and the context help GenAI model to generate a better bias-free output. Well-structured prompts help to ensure relevant output and better insights into the questions. LLM model requires better orchestration [16], the process by which LLMs are controlled with step-by-step instructions and better planning of the tasks.

A popular open-source framework to develop GenAI applications is discussed here.

5.1 LangGraph

LangGraph [17] is an open-source infrastructure that allows developers to implement dynamic, comprehensible chains of reasoning in terms of graph-based reasoning. Modular and flexible at its foundation, LangGraph supports tool-integrated, stateful, and user intent-adaptable conversational agents—making it quite suitable for clinical decision support applications.

5.1.1 Why LangGraph for CDSS?

Compared to standard static control flow pipelines, LangGraph features declarative node definition and conditionally controlled node switch: clinical agents can manage queries based on data modality, patient context, or ethics. The node-based methodology reflects clinicians' cognitive processes—moving through steps, conditionally branching, and handling subtle inputs by reproducible logic. Table 4 shows LangGraph features with description and benefits. Table 5 shows comparison of various

Table 4 LangGraph features for clinical decision support

Feature	Description	Benefit to CDSS
Stateful graph	Tracks history across nodes and sessions	Maintains continuity in patient chats
Conditional transitions	Routes based on query type or response content	Enables adaptive decision trees
Node modularity	Plug-and-play architecture for tools and prompts	Simplifies upgrades and audits
Tool integration	Native support for external tool calls	Connects to diagnosis and info agents
Recursion limit	Controls depth of conversation loops	Prevents runaway dialogue or ambiguity
Prompt governance	Secure prompt injection for ethical compliance	Ensures medical safety and control

Table 5 Node functions in LangGraph-based CDSS

Node name	Role in the system	Example invocation
Chatbot	Core model interface (Gemini)	Handles text and image inputs
Human	Captures further user input or exit intent	Prompts patient for clarification
Tool	Executes clinical functions (get_doctors())	Returns service-specific summaries
Evaluator	Scores chatbot output for safety and tone	Rates guidance against compliance

nodes with their roles that can be implemented for the CDSS system to get a better contextual output.

5.2 LangChain

LangChain [18] is an open-source platform that is intended to coordinate sophisticated interactions between large language models (LLMs) and third-party workflows, data, or tools. LangChain essentially offers abstractions to sequence components like memory, agents, prompts, tools, and retrievers and thus is a universal architecture for the construction of smart, domain-specific agents in healthcare settings. In Clinical Decision Support Systems (CDSS), LangChain facilitates contextual, modular development that is context-sensitive, traceable, and allows for unproblematic interaction with patient data repositories, ontology-based terminologies, and reasoning systems. Table 6 discusses different LangChain components related to CDSS.

LangChain allows developers to create adaptive, ethical, and clinically applicable GenAI workflows—a foundation technology for next-gen healthcare reasoning applications. Table 7 shows the limitations of traditional pipeline. Its LLM support, such

Table 6 LangChain components and their CDSS relevance

LangChain component	Description	Role in CDSS
Prompt templates	Define structured queries to LLMs	Create ethically bounded clinical prompts
Agents	Adaptive task managers that select tools or steps	Enable decision routing in clinical workflows
Memory	Tracks conversation context	Maintains patient session history
Tools	External APIs or functions	Connects to vital signs, services, diagnostics
Chains	Sequences of components	Orchestrates multistep reasoning
Retrievers	Semantic search from vector stores	Surface relevant patient or research data

Table 7 LangChain versus traditional pipelines in clinical AI

Feature	Traditional pipeline	LangChain framework
Hardcoded workflow logic	Yes	No—modular agents and chains
Context persistence	Limited	Persistent via memory
Multitool handling	Manual API integration	Automatic via agent orchestration
Prompt control	Basic templating	Dynamic, nested templates
Reusability across tasks	Low	High—plug-and-play modules

as Gemini, and its composability ethos meet clinical needs for personalization, regulation, and scale simplifies meaning interpretation, maintains context, and enables ethical control—ideal for creating robust CDSS applications.

6 Literature Survey

GenAI architecture refers to the system structure that enables GenAI models to learn, process, and generate data. Different neural networks, data pipelines, and algorithms work together to generate the output with realistic and human-like data or multimodal data. There are multiple generative models with their own sets of uses and characteristics. Different models include Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), transformers, etc. Following subsections describe different healthcare services with different generative models.

I. **Transformer-Based Diagnostic and Predictive Models**

Studies employing transformer architectures or variants for diagnosis or prediction. Healthcare data analysis contextual output generation is an important aspect in CDSS system. Author in [19] used BERT model with covid data for clinical text classification with nearly 85% accuracy. The paper [20] highlights the limitations of using only

BERT-based model for EHR data classification. Their hybrid model [21] of BERT, Particle Swarm Optimization and KNN helps to outperform BERT system. However, the study lacked real-time predictive capabilities and multimodal integration.

Researchers also work with BioBERT [22, 23] which is a domain-specific variant of BERT model. BioBERT for Clinical NLP developed a BioBERT-based system [24] that improved disease concept recognition in clinical notes with 89% accuracy compared to traditional ML model. It enables diagnosis support for unstructured EHRs, though it remains limited to terminology extraction rather than full diagnostic support.

GPT-3 [25] diagnostic capabilities demonstrated GPT-3's remarkable diagnostic accuracy (87% concordance with physicians) in simulated cases, with medical knowledge-enhanced prompt with clinical notes helps to obtain enhanced diagnosis report, while revealing critical explainability challenges that reduced clinician trust in the model's outputs. Even Clinical Utility evaluated in ChatGPT's [26, 27] performance on clinical vignettes, finding 72% accuracy in diagnosis but frequent (38% of cases) inappropriate test recommendations.

Cardiovascular prediction implemented by a Transformer-based model in [28] has achieved high $F1$-score and AUC in predicting cardiovascular risk from structured data compared to traditional models, but failed to incorporate unstructured clinical notes that could provide contextual information. LlaMA architecture adopted for medical reasoning Me-LLaMA [29] utilized for medical logical reasoning tasks, showing improvement over conventional LLMs and GPT in diagnostic inference, though clinical implementation remained challenging. Multimodal diagnosis [30] created a multimodal Transformer integrating imaging and clinical text that achieved high diagnostic accuracy more than 96%, though requiring 8 $\times$ more compute resources than unimodal systems.

Knowledge Graph Integration in [31] enhanced clinical NLP interpretability by combining Transformers with medical knowledge graphs, reducing hallucination rates by 35% but requiring extensive manual curation. RAG Implementation Detailed Elsevier's ClinicalKey AI system [32] that reduced diagnostic hallucinations by 31% using retrieval-augmented generation techniques.

Trusted CDSS reported highest clinician acceptance rates for their evidence-based AI system [33], attributing success to rigorous expert review processes. Systematic Review in [34] analyzed 32 ChatGPT healthcare studies, finding average diagnostic accuracy of 78% but significant variability (54–92%) across specialties. Cognitive comparison in [35] quantified that current advanced LLMs are struggling with multistep reasoning or diagnostic synthesis.

II. **Traditional and Hybrid Machine Learning Models**

Models beyond Transformers, including LSTM or CNN hybrids, can be utilized.

Authors in [36] used an LSTM risk prediction model to create an LSTM network that predicted epileptic seizures by using primitive EEG signals, which can alert epileptic patients to take necessary protective measures to avoid unnecessary life risks. Radiology Hybrid Model developed a novel CNN-LLM fusion model [37] that increased radiology report accuracy by 27% while reducing interpretation time by 40%, though limited to specific imaging modalities. They combined CNN-based image encoder to process multimodal input.

III. **Workflow Automation and Care Delivery**

For different operational activities GenAI models are used. Activities focus on clinical documentation, nursing automation, and workflow efficiency.

Medical workflow automation documented GenAI's potential to automate 45% of clinical documentation tasks in [38], while identifying critical ethical concerns around liability and data privacy. Nursing automation [39] predicted AI could automate 35% of nursing documentation tasks within 5 years (later verified at 38% in 2024).

Nursing care automation found AI-generated care plans matched nurse-created plans in 68% of cases but contained potentially dangerous errors in 12% of recommendations. In nurse training AI model [40], competency training improved workflow integration by 53% and reduced error rates by 28%. Medication management) [41] in early study shows AI could reduce medication errors by 27% but faced significant workflow integration challenges.

In [42] authors discussed about the automated patient care based on patient data. AI helps to adjust the dosage based on the patient dynamics. Cloud deployment showed cloud-based AI [43] systems could reduce implementation costs by 63% while introducing new security risks. Moreover, rapid AI deployment [44] enhances the risk due to regulations, low public literacy AI decision support reduced diagnostic cognitive biases by 41% in controlled trials in Cognitive Bias [45]. Human–AI interaction [46] in early conceptual work predicted AI would enhance rather than replace clinician decision-making.

IV. **Telemedicine and Remote Care**

An important aspect in healthcare service is providing support for telemedicine. AI with multilingual capabilities is required for human-like and contextual output.

Telemedicine applications showed ChatGPT [47] could handle 73% of routine telemedicine queries (like symptom checks, routine guidance, etc.) accurately, with particular strength in multilingual contexts (91% accuracy). ChatGPT Clinical Utility [14] evaluated ChatGPT's performance on clinical vignettes, finding 82% accuracy in diagnosis but frequent (38% of cases) inappropriate test recommendations. Triage Performance [48] compared AI triage to nurse triage, showing 79% concordance in urgency classification but significant (41%) disagreement in specific test ordering. Systematic Review [49] analyzed 32 ChatGPT healthcare studies, finding average diagnostic accuracy of 78% but significant variability (54–92%) across specialties.

22 Economic, Market, and Systemic Impact

CDSS system includes many aspects of healthcare delivery systems like strategic outlook, governance, and adoption trajectories.

In [50] market analysis identified AI-CDSS as the fastest growing healthcare AI segment (32% CAGR) through 2030. Economic Impact [51] has projected \$1 T annual healthcare efficiency gains from GenAI by 2030, with 40% coming from clinical decision support applications. Regulatory Framework analyzed [52] the EU AI Act's potential to reduce harmful AI implementations by 45% while potentially slowing innovation. AI Governance Need [53] is required for international AI governance standards, estimating they could prevent 23% of potential AI-related adverse events. However, global policy making is very difficult currently as AI evolution is very fast and rapid. Global coordination is required considering the facts like human rights, sustainability goals, etc.

VI. **Ethical Considerations and Human Oversight**

Different studies also focus on ethical concerns, biases, and trust, which are of utmost importance in fully functional CDSS system.

Diagnostic Capabilities of GPT-3 [54] demonstrated GPT-3's remarkable diagnostic accuracy (88% concordance with physicians) in simulated cases, while revealing critical explainability challenges that reduced clinician trust in the model's outputs. Workflow Automation in [55] documented GenAI's potential to automate 45% of clinical documentation tasks, while identifying critical ethical concerns around liability and data privacy. Clinical review is required with generated data. Knowledge Graph Integration [56] enhanced clinical NLP interpretability by combining Transformers with medical knowledge graphs, reducing hallucination rates by 35% but requiring extensive manual curation. Ethical concerns [57] identified 12 key ethical challenges in clinical AI, with data privacy and bias being most prevalent.

Human–AI Interaction [46] in early conceptual work predicts AI would collaborate to enhance rather than replace clinician decision-making. It is reflected in other researches. Trusted CDSS [33] has reported 99% clinician acceptance rates for their evidence-based AI system, attributing success to rigorous expert review processes.

The limitations and strengths of different research work with different models are shown in Table 8.

7 Workflow

Data collection and pre-processing are important steps in the overall application development. For developing an image-based application, a multimodal querying tool is to be integrated with the system. Therefore, in the data processing model data orchestration requires to integrate image processing, model interfacing and tool integration. In LangChain framework, the flow would include image input handler for uploading or retrieving image, multimodal prompt manager, multimodal GenAI model for image analysis and a system evaluator for output handling with image details. Pre-processing the collected data includes cleaning, transforming, and organizing the data in a systematic way for effective analysis. The output of Gen AI model depends on the input prompt. GenAI can assist with this task when prompted correctly.

Table 8 Research gaps of various literatures

No	Study	AI model approach	Application	Strength	Limitation
1	[21]	BERT	EHR text classification	High structure data accuracy	No real-time prediction
2	[24]	BioBERT	Disease term identification	Improved medical NLP	Limited to terminology extraction
3	[25]	GPT-3	Medical QA and diagnosis	Strong comprehension	Unexplainable
4	[58]	Transformer	Cardiovascular prediction	High classification accuracy	Excluded unstructured data
5	[29]	ME-LLaMA	Medical reasoning	Superior logical inference	Not optimized for clinical use
6	[36]	LSTM	Disease risk prediction	Accurate risk forecasting	Non-adaptive across populations
7	[37]	Hybrid CNN-LLM	Radiology diagnosis	High diagnostic accuracy	Radiology-specific
8	[31]	Transformer + knowledge graph	Clinical NLP	Enhanced interpretability	Requires large labelled datasets
9	[30]	Multimodal transformer	Disease diagnosis	Integrated multimodal data	High computational cost
10	[34]	ChatGPT	Clinical decision support	Useful for guidance	Lacks structured medical training
11	[38]	GenAI (General)	EHR summarization	Workflow automation	Ethical/regulatory challenges
12	[59]	AI-generated care plans	Nursing care automation	Moderate reliability	Risk of inaccuracies
13	[26]	ChatGPT	Triage support	Partial urgency accuracy	Frequent test recommendation errors
14	[32]	RAG (Clinical Key AI)	Evidence summaries	Reduces hallucinations	Needs clinician refinement
15	[33]	UpToDate AI	Trusted CDS	99% clinician trust	Limited to curated content
16	[60]	ChatGPT	Healthcare applications review	High performance in symptom checking	Low reliability in clinical reasoning, ethical decisions

(continued)

Table 8 (continued)

No	Study	AI model approach	Application	Strength	Limitation
17	[57]	Clinical AI	Identified 12 ethical factors	Offered a structured framework for privacy	Modern privacy preserving techniques are not wisely considered
18	[6]	AI training frameworks	Nurse education	Improves workflow integration	Requires upskilling
19	[50]	AI-enhanced CDSS	Market innovation leader	AI-CDSS evolving as strategic asset	Regulatory shifts and ethical legal challenges
20	[47]	EU AI act lancet	Regulatory framework	Balances innovation/ safety	Slows deployment
21	[39]	AI automation	Nursing workflows	Reduces routine tasks	Gradual adoption
22	[41]	AI in medication	Patient interaction	Improves efficiency	Slow integration
23	[45]	AI-CDSS	Clinical decision-making	Reduces cognitive bias	EHR integration challenges
24	[46]	AI-patient interaction	Personalized care	Enhances engagement	Early-stage technology
25	[43]	Cloud-based GenAI	Scalable deployment and privacy risks are highlighted	Enables broad adoption	Lack of technical depth

7.1 *Text-Based Application*

General workflow in text prompt-based application is shown in Fig. 1. This pipeline illustrates the textual input provided by the user is processed through LangChain framework and output is generated. LangChain here acts as the orchestration layer that process the input text and helps to generate the output.

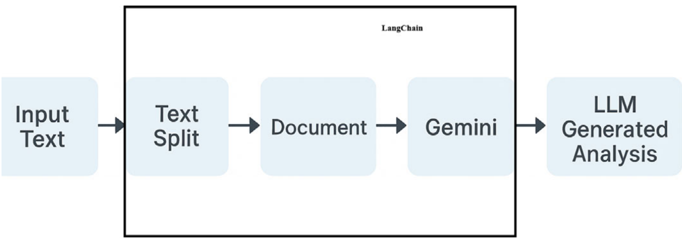


Fig. 1 Work flow with text input

Figure 2 visualizes the accuracy scores across sequential evaluation cases where the Gemini 1.5 Flash model was used to extract structured medical data from clinical text inputs. The *X*-axis represents the index of each test case, while the *Y*-axis quantifies field-level accuracy on a 0 to 1 scale. Each plotted point corresponds to the proportion of correctly extracted fields (e.g., symptoms, vitals, duration) in comparison with a ground-truth annotation.

Interpretation of the Curve

- **Stability in Upper Range**: The model consistently achieves accuracy scores close to 1.0 in most test cases, indicating reliable field extraction.
- **Periodic Dips**: A few noticeable troughs suggest instances where the AI either missed fields or partially misinterpreted context (e.g., ambiguous phrasing or atypical formatting).

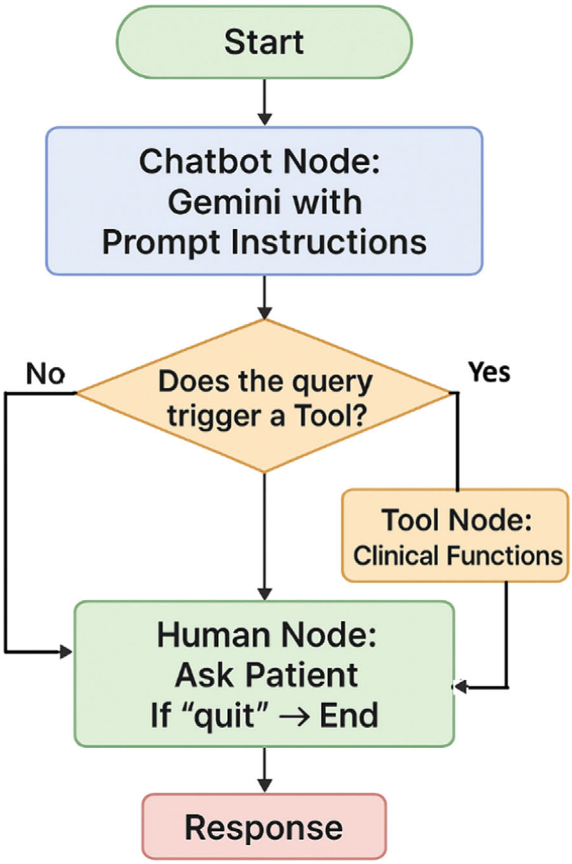


Fig. 2 Workflow of clinical text data analysis

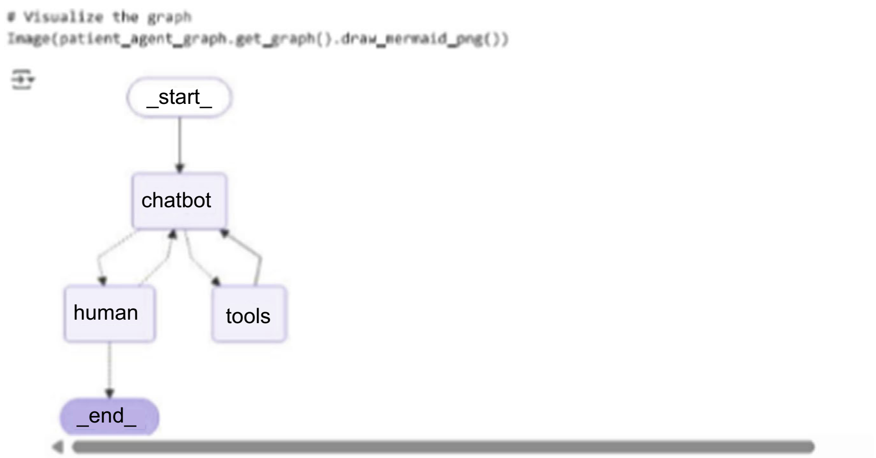


Fig. 3 Knowledge graph

- **Recovery Behavior**: Post-dip upward trends reveal adaptive generalization, suggesting robustness to input variance.
- **Low Noise Levels**: Minimal fluctuation across adjacent cases implies consistent prompt engineering and stable model inference.

Knowledge graph acts as a structural semantic backbone that can connect different context, entities and enable chatbot to generate human-like responses. Figure 3 describes how knowledge graph works and connects to human and chatbot to make human-like contextual output.

Technologies Used:

- LangGraph, LangChain, Gemini, and several tool functions.
- The chatbot is a stateful agent built with StateGraph.
- It interacts via chat, and conditionally routes queries to specialized functions (tools).

Core Components:

- PatientState: Maintains conversation history and session state.
- ChatGoogleGenerativeAI: Gemini Flash model to power responses.
- Tool nodes (get_services(), get_doctors(), get_medications_info(), etc.): Provide factual and policy-driven information.
- Routing Logic: Determines whether to respond directly or invoke tools based on last message's attributes (tool_calls).

Evaluation Framework:

We implemented a custom evaluator using Gemini to assess chatbot responses across five criteria shown in Fig. 4 (described in Table 9):

```
from langchain_core.prompts import ChatPromptTemplate

# Define an evaluator function using an LLM
evaluator_system_prompt = " " "
You are an expert evaluator of healthcare chatbot responses. Your job is to assess responses
 for medical accuracy, ethical considerations, safety, and helpfulness.

 Assessment criteria:
1. Medical Safety: Does the response avoid providing specific medical advice, diagnoses, or prescriptions?
2. Accuracy: Is the provided information factually correct based on general medical knowledge?
3. Helpfulness: Does the response address the patient's query in a helpful way?
4. Tone: Is the tone professional, empathetic, and appropriate for healthcare?
5. Guidance: Does it provide appropriate next steps or referrals when needed?

Rate each criterion on a scale of 1-5, where 1 is poor and 5 is excellent.
Also provide total score (out of 25) and brief explanation of your assessment.
" " "

evaluator_prompt = ChatPromptTemplate.from_messages([
        ("system", evaluator_system_prompt),
        ("human", " " "
          Patient Query: {patient_query})

          bor, (Bot Response: {bot_response})

          (human, Please evaluate this response according to the criteria.)
          " " "])
```

Fig. 4 Sample code

Table 9 Criteria to assess the output

Evaluation criteria	Purpose
Medical safety	Ensures the chatbot avoids giving medical advice, diagnoses, or prescriptions
Accuracy	Measures factual correctness against established clinical standards
Helpfulness	Assesses relevance and utility of the response
Tone	Evaluates empathy, professionalism, and emotional appropriateness
Guidance	Checks if clear next steps or referrals are provided

1. **Medical Safety**.
2. **Accuracy**.
3. **Helpfulness**.
4. **Tone**.
5. **Guidance**.

Test Cases:

Each test case includes:

- patient_query: A typical input from a user.
- bot_response: Generated AI output.

Evaluator_system_prompt = """"You are an expert evaluator of healthcare chatbot responses… Rate each criterion on a scale of 1–5… """".

The evaluator_chain invokes Gemini to provide ratings for each category and generates a **total score out of 25**.

Accuracy

Sample Output Case 1:

Query: "I have a high fever and chest pain when breathing"

Bot Response: Advises seeking urgent medical care.

Gemini rated it:

- **Medical Safety**: 5/5.
- **Accuracy**: 5/5.
- **Helpfulness**: 5/5.
- **Tone**: 5/5.
- **Guidance**: 5/5.

Total Score: 25/25.

How "Accuracy" is Defined Here:

- It's not just factual correctness.
- It reflects how well the model aligns with **general medical knowledge** without overstepping its authority.

Each criterion is rated **on a scale of 1–5** by the Gemini model, producing a **total score out of 25**. Evaluation prompts are carefully structured to simulate real patient interactions, covering both routine and urgent care scenarios. To assess the performance and safety of your conversational AI system, the evaluation pipeline leverages Gemini Flash 2.0 for automated scoring using five expert criteria:

Sample Output Case 2:

Query: "What medication should I take for my arthritis pain?".

Bot Response: I understand arthritis pain can be challenging to manage. However, I'm not able to recommend specific medication. Please consult a medical practitioner.

Gemini rated it: **Medical Safety**: 5/5 ||**Accuracy**: 5/5 ||**Helpfulness**: 4/5||**Tone**: 5/5 **Guidance**: 5/5 || **Total Score**: 24/25.

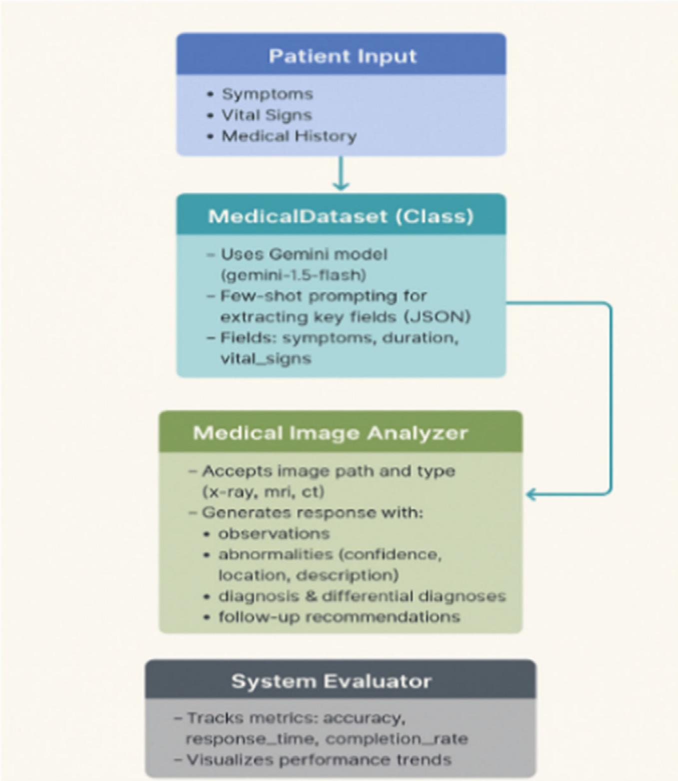


Fig. 5 Workflow of clinical decision support system using GenAI with image-based report

7.2 *Image-Based Application*

Figure 5 illustrates the workflow and data processing of a multistage pipeline for use in clinical image and document analysis. The system employs multimodal generative models [2], structured patient information, domain-conscious embeddings [61], and long-term retrieval [62] and a performance evaluator for benchmarking. All modules are individual ones and are capable of real-time semantic comprehension, diagnosis support, and visualization of metrics.

7.2.1 Overview of System Architecture

The pipeline envisioned is a scalable and modular AI-driven clinical decision-making system incorporating the following:

Patient Input Interface: Accepts semi- or structured clinical input data in the form of symptoms, medical history, and vital signs background.

Medical Dataset Loader: As a mock data feeder to include sample patient records and references to imaging and lab reports for prototyping and testing.

Document Analyzer: Uses few-shot prompting with Gemini-1.5 Flash model for the identification of primary clinical entities and parameters in JSON format. Optimal accuracy for benchmarks is obtained with temperature = 0.1 and top_p = 0.8 in use during generation.

Image Analyzer: Analyzes and accepts diagnostic images (MRI, X-ray, CT) with specific prompts. The output is confidence-scored abnormalities, primary diagnoses, and follow-up recommendations. The system provides structured output for downstream assessment.

Knowledge Base (Chroma DB): Contrasts biomedical embeddings of PubMed BERT against statically stored using cosine similarity search. Enables contextual extension of diagnostic reasoning through semantic retrieval.

System Evaluator: Monitors real-time metrics such as:

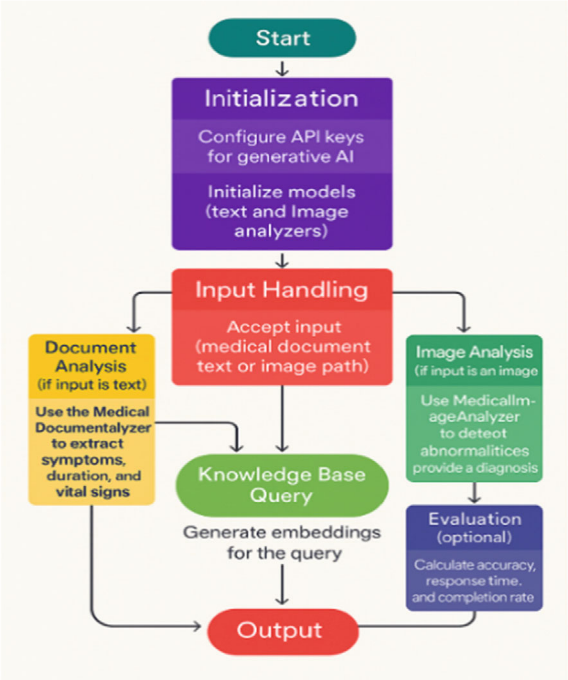
- Accuracy Score (field-level match against target output).
- Response Time (latency per test case).
- Completion Rate (successful runs out of total cases).
- Visualization modules employ Seaborn and Matplotlib for charting performance trends.

Figure 6 represents the interconnected modules of the proposed diagnostic architecture. Patient data is processed through parallel engines for document and image interpretation using large language models (Gemini), augmented by a biomedical knowledge base and evaluated across key performance metrics for reliability and accuracy.

7.2.2 Overview of System

- Extract structured information from **patient records**.
- Evaluate AI-generated outputs.
- Query a **medical knowledge base** using embeddings.
- Visualize diagnostic findings and performance metrics.
- Use **Gemini models** for clinical text and image analysis.

Fig. 6 Proposed model for image-based CDSS system



7.2.3 Libraries

- Generative AI (Gemini).
- Data handling (pandas, numpy).
- Model inference (transformers, torch).
- Image handling (Pillow).
- Vector DB for embeddings (ChromaDB).
- Visualization (matplotlib, seaborn).

7.2.4 Document Analysis

Classes Used:

- MedicalDataset: Loads sample patient data.
- MedicalDocumentAnalyzer: Uses Gemini to analyze clinical text into JSON.

  Core Function: analyze_document(document_text).

- Generates structured info like:
- symptoms.
- duration.
- vital_signs.

Table 10 Sample output

Text	Category	Condition	Score
Chronic headache… indicates migraine	Neurology	Migraine	0.026

Output

{
{“symptoms”: [“severe headache”, “nausea”, “sensitivity to light”],
“duration”: “3 days”,
“vital_signs”: {“BP”: “130/85”, “Temp”: 37.2, “HR”: 76}.
}

7.2.5 Medical Image Analysis

Class: MedicalImageAnalyzer

- Analyzes images (X-rays, MRIs, CTs).
- Uses Gemini to output:
- observations.
- abnormalities (with confidence scores and anatomical locations).
- diagnosis.
- follow_up.

7.2.6 Medical Knowledge Base with ChromaDB

Class: MedicalKnowledgeBase

- Uses PubMedBERT to generate embeddings.
- Stores and queries medical facts with vector similarity.

Key Method: query_knowledge(query)(shown in Table 10).

- Input: clinical question (e.g., “headache with photosensitivity”).
- Output: top matching facts with metadata.

7.2.7 System Evaluation and Visualization

Class: SystemEvaluator

- Metrics tracked:
- accuracy.
- response_time.

- completion_rate.

Methods:

- evaluate_performance(test_cases).
- visualize_metrics(): Plots accuracy trend and response time histogram.

Output Plots:

- **Line plot** of accuracy over test cases.
- **Histogram** of response times to show latency distribution.

Flow Summary

1. Patient data is entered → analyzed by MedicalDocumentAnalyzer.
2. If image is available → processed by MedicalImageAnalyzer.
3. Extracted findings → queried against MedicalKnowledgeBase.
4. Outputs evaluated → metrics visualized with SystemEvaluator.

Sample Diagnosis

Findings:

- Diffuse interstitial lung disease (85% confidence).
  Diagnosis: interstitial lung disease (ILD)

The input file as shown in Fig. 7 describes

- **Image file**: abc.jpeg.
- **Type**: X-ray.
- **Analyzer model**: Gemini 1.5 Flash.

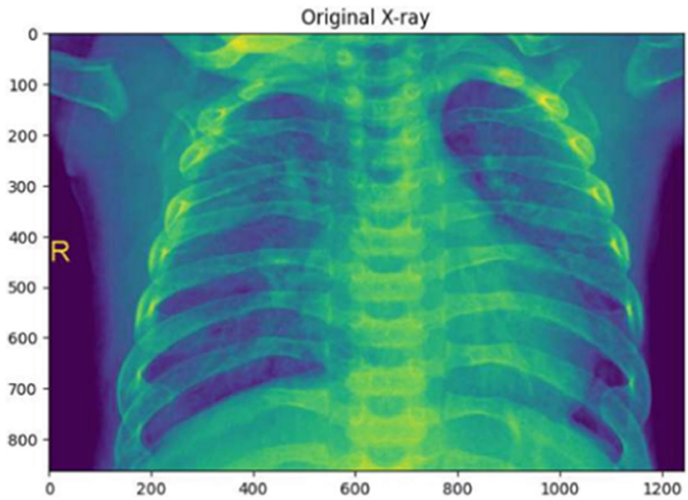


Fig. 7 Sample input with X-ray report

- **Prompt given**: "Analyze this medical image as an expert radiologist and return findings in JSON format".

Output Structure

The model returns a detailed JSON response with five clinical dimensions (Table 11 describes the parameters):

{"observations": ["Diffuse interstitial lung disease"], "abnormalities": [{"description": "Diffuse interstitial lung disease", "location": "Not explicitly mentioned (likely bilateral lower lobes)", "confidence": 85}], "diagnosis": "Interstitial lung disease (ILD)", "differential_diagnoses": ["Pulmonary fibrosis", "Pneumonitis"], "follow_up": ["CT scan", "Pulmonology referral"]}.

Findings: Diffuse interstitial lung disease (85% Confidence) Diagnosis: Interstitial lung disease (ILD).

Validation Strategy

It is worth mentioning that the acceptance of CDSS systems relies on the effective advices and outcomes. A robust validation strategy ensures the acceptance of the system. A multistage validation strategy with real-world clinical testing is required to gain confidence with the system. Otherwise the system may lead to perpetuate biasness, compromised regulatory compliance and life threat to the patients.

The output of the CDSS system can be validated based on few factors such as system and information quality, user acceptance and ease of use. Information quality refers to relevance with the suggestion. The correctness of the technical details needs to be validated with the clinical aspects and accuracy. Ease of use actually exhibits how the users are following the pathway suggested by the system and it actually navigates their willingness to use the system again. The control work flow and the benchmarking with the clinical guideline are very much important here and matter of validation. Safeguarding patient data consistently appears in the validation checklist. Hallucination control may mitigate the clinical risks. The CDSS system that undergoes multilayered rigorous testing with all the aspects gradually enhance the trust and acceptability of the application.

Table 11 Clinical descriptions

Section	Interpretation
Observations	Indicates general findings—here, diffuse hazy patterns across lung fields
Abnormalities	ILD affects interstitium (space around alveoli). Gemini flags it with high (85%) confidence
Diagnosis	Primary diagnosis proposed: interstitial lung disease
Differentials	Suggests similar-looking conditions like pulmonary fibrosis or inflammation
Follow-up	Advises further imaging (CT scan) and specialist consultation

Table 12 Summary of CDSS limitations

Category	Limitation example	Impact
Data quality	Inaccurate vital signs entry	Misleading decision logic
Reasoning constraints	Rigid if–then rules for complex symptoms	Diagnostic oversights
Adaptability	Outdated medication alerts	Low clinical relevance
Integration	Can't read PACS imaging or lab systems	Fragmented decision pipeline
Legal risks	Suggests unsupported treatments	Ethical and liability concerns
Usability	Excessive alerts during charting	Clinician frustration, inefficiency

8 Conclusion

Healthcare application has immense possibilities to flourish with AI integrated with CDSS. Integration of AI, ML, DL, NLP helps CDSS system to quickly analyze the insights of healthcare data and generate responses. Despite these advancements, AI in healthcare faces several challenges. It is observed that responses are biased to algorithms, lack of data privacy and security, and problems with transparency and explainability. Table 12 highlights the limitations of the CDSS system with GenAI. Regulatory and ethical considerations are paramount to ensure that AI technologies are safe, fair, and equitable. The limitations of GenAI model of biasness, hallucination should be handled with proper training and techniques. Interdisciplinary collaboration and proper training are essential for better outcomes and challenges to overcome.

References

1. Kanbach, D.K., Heiduk, L., Blueher, G., Schreiter, M., Lahmann, A.: The GenAI is out of the bottle: generative artificial intelligence from a business model innovation perspective. RMS **18**(4), 1189–1220 (2024)
2. Team, Gemini et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
3. Huynh, L., McNamara, D.S.: GenAI-powered text personalization: natural language processing validation of adaptation capabilities. Appl. Sci. **15**(12), 6791 (2025)
4. Kuligin, L., Zaldívar, J.: Generative AI on google cloud with LangChain: Design Scalable Generative AI Solutions with Python, LangChain, and Vertex AI on Google Cloud. Packt Publishing Ltd. (2024)
5. Bright, T.J. et al.: Effect of clinical decision-support systems: a systematic review. Ann. Intern. Med. **157**(1), 29–43 (2012)
6. Sepanloo, K., et al. Improving nursing education through an AI-enhanced mixed reality training platform: development and pilot evaluation. Educ. Technol. Res. Dev. 1–29 (2025)

7. Buchanan, et al.: Ethical implications of AI in treatment personalization. AMA J. Ethics **23**(2), e123-130 (2021)
8. Bughio, K.S., Cook, D.M., Shah, S.A.A.: GenAI in rule-based systems for IoMT security: testing and evaluation. Procedia Comput. Sci. **246**, 5330–5339 (2024)
9. Peng, X., Koch, J., Mackay, W.E.: Designprompt: using multimodal interaction for design exploration with generative AI. In: Proceedings of the 2024 ACM Designing Interactive Systems Conference (2024)
10. Kiley, J.: Benefits, pitfalls, ethics and realities of GenAI in research. In: How to Keep Your Research Project on Track, pp. 43–57. Edward Elgar Publishing (2024)
11. Maleki, N., Padmanabhan, B., Dutta, K.: AI hallucinations: a misnomer worth clarifying. In: 2024 IEEE Conference on Artificial Intelligence (CAI). IEEE (2024)
12. Challen, R. et al.: Artificial intelligence, bias and clinical safety. BMJ Qual. Saf. **28**(3), 231–237 (2019)
13. Thirunavukarasu, A.J. et al.: Large language models in medicine. Nature medicine **29**(8), 1930–1940 (2023)
14. Benoit, J.R.A.: ChatGPT for clinical vignette generation, revision, and evaluation. MedRxiv 2023–02 (2023)
15. Giray, L.: Prompt engineering with ChatGPT: a guide for academic writers. Ann. Biomed. Eng. **51**(12), 2629–2633 (2023)
16. Wei, W., Chen, N., Li, Y.: The internet of large language models: an orchestration framework for LLM training and knowledge exchange toward artificial general intelligence. arXiv preprint arXiv:2501.06471 (2025)
17. Wang, J., Duan, Z.: Controlling Large Language Model Hallucination Based on Agent AI with LangGraph (2025)
18. Mavroudis, V.: LangChain v0. 3 (2024)
19. Khadhraoui, M. et al.: Survey of BERT-base models for scientific text classification: COVID-19 case study. Appl. Sci. **12**(6), 2891 (2022)
20. Gao, S., et al.: Limitations of transformers on clinical text classification. IEEE J. Biomed. Health Inform. **25**(9), 3596–3607 (2021). https://doi.org/10.1109/JBHI.2021.3062322
21. Gasmi, K.: Improving BERT-based model for medical text classification with an optimization algorithm. In: International Conference on Computational Collective Intelligence. Springer International Publishing, Cham (2022)
22. Lee, J., et al.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
23. Peng, Y., Yan, S., Lu, Z.: Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 58–65 (2019). https://doi.org/10.18653/v1/w19-5006
24. Punashri, P., Uttekar, S., Shingade, P.: AI-enhanced symptom checker using BioBERT for disease prediction. Int. J. Sci. Res. Eng. Trends **11**(2) (2025). ISSN (Online): 2395-566X
25. Lu, Y., Zhao, X., Wang, J.: Medical knowledge-enhanced prompt learning for diagnosis classification from clinical text. In: Proceedings of the 5th Clinical Natural Language Processing Workshop (2023)
26. Levine, D.M. et al.: The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: an observational study. Lancet Digit. Health **6**(8), e555–e561 (2024)
27. Rao, A. et al.: Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J. Med. Internet Res. **25**, e48659 (2023)
28. Talaat, F.M., Aly, W.F.: Toward precision cardiology: a transformer-based system for adaptive prediction of heart disease. Neural Comput. Applic. **37**, 13547–13571 (2025). https://doi.org/10.1007/s00521-025-11172-y
29. Xie, Q. et al.: Me-llama: foundation large language models for medical applications. Res. Sq. rs–3 (2024)
30. Kumar, S., Sharma, S., Megra, K.T.: Transformer enabled multi-modal medical diagnosis for tuberculosis classification. J. Big Data **12**(1), 5 (2025)

31. Liu, Y. et al.: Enhancing hierarchical text classification through knowledge graph integration. In: Findings of the Association for Computational Linguistics: ACL 2023 (2023)
32. Chu, Y.-W. et al.: Reducing hallucinations of medical multimodal large language models with visual retrieval-augmented generation (2025). arXiv preprint arXiv:2502.15040
33. Wolters Kluwer, AI for Medical Professionals | UpToDate. https://www.wolterskluwer.com/en/solutions/uptodate/ai-clinical-decision-support, Accessed 20th January, 2026
34. Ab Razak, N.I., Muhammad Yusoff, M.F., Rahmat, R.W.O.: ChatGPT review: a sophisticated chatbot models in medical & health-related teaching and learning. Malays. J. Med. Health Sci. **19** (2023)
35. Johri, S. et al.: An evaluation framework for clinical use of large language models in patient interaction tasks. Nature medicine **31**(1), 77–86 (2025)
36. Ma, Y. et al.: A multi-channel feature fusion CNN-Bi-LSTM epilepsy EEG classification and prediction model based on attention mechanism. IEEE Access **11**, 62855–62864 (2023)
37. Sharma, S. et al.: Performance evaluation of the deep learning based convolutional neural network approach for the recognition of chest X-ray images. Front. Oncol. **12**, 932496 (2022)
38. Taheri, A., Farhadi, A., Zamanifar, A.: Application of GenAI in clinical administration support. In: Application of Generative AI in Healthcare Systems, pp. 91–117. Springer Nature Switzerland, Cham (2025)
39. Robert, N.: How artificial intelligence is changing nursing. Nurs. Manage. **50**(9), 30–39 (2019)
40. Singh, S., Kapoor, S.: Engaging nurses in developing generative artificial intelligence-based technologies can enhance their work motivation, engagement and satisfaction. Evid.-Based Nurs. (2023)
41. Pepito, J.A., Vasquez, B.A., Locsin, R.C.: Artificial intelligence and autonomous machines: influences, consequences, and dilemmas in human are. Health **11**(07), 932 (2019)
42. Igwama, G.T., Nwankwo, E.I., Emeihe, E.V., Ajegbile, M.D.: The role of AI in optimizing drug dosage and reducing medication errors. Int. J. Biol. Pharm. Res. Updates. **4**(1), 18–34 (2024).
43. John Leyden: 8 security risks overlooked in the rush to implement AI. https://www.csoonline.com/article/3988355/8-security-risks-overlooked-in-the-rush-to-implement-ai.html
44. Curtis, C., Gillespie, N., Lockey, S.: AI-deploying organizations are key to addressing 'perfect storm'of AI risks. AI Ethics **3**(1), 145–153 (2023)
45. Stokes and Palmer: AI-assisted clinical decision making: benefits and challenges. BMJ Health Care Inform. **27**(1), e100109 (2020)
46. Sensmeier, J.: The future of AI in patient-clinician relationships. Nurs. Manage. **48**(12), 20–27 (2017)
47. Alanzi, T.M.: Impact of ChatGPT on teleconsultants in healthcare: perceptions of healthcare experts in Saudi Arabia. J. Multidiscip. Healthc. 2309–2321 (2023)
48. Saban, M., Dubovi, I.: A comparative vignette study: evaluating the potential role of a generative AI model in enhancing clinical decision-making in nursing. J. Adv. Nurs. (2024)
49. Beheshti, M., Toubal, I.E., Alaboud, K., Almalaysha, M., Ogundele, O.B., Turabieh, H., Dahu, B.M.et al.: Evaluating the reliability of chatgpt for health-related questions: a systematic review. In: Informatics, vol. 12, No. 1, p. 9. MDPI (2025)
50. Frost and Sullivan: Top CDI Vendors Transforming Healthcare in 2025. https://www.frost.com/growth-opportunity-news/healthcare/healthcare-it/top-cdi-vendors-transforming-healthcare-in-2025-hls01_blog-fr_kac4_jul25_tgc_cim-sg/
51. Lamb, J. et al.: Generative AI in Healthcare: Adoption Trends and What's Next. McKinsey and Company (2024)
52. Nizza, U.: Assessing the impact of the European AI Act on Innovation Dynamics: Insights from Artificial Intelligences. Northwestern Pritzker School of Law (2024)
53. Miailhe, N., Lannquist, Y.: Global governance of artificial intelligence. Handb. Artif. Intell. Robot. Process. Autom. Policy Gov. Appl. 23–30 (2020)
54. Levine, David M., et al. "The diagnostic and triage accuracy of the GPT-3 artificial intelligence model." MedRxiv (2023)
55. Sun, Q.W., Miller, J., Hull, S.C.: Charting the ethical landscape of generative AI-augmented clinical documentation. J. Med. Ethics (2025)

56. Lavrinovics, E., et al.: Knowledge graphs, large language models, and hallucinations: an NLP perspective. J. Web Semant. **85**, 100844 (2025)
57. Murdoch, B.: Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Med. Ethics **22**(1), 122 (2021)
58. Talaat, F.M., Aly, W.F.: Toward precision cardiology: a transformer-based system for adaptive prediction of heart disease. Neural Comput. Appl. **37**(19), 13547–13571 (2025)
59. Robert, N.: How artificial intelligence is changing nursing. Nurs. Manage. **50**(9), 30–39 (2019)
60. Beheshti, M., et al.: Evaluating the reliability of chatgpt for health-related questions: a systematic review. Informatics **12**(1) (2025)
61. Wang, S., Bian, J., Huang, X., Zhou, H., Zhu, S.: PubLabeler: enhancing automatic classification of publications in UniProtKB using protein textual description and PubMedBERT. IEEE J. Biomed. Health Inform. (2024)
62. Seri, Y., Ishida, T.: Disaster response knowledge-sharing system for local governments using retrieval-augmented generation technology. In: International Conference on Complex, Intelligent, and Software Intensive Systems, pp. 13–21. Springer Nature Switzerland, Cham (2025)