

Review Article

AI-IOT Enabled Health Advisory Chatbot for Diabetics using RAG model

Guda Vanitha¹, Srujana Inturi², M Venkata Krishna Reddy³, T Suvarna Kumari⁴, Golla Manasa⁵

^{1,2,3,4,5}Computer Science Engineering, CBIT, Hyderabad, India.

¹Corresponding Author : gvanitha_cse@cbit.ac.in

Received: 12 March 2026

Revised: 25 July 2026

Accepted: 05 August 2026

Published: 30 September 2026

Abstract - Diabetes is one of the most common chronic diseases that requires regular monitoring to avoid complications. While the IoT devices can only track health data like blood glucose, heart rate, and temperature, they often ignore the Body Mass Index (BMI), which is a major risk factor for Type 2 diabetes. BMI is a powerful measure to understand if a person is underweight, normal, overweight, or obese. A high BMI is closely linked with obesity, and obesity directly contributes to insulin resistance. Without monitoring the BMI, patients may control sugar levels but miss the biggest lifestyle problem. An AI-IoT health monitoring system for diabetic patients has been developed that not only tracks blood glucose, heart rate, and temperature but also calculates the Body Mass Index (BMI). A RAG-powered chatbot is integrated so that patients can ask questions, understand the readings, and receive personalized guidance for diet, exercise, and medication. The Dexcom G6 sensor is used to measure blood glucose, the MAX30100 sensor for heart rate and SpO₂, and the LM35 sensor for body temperature. A smart weighing scale is added to record weight, and the patient's height is stored in the system. By using weight and height, the BMI is calculated. Then all data is processed by Arduino Nano, which will receive raw signals from sensors and then convert them into meaningful health values. The Arduino Nano sends the processed data to Node MCU, which is connected to the internet via Wi-Fi and transfers it to the cloud platform. Cloud platform receives the data and displays it on a web dashboard. A RAG chatbot is used to retrieve trusted medical documents and generate personalized recommendations. Patients can use the web dashboard to view readings and chat with the assistant for explanations. Patients are able to track glucose levels along with BMI in real time. The RAG chatbot explains the results in simple words and suggests lifestyle tips and exercises, helping individuals reduce confusion about diabetes and encouraging self-care.

Keywords - MAX30100, LM35, Arduino Nano, Node MCU, Body Mass Index (BMI), Machine learning algorithm, RAG model, Health chatbot, Large Language Models (LLMs), FAISS, Sentence transformers.

1. Introduction

Diabetes mellitus is viewed as one of the most common chronic conditions on the planet that needs constant monitoring, control over the lifestyle, and self-management during the long term so as to avoid some of the most severe complications, which include cardiovascular disease, neuropathy, nephropathy, and retinopathy [11, 24]. The management of diabetes requires proper levels of glucose monitoring, patient education, health literacy, dietary and exercise compliance with guidelines [2, 15]. Specifically, Type 2 diabetes is closely linked with obesity and insulin resistance, in which Body Mass Index (BMI) is an essential risk factor in the onset of the disease and metabolic disequilibrium [11].

Even though most digital health systems are centered on glucose monitoring, the inclusion of BMI and other physiological indicators can be more effective in demonstrating the actual patient health conditions. The recent

development in Artificial Intelligence (AI), Machine Learning (ML), and Internet of Things (IoT) technologies has drastically changed the way diabetes is monitored and managed [23, 27]. Health monitoring systems that use IoT can provide physiological data like blood glucose, heart rate, oxygen saturation (SpO₂), and body temperature in real-time, which can be transmitted to cloud platforms to enable remote monitoring [25, 30, 31]. In a number of studies, it has been established that low-cost IoT-enabled e-health systems are effective in enhancing accessibility and continuous monitoring in diabetic patients [25]. Simultaneously, diabetes prediction and risk assessment have also extensively employed ML-based models, which are frequently complemented with explainable AI methods to increase transparency and clinical confidence [24, 26, 27].

In more recent times, Large Language Models (LLMs) and conversational AI systems have become a potentially useful tool in diabetes education, triage, and self-management



support [10, 23]. Chatbots supported by AI have demonstrated the potential of enhancing the health literacy of Type 2 diabetes patients and providing additional counseling in gestational and pediatric diabetes care [2, 7, 16]. Nonetheless, standalone LLMs are commonly affected by hallucinations and a lack of reliability, especially when used clinically [6, 21]. To alleviate these shortcomings, Retrieval-Augmented Generation (RAG) has been suggested as a successful framework that improves the output of the LLM using trusted medical documents and guidelines [3, 17, 18]. Systematic reviews point out that RAG enhances factual accuracy, reduces misinformation, and enhances clinical applicability when used in healthcare [4, 12, 13, 22].

RAG-based dialog agents and knowledge-enhanced LLM systems have been built to assist diabetes patients with personalized instructions on nutrition therapy, medication adherence, and risk management [5, 14, 20, 29]. A comparative analysis shows that RAG-enhanced LLMs have a strong effect over baseline models in diabetes education tasks by lessening the number of hallucinations and raising the reliability of responses [1, 13]. Moreover, formalized risk management systems and directed decoding approaches are now implemented to provide safer implementation of generative AI systems in healthcare settings [19, 21].

While there are significant strides in IoT, Artificial Intelligence (AI), Large Language Models (LLMs), and Retrieval-Augmented Generation (RAG), existing research primarily examines each technology individually rather than in concert within a single healthcare system. In most IoT systems, physiological parameters like blood glucose, heart rate, and temperature are being monitored and risk factors like Body Mass Index (BMI) are generally overlooked. Likewise, AI-powered chatbots deliver health data and support for patients, but they are, in most cases, not linked to real-time health monitoring data.

RAG makes the AI responses more accurate and reliable, but it is not fully integrated with IoT based diabetes monitoring. The above restrictions expose a great deal of research gaps in designing an integrated framework between AI and IoT, where real-time health monitoring is integrated with reliable, personalized, and evidence-based recommendations. The need to understand the strengths and weaknesses, as well as the potential of integration, of recent developments in the fields of IoT, AI, LLMs, and RAG technologies is driven by this gap.

The aim of this review is to examine the existing diabetes management systems based on the integration of AI, and IoT, identify the challenges and opportunities, and suggest future research directions to design intelligent healthcare solutions for continuous monitoring, customized recommendations, and enhanced diabetes self-management.

The main contributions of this review paper are summarized as follows:

- This review focuses on recent developments of Artificial Intelligence (AI), Internet of Things (IoT), Machine Learning (ML), Large Language Models (LLMs), and Retrieval-Augmented Generation (RAG) in diabetes monitoring and management.
- It critically evaluates the current systems for monitoring using IoT, existing machine learning models, existing AI-powered chatbots and RAG-based healthcare applications, their methodologies, pros and cons, and clinical applications.
- The review highlights several challenges in existing diabetes management systems, such as limited use of IoT and AI technologies, poor integration of physiological parameters like Body Mass Index (BMI) in LLM systems, hallucination problems with LLMs, and the absence of personalized healthcare recommendations in real-time.
- Unlike the previous review papers, which have mainly focused on individual technologies, this review paper identifies the research gaps and the need for an integrated framework that utilizes AI-IoT-RAG, integrating real-time physiological monitoring with evidence-based conversational intelligence for personalized diabetes management.
- The review highlights future research directions for intelligent, secure, explainable, and patient-centred healthcare systems that enable continuous monitoring, personalised decision support, and better diabetes self-management.

The proposed solution combines the capabilities of the Retrieval-Augmented Generation (RAG) chatbot with real-time monitoring of glucose, heart rate, temperature, and BMI using IoT sensors in a single platform. It is different from the systems currently available because it not only tracks health parameters but also offers individualized, evidence-based recommendations based on medical knowledge. This integration will enhance the reliability of AI responses and aid in improving diabetes self-management.

2. Background and Motivation

Diabetes mellitus is an increasingly expanding health issue worldwide that needs to be monitored regularly, diagnosed early, and managed appropriately by the patient to avoid severe complications like cardiovascular diseases, kidney failure, and nerve damage. Conventional approaches to diabetes treatment are largely based on periodic clinical check-ups and manual observations, which are not always able to deliver continuous evaluation of the health status of a patient [11]. Due to the rise in Type 2 diabetes cases, there

exists an urgent necessity for smart healthcare solutions, which help to monitor patients in real time and provide them with personalized healthcare.

The latest trends in the Internet of Things (IoT) technology have made it possible to have the process of constant health monitoring, which can be achieved using wearable and sensor-based systems. IoT medical solutions enable real-time monitoring and transmission of physiological measurements like blood glucose level, heart rate, and body temperature to the cloud space to enable remote monitoring [25, 30, 31]. Such systems enhance the ability to monitor patients remotely, that is, through glucose monitoring; other important risk factors like Body Mass Index (BMI), which is a leading contributor to obesity and insulin resistance in Type 2 diabetes, are overlooked in the majority of current solutions [11]. AI and ML methods have also contributed to the improvement of diabetes prediction and risk assessment based on medical data and the detection of patterns of a disease [24, 26, 27]. Even though these methods enhance the accuracy of the diagnosis, they tend to work with historical or fixed data and do not integrate with real-time monitoring conditions. Consequently, patients usually obtain forensic information without being provided with constant individualistic information.

Large Language Models (LLMs) and conversational AI-based tools have become effective solutions to enhance diabetes education and patient interaction in recent years. Research has proven that chatbots based on AI have the potential to improve health literacy and can help patients learn about the practice of managing disease management [2, 7, 16]. Behavior changes and self-management is also facilitated using conversational agents that offer interactive advice and lifestyle suggestions [8]. But independent LLMs have disadvantages like hallucinated answers and unreliability due to medicines, which restrict their safe use in healthcare systems [6, 21].

To address these shortcomings, Retrieval-Augmented Generation (RAG) has been proposed as one of the potentially effective methods that can improve AI-generated answers by accessing medical knowledge sources of credible information. The studies indicate that RAG has a considerable positive impact on the factual accuracy, misinformation minimization, and reliability in conversational medical systems [1, 3, 4, 17]. When compared to conventional foot care management devices, RAG-based diabetes assistants have been shown to perform better in providing evidence-based educational and individualized recommendations [13, 20].

Regardless of such technological innovations, the existing literature mostly envisages IoT monitoring systems and AI conversational agents as two separate solutions. Very little has been done to combine real-time body monitoring with RAG-powered intelligent assistants that can analyze

patient data and give context-based recommendations. This weakness drives the necessity of an integrated AI-IoT healthcare system that integrates continual tracking, BMI-based detection of risk, and trustworthy conversational advice to facilitate the overall management of diabetes.

3. Literature Review

The recent developments in the field of Artificial Intelligence (AI), Internet of Things (IoT), and Large Language Models (LLMs) have greatly affected the system of diabetes monitoring and control. A literature review has already been done on machine learning in the diagnosis of diabetes, the use of IoT to monitor real-time glucose, and the use of AI-based chatbots to support self-care, and educate patients. IoT systems can be used to constantly monitor physiological activity by monitoring blood glucose, heart rate, and temperature but most of them do not have smart algorithms to extract value out of the information.

In contrast, conversational agents based on LLM enhance health literacy and patient dialogue but tend to be less reliable because they will experience hallucinations and fail to be grounded in domains. To address these shortcomings, Retrieval-Augmented Generation (RAG) has become a promising method to increase the objectivity of facts and offer evidence-based medical answers. Although these advances have been made, there is little work that combines real-time IoT health monitoring (with vital parameters, like Body Mass Index (BMI) and RAG-powered conversational AI to manage diabetes comprehensively. The paper is a review of the current studies in AI-based diabetes prediction, IoT-enabled systems, the use of LLM in healthcare, and RAG frameworks, outlining the most recent achievements and defining a gap in the studies.

3.1. ML and AI in Diabetes Prediction and Management

AI and ML have played a crucial role in the process of diabetes prediction, diagnosis, and management. Khalifa and Albadawy [11] have written about the revolutionary aspect of AI in the context of improving the prevention, diagnosis, and optimization of diabetes. In the same manner, Bhat et al. [24] conducted performance analysis of Type II diabetes-based ML models with optimized feature selection to establish better prediction accuracy. Tansin Ullah Nabil [26] investigated explainable AI methods to predict diabetes, focusing on the interpretability of clinical decision-making.

There is also the concern of mobile-based and ML-driven systems in predicting diabetes. Omar H. Karam [27] suggested a machine learning diabetes prediction mobile application, which emphasizes accessibility and real-time decision support. Challa Deena Vamsi [29] created complete diabetes management equipment that incorporates both ML models and an AI-based chatbot to offer predictive information and patient advice. This research determine that ML is critically involved in earlier detection and risk evaluation, yet in most

cases it is not integrated with real-time physiological monitoring and conversational intelligence based on confirmable sources of knowledge.

3.2. Internet of Things Health-Monitoring Systems in Diabetes

The IoT can be used to achieve ongoing health care monitoring with the help of interrelated sensors and cloud-computer networks. Okandeji Alexander [25] introduced an e-health monitoring system using the IoT platform and showed that the proposed system is affordable, exhibiting real-time data collection and remote monitoring features. On the same note, the possibilities of wireless health tracking solutions have been pointed out by another IoT-based non-invasive

glucose monitoring system, Glue, which is suggested by Rahmat [30].

New technologies have evolved in the form of non-invasive glucose monitoring systems that incorporate biomedical signal processing methods [31]. These IoT applications make patients more accessible and allow tracking them remotely; nevertheless, the majority of these systems concentrate on glucose measurement without considering the broad risk factors like the body mass index or AI-oriented situation recognition. This divide suggests that there is a need to have smart systems that are integrated to combine physiological monitoring with smart analytics.

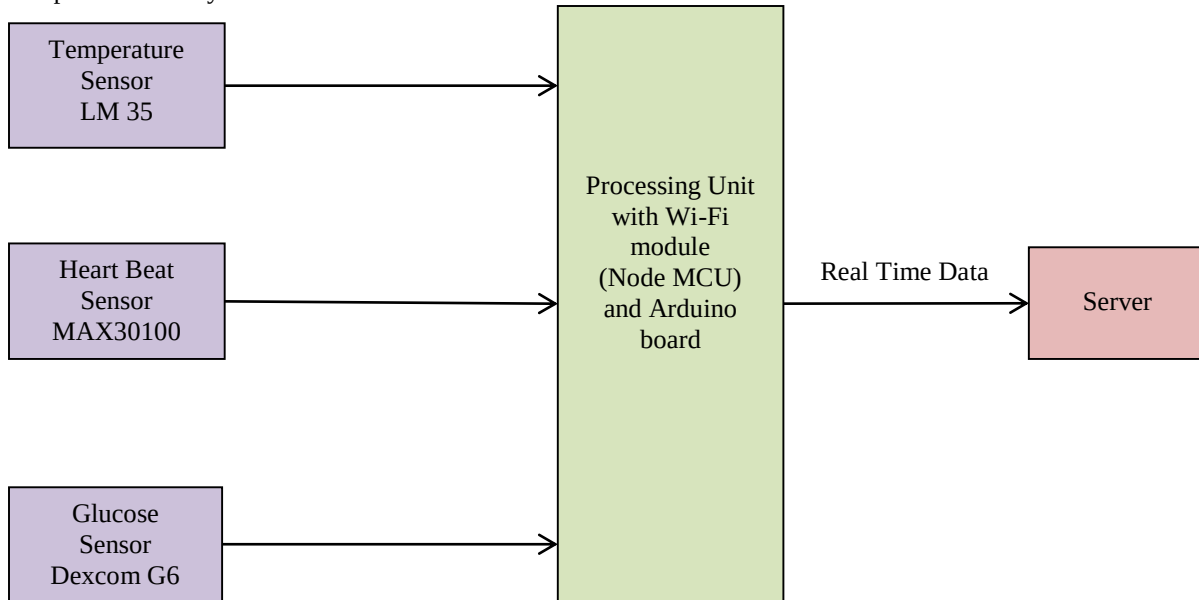


Fig. 1 IoT monitoring system for diabetes

Figure 1 demonstrates a generalized architecture that is commonly taken in the current literature to monitor vital health parameters of the patient with the help of multiple biomedical sensors, such as a temperature sensor (LM35), heart rate sensor (MAX30100), and glucose monitoring device (Dexcom G6). These sensors are connected to a processing unit, which is a microcontroller (Arduino) with a Wi-Fi enabled module (Node MCU). The processing unit combines and works with the sensed data and passes it to a remote server through wireless communication. This allows monitoring, storing, and analysing of patient health information in real-time, facilitating the timely medical intervention and better medical decision-making. This architecture is the starting design followed in most solutions of IoT-based healthcare monitoring reported in recent works.

3.3. Large Language Models (LLMs) in Healthcare

Patient education, triage, and clinical decision support Large Language Models (LLMs) are actively used in the healthcare field. Tushar Mungle [23] presented an overview

of the possibilities of the use of LLM in healthcare with an analysis of opportunities and weaknesses. Järgerstedt and Nilsson [10] tested the performance of LLM in the case of pediatric triage, and the results were encouraging but with a lot of reliability issues.

Conversational AI systems have been studied in diabetes care as a method of enhancing health literacy and patient engagement. Kelly [2] showed that a tailored AI chatbot could help patients with Type 2 diabetes to understand their health conditions better. Krishnamoorthy P [7] explored the use of chatbots to support the self-management of pediatric Type 1 diabetes, whereas Clausen M [16] explored the use of chatbots based on LLM to support gestational diabetes and counseling. According to Shegal [8], a systematic review of conversational agents in diabetes care has proven their potential to increase patient education and adherence. Although these are the benefits, researchers have cited problems like hallucinations, misinformation, and clinical unreliability in generative AI systems [6]. The article by Singh

et al. [21] also addressed risk control strategies that are required when implementing generative health agents safely.

3.4. RAG in Healthcare

In order to overcome challenges of standalone LLMs, Retrieval-Augmented Generation (RAG) has emerged as a powerful solution. RAG improves the reliability of the model by accessing the related medical records and basing answers on the verified knowledge bases. Amugongo et al. [3] of their reviews focused on the RAG applications in healthcare, highlighting factual consistency and clinical safety improvements.

Liu et al. [4] provided the systematical review and clinical guidelines on enhancing the biomedical LLM application through RAG. Ye et al. [18] and Yang et al. [17] also emphasized the significance of the RAG in healthcare generative AI systems. According to Gargari and Habibi [12], RAG is a sentence explaining that hallucinations are reduced through a critical mechanism in medical AI systems.

Hwangbo Y [13] suggested a dual RAG framework using revised guidelines on diabetes to improve reliability and reduce hallucinations. On the same note, Wang et al. [1] showed that RAG is a significant model in enhancing the performance of the LLM in diabetes education as compared to the baseline models. Nawabi et al. [14] presented D-Care, an LLM-based chatbot in diabetes, and Abbasian et al. [20] suggested a knowledge-infused conversational health agent with the use of knowledge graphs and retrieval systems.

The recent achievements like, guided decoding [19] and low-rank adaptation methods of improved RAG [22] only make retrieval-based AI systems more robust and scalable. Kanthathan and Tissera [28] were also keen on semantic search and knowledge-based approaches to enhance accessibility in systems of chronic disease management.

3.5. Analysis of Existing Research

Previous studies have demonstrated that the use of Artificial Intelligence (AI), the Internet of Things (IoT), Machine Learning (ML), Large Language Models (LLMs), and Retrieval-Augmented Generation (RAG) has enhanced diabetes management. Each technology has its own advantages and limitations.

IoT-based health care systems allow for continuous monitoring of various physiological parameters such as blood glucose, heart rate, and body temperature. These systems enhance remote patient monitoring and enable health care workers to have real-time health data. But most current IoT solutions are primarily aimed at collecting and showing sensor data without offering intelligent analysis and personalised suggestions. Also, many systems do not have Body Mass Index (BMI) as an important risk factor for Type 2 diabetes.

The prediction and risk assessment of diabetes have been made with high accuracy by applying Machine Learning models. They can be used to identify patients at risk of developing diabetes. Most prediction models, however, are based on past data sets and not linked to live IoT devices. Because of this, they are unable to monitor the health condition of a patient continually.

LLMs have enhanced patient education by providing straightforward answers to diabetes-related questions. These systems raise the level of health literacy and help patients to manage their disease better. Standalone LLMs, however, can produce inaccurate or hallucinated medical information due to their reliance on pre-trained knowledge.

Many of these issues have been overcome by the use of Retrieval-Augmented Generation (RAG), which retrieves information from trustworthy medical documents and uses it to generate responses. This helps to make healthcare recommendations more reliable and accurate. However, most of the current RAG-based systems focus solely on medical Q&A and cannot be used for the real-time physiological monitoring task.

The literature shows that research into the use of individual technologies and not combined technologies in a single healthcare system. The need of an integrated AI-IoT framework wherein real-time health monitoring, BMI assessment for risk, and RAG-based conversational assistance can be used to offer customized and evidence-based diabetes management solutions still remains.

3.6. Digital Health Interventions and Health Literacy

The use of digital health interventions has shown communication and self-management effects on patients. A systematic review on digital health interventions among adolescents with Type 1 diabetes was done by Tezcan-Güntekin [15], which indicated an improvement in health literacy. The article by Claire Teo [9] discussed the concept of multimodal AI conversations based on the use of the LLM in diabetes prevention, citing the importance of interactive systems in behavioral changes.

On balance, these articles suggest that digital health tools driven by AI can help patients gain much better comprehension, engagement, and adherence. Nevertheless, the majority of the extant literature deals with either the AI-based conversational systems or the IoT-based monitoring in isolation, and does not look into the fusion of both technologies under one system.

As shown in Figure 2, presents full-scale architecture is widely represented in the literature on the prediction and recommendation systems of intelligent diabetes. In this system, a central server is used to receive physiological data sent by sensors through Arduino Nano and Node MCU

devices and process them in real-time. The server unites patient feedback obtained via a chatbot-based user interface and sensor-generated health parameters. A machine learning model is to analyze the inputs which are to predict the type of diabetes. To boost decision support, the Retrieval-Augmented Generation (RAG) engine is utilized to prompt precautions and recommendations depending on the probed condition and

user requests, relying on a body of medical knowledge contextually. The system also keeps a database of user profile, sensor data, prediction, and responses being generated. The presented integrated architecture is the contemporary trend in literature concerning the integration of IoT-based monitoring, AI-based prediction, and advisory systems based on knowledge to support personal healthcare management.

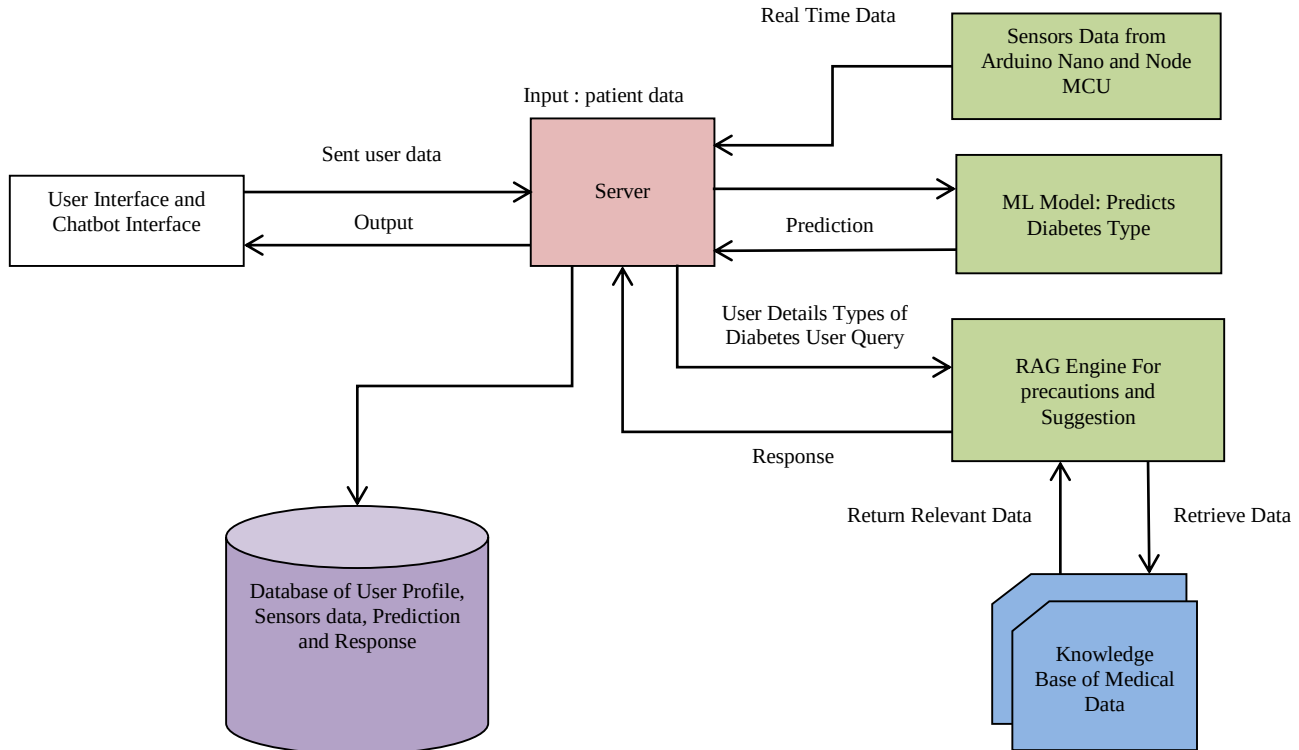


Fig. 2 Architecture of advisory chatbot for diabetics using IoT, ML, and RAG models

3.7. Research Gap

As per the literature: IoT systems are successful in the context of monitoring physiological parameters, but they lack the ability to extract smart insights in context [25, 30, 31]. Chatbots based on LLM have enhanced diabetes education but have challenges of reliability and hallucinations [6, 21]. RAG frameworks improve the accuracy and clinical safety of AI systems in healthcare [1, 4, 13, 17]. Nevertheless, there is little literature that combines real-time IoT-based physiological monitoring (including BMI as a risk factor) and a conversational assistant that runs on RAG to manage diabetes in its entirety. Thus, sensor-based monitoring, cloud connectivity, and the RAG-based and personalized guidance using AI-IoT can be introduced to fill the gap between the raw data on health and the clinical guidance that can be taken into action.

4. Comparative Analysis and Research Insights

The literature reviewed regarding the systems of diabetes management can be generally divided into four areas: (1) the IoT-based monitoring systems, (2) the Machine Learning-

based prediction models, (3) the LLM-based conversational agents, and (4) the Retrieval-Augmented Generation (RAG)-enhanced healthcare system. These approaches are compared and discussed below.

4.1. Internet of Things-based Diabetes Monitors

A number of studies have been directed towards the design of IoT-based health monitoring systems for diabetic patients. Okandeji Alexander [25] designed a low-cost IoT-based e- health monitoring system that has the potential to transmit physiological data in real-time. On the same note, Rahmat [30] presented a non-invasive glucose monitoring system with the use of IoT, and Abdalrhman S [31] proposed a sophisticated non-invasive glucose monitoring system. These systems are good at ensuring uninterrupted glucose monitoring and remote monitoring, but are mainly focused on data retrieval and delivery. They do not have smart decision-support systems and fail to incorporate AI context interpretation. In addition, there are risk factors that are crucial like, Body Mass Index (BMI) that, most of the time are not included in these monitoring systems.

4.2. Prediction Approaches based on Machine Learning

Machine learning methods have been actively used in the prediction of diabetes and risk evaluation. Bhat et al. [24] showed that there was better prediction when optimized feature selection methods were applied. The authors of the study by Tansin Ullah Nabil [26] have integrated explainable AI techniques to increase the level of transparency in diabetes prediction models. H. Karam [27] suggested a mobile-based machine learning system to predict diabetes. Even though such models enhance the accuracy of diagnosis and early detection, they are generally based on pre-existing data as opposed to real-time and physiological information of IoT devices. Moreover, they fail to offer interactive patient instructions or long-term monitoring assistance. Machine Learning model is done based on the following metrics: Accuracy, Precision, Recall, and F1-Score.

$$Accuracy = \frac{TP+TN}{FP+FN+TP+TN} \quad (1)$$

The overall correctness of the prediction model is measured in Equation (1). It displays the number of correct classifications of all the predictions.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Equation (2) represents the number of positive diabetes cases that are predicted as correct. The higher the precision, the less false positive predictions.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

In Equation (3) quantifies how well the model is able to accurately classify real diabetes cases. High recall value indicates that the system can correctly identify more diabetic patients.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

The balance between Precision and Recall is given by Equation (4). When the false positive and false negative predictions must be minimised, this is useful. These are used to evaluate the accuracy and reliability of the system predictions.

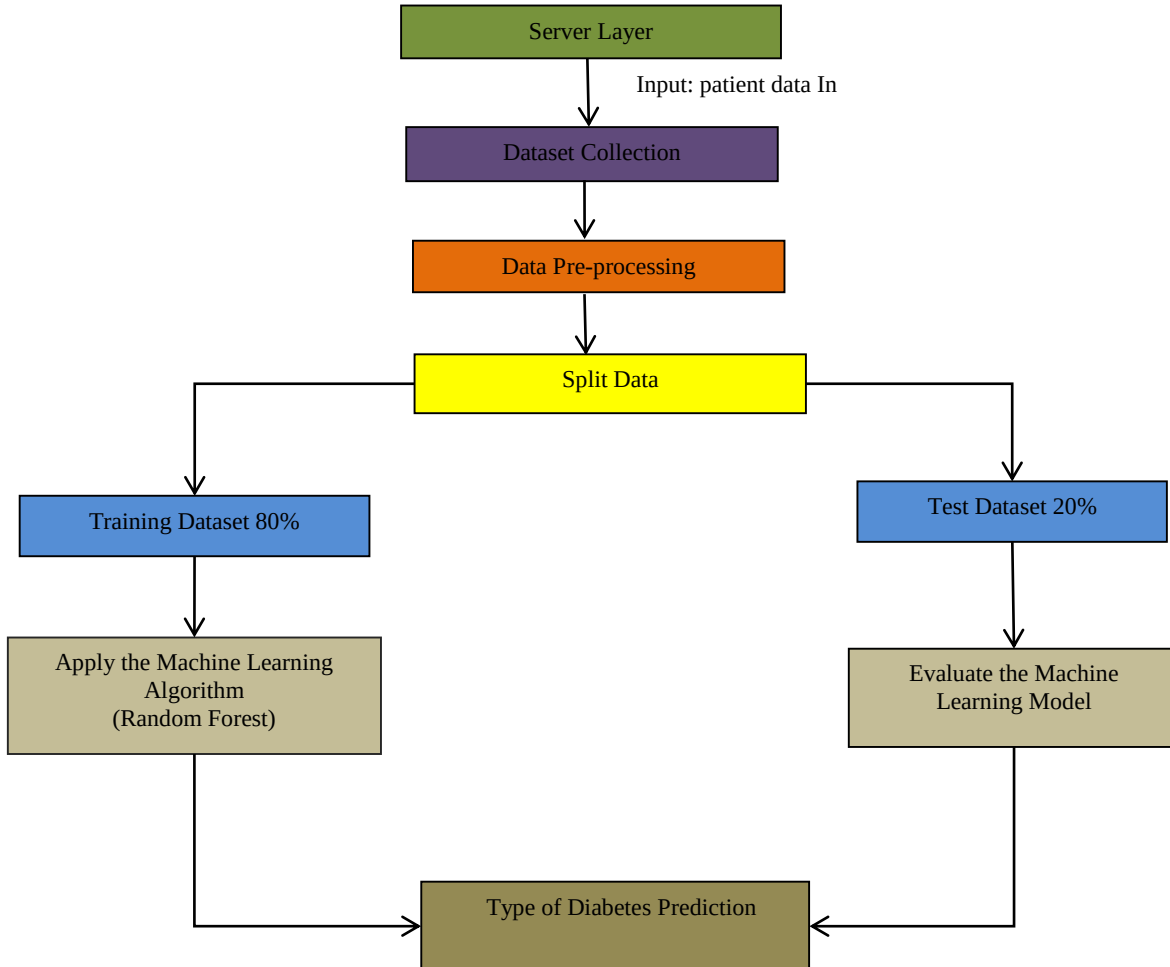


Fig. 3 Architecture of Machine Learning for diabetes type prediction

Figure 3 shows a simplified workflow that is frequently reported and commonly used in the literature to create predictive healthcare models based on information about the patient. The collection of patient information is carried out in this framework by first getting the patient information via a server layer and then putting the data into a structured format. The data obtained is then subjected to preprocessing steps to deal with noise, missing values, and normalization. The finalized dataset is further split into a training and testing set with, the 80:20 ratio being the commonly used. Machine learning algorithms that are being supervised like, Support Vector Machine (SVM) and Random Forest are applied to the training data to create predictive models. The models are tested based on conventional measures such as accuracy, F1-score, recall, ROC curve, and confusion matrix. According to this analysis, the trained system produces forecasts about diabetes. This process is indicative of the standard methodological pipeline used in literature on data driven disease prediction systems.

4.3. Chatbots (LLM-based) to Diabetes Care

Recently, Large Language Models (LLMs) have been used to improve patient engagement and diabetes education. Kelly [2] created a tailor-made AI chatbot, which enhanced health literacy in Type 2 diabetes patients. Krishnamoorthy P. [7] and Clausen M [16] investigated the use of chatbots as a supplement to counseling in the treatment of children and gestational diabetes. The systematic review by Sehgal [8] represents the advantages of conversational agents in the management of diabetes. Although standalone LLM systems have potential in enhancing patient knowledge and attention, hallucinations, misinformation, and clinical reliability are among the challenges associated with the use of these systems [6, 21]. The systems do not have a proper basis in confirmed medical records and are normally not synchronized with the real-time patient health data.

4.4. Retrieval-Augmented Generation (RAG) in Healthcare

In order to improve the reliability and accuracy of facts, Retrieval-Augmented Generation (RAG) has been proposed when used in healthcare. Amugongo et al. [3], Liu et al. [4], and Yang et al. [17] highlighted the fact that RAG could greatly decrease hallucinations because it could base the response on the trusted medical literature. Hwangbo Y [13] showed that the accuracy of responses to queries linked to diabetes was better with a dual RAG framework that is founded on revised diabetes guidelines. Wang et al. [1] demonstrated that RAG-enhanced LLMs are more effective than typical LLMs in the task of diabetes education.

Despite the fact that RAG enhances the reliability of conversations and evidence-based response generation, the majority of the implementations are aimed at improving the functioning of chatbots. They also do not normally incorporate the IoT-based real-time health monitoring data in the retrieval and response generation process.

4.5. Comparative Analysis of Existing Studies

The available studies might be divided into four major domains:

IoT-Based Monitoring Systems: These systems are quite effective to gather physiological data like blood glucose and vital signs in real time. Nevertheless, they are more data transmission platforms with no smart interpretation and individual advisory options.

Machine Learning-Based Prediction Models: ML models enhance the accuracy of the prediction of diabetes risks and their early diagnosis. However, they usually work on stagnant data as opposed to active, real-time observation data.

Chatbots based on LLM in diabetes care: Conversational AI improves patient interaction and health literacy. Nevertheless, standalone LLMs have reliability problems, such as hallucinated or non-evidence-based responses.

RAG-Enhanced Healthcare AI Systems: Retrieval-Augmented Generation RAG is an AI system that enhances the level of factual accuracy and guidelines compliance in AI systems in medical environments. However, the majority of RAG applications are dedicated to fidelity to conversation instead of integration with the IoT-based health monitoring. The comparative analysis of the results in Table 1 demonstrates that the current research in the field of diabetes management is largely concerned with individual elements of technologies as opposed to a systematic framework. IoT-based systems are effective in tracking glucose in real-time and transmitting the data remotely, but practically, they act more as data collection tools, but without introducing an intelligent interpretation tool and critical risk factors like BMI. Machine learning solutions improve the ability to predict diabetes and risk and are generally trained using fixed datasets and do not combine with continuous real-time monitoring systems.

Chatbots based on conversational AI and LLM have demonstrated a high level of improvement in patient engagement, health literacy, and self-management support. However, individual LLM systems have high chances of hallucinations and reliability problems and cannot be directly applied in a clinical setting. The research that discusses safety issues highlights the necessity of systematic protection and validation protocols, but fails to offer an all-encompassing monitoring approach.

One effective solution that has been proposed to improve the accuracy of facts and decrease misinformation in healthcare AI applications is the Retrieval-Augmented Generation (RAG) framework. Although RAG-based systems enhance guideline compliance and reliability during conversations with diabetes-related matters, most applications focus on the performance of the chatbot instead of combining the use of physiological sensors with IoT devices.

In general, the literature suggests a distinct research gap: little has been done to integrate real-time services related to health monitoring, such as calculating BMI, with a conversational assistant, which will be powered by RAG and

give individualized evidence-based advice. This issue inspires the creation of a combined AI-IoT healthcare system that will connect physiological monitoring with smart clinical advice in all-embracing diabetes control.

Table 1. Comparative analysis of the existing study

References	Study Focus	Technology Used	Key Contribution	Limitations
[25, 30, 31]	IoT based monitoring system.	IoT sensors, Cloud platform.	It is used for low-cost remote tracking and real-time glucose monitoring.	It is mainly focused on glucose but not on BMI and AI-based interpretation.
[24, 26, 27]	It is mainly focused on Diabetes prediction.	Machine learning and Explainable AI	It has improved the prediction accuracy and risk assessment.	They have not integrated with the real-time IoT monitoring.
[2, 7, 8, 16]	They have developed the AI Chatbot for diabetes education.	They have used the LLM's and Conversational AI.	It has improved the patient engagement and health literacy.	There is a chance of high hallucination and less verified medical sources.
[6, 21]	It is mainly focused on the safety of generative AI in healthcare.	They have used the LLM evaluation frameworks.	Here they will identify the hallucination risk and need for safeguards.	They do not prove the integrating monitoring system.
[3, 12, 17, 18]	They have included the RAG in health care.	Used the Retrieval-Augmented Generation.	It helps to reduce misinformation and improve factual accuracy.	It is mostly conceptual or systematic reviews.
[14, 20, 29]	It is based on the LLM-based diabetes assistants.	Build the knowledge graphs and LLM chatbots.	It is used for the personalized recommendations and conversational support	Lack of real time sensor data.
[1, 13]	It is an RAG-enhanced diabetes education.	It is dual RAG and Guideline based Retrieval.	It helps to reduce the hallucination and improve the reliability in diabetes responses.	It is mainly focused on the chatbot response but not on physiological monitoring.
[9, 15]	It is focused on Digital health interventions.	Used digital platforms and AI conversation.	It is helpful in behavioral change and health literacy.	There is no integration of the AI-IoT framework.

4.6. Discussion on Integrated AI-IoT-RAG Framework

The comparison indicates the various ways in which different technologies have helped to manage diabetes. Health parameters like blood glucose, heart rate, and body temp can be monitored in real time using IoT-based systems. Most of these systems, however, simply gather and report data, with no intelligent recommendations for their health. Machine Learning models enhance the prediction of diabetes by examining patient information and pinpointing the likelihood of the disease. These models offer good prediction accuracy, but primarily rely on historical data sets and are not linked to monitoring devices. Large Language Model (LLM) chatbots can be used to educate patients about diabetes by addressing their queries and delivering health-related content. But these chatbots can provide wrong or misleading answers as they are not always able to rely on reliable medical information.

Retrieval Augmented Generation (RAG) enhances the trustworthiness of AI-generated content by first pulling information from trusted medical documents before creating

answers. This helps to decrease hallucinations and offer more helpful advice about health and care. Most RAG-based systems, however, are targeted exclusively towards answering questions and are not often able to be integrated with real-time IoT sensor data. In conclusion, comparative results showed the advantages and disadvantages of each technology. Most of the existing systems are based on one technology instead of integrating multiple technologies into a single healthcare solution. Hence, embedding IoT-based real-time monitoring, BMI calculation, and conversation AI powered by RAG can offer a more dependable, customized, and thorough diabetes management system.

5. Challenges and Research Gaps

Most of the IoT-based diabetes monitoring systems tend to have the main aim of gathering and sending physiological information, such as blood glucose. Nevertheless, such systems do not have intelligent AI-based interpretation modules. On the other hand, AI and ML systems are not always based on live sensor data but on fixed datasets. Lack

of seamless connection between the IoT data acquisition and the AI-powered analytics is a concern that lowers the efficiency of the smart healthcare systems.

Most available monitoring systems focus on blood glucose monitoring without attending to other risk indicators like Body Mass Index (BMI), obesity indicators, and lifestyle patterns that are very crucial. As the BMI is closely related to Type 2 diabetes and insulin resistance, this parameter cannot be neglected in the assessment of the risk of diabetes. The behavior of Large Language Models (LLMs) healthcare chatbots creates hallucinated or unproven answers. Unsafe recommendations can come about as a result of misinformation in clinical applications. Even though factual grounding is enhanced with RAG techniques, it is hard to guarantee constant medical reliability.

An IoT system based on chatbots offers generalized recommendations without considering current patient health parameters. Individualized healthcare instructions need a contextual understanding of physiological readings, patient history, and risk factors, which are often lacking in existing solutions. The IoT devices are constantly gathering sensitive health information, which is sent to cloud services. Providing safe communication, data encryption, and adherence to healthcare principles is one of the significant threats to AI-IoT healthcare systems. Some of the suggested AI-based medical systems have only been experimented with or simulated. Their reliability, safety, and effectiveness in the real world are not yet well-validated on a large scale.

Current literature is split into two types, namely, IoT-based monitoring and AI-driven conversational systems. The literature on combining real-time physiological monitoring with RAG-powered intelligent advisory systems is sparse. Although BMI is one of the most important risk factors of Type 2 diabetes, automatic calculation of BMI and risk analysis is not included in many smart health monitoring solutions.

RAG frameworks enhance the reliability of responses but are primarily used with regard to the retrieval of static knowledge. Very little literature integrates RAG with real-time IoT sensor data to produce patient-centered and dynamically changing instructions. The existing systems do not contain state-of-the-art contextual reasoning based on a correlation of various health parameters (glucose, heart rate, temperature, BMI) to give a comprehensive risk assessment.

Despite the suggested safety frameworks, it is necessary to have structured mechanisms that guarantee control of hallucinations and compliance with guidelines, as well as explainability in generative AI healthcare systems. Most of the systems focus on monitoring and prediction and offer limited forms of structure assistance in lifestyle change, diet planning, and exercise advice, depending on patient-specific info.

6. Future Research Directions

In spite of the considerable progress in the areas of AI, IoT, and Retrieval-Augmented Generation (RAG) in diabetes management, several promising research opportunities are still available to study. One of the directions of future study is the development of coherent frameworks that will effectively combine IoT-based physiological monitoring with AI-related analytics and RAG-supported conversational systems. Rather than being individual modules, sensor data acquisition, cloud processing, and intelligent decision support must be a unified ecosystem with the ability to interpret in real time and provide feedback in a personalized way.

The vast majority of existing RAG implementations are based on the work of static document retrieval. The next generation systems must include real-time physiological information (level of glucose, heart rate, temperature, BMI) into the response generation and retrieving pipeline. RAG models that are context-aware and adjust the recommendations based on constantly evolving health parameters will be able to substantially increase individualization and clinical relevance. Because of a high correlation between Type 2 diabetes and obesity, the focus of the research in the future should be extended beyond glucose observance to include changes in BMI, exercise, and diet patterns. Wearable devices and smart scales can be combined with predictive analytics to address lifestyle-related risk factors by its early detection and prevention of interventions.

The new studies ought to investigate multimodal artificial intelligence that incorporates physiological indicators, medical history of the patient, written texts, and potentially medical imaging. Multimodal large language models can facilitate contextual comprehension and enhance holistic risk assessment.

It is important to have transparency and explainability in AI-generated medical recommendations. Further studies are necessary on explainable RAG systems that include citation of sources, references to guidelines, and ways of reasoning in order to enhance confidence on the part of patients and other medical personnel. It is also necessary to develop organized validation frameworks for the generative health agents.

As the use of cloud-based health monitoring continues to rise, data privacy and cybersecurity are two aspects that need more consideration. Future research ought to address edge computing, federated learning, and encrypted data transmission processes such that sensitive patient data can be safely handled without negatively affecting the performance of the system. Most existing AI-based diabetes systems are assessed in controlled experimental conditions. The future studies should focus on longitudinal clinical trials, pilot studies in real life, and partnerships with healthcare facilities to determine effectiveness, usability, and safety in other populations. The new systems that will be set in the future will

no longer be reactive but proactive. Early alerts and tailored behavioral interventions combined with lifestyle analytics can be predicted by AI models to avoid complications and assist patients with managing them instead of treating only the symptoms. The AI healthcare systems are also supposed to include mechanisms of continuous learning, which would update the models with new medical guidelines, patient feedback, and real-time outcomes. Adaptive systems will make sure that the recommendations are up-to-date, precise, and clinical.

7. Conclusion

The management of diabetes should involve constant monitoring and identification of early and proper education of patients in order to avoid serious complications in the long term. The current review explored modern improvements in IoT-based health monitoring systems, machine learning models to predict diabetes, Large Language Model (LLM)-based chatbots, and Retrieval-Augmented Generation (RAG) systems when used in healthcare. It has been analyzed that IoT systems can be used in real-time monitoring of such physiological parameters as blood glucose, heart rate, and temperature, yet they do not necessarily include intelligent

interpretation and a holistic risk analysis. Machine learning methods enhance the accuracy of predictions; they are not usually combined with constant monitoring systems. Chatbots based on the use of LLM's can improve health literacy and patient engagement, but they are associated with difficulties connected to hallucinations and reliability. RAG-based systems enhance the accuracy of facts and compliance with guidelines, although in most cases, real-time sensor data is not introduced.

The key gap in the literature is that there is no entirely coherent AI-IoT system, which would incorporate real-time physiological surveillance, such as Body Mass Index (BMI), and a trusted RAG-driven chatbot. The combination of these technologies can help to close the divide between raw health information, and individualized, evidence-based suggestions. To sum up, future diabetes management systems must aim at the creation of integrated, secure, and understandable AI-IoT systems that can deliver real-time feedback, individualized guidance, and preventive health services. These types of integrated systems can help to enhance patient self-care, increase clinical reliability and improve their long-term health in the future.

References

- [1] Dingqiao Wang et al., "Enhancement of the Performance of Large Language Models in Diabetes Education through Retrieval-Augmented Generation: Comparative Study," *Journal of Medical Internet Research*, vol. 26, pp. 1-12, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [2] Anthony Kelly et al., "The Effectiveness of a Custom AI Chatbot for Type 2 Diabetes Mellitus Health Literacy: Development and Evaluation Study," *Journal of Medical Internet Research*, vol. 27, pp. 1-20, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [3] Lameck Mbangula Amugongo et al., "Retrieval Augmented Generation for Large Language Models in Healthcare: A Systematic Review," *PLOS Digital Health*, vol. 4, no. 6, pp. 1-33, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [4] Siru Liu, Allison B McCoy, and Adam Wright, "Improving Large Language Model Applications in Biomedicine with Retrieval-Augmented Generation: A Systematic Review, Meta-Analysis, and Clinical Development Guidelines," *Journal of the American Medical Informatics Association*, vol. 32, no. 4, pp. 605-615, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [5] Y. Bhanu Sree et al., "Retrieval-Augmented Generation based Large Language Model Chatbot for Improving Diagnosis for Physical and Mental Health," 2024 6th International Conference on Electrical, Control and Instrumentation Engineering, Pattaya, Thailand, pp. 1-8, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [6] Waqar Hussain, and John Grundy, "Advice for Diabetes Self-Management by Chatgpt Models: Challenges and Recommendations," *arXiv preprint arXiv*, pp. 1-13, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [7] Flavia Voiculescu et al., "Natural Language AI Models and Pediatric Type 1 Diabetes: Can Chatbots Help with Diabetes Self-Management and Patient Education?," *JMIR Diabetes*, vol. 10, pp. 1-4, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [8] Moshe Shegal et al., "Conversational Agent Interventions in Diabetes Care: A Systematic Review," *Diabetes Research and Clinical Practice*, vol. 228, pp. 1-13, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [9] Dung Dao et al., "LLM-Powered Multimodal AI Conversations for Diabetes Prevention," *Proceedings of the 1st ACM Workshop on AI-Powered Q&A Systems for Multimedia*, Association for Computing Machinery, New York, NY, USA, pp. 1-6, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [10] Tindra Jargenstedt, and Elin Nilsson, "Applying and Evaluating Large Language Models for Triage at a Paediatric Emergency Department in a Swedish Hospital," Gothenburg University Publications Electronic Archive, pp. 1-132, 2025. [[Google Scholar](#)] [[Publisher Link](#)]
- [11] Mohamed Khalifa, and Mona Albadowy "Artificial Intelligence for Diabetes: Enhancing Prevention, Diagnosis, and Effective Management," *Computer Methods and Programs in Biomedicine Update*, vol. 5, pp. 1-14, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [12] Omid Kohandel Gargari, and Gholamreza Habibi "Enhancing Medical AI with Retrieval-Augmented Generation: A Mini Narrative Review," *Digit Health*, vol. 11, pp. 1-7, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]

- [13] Jaedong Lee et al., “Enhancing Large Language Model Reliability: Minimizing Hallucinations with Dual Retrieval-Augmented Generation based on the Latest Diabetes Guidelines,” *Journal of Personalized Medicine*, vol. 14, no. 12, pp. 1-14, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [14] Awais Khan Nawabi et al., “D-Care: A Multi-Tone LLM-based Chatbot Assistant for Diabetes Patients,” *Biostec Healthinf*, vol. 2, pp. 766-773, 2025. [[Google Scholar](#)]
- [15] Aurélia Naoko Naef et al., “Impact of Digital Health Interventions for Adolescents with Type 1 Diabetes Mellitus on Health Literacy: A Systematic Review,” *BMC Endocrine Disorders*, vol. 23, no. 1, pp. 1-21, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [16] Lukas Lindstrøm et al., “The Potential of Large Language Model-based Chatbot Solutions for Supplementary Counseling in Gestational Diabetes Care,” *Journal of Diabetes Science and Technology*, vol. 18, no. 5, pp. 1247-1248, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [17] Rui Yang et al., “Retrieval-Augmented Generation for Generative Artificial Intelligence in Health Care,” *Npj Health Systems*, vol. 2, no. 1, pp. 1-5, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [18] David J. Bunnell et al., “Bridging AI and Healthcare: A Scoping Review of Retrieval-Augmented Generation-Ethics, Bias, Transparency, Improvements, and Applications,” *medRxiv*, pp. 1-35, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [19] Özgür Uğur et al., “Guided Decoding and its Critical Role in Retrieval-Augmented Generation,” *2025 33rd Signal Processing and Communications Applications Conference (SIU)*, Sile, Istanbul, Türkiye, pp. 1-4, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [20] Mahyar Abbasian et al., “Knowledge-Infused LLM-Powered Conversational Health Agent: A Case Study for Diabetes Patients,” *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Orlando, FL, USA, pp. 1-4, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [21] Matthew R. DeVerna et al., “The Invisible Risks of AI-Generated Health Information,” *arXiv preprint*, 2026. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [22] Yein Choi et al., “Enhanced Retrieval-Augmented Generation using Low-Rank Adaptation,” *Applied Sciences*, vol. 15, no. 8, pp. 1-19, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [23] Mohammed Al-Garadi et al., “Large Language Models in Healthcare,” *arXiv preprint arXiv*, pp. 1-18, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [24] Salliah Shafi Bhat, Gufran Ahmad Ansari, and Mohd Dilshad Ansari, “Performance Analysis of Machine Learning based on Optimized Feature Selection for Type II Diabetes Mellitus,” *Multimedia Tools and Applications*, vol. 84, no. 8, pp. 4945-4964, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [25] Ayodeji Akinsoji Okubanjo et al., “Development of a Low-Cost IoT-based E-Health Monitoring System for Diabetic Patients,” *Journal of Electrical Systems and Information Technology*, vol. 11, no. 1, pp. 1-20, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [26] Isfahuzzaman Tasin et al., “Diabetes Prediction using Machine Learning and Explainable AI Techniques,” *Healthcare Technology Letters*, vol. 10, no. 1-2, pp. 1-10, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [27] Hosam El-Sofany et al., “A Proposed Technique using Machine Learning for the Prediction of Diabetes Disease through a Mobile App,” *International Journal of Intelligent Systems*, vol. 2024, no. 1, pp. 1-13, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [28] Harish Kanthathasan, and Muditha Tissera, “A Knowledge-Driven Approach for Enhancing Chronic Condition Accessibility with Large Language Models and Semantic Search,” *2025 9th International Symposium on Innovative Approaches in Smart Technologies (ISAS)*, Gaziantep, Türkiye, pp. 1-5, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [29] Swathi Buragadda et al., “A Comprehensive System for Diabetes Management and Prediction using AI-Driven Chatbot and Machine Learning Models,” *2025 International Conference on Inventive Computation Technologies (ICICT)*, Kirtipur, Nepal, pp. 97-103, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [30] M.A. Aizat Rahmat et al., “GluQo: IoT-based Non-Invasive Blood Glucose Monitoring,” *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 9, no. 3-9, pp. 71-75, 2017. [[Google Scholar](#)] [[Publisher Link](#)]
- [31] Mohammed H. Al-Jammas et al., “A Non-Invasive Blood Glucose Monitoring System,” *Computers in Biology and Medicine*, vol. 191, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]