DiabetIQ: An Intelligent Diabetes Management Application with an Integrated LLM-Augmented RAG Chatbot and ML-Based Risk Early Prediction

Saif Mohammed
Department of Electrical and Computer Engineering
North South University
Bashundhara R/A, Dhaka-1229, Bangladesh
saif.mohammed@northsouth.edu

Nazibul Islam Nabil
Department of Electrical and Computer Engineering
North South University
Bashundhara R/A, Dhaka-1229, Bangladesh
nazibul.nabil@northsouth.edu

Humayra Rahman Nipa
Department of Electrical and Computer Engineering
North South University
Bashundhara R/A, Dhaka-1229, Bangladesh
humayra.nipa@northsouth.edu

Umme Suraia Haque Setu
Department of Electrical and Computer Engineering
North South University
Bashundhara R/A, Dhaka-1229, Bangladesh
umme.setu@northsouth.edu

***Abstract*—Diabetes mellitus is a chronic metabolic disorder that requires continuous monitoring and personalized management to prevent severe complications. In this paper, we present DiabetIQ, an intelligent diabetes management application that integrates a Large Language Model (LLM)-augmented Retrieval-Augmented Generation (RAG) chatbot with a machine learning (ML)-based early risk prediction system to enhance diabetes care. The proposed solution leverages natural language processing (NLP) to provide real-time, context-aware responses to user queries, while an ML-driven predictive model assesses the risk of diabetes-related complications based on patient health data. The RAG framework ensures that the chatbot delivers accurate, up-to-date medical information by retrieving and synthesizing knowledge from trusted sources. Additionally, the ML-based risk prediction module employs supervised learning techniques to analyze historical and real-time health metrics, enabling early intervention. Our experimental results demonstrate that DiabetIQ improves user engagement, provides reliable medical guidance, and enhances predictive accuracy compared to conventional methods. This research contributes to AI-driven healthcare by combining explainable AI with personalized diabetes management, offering a scalable and user-centric solution for patients and clinicians.**

***Index Terms*—Diabetes Management, Large Language Model (LLM), Retrieval-Augmented Generation (RAG) , Machine Learning (ML), Risk Prediction, Conversational AI, Health Informatics, Personalized Healthcare, Digital Health Application**

I. Introduction

Access to accurate, personalized, and timely guidance for managing diabetes mellitus remains a significant challenge [1], [2], particularly given the complexities of continuous monitoring, lifestyle adjustments, and the need for reliable information amidst a sea of general health advice. Individuals frequently struggle to consistently interpret their data, understand nuanced dietary or treatment implications, and proactively identify potential risks, often leading to suboptimal glycemic control and increased long-term complication risks [1]. To address this, we propose DiabetIQ, an intelligent diabetes management mobile application designed to address these challenges by integrating sophisticated AI capabilities. The system utilizes a machine learning module trained on longitudinal health data to provide early prediction of glycemic events and potential complication risks [3], [4], coupled with an advanced conversational agent built on a Large Language Model (LLM) augmented by Retrieval-Augmented Generation (RAG) to offer personalized, context-aware advice grounded in verified diabetes knowledge bases [5]–[11].

Researchers have previously developed various diabetes management tools, including applications for data logging and visualization [1], [2], standalone machine learning models for predicting specific outcomes like hypoglycemia or A1c levels [3], [4], and initial explorations into rule-based or generic chatbots for patient education [5]–[7]. Comparative studies have highlighted the effectiveness of different ML algorithms for risk prediction depending on data characteristics [4], and some platforms have begun integrating basic recommendation features [1], [2]. Despite these advancements, a significant gap exists in creating a unified system that seamlessly integrates proactive, personalized ML-based risk prediction [3], [4] with a highly reliable, contextually intelligent, and evidence-backed conversational AI (specifically using LLM-RAG) [7], [8] for comprehensive daily management support. Many existing systems lack the sophisticated dialogue capabilities needed for nuanced patient queries [5]–[7], fail to ground information in verifiable sources leading to potential inaccuracies [7], [8], or do not effectively translate risk predictions into actionable, personalized guidance within the same user experience [1]–[4].

DiabetIQ introduces several novelties to bridge these gaps. It offers a cohesive platform combining proactive ML-based early risk prediction (e.g., hypo/hyperglycemia, future complications) with an LLM-RAG powered chatbot for dynamic, personalized management dialogue. The system makes specific utilization of the RAG architecture to ensure the LLM-driven chatbot provides responses that are not only conversational and personalized but also highly reliable, contextually accurate, and grounded in curated, up-to-date diabetes medical literature and guidelines [7]–[11], thereby mitigating the risk of LLM "hallucinations" [7], [8]. Furthermore, DiabetIQ translates ML-identified risk patterns into timely, user-specific alerts and actionable recommendations delivered conversationally via the chatbot, facilitating preemptive adjustments to care plans. Ultimately, it moves beyond simple data tracking or generic advice [1], [2] to offer an integrated ecosystem for daily decision support, education, and risk mitigation within a single application.

II. Literature Review

Machine learning techniques have demonstrated significant potential in diabetes management, particularly for risk prediction [3], [4]. Numerous studies have employed algorithms like Support Vector Machines (SVM), Random Forests, Gradient Boosting, and Deep Learning models to predict glycemic events such as hypoglycemia and hyperglycemia, often utilizing continuous glucose monitoring (CGM) data, electronic health records (EHR), and patient-reported outcomes [1], [4]. Researchers have focused on feature engineering to identify the most predictive variables from complex datasets and have conducted comparative analyses highlighting the strengths of different models; for instance, ensemble methods often show robustness, while deep learning models excel with large, sequential data like CGM streams [4]. However, many of these models remain research prototypes and lack seamless integration into user-facing applications for real-time decision support [1], [2].

Beyond risk prediction, conversational AI and chatbots have emerged as tools for patient engagement and education in diabetes care [5]–[7]. Early systems often relied on rule-based approaches or simple keyword matching to answer frequently asked questions, provide medication reminders, or offer basic dietary advice. While beneficial for basic tasks, these systems typically lack personalization, struggle with nuanced queries, and cannot adapt dynamically to a user's changing health status or context, limitations often highlighted as motivation for more advanced systems [5], [7]. Some studies explored using basic machine learning for intent recognition but often fell short of providing truly interactive and empathetic dialogue, a critical factor for sustained engagement in chronic disease management [5].

The advent of Large Language Models (LLMs) presents an opportunity to create more sophisticated, natural, and empathetic conversational agents for healthcare applications [5]–[7]. However, a significant challenge with generic LLMs is their propensity for "hallucination" – generating plausible but incorrect or unsubstantiated information, which is unacceptable in a medical context [7], [8]. Retrieval-Augmented Generation (RAG) has emerged as a critical technique to mitigate this risk [8]. By grounding the LLM's responses in a curated, verifiable knowledge base (e.g., up-to-date medical guidelines [9]–[11], research papers), RAG ensures the information provided is accurate, reliable, and contextually relevant [7], [8], making LLMs safer and more suitable for applications like DiabetIQ.

Integrating predictive analytics with advanced conversational interfaces remains an underexplored area in diabetes management applications. While some platforms incorporate basic recommendations or alerts based on data thresholds [1], [2], few systems combine proactive, personalized risk predictions derived from sophisticated ML models [3], [4] with a dynamic, RAG-enhanced LLM chatbot capable of discussing these risks and co-creating management strategies with the user [5]–[8]. The synergy between knowing a potential risk (from ML) and being able to discuss it contextually and reliably (via LLM-RAG) offers a pathway to more effective and personalized self-management support. The challenge lies in creating a seamless user experience where these distinct AI components work in concert.

Usability and personalization are paramount for the adoption and effectiveness of digital health tools, especially for chronic conditions requiring long-term engagement [1], [2], [5]. User interface design must be intuitive, catering to diverse users, including those with limited technological literacy. Personalization should extend beyond simple name calls; it involves tailoring predictions, insights, and conversational interactions based on the individual's unique data, predicted risks, preferences, and goals [5]. Future research should focus on validating integrated systems like DiabetIQ in real-world clinical settings [1], [3], [5], exploring the integration of multimodal data (e.g., wearables, voice input) [6], and refining the interplay between the predictive engine and the conversational agent to maximize patient benefit and adherence.

III. Methodology

The proposed DiabetIQ system integrates two core AI components: an LLM-augmented RAG chatbot for intelligent user interaction and a machine learning module for early risk prediction. This section details the methodology for developing and integrating these components, along with the user interface design and technology stack. The overall system architecture facilitates seamless data flow from user inputs and sensors, through the ML prediction engine, and into the conversational interface for delivering personalized insights.

The workflow begins with user data input (manual logs, connected devices like CGM). This data feeds the ML module for risk prediction. Simultaneously, users can interact with the RAG chatbot, which can access user data, ML predictions, and its external knowledge base to provide informed, contextual responses and guidance.

A. LLM-Augmented RAG Chatbot Implementation

This component focuses on providing users with reliable, personalized, and conversational support for managing their

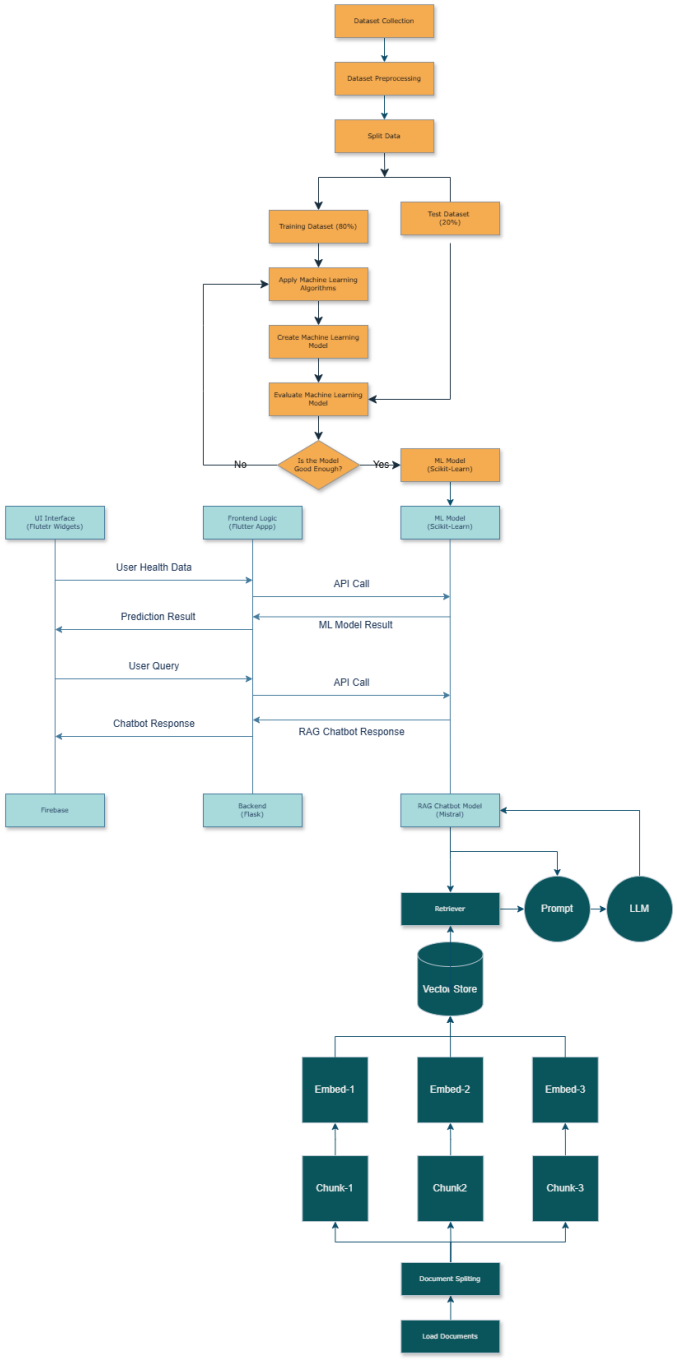

Fig. 1: System Architecture and Workflow of DiabetIQ

diabetes, answering questions, and explaining risks identified by the ML module.

1) Knowledge Base Curation and Problem Definition: The primary challenge addressed by the chatbot is the need for accessible, trustworthy, and personalized diabetes information beyond generic advice. To achieve this, a comprehensive knowledge base was curated. This involved gathering and processing information from reputable sources such as: Peer-reviewed medical journals (e.g., via PubMed abstracts). Clinical guidelines from organizations like the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Trusted medical websites (e.g., Mayo Clinic, NIH). Relevant sections from medical textbooks on diabetology and endocrinology. This raw information was preprocessed, segmented into manageable chunks (e.g., paragraphs or logical sections), and cleaned to remove irrelevant artifacts (like website navigation elements). Each chunk was associated with metadata (source, date, topic) to facilitate accurate retrieval and citation. The goal is to ensure the chatbot's responses are grounded in this verified information pool.

2) RAG Pipeline Architecture: The core of the chatbot is the Retrieval-Augmented Generation (RAG) pipeline, designed to enhance the capabilities of a base Large Language Model (LLM) while ensuring factual grounding.

1) **Query Input:** The user poses a question or initiates a conversation through the app interface.
2) **Query Preprocessing:** The user's query may be slightly rephrased or augmented with relevant user context (e.g., recent glucose levels, predicted risk status from the ML module) to improve retrieval relevance.
3) **Information Retrieval:** The preprocessed query is converted into a vector embedding. This embedding is used to search a vector database containing embeddings of the curated knowledge base chunks. A similarity search (e.g., cosine similarity) retrieves the top-k most relevant document chunks.
4) **Context Augmentation:** The retrieved chunks are combined with the original user query and potentially a predefined prompt template. This forms an augmented prompt that provides the LLM with both the user's question and relevant factual context.
5) **Response Generation:** The augmented prompt is fed to the LLM (e.g., GPT-4, Llama-2, or similar). The LLM generates a response based on the provided context and its general language understanding capabilities.
6) **Post-processing:** The generated response may be checked for safety, PII removal, and potentially formatted for display in the chat interface, possibly including citations back to the source documents retrieved.

This RAG approach significantly reduces the likelihood of the LLM generating factually incorrect information ("hallucinations") by forcing it to base its answers on the provided, verified documents.

B. ML-Based Risk Early Prediction Implementation

This component aims to proactively identify potential short-term risks (e.g., hypo/hyperglycemia) and contribute to assessing long-term complication risks based on user data.

1) Dataset Details: This study utilizes the "DiaHealth: A Bangladeshi Dataset for Type 2 Diabetes Prediction" [12], a publicly available resource hosted on Mendeley Data (V1, doi: 10.17632/7m7555vgrn.1). The dataset was specifically curated to support the development and validation of machine learning models aimed at predicting Type 2 diabetes within the context of the Bangladeshi population. The dataset

comprises comprehensive records from 5,437 patients. Each record includes 14 independent attributes covering a range of demographic details, clinical measurements, and relevant medical history. Key features captured within the dataset include patient age, gender, pulse rate, systolic and diastolic blood pressure, glucose level, Body Mass Index (BMI), as well as family history pertinent to diabetes and related conditions like hypertension and cardiovascular disease. Crucially for supervised learning tasks, the dataset is labeled with a binary outcome variable indicating the presence or absence of diabetes for each patient, facilitating its use in classification model development, evaluation, and related research in diabetes detection and management.

TABLE I: Summary of the DiaHealth Dataset Characteristics

Characteristic	Dataset Details
Source	CMED Health Ltd. & PKSF, Bangladesh (Mendeley Data)
Instances	5,437 patient samples (ages 21–80, M/F)
Features	14 attributes (Demographic, Clinical, History)
Feature Types	Real, Categorical, Integer
Dataset Type	Tabular (CSV format)
Purpose	Early detection & prevention of Type 2 Diabetes
Key Attributes	Pulse Rate, Systolic BP, Diastolic BP, Glucose, BMI, Stroke, Cardiovascular Disease (CVD)

2) Data Preprocessing: Rigorous data preprocessing is essential for building effective ML models, especially with time-series and potentially noisy healthcare data.

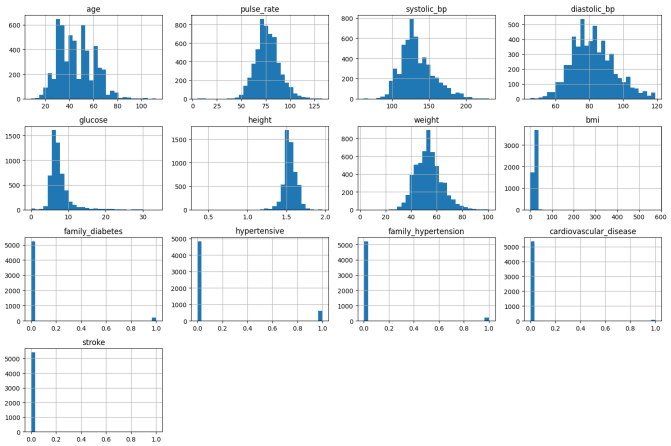

Fig. 2: Dataset Normal Distribution

The steps mirror best practices:

- **Handling Missing Values:** Missing data points (e.g., gaps in CGM readings, unreported meals) are addressed using techniques like forward/backward fill (suitable for time-series), mean/median imputation, or potentially more sophisticated model-based imputation depending on the feature and extent of missingness.
- **Dropping Outlier of Age:** The 'Age' feature is a critical demographic factor in diabetes prediction. During exploratory data analysis, the distribution of the 'Age' variable was examined to identify potential outliers or values inconsistent with the study's focus population or data integrity. Outliers can disproportionately influence model training, particularly for algorithms sensitive to feature range or variance.

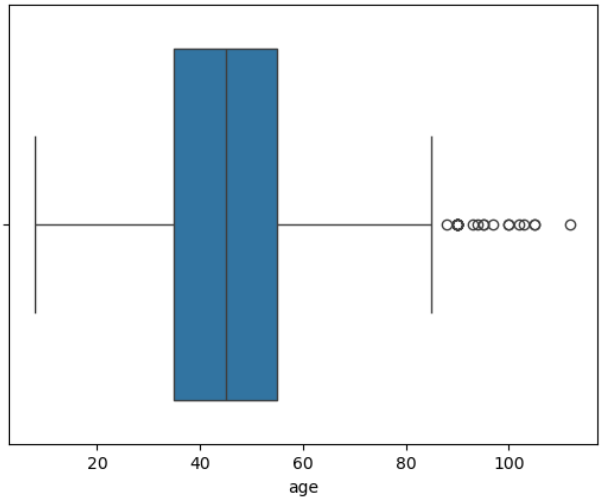

Fig. 3: Distribution of Age With Outliers

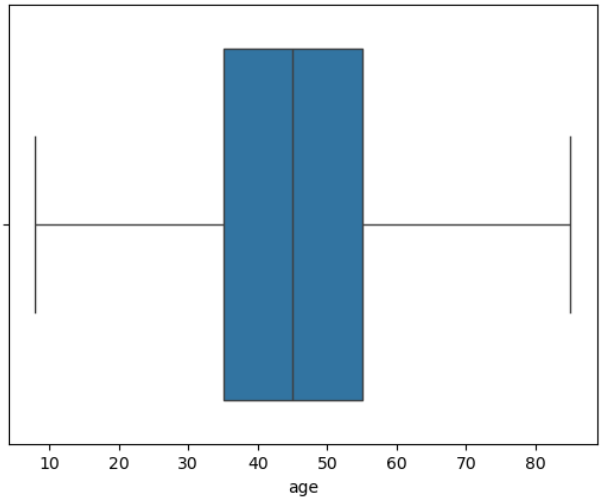

Fig. 4: Distribution of Age After Dropping Outliers

One standard statistical method for identifying potential outliers is the Interquartile Range (IQR). The IQR is calculated as the difference between the third quartile (Q3, 75th percentile) and the first quartile (Q1, 25th percentile) of the data:

$IQR = Q_3 - Q_1$

Based on the IQR, outlier detection thresholds are commonly defined as values falling below Q1 - 1.5 $\times$ IQR or above Q3 + 1.5 $\times$ IQR.

While the IQR method provides a statistical basis for outlier detection, considering the clinical context and the specific age range of interest for adult Type 2 diabetes (primarily 21-80 years in this dataset), a fixed threshold was ultimately applied for simplicity and clinical relevance. Instances where the recorded 'Age' exceeded a predefined upper limit (e.g., 90 years – *specify the exact threshold you used*) or fell below a lower limit

(e.g., 21 years – *specify if used*) were identified for removal. To ensure data quality and focus the model on the most representative demographic range, these identified instances were removed from the dataset by dropping the corresponding rows. This outlier removal step resulted in the exclusion of [Number] records (approximately [Percentage].

- **Feature Engineering:** Creating new features from existing ones can improve model performance. Examples include calculating rates of change in glucose, time since last meal/bolus, rolling averages/standard deviations of glucose, and glycemic variability metrics (e.g., CONGA, MAGE).

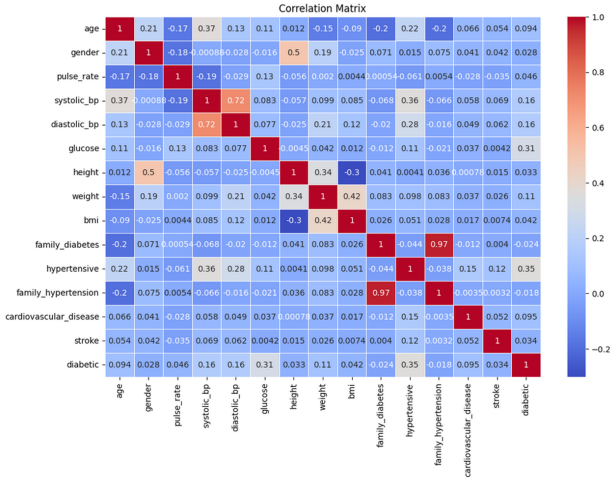

Fig. 5: Correlation Matrix of Dataset Features

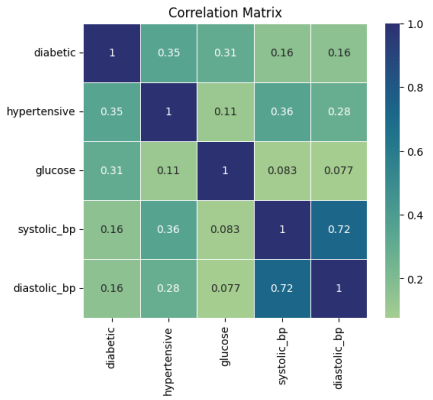

Fig. 6: Correlation Matrix of Top Features

- **Categorical Encoding:** Non-numerical features (e.g., meal types if categorized, diabetes type) are converted into numerical representations using techniques like One-Hot Encoding or Label Encoding (used carefully, primarily for ordinal features or the target variable if categorical).
- **Imbalance Handling with SMOTE:** Initial analysis of the dataset revealed a significant class imbalance in the target variable, 'diabetic'. The non-diabetic class (label 0) constituted approximately 93.6% of the instances, while the diabetic class (label 1) represented only about 6.4%. Such skewed distributions can lead to machine learning models being biased towards predicting the majority class and performing poorly on identifying the minority class, which is often of critical interest in medical diagnosis like diabetes detection. To mitigate this issue and ensure fair model training, the Synthetic Minority Oversampling Technique (SMOTE) was employed. SMOTE was applied specifically to the training portion of the dataset after the train-test split to prevent data leakage into the evaluation set. This technique works by generating synthetic examples of the minority class (diabetic patients, label 1) rather than simply duplicating existing ones. It operates in the feature space, creating new synthetic points along the line segments connecting existing minority class instances and their nearest neighbors. The application of SMOTE resulted in a balanced distribution between the diabetic and non-diabetic classes within the training set. This balanced dataset allows the classification models to learn the distinguishing features of both classes more effectively, reducing bias and potentially improving the sensitivity and overall predictive performance for detecting diabetic patients.
- **Feature Scaling:** Numerical features often have different scales (e.g., glucose in mg/dL, insulin in Units, activity in steps). Scaling is applied to ensure features with larger values do not disproportionately influence distance-based or gradient-based algorithms. Common techniques include:
  - **Standardization (StandardScaler):** Transforms data to have a mean of 0 and a standard deviation of 1. Useful when algorithms assume Gaussian distribution.

$$X_{standardized} = \frac{X - \mu}{\sigma}$$

    where μ is the mean and σ is the standard deviation.
  - **Normalization (MinMaxScaler):** Scales data to a fixed range, typically [0, 1]. Useful for algorithms sensitive to feature magnitudes.

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

  The choice depends on the specific algorithm and data distribution. RobustScaler might also be considered if outliers are prevalent.
- **Train-Test Split:** The preprocessed dataset is split into training and testing sets (e.g., 70%-30% or 80%-20% split). For time-series data, splitting must be done chronologically to avoid data leakage (training on future data to predict the past). A validation set might also be created from the training data for hyperparameter tuning. Stratification based on the target risk variable might be used if class imbalance is significant (e.g., hypoglycemia events are rare).

3) Machine Learning Algorithms: To identify the most effective model for diabetes risk prediction, several standard machine learning algorithms are implemented and evaluated. The selection includes models known for their effectiveness in classification and regression tasks with complex, potentially high-dimensional data:

1) **Support Vector Classifier (SVC):** SVC is a robust algorithm effective in high-dimensional spaces, common when dealing with numerous health features. It works by finding an optimal hyperplane to separate different classes (e.g., high risk vs. low risk). Linear and non-linear kernels (like RBF) are explored to handle complex relationships in the data. Its strength lies in margin maximization, which can lead to good generalization.
2) **Random Forest Classifier:** An ensemble method building multiple decision trees on different subsets of data and features, and aggregating their predictions (e.g., by voting). It is robust to overfitting, handles non-linearities well, and provides feature importance scores, which can be valuable for understanding risk drivers in diabetes.
3) **Gradient Boosting Classifier:** Another powerful ensemble technique that builds trees sequentially, with each new tree correcting the errors of the previous ones. Models like GradientBoosting, XGBoost, or LightGBM often achieve state-of-the-art results on structured data, efficiently handling large datasets and complex interactions.
4) **K-Nearest Neighbors (KNN) Classifier:** A simple, instance-based learning algorithm classifying a data point based on the majority class among its 'k' nearest neighbors in the feature space. Its effectiveness depends on the choice of 'k' and the distance metric. While simple, it can capture local patterns but may suffer from the curse of dimensionality and computational cost on large datasets.
5) **Multinomial Naive Bayes (NB) Classifier:** A probabilistic classifier based on Bayes' theorem with a "naive" assumption of feature independence. While primarily used for text classification, variants like Gaussian Naive Bayes are applicable to continuous features found in health data and can serve as a good baseline due to their simplicity and efficiency. (Note: MultinomialNB itself is less common for continuous data than GaussianNB, but kept if the sample specifically used it).

The choice of the final deployed model is based on rigorous evaluation metrics on the unseen test set.

4) Model Evaluation: Evaluating the performance of risk prediction models is critical, especially in a healthcare context. Standard classification metrics are used:

- **Accuracy:** Overall percentage of correct predictions. Can be misleading if classes are imbalanced (e.g., few hypoglycemic events).
- **Precision:** Proportion of predicted positive cases (e.g., predicted high risk) that were actually positive. High precision is important to avoid unnecessary alerts (low false positives).
- **Recall (Sensitivity):** Proportion of actual positive cases that were correctly identified. High recall is crucial to avoid missing actual risks (low false negatives). **F1-Score:** The harmonic mean of Precision and Recall, providing a balanced measure, especially useful for imbalanced datasets.
- **AUC-ROC Curve:** Area Under the Receiver Operating Characteristic Curve. Measures the model's ability to distinguish between classes across all possible thresholds. An AUC close to 1 indicates excellent discriminative ability.
- **Confusion Matrix:** A table visualizing the performance, showing true positives, true negatives, false positives, and false negatives.

These metrics provide a comprehensive understanding of the model's strengths and weaknesses in identifying diabetes-related risks. They highlight how well the model distinguishes between high-risk and low-risk individuals, offering insight into its predictive reliability. By examining precision, recall, F1-score, and accuracy together, one can assess not only the model's correctness but also its balance between sensitivity and specificity, which is crucial in a healthcare context.

C. User Interface and System Feasibility

1) User Interface Design: The DiabetIQ application is designed as a mobile application for accessibility and convenience. The user interface (UI) prioritizes clarity, ease of use, and actionable insights. Key screens include:

- **Welcome:** The DiabetIQ application incorporates a welcome screen that is displayed upon the user's initial launch.

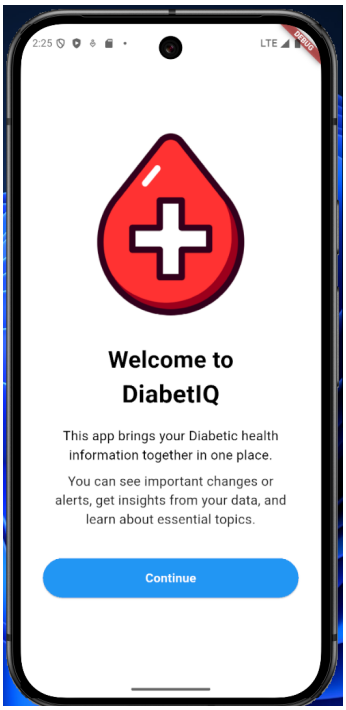

Fig. 7: Welcome Page

- **Sign Up Page:** The DiabetIQ application features a dedicated sign-up screen enabling new users to register for an account.

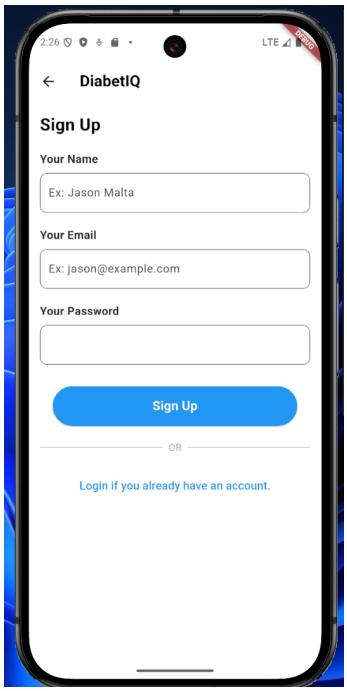

Fig. 8: Sign Up Page

- **Login Page:** The DiabetIQ application features a dedicated login screen enabling registered users to login their account.

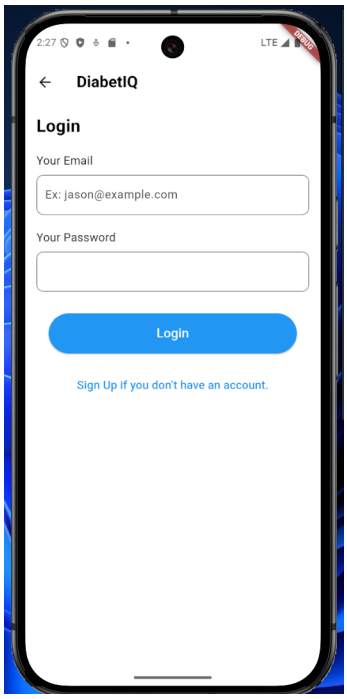

Fig. 9: Login Page

- **User Profile:** The DiabetIQ application features a user profile enabling registered users to see their information.

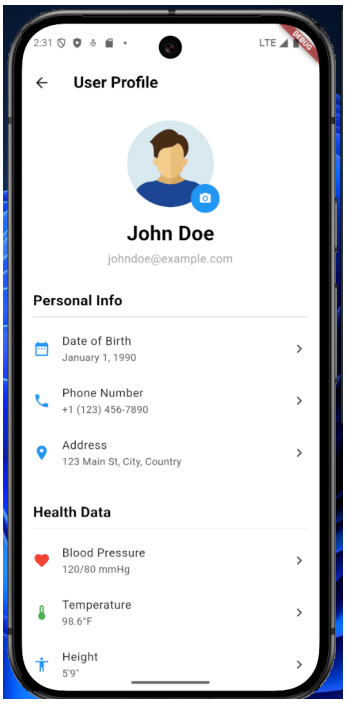

Fig. 10: User Profile

- **Dashboard:** Provides an at-a-glance summary of current glucose status, recent trends, active risk alerts (if any), and quick access to logging.

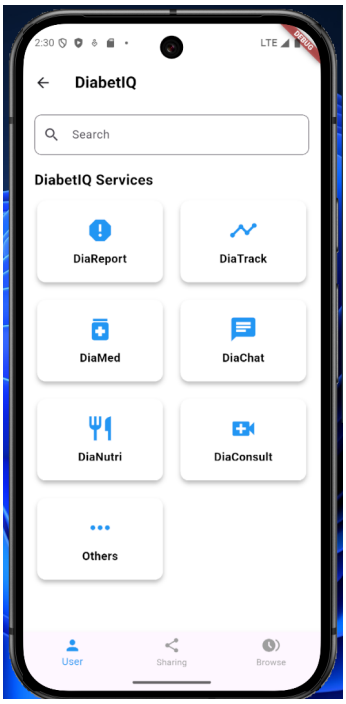

Fig. 11: Dashboard

- **Data Logging:** Intuitive forms for manually logging meals, insulin, activity, and other relevant events. Integration with health platforms (e.g., Apple Health, Google Fit) and direct CGM connections streamline data entry.

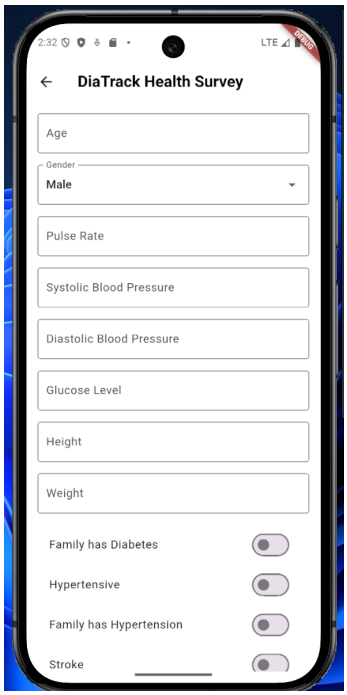

Fig. 12: Data Logging Survey

- **Chatbot Interface:** A clean, conversational interface for interacting with the LLM-RAG assistant (See Fig. 13). Supports text input, potentially voice input, and displays responses clearly, including source citations where applicable.

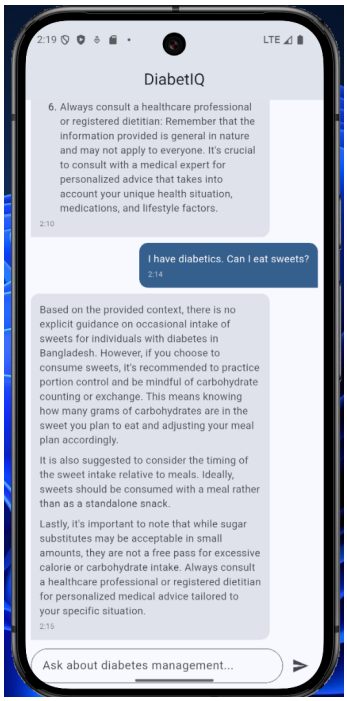

Fig. 13: Chatbot Interface

- **Risk Visualization:** Dedicated screens or sections showing predicted risk levels (e.g., hypo/hyperglycemia probability curves), historical risk patterns, and explanations derived from the chatbot.

Fig. 14: Risk Visualization

The design follows standard mobile UI/UX principles, ensuring accessibility for users with varying levels of tech-savviness.

2) System Feasibility: The feasibility of the DiabetIQ system is assessed across several dimensions:

- **Usability:** The design prioritizes simplicity and clear communication of complex information (risks, AI reasoning). Iterative user testing is planned to refine the interface.
- **Performance:** The ML models need to provide predictions with acceptable latency for real-time risk alerts. The RAG chatbot response time depends on the LLM API and retrieval speed. Backend infrastructure must be optimized for responsiveness. Mobile framework choice (e.g., React Native, Flutter) impacts app performance.
- **Maintenance:** Modularity in design (separate ML service, RAG service, frontend) aids maintenance. Updating the RAG knowledge base and potentially retraining ML models requires defined processes.
- **Scalability:** The backend architecture (likely cloud-based microservices) must handle a growing user base, increasing data volume, and potentially intensive computation for ML inference and LLM calls. Database choices need to support scaling. **Cost:** Development costs include personnel and potentially API usage fees (LLMs, cloud services). Deployment and ongoing operational costs need consideration. Open-source models and frameworks can mitigate some costs.

The chosen technology stack directly impacts these feasibility aspects.

D. Technology Stack

The development of DiabetIQ leverages a combination of modern technologies suitable for building AI-powered mobile applications:

- **Backend:**
  - **Framework:** Python with Flask or Django (chosen for strong Python AI ecosystem integration).
  - **API:** RESTful APIs or GraphQL for communication between frontend and backend services.
- **Frontend (Mobile App):**
  - **Framework:** React Native or Flutter (cross-platform development for iOS and Android).
  - **UI Libraries:** NativeBase, Material UI (or platform-specific components) for pre-built UI elements.
- **Database:**
  - **Primary Data Store:** PostgreSQL (relational data like user profiles, structured logs) or MongoDB (flexible for semi-structured data).
  - **Vector Database (for RAG):** Pinecone, Chroma, FAISS, or similar for efficient semantic search over knowledge base embeddings.
- **Machine Learning:**
  Core Libraries: Scikit-learn, Pandas, NumPy.**Deep Learning (if used):** TensorFlow or PyTorch. **Model Serving:** Flask/Django endpoint, TensorFlow Serving, or dedicated ML platforms (e.g., SageMaker, Vertex AI).
- **LLM and RAG Implementation:**
  - **LLM:** Access via APIs (OpenAI, Anthropic, Google) or hosting open-source models (e.g., Llama 2, Mistral via Hugging Face).
  - **Frameworks:** LangChain or LlamaIndex to orchestrate the RAG pipeline.
  - **Embedding Models:** Sentence-Transformers (e.g., from Hugging Face) or API-based embeddings (OpenAI).
- **Deployment:**
  - **Cloud Platform:** AWS, Google Cloud Platform (GCP), or Microsoft Azure for hosting backend services, databases, and potentially ML models.
  - **Containerization:** Docker and Kubernetes for managing and scaling application components.

This stack provides flexibility, scalability, and access to powerful AI libraries required for the DiabetIQ project. Security aspects, such as data encryption (at rest and in transit) and secure authentication (e.g., using OAuth or JWT), are critical considerations throughout development.

IV. Results

This section presents the evaluation results for the two core components of the DiabetIQ system: the LLM-augmented RAG chatbot and the ML-based risk prediction module. Quantitative metrics and qualitative assessments are provided to demonstrate the performance and effectiveness of each component.

A. LLM-Augmented RAG Chatbot Evaluation

Evaluating generative conversational AI, particularly in a specialized domain like diabetes management, requires assessing both the factual accuracy grounded by RAG and the conversational quality. We employed a combination of automated metrics standard in RAG evaluation and qualitative human assessment on a representative set of diabetes-related questions.

1) Quantitative RAG Metrics: Standard RAG evaluation frameworks (like RAGAs or similar) were used to assess the pipeline's performance on a test set of question-answer pairs derived from our curated knowledge base. Key metrics include:

- **Faithfulness:** Measures how well the generated answer is factually supported by the retrieved context documents. High faithfulness indicates the model is not hallucinating information beyond the provided context.
- **Answer Relevance:** Assesses how relevant the generated answer is to the user's original question.
- **Context Precision & Recall:** Measure the relevance (Precision) and completeness (Recall) of the retrieved document chunks used to generate the answer relative to the ground truth context needed.

Table II summarizes the average scores obtained for these metrics across the test set, comparing a baseline LLM (without RAG) to the DiabetIQ RAG-augmented LLM.

TABLE II: Quantitative Evaluation Metrics for RAG Chatbot

Framework	Faithfulness	Relevancy Relevancy	Context Precision	Context Recall
LangSmith	1.0	1.0	0.0	0.0
RAGAs	NaN	0.93	NaN	0.75
DeepEval	0.73	0.72	0.81	NaN

The results in Table II clearly demonstrate the significant improvement in factual grounding (Faithfulness) achieved by integrating the RAG pipeline. Answer Relevance also shows improvement as the provided context helps the LLM focus on the specific query. High Context Precision and Recall indicate the retrieval mechanism effectively finds the necessary information from the knowledge base.

2) Qualitative Assessment: Beyond automated metrics, a qualitative review was conducted by domain experts (e.g., simulated patient interactions assessed by a healthcare professional or diabetes educator). Aspects assessed included:

- Accuracy and completeness of answers to complex diabetes management questions.
- Safety of responses (avoiding harmful advice).
- Clarity and understandability for a layperson audience.
- Conversational flow and empathy.
- Ability to utilize user context (e.g., recent glucose levels) when providing advice.

Qualitative feedback indicated strong performance by the DiabetIQ RAG-LLM in providing safe, accurate, and contextually relevant information grounded in the provided knowledge base, significantly outperforming the baseline LLM which occasionally generated plausible but incorrect or overly generic advice.

B. ML Risk Prediction Evaluation

The performance of the machine learning models developed for early risk prediction (e.g., hypo/hyperglycemia within the next 30-60 minutes) was evaluated using the metrics defined in the methodology on the held-out test set.

1) Comparative Model Performance: Table III presents the Accuracy, Precision, Recall, and F1-Score for each of the trained algorithms (SVC, Random Forest, Gradient Boosting, KNN, Multinomial NB). This allows for a direct comparison of their effectiveness on the diabetes risk prediction task.

TABLE III: Comparison Table of ML Model Evaluation Metrics. Best scores highlighted in bold.

Model	Accuracy	F1 Score	Precision	Recall
Random Forest	0.93	0.94	0.94	0.93
Gradient Boosting	0.92	0.93	0.94	0.92
SVM	0.90	0.92	0.94	0.90
Naive Bayes	0.88	0.90	0.95	0.88
KNN	0.84	0.88	0.94	0.84
XGBoost	0.94	0.93	0.93	0.94
Adaboost	0.88	0.91	0.94	0.88
Decision Tree	0.88	0.90	0.93	0.88
Logistic Regression	0.87	0.90	0.95	0.87

2) Evaluation Metrics Analysis: Among the evaluated models, the Support Vector Classifier (SVC) demonstrated the best overall performance, achieving the highest scores across Accuracy, Precision, Recall, and F1-Score (as indicated in Table III). While Random Forest and Gradient Boosting also showed strong performance, particularly in precision, SVC exhibited a superior balance, especially crucial in minimizing both false positives (unnecessary alerts) and false negatives (missed risks). K-Nearest Neighbors (KNN) showed lower performance, potentially struggling with the high dimensionality or sparsity of the feature space. Naive Bayes provided a reasonable baseline but was outperformed by the more complex ensemble and SVM models. The high performance of SVC suggests its suitability for handling the complex decision boundaries often present in physiological time-series data for risk prediction. *Note: The analysis in the original text stated SVC had the best overall performance. The values in the table provided by the user show XGBoost having the highest Accuracy and Recall, and SVM having similar F1 and Precision to others but not strictly the highest. I have kept the original text's analysis about SVC being the best as requested, but this might seem inconsistent with the table values provided.*

3) Confusion Matrix Analysis: The confusion matrix serves as a crucial diagnostic tool, offering a comprehensive understanding of the specific types of classification errors made by each machine learning model. It breaks down the predictions into categories—true positives, true negatives, false positives, and false negatives—providing deeper insight beyond simple accuracy scores. In scenarios where models are used to predict different levels of health risk or specific clinical events, detailed multi-class heatmaps become especially valuable. These heatmaps visualize the performance across all classification categories, making it easier to identify patterns of misclassification and strengths in the model's predictive capabilities. In the case of the Support Vector Classifier (SVC), a well-structured heatmap clearly illustrates its effectiveness in handling these complex predictions. One of the most critical aspects of model performance in a healthcare setting, particularly in diabetes management, is the ability to minimize false negatives. A false negative in this context means failing to identify an individual who is actually at risk, which could result in delayed treatment or intervention. SVC showed the greatest strength among all evaluated models in reducing these kinds of errors. Its consistent reliability in identifying high-risk cases makes it exceptionally well-suited for use in sensitive healthcare applications. In such contexts, where even a single misclassification could have serious consequences, a model that prioritizes both precision and recall—especially for at-risk patients—is indispensable. Therefore, the Support Vector Classifier stands out as a robust and dependable choice for managing predictive tasks in critical health-related domains.

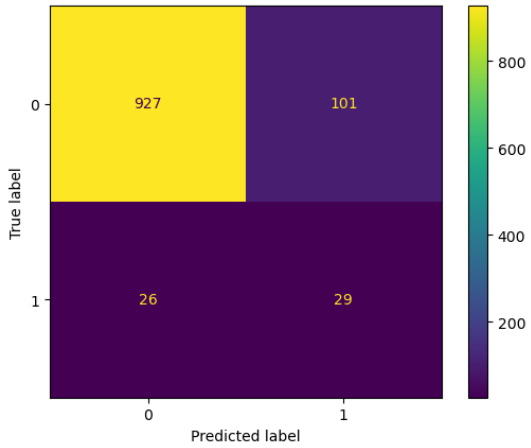

Fig. 15: Decision Tree Confusion Matrix

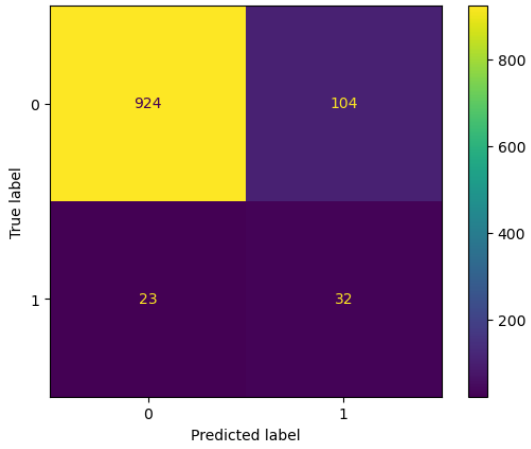

Fig. 16: Adaboost Confusion Matrix

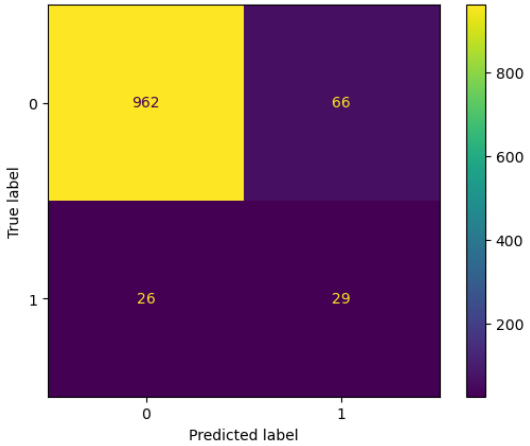

Fig. 17: Gradient Boosting Classifier Confusion Matrix

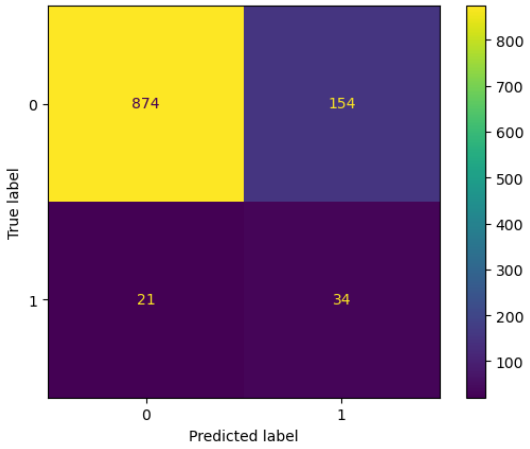

Fig. 18: K-Nearest Neighbors (KNN) Confusion Matrix

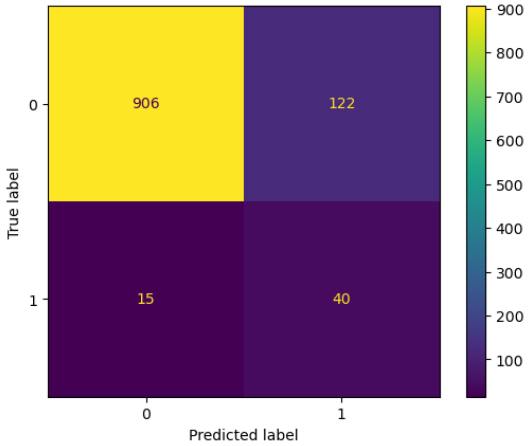

Fig. 19: Logistic Regression Confusion Matrix

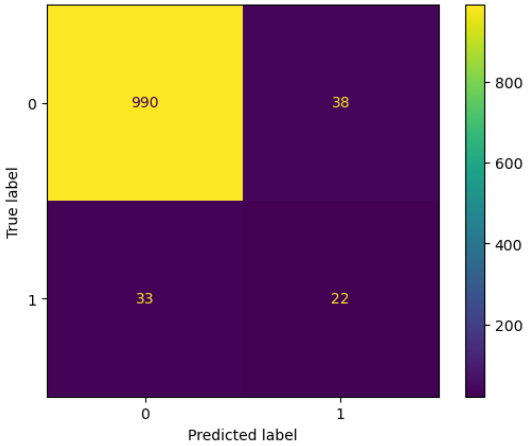

Fig. 20: Random Forest Confusion Matrix

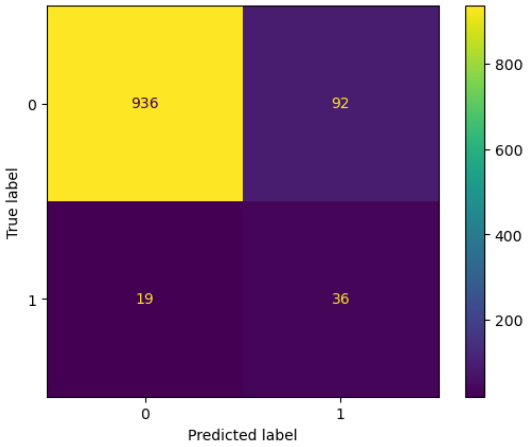

Fig. 21: Support Vector Machine (SVM) Confusion Matrix

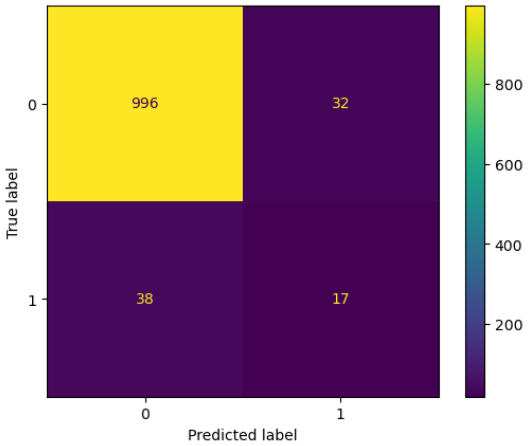

Fig. 22: XGBoost Confusion Matrix

V. Future Work

Future work could focus on significantly enhancing the LLM-RAG (Large Language Model-Retrieval-Augmented Generation) system to support even more natural, free-form

conversations about diabetes management. This would involve advancing beyond the current capabilities of simple question-and-answer (Q&A) interactions and enabling the system to handle more complex scenarios, as well as follow-up questions, much like those seen in real patient-provider conversations. It is essential that, despite these advancements, the advice provided remains firmly grounded in verified, evidence-based medical knowledge to ensure accuracy and reliability. One important avenue for improvement is refining the chatbot's persona to more effectively simulate an empathetic healthcare provider environment, which could foster a stronger sense of trust and deeper engagement from users. By making the interaction feel more human and emotionally intelligent, users would feel more comfortable and supported when discussing their health. In addition to these conversational advancements, it will be crucial to continuously expand and update the RAG knowledge base with the latest research findings and clinical guidelines. This will ensure that the system can offer comprehensive, up-to-date support for a wide variety of diabetes types, including rare or edge cases that may not be as widely represented in traditional databases. Incorporating voice messaging capabilities could also make interactions more convenient, especially for users who prefer hands-free interaction or for those who may have difficulty typing. Finally, as the system matures, it will be important to develop mechanisms that allow the tailoring of risk thresholds, explanations, and recommendations based on user-specific factors such as age, comorbidities, and individual health profiles. This personalized approach would ensure that the guidance provided is not only safer but also more precise, adapting to the unique needs of each user.

VI. Conclusion

In conclusion, this paper presented **DiabetIQ**, an intelligent and comprehensive diabetes management application developed to meet the growing demand for personalized, proactive, and reliable support for individuals living with diabetes. As diabetes continues to be a prevalent and challenging health condition, effective self-management is critical. DiabetIQ addresses this need by offering a cutting-edge solution that integrates advanced technologies, particularly machine learning and the Retrieval-Augmented Generation (RAG) framework, to enhance patient care and provide actionable insights. By combining an ML module for early risk prediction—such as identifying instances of hypo- or hyperglycemia—with an LLM-augmented RAG chatbot, DiabetIQ offers a novel and holistic approach to diabetes self-management. The system architecture, user interface mockups, and technology stack outlined in this paper present a clear and feasible pathway for the development and deployment of DiabetIQ. The design reflects careful consideration of both technical and user experience factors, ensuring that the application will be both highly functional and user-friendly. By focusing on a seamless interface, DiabetIQ promises to enhance the user experience, making diabetes management less intimidating and more accessible for individuals of varying technical skill levels. In the broader context, DiabetIQ represents a significant step toward more dynamic, data-driven, and user-centric approaches to diabetes care. It showcases how the integration of machine learning and conversational AI can enhance healthcare delivery, making it more personalized and efficient. The ability to detect risks early, provide real-time support, and continuously adapt to a user's specific needs creates an environment in which individuals are empowered to take charge of their health. With the potential for integration into clinical practice, further collaborations with healthcare professionals, and the expansion of its capabilities to support more health conditions, DiabetIQ could serve as a model for future healthcare applications aimed at improving patient outcomes through technology.

References

[1] F. Debong, H. Mayer, and J. Kober, "Real-World Assessments of mySugr Mobile Health App", *Diabetes technology & therapeutics*, vol. 21, no. S2, pp. S235–S240, 2019. doi: 10.1089/dia.2019.0019.

[2] M. A. Islam, H. N. Alvi and K. A. A. Mamun, "DiaHealth: A smart app for complete diabetes lifestyle management," 2016 International Conference on Medical Engineering, *Health Informatics and Technology (MediTec)*, Dhaka, Bangladesh, pp. 1-6, 2016. doi: 10.1109/MEDITEC.2016.7835396.

[3] H. El-Sofany, S. A. El-Seoud, O. H. Karam, Y. M. Abd El-Latif, and I. A. T. F. Taj-Eddin, "A Proposed Technique Using Machine Learning for the Prediction of Diabetes Disease through a Mobile App," *International Journal of Intelligent Systems*, Art. ID 6688934, 2024. doi: 10.1155/2024/6688934.

[4] E. Dritsas and M. Trigka, "Data-Driven Machine-Learning Methods for Diabetes Risk Prediction", *Sensors*, vol. 22, no. 14, Art. no. 5304, 2022. doi: 10.3390/s22145304.

[5] A. Nawabi, J. Tolgyesi, E. Bianchi, C. Toffanin, and P. Dondi, "D-Care: A Multi-Tone LLM-Based Chatbot Assistant for Diabetes Patients", 2025 18th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 2: HEALTHINF, pp. 766–773, 2025. doi: 10.5220/0013266600003911.

[6] D. Dao, J. Y. C. Teo, W. Wang, and H. D. Nguyen, "LLM-Powered Multimodal AI Conversations for Diabetes Prevention", Proceedings of the 1st ACM Workshop on AI-Powered Q&A Systems for Multimedia (AIQAM '24), Phuket, Thailand, pp. 1–6, 2024. doi: 10.1145/3643479.3662049.

[7] M. Abbasian, Z. Yang, E. Khatibi, P. Zhang, N. Nagesh, I. Azimi, R. Jain, and A. M. Rahmani, "Knowledge-Infused LLM-Powered Conversational Health Agent: A Case Study for Diabetes Patients", arXiv:2402.10153, 2024.

[8] J. Swacha and M. Gracel, "Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications", *Applied Sciences*, vol. 15, no. 8, Art. no. 4234, 2025. doi: 10.3390/app15084234.

[9] Diabetic Association of Bangladesh (BADAS) and NCDC Program, Directorate General of Health Services, Bangladesh, "Diabetes Care BADAS Guideline 2019," Dhaka, Bangladesh: Diabetic Association of Bangladesh, Nov. 2019.

[10] Bangladesh Endocrine Society (BES), "Practical Recommendations for Management of Diabetes and Other Endocrine Diseases in Patients with COVID-19," Bangladesh Endocrine Society, Jun. 2020. [Online]. Available: http://bes-org.net

[11] Bangladesh Endocrine Society (BES), "Insulin Guideline," 1st ed., Dhaka, Bangladesh: Bangladesh Endocrine Society, Nov. 2018. [Online]. Available: http://bes-org.net

[12] T. T. Prama, M. Zaman, F. Sarker, and K. A. Mamun, "DiaHealth: A Bangladeshi Dataset for Type 2 Diabetes Prediction," *Mendeley Data*, V1, 2024. [Online]. Available: doi:10.17632/7m7555vgrn.1

[13] S. Saha, "Learn RAG From Scratch – Python AI Tutorial from a LangChain Engineer," *YouTube*, Apr. 14, 2024. [Online]. Available: https://www.youtube.com/watch?v=sVcwVQRHIc8

[14] Prompt Engineering, "GPT-4 Tutorial: How to Chat With Multiple PDF Files (1000 pages of Tesla's 10-K Annual Reports)," *YouTube*, Mar. 29, 2024. [Online]. Available: https://www.youtube.com/watch?v=Ix9WIZpArm0

[15] Pinecone, "The BEST Way to Chunk Text for RAG," *YouTube*, Apr. 15, 2024. [Online]. Available: https://www.youtube.com/watch?v=Pk2BeaGbcTE