GlucoSense: A RAG-LLM System for Personalized Diabetes Risk Prediction and Explanation

TEAM
AKSHATA DESAI
22MIC7059
AMAN LODHA
22MIC7062

FACULTY GUIDE
AZEES MARIA JOHN FRANCIS
SCOPE, VIT AP UNIVERSITY

INTRODUCTION

PROBLEM STATEMENT

Diabetes is a chronic condition affecting millions worldwide.
Globally, approximately 589 million adults are living with diabetes. This number is projected to rise to 853 million by 2050.
In India alone, an estimated 101 million people are living with diabetes, with another 136 million considered pre-diabetic.
More than 90% of people with diabetes have type 2, which is driven by socioeconomic, environmental, and genetic factors.
Early detection and continuous, personalized care are essential for managing the disease and preventing complications.
However, current healthcare systems are limited by accessibility and a shortage of human resources.

Our Solution: GlucoSense: A Three-Layered Intelligent System

An integrated platform designed to provide accurate, transparent, and actionable diabetes care.

Layer 1: Accurate Risk Prediction

Utilizes a powerful, structured Machine Learning model for high-accuracy risk scoring.

Layer 2: Transparent Patient Comparisons

Employs FAISS for real-time semantic search to find similar patient profiles, providing case-based context.

Layer 3: Expert-Style Recommendations

Leverages a Large Language Model (LLM) to generate clear explanations and personalized care plans.

RAG (Retrieval-Augmented Generation)

What is RAG?

RAG combines **retrieval** of relevant knowledge and **generation** of responses using a language model.
It enhances accuracy and context-awareness by grounding the response in **real data**.

How We Use RAG in GlucoSense:

After the ML model predicts diabetes risk, we **retrieve similar patients** using FAISS and sentence embeddings.
These patient records serve as **context** for the LLM.

LLM Wrapper – Generating Explanations and Care Plans

What is an LLM (Large Language Model)?
An LLM is a **neural network** trained on **massive text corpora** to understand, generate, and reason in **natural language**.
LLMs can summarize information, explain complex concepts and answer questions conversationally.

Why Are We Using an LLM?
Interpretability: Explains the ML prediction in natural language, helping doctors and patients understand *why* a person is at risk.
Personalized Care: Generates custom health advice (diet, tests, follow-ups) tailored to the patient's profile.
Context-Aware Reasoning: Uses retrieved similar cases (via FAISS) to provide grounded, case-based explanations and recommendations.

PROPOSED APPROACH

Step 1: Data Preprocessing and Feature Engineering

- Load and clean patient dataset (demographics, vitals, medical history).
- Encode categorical variables (e.g., gender, smoking history).
- Apply StandardScaler to normalize numerical features.
- Handle class imbalance using **SMOTE** to oversample the minority class.
- Select important features using SelectFromModel.

Step 2: Diabetes Risk Prediction using Ensemble Learning

- Train base models:
  - **AdaBoostClassifier**
  - **XGBoostClassifier**
- Use **CatBoostClassifier** as the **meta-learner** in a StackingClassifier setup.
- Optimize hyperparameters for learning rate, estimators, subsample ratio.

Step 3: Patient Profile Transformation for Similarity Search

- Convert structured patient data into a **natural language summary**:

arduino

"Gender: Female, Age: 52, BMI: 28.6, Glucose: 155, Family Diabetes: Yes..."

- Use a **Sentence Transformer** model (all-MiniLM-L6-v2) to generate embeddings for each patient row.
- Normalize embeddings to unit vectors for efficient similarity computation.

Step 4: Semantic Similarity Search with FAISS

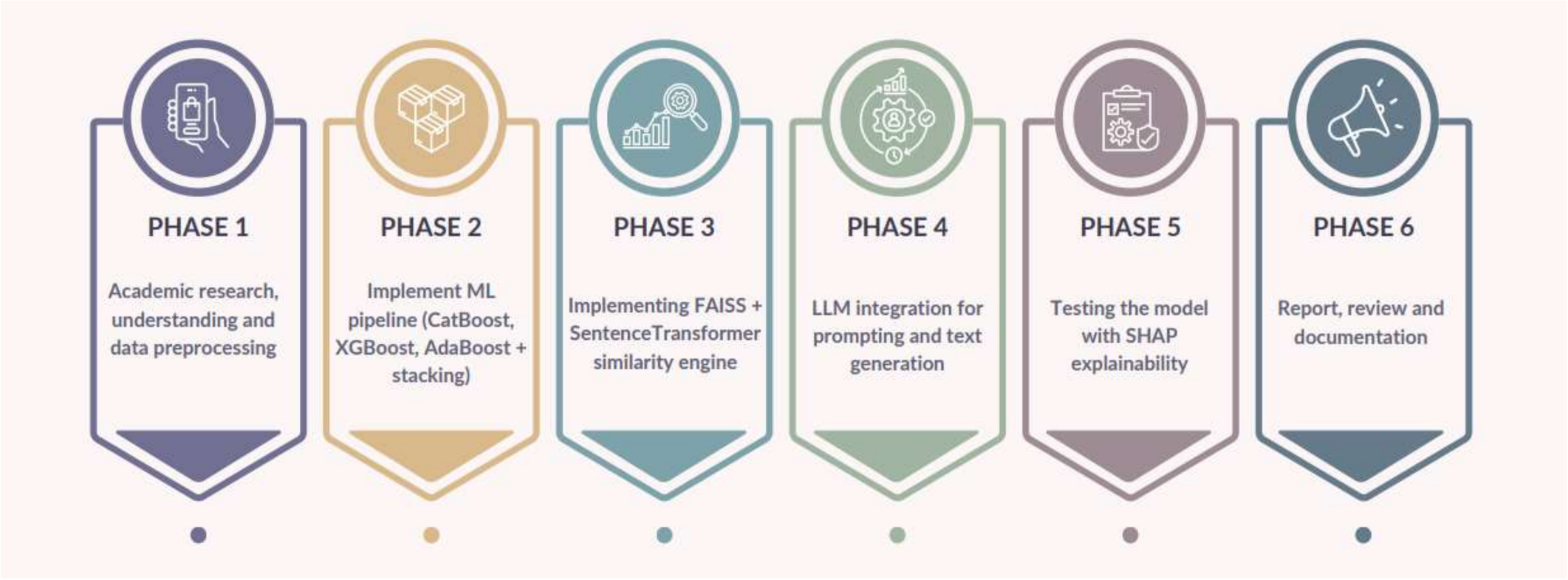
- Build a **FAISS index** on the embedded patient vectors.
- For a new patient, embed their profile summary using the same model.
- Search the index to retrieve **top-k most similar patients** based on L2 distance.
- Filter similar patients to match the **same predicted class** (diabetic/non-diabetic) for relevance.

Step 5: LLM-Based Explanation & Personalized Care Plan

- Construct a **prompt** using:
  - Patient's vitals and predicted class
  - Probability score
  - Top-5 similar patient summaries

- Use a **Large Language Model (LLM)** like GPT-3.5 to generate:
  - A **natural language explanation** justifying the prediction
  - A **customized care plan** including:
    - Lifestyle/dietary changes
    - Test follow-ups (e.g., HbA1c, glucose monitoring)
    - Risk mitigation tips

PROJECT APPROACH TIMELINE



SURVEY – EXISTING WORK

Personalized diabetes risk prediction is critical for early intervention and targeted management strategies. Traditional approaches often rely on population-level data, which may not account for individual-specific factors such as genetic predisposition, lifestyle, and environmental influences. Recent studies have emphasized the need for systems that can ingest diverse data modalities, including demographic, clinical, and time-series information, to provide individualized risk assessments (Belyaeva et al., 2023)

The development of systems like HeLM (Health Large Language Model for Multimodal Understanding) has demonstrated the potential of LLMs in leveraging high-dimensional clinical data to estimate disease risk.

- Supervised models like **Random Forest**, **SVM**, and **Gradient Boosting** have been widely used for diabetes risk prediction.
- Mamun et al. (2024) showed that **XGBoost with polynomial features** achieved **99.22% accuracy**, outperforming AdaBoost and LGBM.
- **Ensemble methods**, especially Gradient Boosting, handle complex feature interactions better, leading to high performance.
- However, traditional models lack **interpretability**, making them less suitable for clinical use.
- To improve trust, **XAI techniques** like **SHAP** and **LIME** are used to explain model decisions and feature importance.

Role of RAG LLMs in Diabetes Risk Prediction

RAG LLMs have emerged as a powerful tool for personalized diabetes risk prediction. These models combine the strengths of large language models with external knowledge retrieval systems, enabling them to access and synthesize information from diverse sources. (Belyaeva et al., 2023)

The ability of RAG LLMs to process multimodal data makes them particularly suitable for diabetes prediction. For example, a system developed by demonstrated that combining demographic and clinical features with time-series data significantly improves the accuracy of diabetes risk prediction.

Paper (Venue / Year)	Advantages	Limitations	Key Findings / Results
DiabetIQ – North South University technical report (**2025**)	• Combines ML-based risk prediction with an LLM-augmented RAG chatbot in one platform.• Provides **context-aware answers** grounded in curated medical content.• Improved user engagement through conversational interface.	• Explanation not linked to actual similar-patient records, only general guidelines.• Prediction engine and chatbot are loosely integrated.• Case grounding is based on medical literature, not clinical analogies. (ResearchGate)	• Demonstrated **better predictive accuracy** and **higher user engagement** than non-hybrid methods.
Knowledge-Infused LLM Health Agent (arXiv / Workshop, **2024**)	• Integrates **ADA dietary guidelines** and **Nutritionix analytics** for personalized nutritional advice.• Responses are both **contextual and quantitatively accurate.**	• Evaluation limited to **100 dietary-related questions**.• Not part of an end-to-end risk prediction or care pipeline. Focused solely on nutrition rather than full risk reasoning. (arxiv.org)	• Agent **outperformed GPT-4** on managing essential nutrient queries for diabetic patients. (arxiv.org)
RISE: Retrieval-Augmented Information System Enhancement – J. Med. Internet Res. (**2024**)	• Plug-in for LLMs to retrieve domain-specific content and augment response generation.• Works with multiple LLMs (GPT-4, Claude, Bard) to enhance accuracy.	• Evaluated only on 43 diabetes-related Q&A pairs.• Not tied to structured patient data or risk models.• Limited to educational NLP tasks, not predictive pipelines.	• ~12 pp increase in accuracy; comprehensiveness scores rose ~0.44; understandability improved ~0.19. (jmir.org)
Mohsen et al., "Scoping review of AI methods for T2DM risk prediction" – npj Digital Medicine (**2023**)	• Systematic review of 40 studies (1991–2022) mapping AI usage in diabetes prediction.• Identifies superiority of **multimodal models** (EHR + omics or imaging).• Notes gaps in validation, interpretability, and reporting.	• Only **5 of 40** total studies included **external validation**.• Less than half used **interpretability tools** (e.g., SHAP).• Limited multimodal, real-time, or personalized pipelines.	• Highlighted that **multimodal predictive models** outperform unimodal ones.• Only ~50 % of studies applied interpretability; calibration often missing. (nature.com)

Most systems use general medical literature for explanations, not actual, similar patient cases. There is a lack of true case-based similarity search to ground and contextualize predictions.
Prediction and explanation modules are often separate, which reduces overall interpretability and personalization.

Our goal is to build an intelligent, explainable, and patient-centric system. We contribute by:
High-Accuracy Ensemble Model: A stacked model for superior prediction.
Semantic Similarity Retrieval: Using FAISS for real-time retrieval of similar patient profiles.
True RAG Integration: Grounding LLM explanations in the data of these similar cases.
Interpretability Layer: Using SHAP to provide clinical-grade feature importance.

OUR OBJECTIVE

To develop an intelligent, explainable, and patient-centric diabetes prediction system by integrating structured machine learning, semantic patient similarity search, and large language models (LLMs) for actionable care planning.

OUR CONTRIBUTION

- High-Accuracy Diabetes Prediction Engine using a stacked **ensemble model** (AdaBoost + XGBoost + CatBoost).

- **Semantic Similarity Retrieval** using FAISS + Embeddings for real-time retrieval of top-k similar patients

- **LLM Wrapper for Explanation** & Personalized Care which provides risk explanation and custom care plan based on patient profile and similar cases.

- Use **Retrieval-Augmented Generation (RAG)** to ground output in actual data.

- Interpretability & Trust Layer using **SHAP** to enhance clinical decision-making

CONCLUSION

Diabetes is a significant and growing public health challenge that demands innovative, personalized solutions.

•**GlucoSense** addresses the critical limitations of existing systems by creating a tightly integrated pipeline that delivers **accurate predictions**, **transparent reasoning**, and **actionable care plans**.

•Our core innovation lies in the use of a **true case-based RAG system**, which grounds AI-generated explanations in the data of actual, similar patients—a significant step forward in building trust and providing context.

•By combining the predictive power of ensemble machine learning with the explanatory capabilities of LLMs, GlucoSense has the potential to empower both clinicians and patients, facilitating earlier intervention and more effective, personalized diabetes management.

THANK YOU