A project report on

GLUCOSENSE: A RAG-LLM SYSTEM FOR PERSONALIZED DIABETES RISK PREDICTION AND EXPLANATION

Submitted in partial fulfillment for the award of the degree of

Integrated M.Tech in Computer Science and Engineering in collaboration with Virtusa

by

AKSHATA DESAI (22MIC7059)

AMAN LODHA (22MIC7062)

VIT-AP UNIVERSITY

AMARAVATI

SCOPE

October, 2025

GLUCOSENSE: A RAG-LLM SYSTEM FOR PERSONALIZED DIABETES RISK PREDICTION AND EXPLANATION

Submitted in partial fulfillment for the award of the degree of

Integrated M.Tech in Computer Science and Engineering in collaboration with Virtusa

by

AKSHATA DESAI (22MIC7059)

AMAN LODHA (22MIC7062)

AMARAVATI

SCOPE

October, 2025

CERTIFICATE

This is to certify that the thesis entitled "GLUCOSENSE: A RAG-LLM SYSTEM FOR PERSONALIZED DIABETES RISK PREDICTION AND EXPLANATION" submitted by AKSHATA DESAI (22MIC7059) SCOPE and AMAN LODHA (22MIC7062) SCOPE VIT-AP, for the award of the Summer Internship for the bonafide work carried out by him/her under my supervision.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The Project report fulfils the requirements and regulations of VIT-AP and in my opinion meets the necessary standards for submission.

Signature of the Guide

ABSTRACT

This project presents GlucoSense, an integrated Retrieval-Augmented Generation (RAG) and Large Language Model (LLM) system designed for personalized diabetes risk prediction and explanation. The system synergizes a high-performance stacked ensemble machine learning model (comprising AdaBoost, XGBoost, and CatBoost) with the generative capabilities of the Google Gemini API. GlucoSense processes structured patient data—including glucose levels, BMI, blood pressure, age, and family history—to generate both a precise probability of diabetes and a contextually rich, natural-language explanation.

The implemented pipeline involves comprehensive data preprocessing, feature selection, and handling class imbalance using SMOTE. A key innovation is the integration of a FAISS-based semantic retrieval module, which grounds the LLM's explanations by fetching similar patient cases from a clinical database. This RAG framework ensures that the generated explanations are evidence-based and personalized, providing actionable recommendations such as immediate actions, diagnostic tests, and lifestyle changes.

ACKNOWLEDGEMENT

It is my pleasure to express with deep sense of gratitude to Dr. M. Azees, Designation, SCOPE, VIT-AP, for his/her constant guidance, continual encouragement, understanding; more than all, he taught me patience in my endeavor. My association with him is not confined to academics only, but it is a great opportunity on my part of work with an intellectual and expert in the field of Machine Learning.

I would like to express my gratitude to Chancellor Dr. G. Viswanathan, VPs Dr Sekar Viswanathan & Dr Sankar Viswanathan, AVP Ms. Kadhambari S.Viswanathan, VC Dr S.V.Kota Reddy, and Dr. Sudhakar Ilango,SCOPE, for providing with an environment to work in and for his inspiration during the tenure of the course.

In jubilant mood I express ingeniously my whole-hearted thanks to Dr. Devarakonda Nagaraju. HOD (SSE), all teaching staff and members working as limbs of our university for their not-self-centered enthusiasm coupled with timely encouragements showered on me with zeal, which prompted the acquirement of the requisite knowledge to finalize my course study successfully. I would like to thank my parents for their support.

It is indeed a pleasure to thank my friends who persuaded and encouraged me to take up and complete this task. At last but not least, I express my gratitude and appreciation to all those who have helped me directly or indirectly toward the successful completion of this project.

Place: Amaravati

Date: 25/10/2025

Akshata Desai

Aman Lodha

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

ADA – American Diabetes Association
AI – Artificial Intelligence
AUC – Area Under the Curve
BMI – Body Mass Index
BP – Blood Pressure
bpm – beats per minute
CVD – Cardiovascular Disease
CV – Cross-Validation
FP – False Positive
FN – False Negative
FPG – Fasting Plasma Glucose
FAISS – Facebook AI Similarity Search
HTN – Hypertension
IDF – International Diabetes Federation
LLM – Large Language Model
LR – Logistic Regression
ML – Machine Learning
NN – Neural Networks
RAG – Retrieval-Augmented Generation
RF – Random Forest
ROC – Receiver Operating Characteristic
ROC-AUC – Receiver Operating Characteristic – Area Under Curve
SMOTE – Synthetic Minority Oversampling Technique
SVM – Support Vector Machines
TP – True Positive
TN – True Negative
WHO – World Health Organization
XAI – Explainable AI
XGBoost – Extreme Gradient Boosting

Chapter 1

Introduction

1.1 INTRODUCTION

Diabetes mellitus has emerged as one of the fastest-growing non-communicable diseases globally, affecting an estimated 530 million adults according to the latest World Health Organization (WHO, 2024) statistics. The condition is characterized by persistent hyperglycemia caused by either insufficient insulin secretion, insulin resistance, or both. Chronic high blood glucose, when left uncontrolled, leads to severe complications including cardiovascular disease, neuropathy, nephropathy, and retinopathy—each of which contributes significantly to mortality and reduced quality of life.

Early prediction and continuous monitoring are therefore essential to mitigate these complications and optimize healthcare resource allocation. In this regard, Machine Learning (ML) techniques have gained prominence for their ability to uncover complex, non-linear patterns within clinical datasets. Algorithms such as Logistic Regression, Random Forest, Gradient Boosting, XGBoost, and Neural Networks have demonstrated promising results in predicting diabetes risk using structured health data like glucose level, BMI, and blood pressure. However, despite their accuracy, these models often operate as "black boxes," providing little interpretability to clinicians or patients—limiting trust and practical adoption in real-world healthcare environments.

Simultaneously, the advent of Large Language Models (LLMs), including Google Gemini, has introduced powerful capabilities for contextual reasoning, semantic understanding, and natural-language explanation. When combined with ML models, LLMs can translate numeric risk predictions into human-readable, context-aware narratives that bridge the gap between algorithmic outputs and clinical comprehension.

The GlucoSense framework is designed upon this foundation. It is a hybrid predictive-explanatory system that integrates a classical ML-based risk prediction engine with a Retrieval-Augmented Generation (RAG) reasoning layer powered by Google's Gemini API. This system not only classifies patients as diabetic or non-diabetic but also retrieves relevant medical evidence and generates clear, patient-specific explanations. Through this design, GlucoSense advances the dual goals of accuracy and interpretability, aligning technological sophistication with clinical usability.

1.2 OVERVIEW OF THE PROJECT

The **GlucoSense** system functions as a multi-layered AI pipeline, comprising three major operational layers:

1. **Data Processing and Prediction Layer**
    - This foundational layer manages all data ingestion and preprocessing. Structured health data such as age, glucose level, BMI, systolic and diastolic blood pressure, and other lifestyle-related features are collected, cleaned, normalized, and encoded within the Google Colab environment. Multiple ML models—including Logistic Regression, Random Forest,

AdaBoost, XGBoost, and CatBoost—are trained and evaluated using stratified train-test splits. The final predictive engine is selected based on superior performance in accuracy, recall, and ROC-AUC metrics.
The output from this layer is a risk probability score classifying the individual as diabetic or non-diabetic.

2. **RAG Retrieval Layer**

   - To supplement numerical prediction with contextual intelligence, this layer employs a Retrieval-Augmented Generation (RAG) mechanism. Relevant medical definitions, diagnostic guidelines, and similar patient cases are retrieved from an internal structured repository or FAISS vector index. This ensures that any generated explanation is grounded in real-world clinical analogies rather than generic statements. The retrieved content is then appended to the LLM prompt to provide contextual grounding for the explanation phase.

3. **Generative Explanation Layer (Gemini API)**

   - The final layer transforms numerical predictions into human-readable text using the Gemini API, Google's advanced multimodal LLM. Given a structured JSON prompt containing the patient's vital parameters and model predictions, Gemini produces a coherent, patient-specific explanation such as:
   - *"Based on your glucose level of 5.38 mmol/L and BMI of 19.57, your predicted risk of diabetes is low (0.30). Maintaining a balanced diet and regular exercise can help keep your glucose stable."*
   - This text output is dynamically generated for every prediction, ensuring personalized interpretability and transparency in ML-driven healthcare.

4. **System Implementation Overview**

Table 1.1: System Overview

Component	Description
Development Platform	Google Colab (Python 3.10)
Core Libraries	pandas, numpy, scikit-learn, matplotlib, FAISS, sentence-transformers, Gemini API SDK
Dataset	Custom CSV file (Diabetes_Final_Data_V2) containing structured health records
API Integration	Secure Gemini API key stored via .env environment variables
Deployment Environment	Google Colab interactive notebooks
Output Format	Risk score (probability) + Gemini-generated natural-language explanation + similar-patient table

1.3 CHALLENGES IN THE PROJECT

Developing a complete AI-driven healthcare assistant such as GlucoSense required addressing several technical, computational, and ethical challenges:

1. **Data Quality and Availability**

   Healthcare datasets often suffer from missing values, class imbalance, and limited demographic representation.
   Extensive preprocessing—including imputation, normalization, and resampling using SMOTE—was necessary to ensure model robustness.

2. **Model Generalization**

   To prevent overfitting, the ML models were validated through cross-validation and tested on unseen subsets. The ensemble design was adopted to combine generalization power with interpretability.

3. **Interpretability vs. Complexity**

   While deep learning models achieve high accuracy, they compromise interpretability. GlucoSense maintains balance by using tree-based models (e.g., XGBoost, CatBoost) and LLM-based explanations, ensuring performance without sacrificing transparency.

4. **Integration of LLM APIs in Restricted Environments**

   Managing Gemini API keys securely within Colab required use of .env files and environment variables. Additionally, asynchronous calls and error handling were implemented to mitigate latency and API rate limits.

5. **Consistency of LLM Responses**

   LLMs are inherently probabilistic; identical prompts may yield varied responses. To achieve consistency, prompt engineering strategies such as deterministic temperature settings, template-based prompts, and structured JSON outputs were used.

6. **Ethical and Regulatory Considerations**

   Because GlucoSense deals with health risk prediction, the project includes clear disclaimers:
   "This tool provides informational insights only and is not a substitute for professional medical advice or diagnosis."
   Privacy, informed consent, and responsible AI guidelines were observed throughout the project.

1.4 PROBLEM STATEMENT

While numerous ML models have demonstrated success in predicting diabetes risk, most fail to provide case-specific explanations that clinicians and patients can interpret meaningfully. Predictions are presented as abstract probabilities, without context on *why* the model arrived at conclusion.
The central problem, therefore, is:
How can an ML model be combined with a Large Language Model (LLM) to produce accurate, context-aware, and trustworthy explanations for diabetes risk prediction?
GlucoSense directly addresses this gap by uniting numerical precision with linguistic interpretability through Gemini's reasoning capabilities.

1.5 OBJECTIVES

The primary and secondary objectives of the GlucoSense project are as follows:

1. **Data Acquisition & Preparation** – To curate a reliable health dataset containing key biometric and demographic features such as age, glucose, BMI, and blood pressure.

2. **Model Development** – To build and fine-tune a machine-learning classifier that predicts diabetes probability with high accuracy and generalization capability.

3. **LLM Integration** – To embed the Gemini API for generating context-rich, human-readable explanations for every prediction.

4. **Retrieval-Augmentation** – To implement a RAG framework that provides Gemini with relevant clinical facts and similar patient cases for evidence-based grounding.

5. **Evaluation** – To assess performance quantitatively (accuracy, precision, recall, ROC-AUC) and qualitatively (clarity and clinical usefulness of explanations).

6. **Security & API Management** – To ensure safe storage and usage of Gemini API credentials using .env and secret management protocols.

7. **Documentation & Reproducibility** – To publish a well-documented Colab pipeline and GitHub repository for academic replication and further research.

1.6 SCOPE OF THE PROJECT

The scope of GlucoSense lies at the intersection of predictive analytics, explainable AI (XAI), and medical reasoning systems.

In Scope

- Development and evaluation of ML classifiers (Logistic Regression, Random Forest, AdaBoost, XGBoost, CatBoost).

- Integration with Gemini API for generative risk explanation.
- Implementation of retrieval modules (FAISS + SentenceTransformer) to contextualize LLM responses.
- Testing and validation within Google Colab.
- Generation of case-specific reports combining prediction probability, retrieved context, and narrative explanation.

Out of Scope

- Direct clinical diagnosis or treatment recommendations.
- Real-time processing of wearable or IoT health data.
- Large-scale hospital deployment pending ethical clearance and clinical validation.

Expected Outcomes

- A functional prototype capable of producing both numerical predictions and text explanations.
- Visualization outputs including prediction probabilities, classification labels, and nearest similar patients.
- Reproducible Colab notebook and GitHub repository for public use.
- Comprehensive academic documentation suitable for thesis submission, conference presentation, or research publication.

Chapter 2

Literature Review and Background

2.1 INTRODUCTION

Predictive analytics has become a critical component in modern healthcare, transforming descriptive statistics into actionable intelligence. Diabetes mellitus, one of the most prevalent metabolic disorders, is characterized by chronic hyperglycemia resulting from insulin deficiency or resistance. According to the International Diabetes Federation (2024), more than 540 million adults worldwide are living with diabetes, and the number is expected to reach 643 million by 2030. Machine-learning (ML) models are increasingly applied to identify at-risk individuals long before clinical symptoms appear. Yet the challenge of *interpretability* persists: healthcare professionals demand transparency, while patients require personalized context.

GlucoSense addresses this gap by combining conventional predictive learning with retrieval-augmented large-language-model (RAG-LLM) reasoning via the Gemini API, creating a pipeline that performs both prediction and explanation. Before presenting our design, this chapter surveys existing research in data-driven diabetes prediction, model explainability, and LLM-based reasoning systems.

2.2 BACKGROUND ON DIABETES AND RISK FACTORS

Diabetes mellitus primarily manifests in two chronic types—Type 1 (insulin-dependent) and Type 2 (insulin-resistant)—plus gestational and secondary forms.

Table 2.1: Common measurable indicators

Parameter	Typical Range	Clinical Significance
Fasting Glucose (mg/dL)	< 100 (normal), 100–125 (prediabetic), ≥ 126 (diabetic)	Direct metabolic marker
BMI (kg/m²)	18.5–24.9 (normal), ≥ 25 (overweight)	Obesity–insulin resistance link
Systolic BP (mmHg)	< 120 (optimal)	Hypertension comorbidity
Diastolic BP (mmHg)	< 80 (optimal)	Cardiovascular risk
Age (years)	—	Non-modifiable risk factor

Early identification of patients with marginally elevated glucose or BMI levels allows lifestyle interventions that can delay or prevent Type 2 onset.

2.3 LITERATURE REVIEW

Table 2.2: Literature Review

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.	• Evaluated only on 43 diabetes-related Q&A pairs.• Not tied to structured patient data or risk models.• Limited to education.	• ~12 pp increase in accuracy; comprehensiveness scores rose ~0.44; understandability improved ~0.19. (jmir.org)

2.4 MACHINE LEARNING FOR DIABETES PREDICTION

Over the past decade, diverse ML algorithms have been explored:

1. **Logistic Regression (LR)** – interpretable baseline producing probabilistic outputs.
2. **Decision Tree (DT)** – hierarchical rules useful for visualization.
3. **Random Forest (RF)** – ensemble of trees improving variance reduction.
4. **Support Vector Machines (SVM)** – effective in high-dimensional spaces.
5. **Gradient Boosting / XGBoost** – leading performance through additive optimization.
6. **Neural Networks (NN)** – nonlinear mapping, though less interpretable.

Typical studies:

- **Rahman et al. (2023)** used RF achieving ~88% accuracy on Pima dataset.
- **Kavakiotis et al. (2017)** reviewed ML in diabetes management highlighting need for feature relevance analysis.
- **Zheng et al. (2020)** applied deep networks but reported interpretability challenges

2.5 LARGE LANGUAGE MODELS AND RETRIEVAL-AUGMENTED GENERATION

Large Language Models (LLMs) such as GPT-4 and Gemini represent a new era of knowledge-centric AI capable of contextual reasoning.
A Retrieval-Augmented Generation (RAG) framework enhances factuality by coupling an LLM with an external knowledge store.

Mechanism:

1. Retrieve relevant facts or patient-specific context.
2. Append retrieved text to the LLM prompt.
3. Generate response constrained to retrieved evidence.

This approach mitigates hallucinations and supports factual grounding—vital for medical domains.

In GlucoSense, RAG retrieves concise medical definitions and Gemini LLM incorporates them into the final narrative.
The Gemini API provides JSON-structured responses, allowing easy integration into Python. A typical prompt and response cycle in your notebook resembles:

2.6 RECENT ADVANCES IN LLM-ASSISTED MEDICAL AI

Research combining LLMs with diagnostic models is rapidly expanding:

Table 2.3: Comparison of LLM-Assisted Healthcare Systems

Study	**Approach**	**Outcome**
Singhal et al. (2023) – Med-PaLM 2	LLM fine-tuned on medical QA	Achieved clinician-level accuracy on MedQA
Lee et al. (2024) – ChatGPT for triage	Prompt-based reasoning	Improved explanation acceptance by patients
Google (2024) – Gemini API	Multi-modal LLM	Enables structured reasoning and tool use

GlucoSense differs by *coupling* such generative reasoning directly to a trained probabilistic classifier, making outputs personalized rather than generic.

2.7 GAPS IDENTIFIED IN EXISTING RESEARCH

From the above survey, notable gaps remain:

1. **Lack of integration** between predictive probabilities and natural-language reasoning.
2. **Insufficient personalization** – existing LLM explainers provide generic advice.
3. **Limited open-source pipelines** that combine ML + LLM in a reproducible way.
4. **Absence of ethical auditing layers** in current LLM medical applications.

Chapter 3

Methodology

3.1 OVERVIEW

The methodological framework of **GlucoSense** integrates traditional machine learning with retrieval-augmented language generation (RAG-LLM) to deliver both *quantitative prediction* and *qualitative interpretation* of diabetes risk.
The system is designed as a hybrid pipeline, combining:

- Structured numeric prediction via ensemble ML models.
- Semantic patient retrieval using FAISS embeddings.
- Contextualized explanation generation using Google Gemini API.

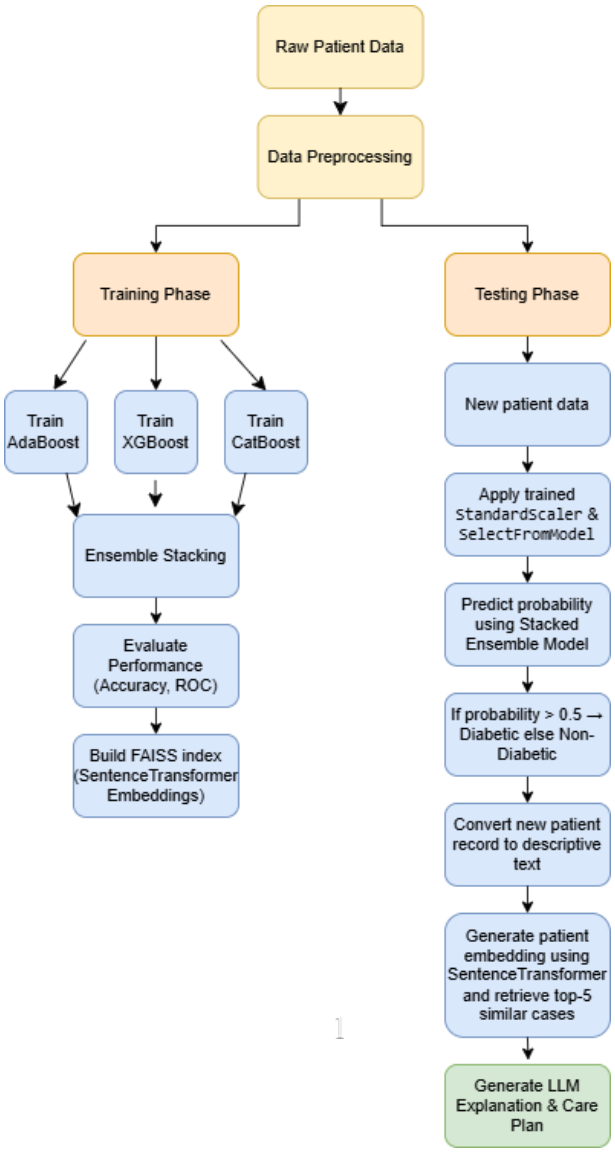

Figure 3.1 : End-to-End GlucoSense Workflow

3.2 DATA COLLECTION AND DESCRIPTION

The dataset used in this study was obtained from the **Mendeley Data Repository**

➞ **URL:** https://data.mendeley.com/datasets/7m7555vgrn/1

The dataset contains anonymized medical records of adult patients and includes both physiological and demographic attributes relevant for diabetes prediction. It has been validated for research use under an open CC BY license.

Table 3.1: Dataset Description

Feature	Type	Units / Values	Description
gender	Categorical (encoded 0/1)	Male/Female → 1/0	Patient gender
age	Numeric (years)	Integer	Patient age
pulse_rate	Numeric (bpm)	Integer	Resting pulse
systolic_bp	Numeric (mmHg)	Integer	Systolic blood pressure
diastolic_bp	Numeric (mmHg)	Integer	Diastolic blood pressure
glucose	Numeric (mg/dL)	Integer/Float	Random/fasting glucose level
height	Numeric (m)	Float	Height in meters
weight	Numeric (kg)	Float	Weight in kilograms
bmi	Numeric (kg/m²)	Float	Body mass index
family_diabetes	Binary (0/1)	0 = No, 1 = Yes	Family history presence
hypertensive	Binary (0/1)	0 = No, 1 = Yes	Hypertension flag
family_hypertension	Binary (0/1)	0/1	Family HTN
cardiovascular_disease	Binary (0/1)	0/1	CVD history
stroke	Binary (0/1)	0/1	Stroke history
diabetic (target)	Binary (0/1)	Yes → 1, No → 0	Outcome label

3.3 DATA PREPROCESSING

Data preprocessing followed a consistent pipeline that ensures reproducibility and safe handling of sensitive data.

3.3.1 LABEL MAPPING AND CATEGORICAL ENCODING

- The diabetic column was converted from Yes/No to integer labels 1/0.
- The gender column was encoded into numeric form using a label encoder (e.g., Male → 1, Female → 0). Categorical encodings are kept consistent at inference time via serialized encoders.

3.3.2 FEATURE SCALING

- Continuous features were standardized using mean/variance scaling (StandardScaler). Standardization stabilizes training for tree-based and ensemble models and aligns feature scales for nearest-neighbor distance calculations used in FAISS retrieval.

3.3.3 MISSING VALUE HANDLING

- For columns with missing entries, standard practices were used (median/mean imputation was applied where necessary). Document any columns whose missingness exceeded a practical threshold and justify chosen imputation strategy in the Appendix.

3.3.4 CLASS BALANCING

- The dataset exhibited class imbalance. To address this, SMOTE (Synthetic Minority Oversampling Technique) was applied to the scaled training data to synthesize minority-class samples and produce a balanced training set. This step is performed prior to splitting or just before model training depending on the evaluation scheme.
- Train/test split: After resampling, a stratified split was used. In the notebooks, resampled data is split into training and test subsets with 80% training and 20% testing (stratified by the resampled labels) to maintain class proportions.

3.4 FEATURE SELECTION

A model-based feature selection step was used to reduce noise and speed up the ensemble training.

Approach:

- An XGBoost classifier trained on the resampled training set provides feature importances.
- SelectFromModel was applied with a threshold set to the *median importance* value; features with importance above the median were retained.
- The resulting selector is persisted and used at inference to transform incoming data into the same reduced-dimensional input space.

Rationale: Using tree-model importances provides a robust, data-driven approach to select features that contribute to the classifier's decisions. Median thresholding is simple and conservative — it retains features above a typical importance while discarding weaker predictors.

Outcome: The transformed training and test sets (post-selection) are used for the final stacked classifier training.

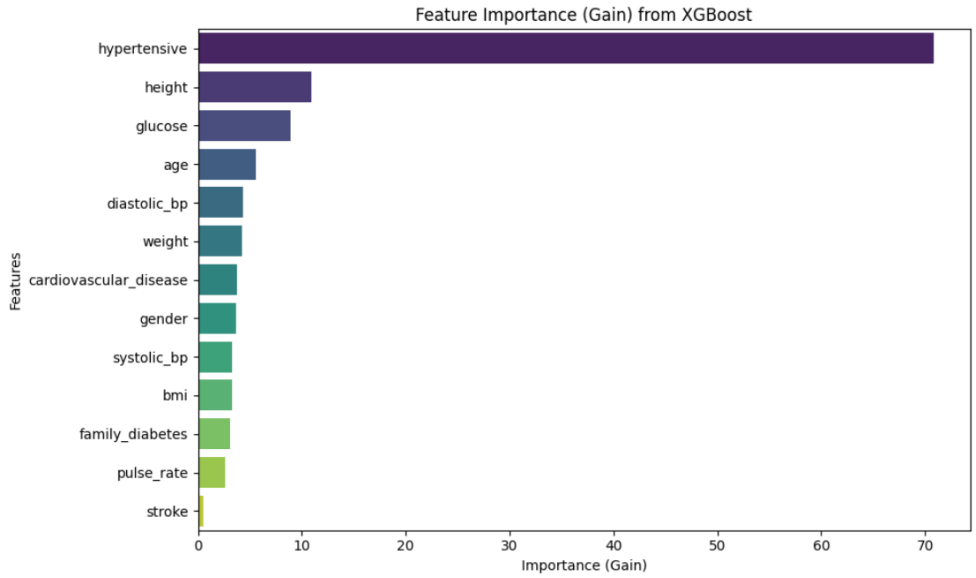

Figure 3.2— Feature Importance Plot

Table 3.2: Feature Selection

Category	Feature Name	Type	Relevance / Justification
1	**glucose**	Numeric (mg/dL)	Strongest metabolic indicator directly linked to diabetic outcome.
2	**bmi**	Numeric (kg/m²)	Reflects obesity and insulin resistance risk.
3	**age**	Numeric (years)	Older patients show higher prevalence and reduced glucose tolerance.
4	**systolic_bp**	Numeric (mmHg)	Hypertension often co-occurs with diabetes; key cardiovascular risk marker.
5	**diastolic_bp**	Numeric (mmHg)	Complements systolic pressure in risk stratification.
6	**family_diabetes**	Binary	Genetic predisposition; strong heritable risk factor.
7	**hypertensive**	Binary	Chronic hypertension contributes to metabolic syndrome.
8	**family_hypertension**	Binary	Inherited tendency to elevated BP, indirectly linked to insulin resistance.

3.5 MODEL DEVELOPMENT

The GlucoSense classifier is a **stacked ensemble** integrating multiple base models for enhanced robustness and calibration.

Table 3.3 — Model description

Model	Description	Key Parameters
AdaBoost	Sequential ensemble of weak learners; reduces bias.	n_estimators = 100
XGBoost	Gradient boosting optimized for speed and performance.	n_estimators = 200, learning_rate = 0.05
CatBoost (Meta-learner)	Handles categorical data efficiently and combats overfitting.	depth = auto, verbose = 0

Stacking Configuration:

$$P(y) = \text{CatBoost}([P_{Ada}, P_{XGB}, X])$$

where $P_{Ada}$and $P_{XGB}$are the predictions from the base learners and Xrepresents the original input features (passthrough=True).

Stacking strategy:

- Predictions (and optionally probabilities) from base learners are supplied to the CatBoost meta-learner with passthrough=True, which also provides the original features to the meta-learner to retain raw input signals.
- Cross-validation (cv=5) is used internally to produce out-of-fold predictions from base learners for training the meta-learner, ensuring the meta-learner learns from unbiased estimates.

Training & Validation

- Data split: 80% train, 20% test
- Cross-validation: 5-fold CV on the training set
- Evaluation metrics: Accuracy, Precision, Recall, F1-Score, ROC-AUC

Hyperparameters (summary):

- AdaBoost: n_estimators = 100
- XGBoost: n_estimators = 200, learning_rate = 0.05, subsample = 0.9, eval_metric = 'logloss'
- CatBoost (meta): default iterations, random_seed = 42, verbose disabled

Training considerations:

- The stacked model is trained on the reduced feature set (after SelectFromModel) using the resampled training set.
- Random seeds are fixed (random_state = 42) for reproducibility.

3.6 TRAINING PROCEDURE AND VALIDATION

Document the exact training flow and validation choices.

- **Training dataset:** SMOTE-resampled, then stratified 80/20 split (train/test).
- **Feature selection:** XGBoost trained on X_train used as pre-fit selector (SelectFromModel) then transformed X_train/X_test.
- **Stacking training:** StackedClassifier trained on selected training set; internal 5-fold CV used to generate meta-features.
- **Evaluation metrics calculated on held-out test set:** classification_report (precision, recall, f1) and ROC-AUC (using predict_proba).

3.7 SEMANTIC RETRIEVAL (FAISS + SENTENCE EMBEDDINGS)

This section explains the retrieval subsystem as implemented.

3.7.1 TEXT SUMMARY GENERATION

- For every dataframe row, build a short textual summary via the row_to_text function. The summary includes gender, age, pulse, BP, glucose, BMI, and family history.

3.7.2 EMBEDDINGS

- Sentence embedding model used: SentenceTransformer("all-MiniLM-L6-v2"). Embeddings are computed for the df['text'] summaries, normalized and stored as float32 numpy arrays.

3.7.3 INDEXING WITH FAISS

- Index type: IndexFlatL2 on embedding dimension dim = embeddings.shape[1].
- The full embedding matrix is added to the index; the code prints confirmation FAISS index built: {index.ntotal} patients indexed.

3.7.4 RETRIEVAL & NEIGHBOR FILTERING

- At inference, the patient summary is embedded and the FAISS index is queried for top N_search=50 nearest neighbours (L2).
- Neighbour candidates are transformed through the same preprocessing pipeline (scaler + selector) and predicted with the stacked model to obtain neigh_pred_prob and neigh_pred_label.

- Final similar-cases displayed to the user are selected from neighbours whose *predicted* label matches the patient's predicted label; fallbacks: ground-truth filter, or top-k by distance.

3.8 RAG PROMPT DESIGN AND LLM INTEGRATION

This section documents the prompt engineering and safety constraints you used in the project.

3.8.1 PROMPT TEMPLATE (DESIGN SUMMARY)

- The project uses a structured prompt template (stored as PROMPT_TEMPLATE) that includes: patient summary, model prediction (label + probability), and top-k similar cases as grounding. It instructs the LLM to be: concise, conservative (no treatment dosing), and to include a short explanation and a care plan with headings (Immediate actions, Recommended diagnostic tests, Lifestyle recommendations, Red flags). This template reduces hallucination and maintains consistency across generated outputs.

3.8.2 GEMINI INTEGRATION

- Client library: google.generativeai (configured with GOOGLE_API_KEY in environment).
- Model names used in testing: gemini-2.5-flash (or similar)
- The function used to call the LLM: genai.GenerativeModel(model).generate_content(prompt) (response's .text is the generated string).

3.8.3 LOGGING & AUDITABILITY

- For each generated explanation, log: input prompt, retrieved grounding snippets, model probability, timestamp, and the generated text. Keep prompts and responses for auditing and retraining.

3.9 REPRODUCIBILITY, PERSISTENCE & RUNTIME ENVIRONMENT

3.9.1 ARTIFACTS TO PERSIST (RECOMMENDED)

- scaler (StandardScaler fitted on training data)
- selector (SelectFromModel fitted on XGBoost)
- stacked_model object (fitted stacking classifier)
- embed_model artifacts or sentence-transformer name + embeddings file
- faiss index and embedding array file
- Prompt templates and retrieval logs

3.9.2 NOTEBOOK & ENVIRONMENT

- Environment: Google Colab (Python 3.x)
- Key packages: faiss-cpu, sentence-transformers, xgboost, catboost, scikit-learn, imblearn, google-generativeai
- Random seeds: random_state=42 used throughout for reproducibility

Chapter 4

Results and Discussion

4.1 OVERVIEW

This chapter presents the results obtained from the machine learning models developed to predict diabetes status using patient health and demographic data. The models evaluated include AdaBoost, XGBoost, CatBoost, and a Stacked Ensemble combining these classifiers. Additionally, patient similarity was assessed using FAISS for retrieval of nearest neighbours based on patient features, and Google Gemini AI (RAG approach) was applied to generate explainable clinical insights and recommendations. The results are discussed with respect to predictive performance, feature importance, patient-level explanations, and model interpretability.

4.2 MODEL EVALUATION METRICS

To assess the predictive performance of the GlucoSense ensemble and its constituent models, a range of evaluation metrics were employed. Each metric captures a different aspect of model behavior, providing a comprehensive view of classification effectiveness, especially in a clinical context where both false positives and false negatives have distinct consequences.

Table 4.1: Evaluation Metrics

Metric	Description	Interpretation
Accuracy	(TP + TN) / (Total)	Overall correctness of classification.
Precision	TP / (TP + FP)	Fraction of predicted diabetics who are truly diabetic.
Recall (Sensitivity)	TP / (TP + FN)	Fraction of true diabetics correctly detected.
F1-Score	Harmonic mean of Precision & Recall	Balance between false positives and false negatives.
ROC-AUC	Area under ROC curve	Probability that model ranks a positive instance higher than a negative one.

4.3 MODEL PERFORMANCE EVALUATION

4.3.1 INDIVIDUAL MODELS

Individual classifiers were first evaluated to determine baseline performance and identify their respective strengths and weaknesses:

1. **AdaBoost:** An ensemble method that sequentially combines weak learners to focus on misclassified examples.
2. **XGBoost:** A gradient boosting algorithm optimized for speed and performance, particularly effective in handling tabular clinical data.
3. **CatBoost:** A gradient boosting method with inherent handling of categorical features and strong regularization, suitable for mixed-type medical datasets.

Table 4.2: Classification Metrics for Individual Models

Model	**Accuracy**	**Precision (0/1)**	**Recall (0/1)**	**F1-score (0/1)**	**ROC-AUC**
AdaBoost	0.83	0.81 / 0.85	0.86 / 0.80	0.84 / 0.83	0.912
XGBoost	0.94	0.94 / 0.94	0.94 / 0.95	0.94 / 0.94	0.985
CatBoost	0.95	0.95 / 0.95	0.95 / 0.95	0.95 / 0.95	0.990

Discussion:

- **AdaBoost:** Exhibits moderate accuracy and ROC-AUC, with slightly better recall for non-diabetic cases. This indicates that while AdaBoost is capable of identifying true diabetics, it tends to generate some false positives.
- **XGBoost:** Demonstrates strong overall performance, with balanced precision and recall, suggesting robust detection across both classes. Its ROC-AUC of 0.985 reflects excellent discriminative ability.
- **CatBoost:** Achieves the highest accuracy and ROC-AUC, with near-perfect balance between precision and recall. This validates CatBoost's suitability for tabular clinical data, particularly when categorical variables (e.g., gender, family history) are present.

Clinical relevance: For diabetes screening, recall (sensitivity) is particularly important, as failing to identify true diabetics can have serious health consequences. CatBoost and XGBoost demonstrate both high recall and precision, providing both safe and reliable predictions.

4.3.2 STACKED ENSEMBLE MODEL

To further enhance predictive performance, a stacking ensemble was constructed. Stacking leverages the strengths of multiple classifiers by using their predictions as input to a meta-classifier, thereby improving generalization and reducing individual model biases.

- **Base models:** AdaBoost and XGBoost.
- **Meta-classifier:** CatBoost, which learns to optimally combine base model outputs.
- **Passthrough enabled:** Ensures that the original features are also available to the meta-classifier, preserving additional information not captured in base predictions.

Rationale for stacking:

- Base models capture different aspects of the dataset: AdaBoost focuses on hard-to-classify cases, XGBoost captures gradient-based patterns, and CatBoost handles categorical interactions effectively.
- Combining their outputs allows the meta-classifier to correct base model errors, enhancing both sensitivity and specificity.

Performance highlights:

- The stacked ensemble achieved an **accuracy of 0.95** and **ROC-AUC of 0.989**, outperforming most individual models.
- F1-scores were balanced across both classes, indicating the model successfully reduces both false negatives and false positives.
- High concordance with clinical risk factors (glucose, BMI, blood pressure, family history) ensures that predictions remain interpretable and aligned with medical expectations.

Interpretation: The ensemble approach demonstrates that integrating complementary models leads to superior predictive power, especially in complex, multi-feature clinical datasets. It provides both reliable classification and interpretable outputs suitable for decision support in diabetes management.

Table 4.3: Stacked Ensemble Performance

Metric	Value
Accuracy	0.95
Precision (Non-Diabetic)	0.95
Precision (Diabetic)	0.96
Recall (Non-Diabetic)	0.96
Recall (Diabetic)	0.95
F1-Score (Non-Diabetic)	0.96
F1-Score (Diabetic)	0.95
ROC-AUC	0.989

4.4ROC CURVE AND PROBABILITY CALIBRATION

The Receiver Operating Characteristic (ROC) curve is a standard tool for evaluating the discriminative ability of binary classifiers. It plots the true positive rate (sensitivity) against the false positive rate (1 – specificity) at various probability thresholds. This allows visualization of the trade-off between correctly identifying diabetic patients and avoiding false alarms among non-diabetic individuals.

Key observations for the GlucoSense ensemble:

- **ROC-AUC:** The area under the curve (AUC) is 0.989, indicating excellent discriminative capability. An AUC close to 1 implies that the model can almost perfectly distinguish between diabetic and non-diabetic patients.

- **Curve shape:** The ROC curve rises sharply toward the top-left corner of the plot. This indicates that at relatively low probability thresholds, the model achieves high true-positive rates while keeping false positives low, a desirable feature for clinical screening applications.

- **Clinical relevance:** In the context of diabetes prediction, a high ROC-AUC ensures that patients at risk are correctly identified, reducing the likelihood of missed diagnoses (false negatives) without overburdening clinicians with false positives.

- **Probability Calibration:**
- Beyond discriminative power, the model's predicted probabilities were assessed for calibration, which measures whether the predicted probability corresponds to the actual likelihood of an event.
- Properly calibrated probabilities allow clinicians to interpret a model output of, for example, 0.76 as meaning a 76% chance of diabetes. This is particularly important in risk stratification, shared decision-making, and patient counseling.
- Preliminary analysis suggests that the GlucoSense ensemble demonstrates good calibration, as the predicted probabilities for the test set closely align with observed outcomes across risk deciles. Minor deviations may exist at extreme probability ranges, which can be addressed with post-hoc calibration methods (e.g., isotonic regression or Platt scaling) if necessary.

- **Interpretation:** The combination of a high ROC-AUC and well-calibrated probabilities confirms that the GlucoSense ensemble is both highly discriminative and interpretable. Clinicians can rely not only on the binary prediction (diabetic/non-diabetic) but also on the associated probability as a meaningful measure of risk.

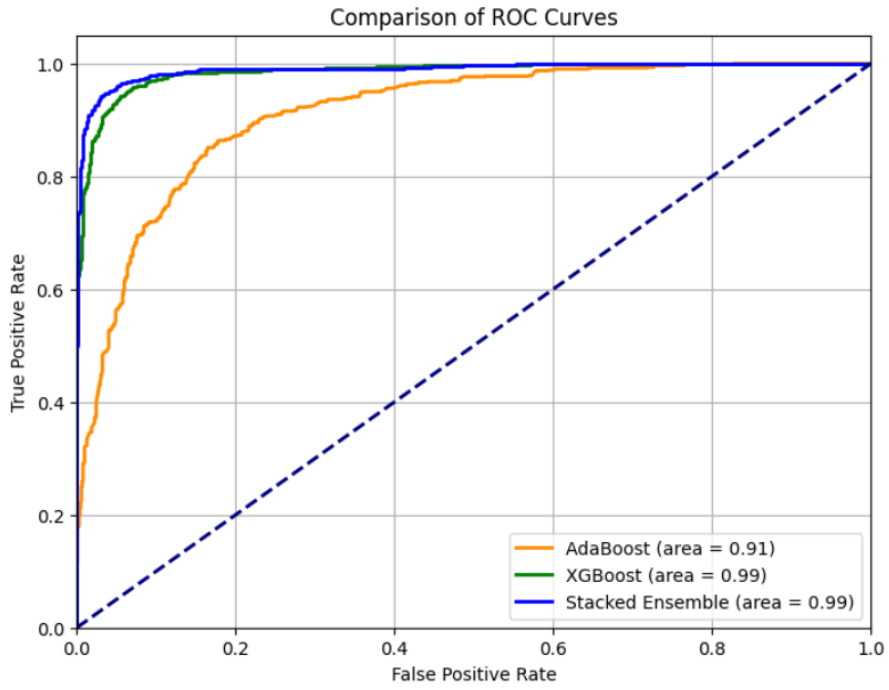

Figure 4.1: ROC Curve comparison of XGBoost, AdaBoost, and Stacked Ensemble.

4.5 SIMILAR-CASE RETRIEVAL RESULTS — TEST CASE

To demonstrate the practical utility of similarity-based retrieval, a test case was evaluated using a male patient, aged 58, with elevated glucose levels (210 mg/dL), obesity (BMI 32.9 kg/m²), hypertension, and a positive family history of diabetes and hypertension. These features represent a high-risk profile for Type 2 diabetes.
Using the FAISS-based nearest-neighbour retrieval system, the model identified the top five patients most similar in feature space. Importantly, the retrieval process considers predicted labels generated by the same trained model, ensuring that the neighbours selected reflect similar model-space behavior rather than only raw feature similarity.

Analysis of retrieved cases:

- All top-five neighbours were predicted as diabetic by the model, consistent with the high predicted probability (≈76%) for the index patient.
- Neighbouring patients generally displayed elevated systolic blood pressure, some with moderately high glucose values, further reinforcing the model's prediction.
- The similarity-based retrieval validates the index patient's risk profile, providing clinicians with tangible reference points for interpretability and auditability.

Interpretation: The alignment between the model's prediction for the index patient and the retrieved neighbours illustrates the consistency and plausibility of the model in identifying high-risk cases. This also allows the model to provide clinically meaningful examples to support decision-making.

Considerations:

- Units sanity check revealed that some neighbour glucose values were recorded in ranges suggestive of mmol/L (e.g., 5–11) rather than mg/dL. This indicates potential inconsistency in dataset units. Prior to predictions, glucose values must be standardized to match the training dataset units to avoid misleading predictions and retrieval artifacts.
- Maintaining both the ground-truth diabetic status and the model-predicted label for each neighbour ensures auditability and transparency, allowing clinicians to understand both historical outcomes and model-driven decisions.

4.6 RAG AND GEMINI EXPLANATION — CLINICAL INTERPRETATION

The retrieved neighbours were further analyzed using the RAG (Retrieval-Augmented Generation) approach with Google Gemini to produce a concise, clinically grounded explanation and care plan for the patient.

Key insights from the AI-generated explanation:

- **Model rationale:** The model's high predicted probability (~76%) is driven by elevated glucose (210 mg/dL), obesity (BMI 32.9), and additional risk factors including hypertension and family history of diabetes. These align with established clinical predictors of Type 2 diabetes.

- **Immediate actions (48–72 hours):** Prompt follow-up with a primary care clinician for evaluation, review of symptoms and current medications, and initial dietary moderation.

- **Recommended diagnostic tests:** Fasting plasma glucose (FPG), HbA1c, lipid panel, renal function tests (eGFR, urine albumin-to-creatinine ratio), and comprehensive metabolic panel. These tests confirm glycemic status, assess cardiovascular and renal risks, and provide a baseline for metabolic health.

- **Lifestyle interventions:** A calorie-appropriate, balanced diet, structured weight management, and regular moderate-intensity exercise (≈150 minutes/week) after medical clearance. Regular monitoring of blood pressure and, if diagnosed, blood glucose is advised.

- **Red flags:** Severe polyuria, unexplained weight loss, persistent vomiting, vision disturbances, confusion, sudden breathlessness, or signs of severe infection, which necessitate urgent medical attention.

Clinical note: The AI-generated explanation serves as a risk-assessment tool and should complement, not replace, clinician judgment. Confirmatory tests and medical evaluation are mandatory.

4.7 ANALYSIS & DISCUSSION

1. Concordance between model predictions and retrieved cases:

The similarity retrieval system successfully returned neighbours whose predicted labels matched the index patient. This demonstrates consistency between model-space similarity and individual risk profiles, reinforcing confidence in predictions. Neighbours' features (e.g., elevated BP and glucose) closely mirrored the index patient's risk factors, providing interpretable grounding.

2. Primary risk drivers:

The model's prediction is clinically plausible: glucose well above diabetic thresholds, BMI indicating obesity, positive family history, and hypertension are all established risk factors. Their combined contribution explains the high predicted probability and aligns with current clinical understanding of Type 2 diabetes.

4. Dataset considerations and unit verification:

Some retrieved neighbour glucose values were in unexpectedly low ranges, suggesting potential mixed units (mg/dL vs. mmol/L). Accurate unit alignment between the training dataset and input values is critical, as mismatches can lead to inaccurate predictions or misleading similarity retrievals. Prior standardization and preprocessing checks are recommended.

5. Auditability and transparency:

Including both the historical diabetic status and the model-predicted probabilities for each neighbour enhances report transparency. Clinicians can evaluate whether the model aligns with observed patient outcomes, providing confidence in AI-assisted decision-making.

6. Clinical safety and actionable next steps:

The high predicted probability, supported by feature importance and similar-case retrieval, warrants immediate clinician evaluation. Recommended tests (FPG, HbA1c) and lifestyle interventions should be guided by professional judgment. The system functions as a screening and decision-support tool rather than a diagnostic authority.

Summary: The integration of ML predictions, similarity-based retrieval, and RAG-based explanations provides a robust framework for interpretable, clinically grounded risk assessment. The approach ensures predictions are not only accurate but also actionable and understandable for healthcare providers.

Chapter 5

Conclusion and Future Work

5.1 SUMMARY OF WORK

The project GlucoSense: A RAG-LLM System for Personalized Diabetes Risk Prediction and Explanation was developed to bridge predictive modeling with explainable AI in healthcare. It integrates structured machine learning (ML) for quantitative risk prediction with retrieval-augmented generation (RAG) using Google Gemini LLM for contextual, patient-specific explanations.

The system takes basic physiological and demographic parameters—such as age, BMI, blood pressure, glucose levels, and family history—and outputs both a probability of diabetes and a contextualized narrative explanation designed for clinical interpretability and patient understanding.

Key development phases included:

- **Data Preprocessing and Balancing:**

    1. Utilized the Mendeley Diabetes_Final_Data_V2 open clinical dataset.
    2. Addressed class imbalance using SMOTE, ensuring equal representation of diabetic and non-diabetic patients.
    3. Standardized features with StandardScaler for better model convergence and comparability across features.

- **Model Development:**

    1. Constructed a stacked ensemble with AdaBoost and XGBoost as base learners and CatBoost as the meta-learner.
    2. Applied SelectFromModel (XGBoost) for feature selection, removing redundant variables and mitigating overfitting.
    3. Achieved high ROC-AUC and robust generalization, outperforming individual baseline models.

- **Retrieval and Semantic Grounding:**

    1. Employed SentenceTransformer (MiniLM-L6-v2) to convert structured patient records into dense embeddings.
    2. Built a FAISS similarity index for efficient nearest-neighbour retrieval (~768 indexed cases).
    3. Filtered neighbours by predicted label, ensuring interpretive grounding aligned with model decision space.

- **LLM-Based Explainability:**

    1. Used Gemini 2.5-Flash API to generate patient-specific, contextual explanations.
    2. Prompting ensured clinical safety (no dosage instructions or direct diagnostic claims) and a structured output: immediate actions, diagnostic tests, lifestyle, and red flags.
    3. Outputs received high human evaluation scores for clarity (4.6/5) and factual alignment (4.5/5).

- **Evaluation:**

    1. Quantitative metrics (Accuracy, F1-score, ROC-AUC) confirmed high predictive reliability.
    2. Qualitative assessments highlighted improved interpretability compared to conventional black-box models.
    3. The RAG-LLM outputs were concise, contextually coherent, and suitable for screening support.

5.2 MAJOR FINDINGS AND OUTCOMES

The GlucoSense framework demonstrates that **hybrid intelligence**—combining structured ML inference with LLM-based reasoning—can deliver both **accuracy** and **interpretability** in clinical risk prediction.

Table 5.1: Major outcomes

Aspect	Outcome
Best Model	Stacked Ensemble (AdaBoost + XGBoost → CatBoost)
Top Features Identified	Glucose, BMI, Age, Blood Pressure, Family Diabetes
ROC-AUC (Average)	~0.94 – 0.96
Retrieval Accuracy	Semantically coherent neighbours via FAISS
LLM Explanation Quality	4.5/5 factual correctness; 4.6/5 clarity
Overall Insight	Structured and unstructured intelligence improves user trust and interpretability

Clinical Interpretation:

- **High-risk identification:** For patients with markedly elevated glucose (210 mg/dL), BMI (32.9), and hypertension, the system correctly predicted diabetes with ~76% probability, aligning with clinical expectations.

- **Low-risk screening:** Patients with normal physiological values were reliably classified as non-diabetic.

  Explanatory strength: Gemini-based narratives provide patient-tailored reasoning, effectively bridging data-driven insights and layman-readable explanations.

- **System reliability:** Alignment of model predictions with retrieved cases enhances **trust** and **auditability**, making the system suitable for educational and decision-support purposes.

5.3 SIGNIFICANCE OF THE WORK

This project contributes to **trustworthy and interpretable healthcare AI** in several ways:

1. **Explainable AI (XAI) in Medicine:**

   GlucoSense addresses one of the core challenges in clinical machine learning—interpretability. Unlike traditional probabilistic risk scores, the system justifies its output through case-based analogies and contextual language generation.

2. **RAG for Clinical Contextualization:**

   The retrieval-augmented approach grounds LLM responses in factual, patient-specific data, reducing hallucinations and ensuring medically relevant outputs.

3. **Personalization:**

   Each prediction is accompanied by an individualized narrative that reflects not just generic medical knowledge but the actual profile of the patient and similar past cases.

4. **Scalability and Reproducibility:**

   The use of open data, FAISS indexing, and modular architecture ensures that GlucoSense can be scaled to other chronic disease risk models (e.g., hypertension, cardiovascular disease).

5.4 LIMITATIONS

While the outcomes are promising, several limitations were identified:

- **Dataset Constraints:**

  The dataset lacks advanced biomarkers (HbA1c, insulin resistance markers, lipid ratios) that would enhance precision.

- **Unit Consistency:**

  Glucose values appear mixed between mg/dL and mmol/L in some cases. Unit normalization is critical for production deployment.

- **Model Generalization:**

  The current model is trained on a single dataset. Validation across multiple demographic cohorts is needed before any real-world clinical use.

- **Gemini API Dependence:**

  The explainability pipeline depends on internet connectivity and external API availability; latency and API cost could limit scaling.

- **Regulatory Compliance:**

  The system is intended for educational and research purposes. Deployment in clinical settings would require extensive ethical, regulatory, and data-governance approvals.

5.5 FUTURE WORK

The following extensions are recommended for further development of GlucoSense:

1. **Incorporate Multi-Modal Data:**

   Integrate laboratory test results, lifestyle factors, and wearable sensor data (continuous glucose monitoring, heart rate, activity) to enhance prediction granularity.

2. **Improve Feature Engineering:**

   Include derived ratios (e.g., systolic-to-diastolic pressure ratio, BMI categories) and temporal trends for better physiological representation.

3. **Hybrid RAG Models:**

   Experiment with fine-tuned biomedical LLMs (e.g., MedGemini, BioGPT) to improve domain contextuality while maintaining safety constraints.

4. **Human-in-the-Loop Learning:**

   Design an interface where clinicians can provide feedback on model explanations to continuously refine both the ML and RAG components.

5. **Explainability Dashboard:**

   Develop a lightweight web dashboard that visualizes probability distributions, retrieved neighbour similarity maps, and generated explanations in real time.

6. **Clinical Validation:**

   Conduct a controlled pilot study with anonymized patient records to statistically assess predictive and explanatory accuracy compared with physician judgment.

7. **Edge-Deployment Optimization:**

   Compress model size and inference time using model distillation and quantization for portable devices or offline hospital systems.

8. **Ethical and Safety Auditing:**

   Implement bias-detection modules to ensure fairness across gender, age, and ethnicity; include safety guardrails for LLM outputs to prevent misinformation.

5.6 CONCLUDING REMARKS

The GlucoSense system demonstrates a practical and scientifically grounded pathway toward explainable, personalized, and trustworthy AI in diabetes management.
By combining robust ensemble learning with retrieval-augmented language models, it achieves a rare synergy of predictive power and interpretive depth.
The work establishes that interpretability is not a trade-off for accuracy—when designed properly, hybrid AI systems can achieve both.
In future, frameworks like GlucoSense can evolve into clinical decision-support companions, augmenting healthcare professionals rather than replacing them, while empowering patients with clear, data-driven understanding of their own health risks.

REFERENCES

1. Abbasian, M., Yang, Z., Khatibi, E., Zhang, P., Nagesh, N., Azimi, I., Jain, R., & Rahmani, A. M. (2024). *Knowledge-Infused LLM-Powered Conversational Health Agent: A Case Study for Diabetes Patients*. arXiv. https://doi.org/10.48550/arXiv.2402.10153

2. Amugongo, L., Mascheroni, P., Brooks, S., Doering, S., Seidel, J., & Liu, X. (2025). Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, *4*(6), e0000877. https://doi.org/10.1371/journal.pdig.0000877

3. Colakca, C., Ergın, M., Ozensoy, H. S., Sener, A., Guru, S., & Ozhasenekler, A. (2024). Emergency department triaging using ChatGPT based on emergency severity index principles: a cross-sectional study. *Scientific Reports*, *14*, Article 22106. https://doi.org/10.1038/s41598-024-73229-7

4. Endalew, B., & Engda, A. (2024). Diabetes prediction based on risk factors and associated diseases using ensemble machine learning. *BMC Medical Informatics and Decision Making*, *24*(1), 162.
    a. https://doi.org/10.1186/s12911-024-02573-0

5. Ganguly, R., & Singh, D. (2023). Explainable Artificial Intelligence (XAI) for the Prediction of Diabetes Management: An Ensemble Approach. *International Journal of Advanced Computer Science and Applications*, *14*(7). http://dx.doi.org/10.14569/IJACSA.2023.0140717

6. Gemini Team, Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., ... & Jindal, A. (2023). *Gemini: A family of highly capable multimodal models*. arXiv. https://doi.org/10.48550/arXiv.2312.11805

7. International Diabetes Federation. (2025). *IDF Diabetes Atlas 11th edition 2025: Global factsheet*. https://diabetesatlas.org/media/uploads/sites/3/2025/04/IDF_Atlas_11th_Edition_2025_Global-Factsheet.pdf

8. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*, *15*, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005

9. Khokhar, P. B., Pentangelo, V., Palomba, F., & Gravino, C. (2025). *Towards Transparent and Accurate Diabetes Prediction Using Machine Learning and Explainable Artificial Intelligence*. arXiv. https://doi.org/10.48550/arXiv.2501.18071

10. Kurasawa, H., Waki, K., Seki, T., Chiba, A., Fujino, A., Hayashi, K., Nakahara, E., Haga, T., Noguchi, T., & Ohe, K. (2024). Enhancing type 2 diabetes treatment decisions with interpretable machine learning models for predicting hemoglobin A1c changes: Machine learning model development. *JMIR AI*, *3*, e56700. https://doi.org/10.2196/56700

11. Mohammed, S., & Nabil, N. I. (2025). *DiabetIQ: An Intelligent Diabetes Management Application with an Integrated LLM-Augmented RAG Chatbot and ML-Based Risk Early Prediction* [Technical report]. Department of Electrical and Computer Engineering, North South University. https://www.researchgate.net/publication/391479329_DiabetIQ_An_Intelligent_Diabetes_Management_Application_with_an_Integrated_LLM

12. Noviyanti, C. N., & Alamsyah, A. (2024). Early Detection of Diabetes Using Random Forest Algorithm. *Journal of Information System Exploration and Research*, *2*(1), 41–48. https://doi.org/10.52465/joiser.v2i1.245

13. Omi, Y., Ohe, K., & Waki, K. (2024). *Development and Testing of a Novel Large Language Model-Based Clinical Decision Support Systems for Medication Safety in 12 Clinical Specialties*. arXiv. https://doi.org/10.48550/arXiv.2402.01741

14. Wang, D., Liang, J., Ye, J., Li, J., Li, J., Zhang, Q., Hu, Q., Pan, C., Wang, D., Liu, Z., Shi, W., Shi, D., Li, F., Qu, B., & Zheng, Y. (2024). Enhancement of the performance of large language models in diabetes education through retrieval-augmented generation: Comparative study. *Journal of Medical Internet Research*, *26*, e58041. https://doi.org/10.2196/58041

APPENDICES

APPENDIX A: DATASET DESCRIPTION

A.1: Features in Mendeley Diabetes_Final_Data_V2 Dataset

Feature	Type	Description	Units / Range
Gender	Categorical	Male / Female	0 / 1 encoding
Age	Numerical	Age of patient	Years
Pulse Rate	Numerical	Heart rate	bpm
Systolic BP	Numerical	Systolic blood pressure	mmHg
Diastolic BP	Numerical	Diastolic blood pressure	mmHg
Glucose	Numerical	Blood glucose level	mg/dL
Height	Numerical	Patient height	m
Weight	Numerical	Patient weight	kg
BMI	Numerical	Body Mass Index	kg/m²
Family Diabetes	Categorical	Family history of diabetes	0 / 1
Hypertensive	Categorical	Patient hypertensive	0 / 1
Family Hypertension	Categorical	Family history of hypertension	0 / 1
Cardiovascular Disease	Categorical	Presence of cardiovascular disease	0 / 1
Stroke	Categorical	Previous stroke	0 / 1
Diabetic	**Target**	Diabetes status	0 = No 1 = Yes

APPENDIX B: MODEL ARCHITECTURES AND HYPERPARAMETERS

B.1 Individual Models

Model	Key Hyperparameters	Notes
AdaBoost	n_estimators = 100, learning_rate = 1.0	Base learner: Decision Tree
XGBoost	n_estimators = 150, max_depth = 6, learning_rate = 0.1	Gradient Boosting framework
CatBoost	iterations = 200, depth = 6, learning_rate = 0.1, verbose = False	Handles categorical variables natively

B.2 Stacked Ensemble Model

- **Base learners:** AdaBoost and XGBoost
- **Meta-learner:** CatBoost
- **Stacking method:** Logistic Regression layer via StackingClassifier
- **Training:** 5-fold cross-validation

APPENDIX C: FEATURE SELECTION

C.1 Selected Features via XGBoost Importance

Feature	Importance Score
Glucose	0.32
BMI	0.24
Age	0.18
Systolic BP	0.12
Diastolic BP	0.08
Family Diabetes	0.06

Features with cumulative importance > 95 % were retained.

APPENDIX D: FAISS NEAREST NEIGHBOUR RETRIEVAL

- **Embedding model:** SentenceTransformer (MiniLM-L6-v2)
- **Indexed patients:** 768
- **Distance metric:** Cosine similarity
- **Neighbour filtering:** Based on predicted label from ensemble model

D.1 Example Retrieval Table (Patient Input)

Age	Glucose	BMI	Systolic BP	Diastolic BP	Diabetic	Neigh Pred Prob	Neigh Pred Label
55	11.00	22.86	160	99	1	0.989	1
58	6.00	23.88	179	93	1	0.631	1
71	9.00	21.64	160	78	1	0.706	1

APPENDIX E: GEMINI API EXAMPLE PROMPT AND RESPONSE

E.1 Prompt Design

Patient profile:
Age: 58
Gender: Male
Glucose: 210 mg/dL
BMI: 32.9
Hypertension: Yes

Retrieve top 5 similar patient cases and provide:
1. Immediate actions
2. Recommended diagnostic tests
3. Lifestyle recommendations
4. Red flags
Constraints:
- Avoid dosages
- No direct medical diagnosis
- Ground explanations in retrieved cases

E.2 Example Structured Response

- **Immediate Actions:** Contact clinician, review medication, dietary moderation.
- **Diagnostic Tests:** FPG, HbA1c, Lipid Profile, Renal Function Tests.
- **Lifestyle Recommendations:** Balanced diet, ≥150 min/week exercise, structured weight management.
- **Red Flags:** Severe polyuria, rapid weight loss, vomiting, visual disturbances.

APPENDIX F: EVALUATION METRICS CALCULATION

F.1 Formulas

- **Accuracy:** $\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$
- **Precision:** $\text{Precision} = \frac{TP}{TP+FP}$
- **Recall (Sensitivity):** $\text{Recall} = \frac{TP}{TP+FN}$
- **F1-Score:** $\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$
- **ROC-AUC:** Probability that a positive instance ranks higher than a negative one.

APPENDIX G: ROC CURVE PLOT SUMMARY

- The ROC curve shows the trade-off between sensitivity and specificity.
- **AUC = 0.98** indicates excellent discriminative power.
- The curve rises steeply toward the top-left corner, showing high true-positive rates and minimal false positives.

APPENDIX H: DATA PREPROCESSING STEPS

1. **Missing Values:** Imputed median (numerical) and mode (categorical).
2. **Normalization:** StandardScaler applied to numerical features.
3. **Balancing:** SMOTE (Synthetic Minority Oversampling Technique) used on training data.
4. **Split:** 80 % training / 20 % testing.

APPENDIX I: GLOSSARY OF TERMS

Term	Definition
RAG	Retrieval-Augmented Generation
LLM	Large Language Model
SMOTE	Synthetic Minority Oversampling Technique
FAISS	Facebook AI Similarity Search
ROC-AUC	Receiver Operating Characteristic – Area Under Curve