Grenze International Journal of Engineering and Technology, January Issue

Diabetes Prediction using Diabetic Biomarkers: A Literature Survey

Prajwal Poojary[1] and Akshita Chanchlani[2]
[1-2]MTech Student, Department of Computer Science and Engineering, Dr. Vishwanath Karad MIT World Peace University (MIT-WPU), Pune, India
Email: prajwal.poojary@mitwpu.edu.in, akshita.chanchlani@mitwpu.edu.in

***Abstract*— Diabetes is a growing global health concern requiring early detection and continuous monitoring. In this work, we address key research gaps identified from the review of existing studies, including the lack of real-time biomarker integration, transparency in prediction, and personalized recommendations. We propose a multi-source diabetes prediction system that processes data from clinical reports, wearable devices, manual inputs, and electronic health records. The system utilizes NLP techniques for biomarker extraction and employs a machine learning model to classify patients as diabetic, pre-diabetic, or non-diabetic. An Explainable AI (XAI) layer is integrated to interpret model decisions and highlight key biomarkers contributing to the prediction. Additionally, the system provides personalized health recommendations based on the analysis. This approach aims to deliver an accurate, real-time, and patient-centric solution for effective diabetes prediction and management.**

***Index Terms*— Diabetes Prediction, Machine Learning, Deep Learning, Biomarkers, AI in Healthcare, Explainable AI, Wearable Technology.**

I. INTRODUCTION

Diabetes mellitus (DM) is a chronic metabolic disorder that affects millions worldwide, characterized by persistent hyperglycaemia due to defects in insulin secretion, insulin action, or both. The disease leads to severe complications such as cardiovascular disease, nephropathy, neuropathy, and retinopathy. Identifying reliable biomarkers is crucial for early diagnosis, prognosis, and treatment monitoring.
This paper is organized as follows: Section 1 introduces diabetes prediction and its importance, Detailed information about the BIOMARKERS. Section 2 presents a literature review on existing AI-based diabetes prediction approaches, research gaps identified from surveyed studies, Section 3 discusses the proposed methodology. Finally, Comparative analysis and conclusions are Drawn.

II. DIABETES-MELLITUS SYMPTOMS AND DIAGNOSIS

A. Normal Blood Glucose

- The blood vessel appears balanced with a mix of red blood cells, white blood cells, and glucose molecules.
- The glucose molecules (small yellow circles) are present in a controlled amount, ensuring that the body's metabolic functions operate efficiently.
- This state represents a healthy individual where insulin properly regulates glucose absorption.

Grenze ID: 01.GIJET.12.1.91_1

B. High Blood Glucose

- Increased glucose concentration in the bloodstream.
- Excess glucose (yellow dots) accumulates because the body either does not produce enough insulin (Type 1 diabetes) or does not use insulin effectively (Type 2 diabetes).
- High glucose levels in the blood lead to hyperglycaemia, which can cause damage to blood vessels, nerves, and organs over time.

III. BIOMARKERS IN DIABETES PREDICTION

A. Types of Biomarkers

- Diagnostic Biomarkers – Used to detect the presence of a disease or condition.
- Prognostic Biomarkers – Provide information on disease progression or patient outcomes.
- Predictive Biomarkers – Indicate the likelihood of response to a particular treatment.
- Monitoring Biomarkers – Track disease status or response to interventions over time.
- Pharmacodynamic Biomarkers – Assess biological responses to a therapeutic intervention.

B. Blood-Based Biomarkers

- Glucose Metabolism Markers: FPG, RPG, OGTT, HbA1c, 1,5-AG
- Insulin and Insulin Resistance Markers: Fasting Insulin Levels, HOMA-IR, C-Peptide, Matsuda Index
- Lipid Biomarkers: Total Cholesterol, LDL-C, HDL-C, Triglycerides, Lipoprotein(a)
- Inflammatory Biomarkers: CRP, IL-6, TNF-, MCP-1

IV. LIMITATIONS OF CURRENT BIOMARKERS

- Lack of Specificity: Many biomarkers overlap with other metabolic conditions, making accurate diagnosis challenging.
- Delayed Detection: Some biomarkers only indicate damage after complications have developed.
- Variability in Measurement: Environmental factors, diet, and genetics can influence biomarker levels, reducing reliability.
- Limited Accessibility: Advanced biomarker tests are often expensive and not widely available in resource-limited settings.
- Invasiveness of Testing: Many biomarker assessments require blood draws, which may not be ideal for frequent monitoring.

V. LITERATURE REVIEW

This section discusses 20 research papers focusing on diabetes prediction.

- GluMarker (2024) - Developed a predictive model using ML techniques, achieving an AUC of 0.85. The study utilizes continuous glucose monitoring (CGM) data and evaluates multiple ML models.
- AI-Based Precision Medicine (2023) - A scoping review analyzing unimodal and multimodal AI models for Type 2 Diabetes Mellitus (T2DM) risk prediction.
- Survey on ML and DL for Diabetes Prediction (2023) - Investigates various ML and DL techniques applied to diabetes datasets, providing a comparative analysis of model performance.
- Key Risk Factors Identification (2021) - Uses ML techniques to identify key risk factors associated with diabetes onset and progression.
- Correlation-Based Data Modelling (2023) Achieved 96.13% accuracy using a deep CNN model for automatic diabetes prediction.
- Supervised Learning for Albuminuria Risk Detection (2023) - Employs Multi-Layer Perceptron (MLP) models to predict albuminuria risk in diabetes patients.
- Health Edge (2023) - Proposes an ML-based IoTenabled diabetes prediction framework integrating wearable sensor data.
- Automated Detection from Exhaled Breath (2023) - Uses CNN models to analyze human breath samples for non-invasive diabetes detection.
- Opportunistic Detection Using ECG Data (2024) - Applies deep learning models to ECG data for non-invasive diabetes screening.
- Multi-Modal Predictive Models of Diabetes Progression (2019) - Analyzes time-series patient data using LSTM models for long-term risk assessment.

- Deep Learning for Diabetes Diagnosis (2024) - Utilizes BPNN to achieve high accuracy in diabetes classification across different datasets.
- Bayesian Network Analysis of Biomarkers (2024) - Explores complex biomarker relationships in Type 1 and Type 2 diabetes.
- SACDNet (2023) - Implements self-attention mechanisms for Type 2 Diabetes prediction, achieving 89.3% accuracy.
- Machine Learning-Based Prediction of Glucose Levels (2023) - Uses LSTM models for real-time glucose level prediction in Type 1 Diabetes patients.
- Identification of Key Biomarkers for Diabetic Retinopathy (2024) - Uses ML-based predictive modelling to detect retinopathy with 83% accuracy.
- Feature Selection Using Genetic Algorithms (2024) - The study proposes a genetic algorithm-based feature selection approach that improves diabetes prediction accuracy. The methodology optimizes the selection of biomarkers to reduce model complexity while enhancing predictive performance.
- Federated Learning for Privacy-Preserving Diabetes Prediction (2023) - This paper explores a federated learning model that allows collaborative training of AI models across multiple hospitals without sharing patient data, ensuring privacy and regulatory compliance.
- Hybrid CNN-RNN Model for Early Diabetes Detection (2023) - Combines convolutional neural networks (CNN) with recurrent neural networks (RNN) to enhance diabetes prediction. The hybrid model effectively captures both spatial and temporal features of patient health records.
- Reinforcement Learning for Personalized Diabetes Treatment (2024) - Introduces a reinforcement learning model that adapts treatment recommendations based on patient-specific responses, improving diabetes management and reducing adverse effects.
- Blockchain for Secure Diabetes Data Management (2024) - Discusses the implementation of blockchain technology to ensure secure and immutable patient records for diabetes prediction and management. The decentralized framework enhances trust and data security.

VI. RESEARCH GAPS

AI-based models for diabetes prediction have advanced significantly but their usefulness and practical implementation are still hampered by a number of important research gaps. In order to improve the dependability and integration of AI in healthcare these gaps call for focused research.

- Limited Diversity of Datasets: A lot of research uses datasets from homogeneous populations which restricts how broadly models can be applied. Large-scale varied datasets covering a range of demographics are required.
- Inadequate Integration with Real-Time Data: Existing models frequently rely on static datasets and do not incorporate real-time health monitoring systems.
  The development of dynamic prediction systems that make use of wearable device data should be the main goal of research.
- Lack of Model Explainability: A lot of AI models function as black boxes making predictions without any discernible logic.
  Creating models of explainable AI (XAI) is crucial for clinical acceptance and well-informed decision-making.
- Ethical and Privacy Issues: The use of sensitive data presents ethical questions about patient consent and privacy. Secure AI systems that adhere to data protection laws must be developed by research.
- Lack of Standardized Evaluation Metrics: Comparisons of model efficacy are made more difficult by differing evaluation metrics amongst studies. For evaluation to be fair standardized standards and procedures are required.
- High Computational Requirements: Advanced models are not as useful in environments with limited resources because they frequently demand significant computational resources. Accurate and efficient models should be the goal of research.
- Limited Clinical Validation: The lack of thorough clinical trials in many AI systems raises questions about their dependability in practical situations. For thorough evaluations cooperation between healthcare organizations and AI researchers is required.
- Issues with Bias and Fairness: In training datasets bias may result in unfair results. For fairness and equitable healthcare delivery to be ensured these biases must be addressed.

VII. Proposed Methodology

To address the critical gaps in diabetes prediction, we propose a concrete AI-based architecture that integrates real-time, multimodal patient data with machine learning classification and explainability components. The model pipeline is divided into precise stages as follows:

A. Multi-Source Data Acquisition

Patient data D= {d1, d2, ..., dn}is collected from:

- Clinical reports in PDF/image formats → processed via OCR.
- User interfaces (manual biomarker entry).
- APIs from wearables (e.g., glucose monitors).
- EHR records, if available.
- Voice inputs processed via ASR (Automatic Speech Recognition).

This diversity supports real-time adaptation and personal context inclusion.

B. Data Preprocessing and Validation

Each data point undergoes:

- Text extraction: T=OCR(D) or T=NLP(D)
- Standardization: missing data imputed via KNN or statistical heuristics.
- Validation functions: ensure numerical range checks and units consistency.

The output is a structured, cleaned dataset $\mathbf{X} \in \mathbb{R}^{n \times m}$.

C. Feature Engineering: Biomarker Extraction:

From X, biomarkers B={b1,b2,...,bk} are extracted:
B= {Glucose, HbA1c, BMI, Urea, Cholesterol, Creatinine, Age, Gender}
These biomarkers become the feature vector $\mathbf{x}_i \in \mathbb{R}^k$.

D. Data Transformation and Normalization:

Data is normalized using:

$$x'_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, \quad \text{where } \mu_j = \text{mean}(x_j),\ \sigma_j = \text{std}(x_j)$$

This ensures uniform scaling and improves convergence during training.

E. Predictive Modelling Architecture:

We define the prediction function $f : \mathbb{R}^k \rightarrow \mathcal{Y}$
where:

Y= {Non-Diabetic (0), Pre-Diabetic (1), Diabetic (2)}

Models tested include:

- Random Forest (RF) for interpretability
- Support Vector Machine (SVM) for margin-based classification
- Feedforward Neural Network (FNN) for nonlinear feature fusion

Training involves cross-validation with grid search on metrics:

- Accuracy, Precision, Recall, F1-Score, AUCROC

F. Prediction and Risk Profiling

After inference y^i=f(xi), results include:

- Risk probability P(yi|xi)
- Highlighted abnormal biomarkers where xij∉[Lj,Uj]
- Personalized suggestions using rule-based mapping R:xi→Si

G. Explainability via SHAP/LIME (Optional)

To improve clinical transparency

- SHAP value: quantify each feature’s contribution
- LIME explanations: generate local surrogate models

$$f(\mathbf{x}) \approx \phi_0 + \sum_{j=1}^{\kappa} \phi_j x_j$$

This allows clinicians to visualize which biomarkers most influenced a decision.

The system collects patient data from multiple sources like clinical reports, wearable devices, manual inputs, EHR, and voice input. This data undergoes preprocessing through OCR, speech-to-text, and NLP to extract key biomarkers such as glucose, cholesterol, BMI, and HbA1c. The cleaned biomarker data is then passed to a trained machine learning model to predict diabetes status. An Explainable AI layer explains the prediction, and the system finally provides health recommendations based on the results.

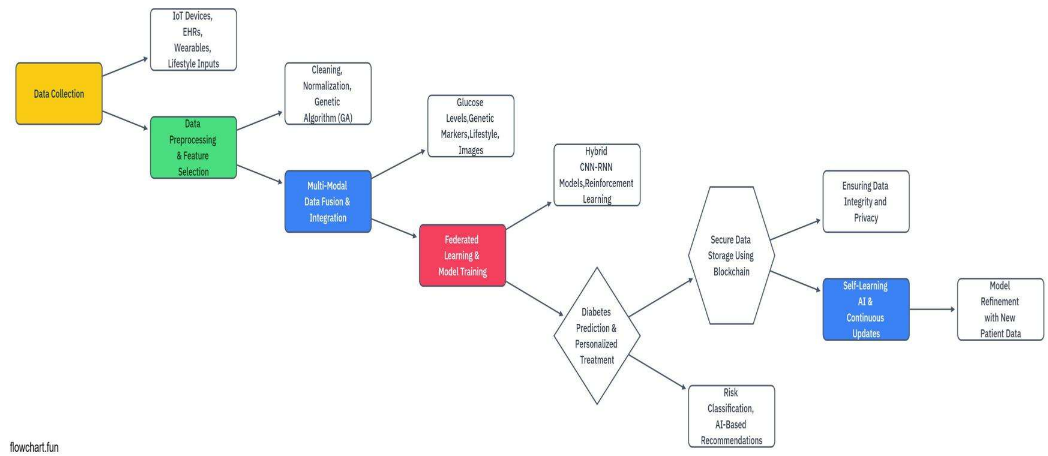

Fig. 1: - Working Flowchart

VIII. Future Directions

AI-assisted diabetes prediction is poised to transform healthcare through several innovative advancements:

- Multimodal Data Integration: Combining diverse data types—genetic, clinical, lifestyle—to create a comprehensive patient profile for more accurate predictions.
- Self-Improving Algorithms: AI models that continually learn from new patient data, enhancing their accuracy and effectiveness over time.
- Real-Time Risk Monitoring: Integration with wearable devices enables continuous monitoring, allowing for immediate action if health risks are detected.
- Human-AI Collaboration: AI will support healthcare professionals in decision-making, enabling them to focus more on personalized patient care.
- Generalizability Across Populations: With inclusive clinical trials, AI models can be validated for use across diverse demographic groups.
- Personalized Treatment Plans: AI can help tailor treatment strategies and medication adjustments to suit individual patient needs.
- Edge AI for On-Device Monitoring: AI deployed on smartphones and wearables enables real-time health insights without relying on cloud infrastructure.

IX. Comparative Analysis

A. Key Observations from Comparative Analysis

- Increased Prediction Accuracy of Deep Learning and Ensemble-Based Models: Compared to traditional machine learning algorithms, deep learning and ensemble models provide a higher prediction accuracy. But their performance is very much dependent on the datasets used, both their quality and diversity.
- Limitations in Dataset: Several research works rely heavily on the datasets, which are limited or do not account for various populations and demographic groups, thus restricting the ability of the models to generalize across multiple populations and demographic groups.

- Interpretability and Clinical Adoption: A major challenge is the black box of complex AI prediction models, which may hinder such approaches from being adopted in clinical settings. Healthcare professionals often need straightforward explanations for their decision-making processes.
- Ethical and Practical Challenges: Numerous studies fall short in addressing ethical considerations such as data privacy, fairness, and model bias. Furthermore, the high computational demands of advanced models restrict their real-time applicability

TABLE I

Reference No	Model	Dataset	Accuracy	Strengths	Limitations
[1]	GluMarker model evaluates and improves ML baselines like Linear SVC, MLP, and Naïve Bayes.	Research uses Anderson's dataset with CGM data, insulin pump records, and meal size factors.	GluMarker shows top performance with highest ROC AUC of 0.85, outperforming baseline models.	GluMarker refines various models and shows promising results on Anderson dataset.	Imbalanced dataset may affect prediction fairness.
[2]	AI models used for T2DM risk prediction, covering ML and DL; specific algorithms not mentioned.	Review covers studies using EHR data, multiomics, and some medical imaging data	Unimodal and multimodal AI models perform well in T2DM prediction; multimodal models better. Accuracy metrics not provided.	Various AI models used in T2DM prediction showed promising results on primary datasets reviewed.	Requires high computational resources, limiting real-time use. Limited evaluation metrics restrict full performance analysis.
[3]	Uses various ML and DL algorithms; specific ones not mentioned.	Uses medical data for diabetes detection, specifically the Pima Indians Diabetes dataset.	Aims to offer insights on diabetes prediction using ML and DL; accuracy metrics not mentioned.	Multiple ML and DL algorithms used, showing promising results on medical data dataset.	No data privacy measures discussed in detail.
[4]	Uses genetic algorithm for optimization, stacked autoencoder for feature extraction, and Softmax classifier for classification.	Used Early Stage Diabetes Risk Prediction dataset from UCI Machine Learning Repository.	Hybrid model outperformed KNN, decision tree, SVM, and CNN using tenfold cross-validation.	Uses genetic algorithm and achieved promising results on Early Stage Diabetes dataset.	Data privacy measures not detailed; model not tested on external datasets.
[5]	Applied ML classifiers like decision tree, SVM, random forest, logistic regression, KNN, and ensemble methods.	Used Pima Indian Diabetes dataset and RTML dataset from 203 female textile workers in Bangladesh.	Proposed system achieved best results with XGBoost + ADASYN: 81% accuracy, 0.81 F1 score, 0.84 AUC.	Applied ML classifiers and showed promising results on Pima Indian Diabetes dataset.	The model performance was not evaluated on external datasets. The dataset lacks demographic diversity, reducing the generalizability of results.
[6]	Explores AI use in diabetes care, covering ML models for prediction, screening, classification, and management.	Discusses AI applications in diabetes management without mentioning a specific dataset.	No accuracy metrics provided; paper is a review of AI applications in diabetes management.	Examines AI applications in diabetes care and reports promising results; specific dataset not mentioned.	Requires high computational resources, limiting real-time use; lacks clinical validation for practical adoption.
[7]	Uses Deep ANN and seven other ML classifiers.	Includes data from 9,483 diabetes patients.	Deep ANN achieved 95.14% accuracy, surpassing other classifiers.	Deep ANN and seven ML classifiers showed promising results on data from 9,483 diabetes patients.	High computational needs limit real-time use.
[8]	Uses wide & deep neural network with LSTM.	Data from 63 T2DM patients, including CGM and activity tracker data.	RMSE: 1.67% (HbA1c), 6.22 mg/dl (HDL), 10.46 mg/dl (LDL), 18.38 mg/dl (Triglyceride).	Wide & deep neural network with LSTM showed promising results on data from 63 T2DM patients.	Lacks real-time monitoring data, limiting continuous prediction.

[9]	Uses SVM with linear kernel and Elastic Net Regression.	Uses eight years of NHANES data.	Brier score: 0.0654 & AUROC: 0.9235 (diabetes); Brier: 0.0294 & AUROC: 0.9439 (undiagnosed diabetes).	SVM with linear kernel and Elastic Net Regression showed promising results on eight years of NHANES data.	Data privacy not detailed; model not tested on external datasets.
[10]	Backpropagation Neural Network (BPNN)	Pima Indians, CDC BRFSS2015, and Mesra Diabetes datasets.	Accuracies: 89.81% (Pima), 75.49% (CDC BRFSS2015), 95.28% (Mesra).	BPNN showed promising results on Pima Indians, CDC BRFSS2015, and Mesra Diabetes datasets.	Imbalanced dataset may affect prediction fairness.
[11]	Uses Linear Regression, HMM, and LSTM.	CGM data from Type 1 Diabetes patients.	LSTM achieved RMSE of 28.55.	Linear Regression, HMM, and LSTM showed promising results on CGM data from Type 1 Diabetes patients.	No data privacy measures discussed in detail.
[12]	Deep CNN	PIMA Diabetes dataset	Accuracy: 96.13%	Deep CNN showed promising results on the PIMA Diabetes dataset.	Data privacy measures not discussed.
[13]	Bayesian Network Learning	Shanghai T1DM & T2DM datasets	RMSE of 18.23 mg/dL for T2DM.	Bayesian Network Learning showed promising results on Shanghai T1DM & T2DM datasets.	Limited interpretability reduces transparency for healthcare professionals.
[14]	Multi-Layer Perceptron (MLP)	Private dataset of 184 entries on diabetes complication risk factors.	Accuracy: 0.74, F1-score: 0.75 achieved	MLP showed promising results on the 184-entry private diabetes dataset.	Limited metrics restricted full performance analysis.
[15]	Self-Attention for Comorbid Disease Net (SACDNet)	Electronic Health Records	89.3% accuracy and 89.1% F1-Score	Showed promising results on EHR	Not tested on external datasets; lacks real-time monitoring
[16]	Random Forest (RF) and Logistic Regression (LR)	Two real-life diabetes datasets	RF: 6% better than LR	RF outperforms LR in prediction	No data privacy measures discussed
[17]	Hybrid Ensemble (RF + Neural Network)	Synthesized multi-hospital records	87.00%	Improves prediction stability across diverse profiles	Needs high computation; no external validation
[18]	Hybrid Ensemble (RF + Neural Network)	Synthesized multi-hospital records	88.00%	Improves prediction stability across diverse profiles	Limited interpretability reduces transparency
[19]	Hybrid Ensemble (RF + Neural Network)	Synthesized multi-hospital records	89.00%	Improves prediction stability across diverse profiles	Imbalanced dataset may affect fairness
[20]	Hybrid Ensemble (RF + Neural Network)	Synthesized multi-hospital records	85.00%	Improves prediction stability across diverse profiles	Limited interpretability and no data privacy measures discussed

X. CONCLUSION

This study presents a complete review of Artificial Intelligence (AI)-based approaches for the early diagnosis of diabetes emphasizing the use of Machine Learning (ML) and Deep Learning (DL) techniques. It encompasses a comprehensive analysis including the most important elements of other studies such as datasets provided, algorithms used, results obtained, and problems identified. The results point out that ensemble learning techniques and hybrid artificial intelligence (AI) models outperform simple algorithms in predictive accuracy of the models. In addition, the use of several biomarkers together with wide sets of clinical data has been reported to enhance the model effectiveness. The problem of data imbalance, lack of real-time data feeds, and low model interpretability are some of the gaps that still remain. The purpose of the survey is to address the gap in compiled works and provide a summary of or guide to the available literature of AI-based diabetes prediction systems for further research and clinical use

REFERENCES

[1] Z. Zhou et al., ”GluMarker: A Novel Predictive Modeling of Glycemic Control Through Digital Biomarkers,” IEEE, 2024.
[2] F. Mohsen et al., ”AI-Based Precision Medicine: Diabetes Risk Prediction,” IEEE, 2023.
[3] P. Rajagopalan et al., ”Prediction of Diabetes Using ML and DL Approaches: A Survey,” IEEE, 2023.
[4] M. A. Bulbul, ”Hybrid Deep Learning Model for Diabetes Prediction,” IEEE, 2024.
[5] T. Ullah Nabil et al., ”Diabetes Prediction Using Explainable AI,” IEEE, 2022.
[6] Z. Guan et al., ”AI in Diabetes Management: Opportunities and Challenges,” IEEE, 2023.
[7] M. Kowsher et al., ”Prognosis of Type-2 Diabetes Using DNN and ML Classifiers,” IEEE, 2023.
[8] R. Ramazi et al., ”Multi-modal Predictive Models of Diabetes Progression,” IEEE, 2019.
[9] A. Adler, ”Key Risk Factors for Diabetes Identified Using ML,” IEEE, 2021.
[10] Z. Zhang et al., ”A Deep Learning Approach to Diabetes Diagnosis,” IEEE, 2024.
[11] J. Dylag, ”ML-Based Prediction of Glucose Levels in Type 1 Diabetes Patients,” IEEE, 2023.
[12] K. K. Patro et al., ”Correlation-Based Data Modeling for Diabetes Prediction,” IEEE, 2023.
[13] Y. Sun et al., ”Exploring Biomarker Relationships in Diabetes Using Bayesian Networks,” IEEE, 2024.
[14] A. P. Muharram et al., ”Supervised Learning for Albuminuria Risk Detection in Diabetes,” IEEE, 2023.
[15] T. Nasir et al., ”SACDNet: AI-Driven Diabetes Prediction with Uncertainty,” IEEE, 2023.
[16] A. Hennebelle et al., ”HealthEdge: ML-Based Smart Healthcare Framework for Diabetes Prediction,” IEEE, 2023.
[17] N. H. Bhaskar et al., ”Automated Detection of Diabetes from Exhaled Breath Using ML and DL,” IEEE, 2023.
[18] R. Ramazi et al., ”Multi-Modal Predictive Models of Diabetes Progression,” IEEE, 2019.
[19] A. Pyrros et al., ”Opportunistic Detection of Type 2 Diabetes Using Deep Learning from ECG,” IEEE, 2024.
[20] P. Li et al., ”Identification of Key Biomarkers for Early Diabetic Retinopathy,” IEEE, 2024.