



**MASTER'S DEGREE IN WEB TECHNOLOGY AND  
SYSTEMS ENGINEERING**

**AI-Driven Predictive Model for Diabetes Risk  
Assessment**

**AMINA TOUMIA**

**FIRMINO OLIVEIRA DA SILVA**

Master's Dissertation submitted for partial satisfaction of the requirements for the Master's degree carried out under the supervision of Professor Firmino da Silva and Jorge Duque presented to ISLA - Polytechnic Institute of Management and Technology of Vila Nova de Gaia to obtain the Master's degree in WEB TECHNOLOGY AND SYSTEMS ENGINEERING, in accordance with the Order nº 9371/2020.

## Acknowledgements

I would like to express my sincere gratitude to all those who supported and guided me throughout this research journey.

First and foremost, I extend my deepest appreciation to my supervisors, Professor Firmino Da Silva and Professor Jorge Duque, for their invaluable guidance, expertise, and continuous encouragement throughout the development of this dissertation. Their insightful feedback and dedication to academic excellence have been instrumental in shaping this work.

I am grateful to ISLA - Polytechnic Institute of Management and Technology of Vila Nova de Gaia for providing the academic environment and resources that made this research possible. Special thanks to the faculty of the Master's program in Web Technology and Systems Engineering for their support and knowledge sharing throughout my studies.

My heartfelt thanks go to my family for their unwavering support, patience, and encouragement during this challenging yet rewarding journey. Their belief in my abilities has been a constant source of motivation.

I would also like to acknowledge my colleagues and friends who provided valuable discussions, feedback, and moral support throughout the research process. Your companionship made this journey less solitary and more enjoyable.

Finally, I acknowledge the open-source community and the creators of the BRFS 2015 dataset made available through Kaggle, as well as the developers of scikit-learn, Django, and React frameworks that were essential to this project's implementation.

To all who contributed in any way to the completion of this work, I extend my sincere gratitude.

## Resumo

Este estudo desenvolveu um modelo preditivo baseado em IA para avaliação de risco de diabetes utilizando algoritmos de Random Forest e Regressão Logística. A análise utilizou o conjunto de dados BRFSS 2015 contendo 253.680 registros de saúde do Sistema de Vigilância de Fatores de Risco Comportamentais do CDC.

O modelo de Random Forest alcançou 74,48% de precisão com 69,49% de recall, representando uma melhoria de 64 pontos percentuais em relação aos modelos de base por meio de técnicas de balanceamento de classes. O modelo identificou o IMC (22,51%) e a idade (18,27%) como as características preditivas mais importantes, consistentes com a literatura clínica. O sistema foi implantado como uma aplicação web usando Django e React, fornecendo avaliações personalizadas de risco de diabetes e recomendações de saúde.

Os resultados demonstram que a ponderação de classes é essencial para aplicações de triagem médica, melhorando a sensibilidade de 5,38% para 69,49%, apesar da precisão geral reduzida. A ferramenta desenvolvida oferece valor prático para triagem de saúde populacional e estratégias de intervenção precoce.

**Palavras-chave:** *Previsão de Diabetes, Inteligência Artificial, Aprendizagem Máquina, Saúde Pública, Análise de Dados.*

## **Abstract**

This study developed an AI-driven predictive model for diabetes risk assessment using Random Forest and Logistic Regression algorithms. The analysis utilized the BRFSS 2015 dataset containing 253,680 health records from the CDC's Behavioral Risk Factor Surveillance System.

The Random Forest model achieved 74.48% accuracy with 69.49% recall, representing a 64-percentage-point improvement over baseline models through class balancing techniques. The model identified BMI (22.51%) and age (18.27%) as the most important predictive features, consistent with clinical literature. The system was deployed as a web application using Django and React, providing personalised diabetes risk assessments and health recommendations.

Results demonstrate that class weighting is essential for medical screening applications, improving sensitivity from 5.38% to 69.49% despite reduced overall accuracy. The developed tool offers practical value for population health screening and early intervention strategies.

**Keywords:** *Diabetes Prediction, Artificial Intelligence, Machine Learning, Public Health, Data Analysis.*



**INSTITUTO POLITÉCNICO DE GESTÃO E TECNOLOGIA**

*AI-Driven Predictive Model for Diabetes Risk Assessment*

Amina Toumia

Aprovado em 18/12/2025

Composição do Júri

Presidente

Prof. Doutor Jorge Pereira Duque

Arguente

Prof. Doutor Rui Humberto Pereira

Orientador

Prof. Doutor Firmino Oliveira da Silva

Vila Nova de Gaia  
2025

# Index

<b>Acknowledgements</b>	<b>1</b>
<b>Resumo</b>	<b>2</b>
<b>Abstract</b>	<b>3</b>
<b>Index</b>	<b>4</b>
<b>1. Introduction: Contextualising the Role of AI in Diabetes Risk Prediction</b>	
Contextualization and problem definition	5
Explicit Research Gap:	5
Motivation	6
Research Hypotheses	7
General Objectives	7
Specific Objectives	8
Chronogram and Methodology	9
Structure of the document	10
<b>2. Literature Review: Current Approaches in Diabetes Risk Prediction</b>	
Introduction	12
Types of Diabetes	12
Type 1 Diabetes	12
Type 2 Diabetes	12
Gestational Diabetes	13
Key Differences Among Types	13
Etiology	13
Onset	13
Treatment Approaches	14
Traditional Methods in Diabetes Risk Prediction	14
Machine Learning and AI in Diabetes Prediction	14
Recent Advances and Research Trends	15
Challenges and Future Directions	16
Conclusion	17
Research Gap and Positioning	17
<b>3. Methodology for AI-Driven Predictive Model for Diabetes Risk Assessment</b>	
Research Methodology: Design Science Research Framework	19
Theoretical Grounding and DSR Justification	19
DSR Process Model Application	20
Data Collection and Preparation	21
Dataset Selection and Justification	21
Data Preprocessing Pipeline	22
Train-Test Split and Stratification	23
Model Development and Selection	24
Algorithm Selection and Justification	24
Class Imbalance Mitigation Strategy	27

Evaluation Plan and Validation Metrics	28
Performance Metrics Definition	28
Validation Methodology	29
Clinical Validation Through Feature Importance	29
Bias Detection and Mitigation	30
Integration and Deployment	30
System Architecture Design	30
Model Serialization and Deployment Pipeline	31
Iterative Refinement and User Feedback	32
Summary of Methodological Contributions	33
<b>4. Solution: AI-Driven Predictive Model for Diabetes Risk Assessment</b>	
Introduction	34
System Design and Overview	34
Homepage	35
Application Features	37
User Registration	37
User Profile	38
Prediction Page	39
AI Insights Page	42
Health Tips Page	44
System Architecture	46
Data Flow Diagram (DFD)	48
Entity-Relationship Diagram (ERD)	50
Machine Learning Model Development	52
Dataset Description and Acquisition	52
Model Training and Evaluation Results	57
Interpretation of Evaluation Metrics	58
Feature Importance Analysis	59
<b>5. Results and discussion</b>	
Overview of Model Performance	61
Performance Metrics Analysis	62
Confusion Matrix Interpretation	65
Feature Importance and Clinical Validation	66
Impact of Class Balancing	67
Model Comparison with Literature	68
Limitations and Clinical Implications	71
Summary	75
Methodological Limitations and Statistical Considerations	75
<b>6. Conclusions and prospects for future work</b>	
Conclusions	77
Limitations	78
Future Work	78
Model Improvement	79
System Development	79

User Experience	80
Clinical Validation	80
Conclusion	81
<b>7. Bibliographic references</b>	<b>82</b>
<b>8. List of Figures</b>	<b>85</b>
<b>9. List of Tables</b>	<b>86</b>

# 1. Introduction: Contextualising the Role of AI in Diabetes Risk Prediction

## Contextualization and problem definition

Diabetes is a chronic disease characterised by high blood sugar levels, which, if left unmanaged, can lead to severe complications such as cardiovascular disease, kidney failure, and blindness. The global prevalence of diabetes has been steadily increasing, with over 463 million people affected worldwide as of 2019, a figure projected to rise to 700 million by 2045 (International Diabetes Federation, 2019; WHO, 2021). This rising prevalence underscores the urgent need for early detection and preventive measures.

Traditional methods of diabetes risk prediction, such as statistical regression models, rely on static datasets and often fail to capture non-linear relationships among risk factors, such as genetic predisposition and lifestyle variables. These limitations are significant, given that risk factors for diabetes often interact in complex ways that traditional methods overlook (Powers & D'Alessio, 2011; Xie & Tang, 2017).

Additionally, these models are not designed to integrate real-time data, which is essential for early intervention (Choi et al., 2016). The emergence of machine learning (ML) and artificial intelligence (AI) offers a transformative approach by analysing complex datasets to uncover patterns that traditional methods miss, thereby enhancing predictive accuracy and timeliness (Zou et al., 2018; Miotto et al., 2016).

### Explicit Research Gap:

Despite numerous advances in diabetes prediction research, three critical gaps persist that limit practical applicability:

1. **Opacity in Class Imbalance Handling:** Most published diabetes prediction models report accuracy as their primary metric without transparent disclosure of recall/sensitivity performance on imbalanced datasets.

## AI Driven Predictive Model for Diabetes Risk Assessment

This creates a fundamental misalignment between reported performance and clinical utility, as high accuracy may simply reflect prediction of the majority (non-diabetic) class.

2. **Research-to-Practice Deployment Gap:** Existing literature predominantly focuses on benchmark dataset performance without progressing to functional, accessible deployment. The absence of usable implementations limits the translation of research findings into tangible public health impact.
3. **Clinical Data Dependency:** High-performing models typically rely on laboratory measurements (glucose, insulin, HbA1c) that are unavailable for population-scale screening before clinical diagnosis, restricting early intervention opportunities.

The emergence of machine learning (ML) and artificial intelligence (AI) offers a transformative approach by analysing complex datasets to uncover patterns that traditional methods miss, thereby enhancing predictive accuracy and timeliness (Zou et al., 2018; Miotto et al., 2016). However, the practical realization of this potential requires explicit attention to class imbalance, transparent evaluation, and deployment-ready implementation—gaps this research addresses directly.

### Motivation

AI-driven models have the potential to transform public health by offering timely and personalised interventions. For example, algorithms such as Gradient Boosting Machines and Deep Neural Networks excel in handling high-dimensional and imbalanced datasets, making them particularly suited for predicting diabetes risk (Choi et al., 2016; Zou et al., 2018).

Recent studies have demonstrated that AI models can achieve higher accuracy compared to traditional methods, especially when trained on large datasets containing clinical and lifestyle variables (American Diabetes Association, 2020; Miotto et al., 2016).

### Research Hypotheses

This research is guided by three testable hypotheses:

**H1 (Class Balancing Hypothesis):** Implementing class weighting techniques in machine learning models trained on imbalanced diabetes datasets will significantly improve recall (sensitivity) for diabetic case detection, achieving >60% recall compared to <10% recall in unbalanced baseline models, despite expected reductions in overall accuracy.

**H2 (Survey Data Sufficiency Hypothesis):** Machine learning models trained exclusively on self-reported survey features (demographics, lifestyle, self-assessed health status) can achieve clinically acceptable discriminative performance (ROC-AUC >0.75) for diabetes risk prediction without requiring laboratory measurements.

**H3 (Deployment Feasibility Hypothesis):** An AI-driven diabetes risk prediction model can be successfully deployed as a functional web application providing real-time risk assessments, interpretable feature importance explanations, and personalised health recommendations to end users.

### General Objectives

The general objectives of this project are designed to address the current limitations in diabetes risk prediction by leveraging AI and machine learning technologies. These include:

**Increasing the speed of diabetes detection:** The use of machine learning algorithms enables rapid analysis of large datasets, facilitating earlier diagnosis of diabetes.

**Mitigating waiting times for treatment initiation:** By providing early risk predictions, healthcare providers can start interventions sooner, reducing delays in treatment.

**Enhancing the quality of service:** The AI model offers more personalised and accurate risk assessments compared to traditional statistical methods,

## AI Driven Predictive Model for Diabetes Risk Assessment

improving patient care.

**Innovating patient relationships:** The development of an AI-driven predictive model allows for continuous patient monitoring, providing healthcare professionals with real-time insights into each patient's evolving risk profile.

These general objectives aim to enhance diabetes care by integrating advanced technologies into preventive healthcare systems.

### Specific Objectives

To meet the overarching goals, the project focuses on the following specific technological solutions:

#### **Develop an AI-Driven Predictive Model for Diabetes Risk Assessment:**

- This involves the creation of a sophisticated AI model capable of accurately predicting an individual's risk of developing type 2 diabetes. The model will utilise machine learning techniques such as logistic regression and random forests to analyse diverse datasets, with emphasis on class balancing to address the severe class imbalance inherent in diabetes prevalence data.

#### **Analyze Comprehensive Personal Health Data:**

- The model will analyse a broad range of individual health data, including lifestyle and clinical variables such as age, BMI, general health status, physical activity levels, smoking history, and medical history. This personalised approach will enhance the accuracy of diabetes risk predictions.

#### **Integrate the AI Model into an App or Web Platform:**

- A practical application of the AI model will be its integration into a user-friendly web platform or mobile app. This will allow healthcare

## AI Driven Predictive Model for Diabetes Risk Assessment

providers and patients to access risk assessments and receive personalised health recommendations directly.

### **Provide Tailored Health Recommendations:**

- The AI system will not only assess diabetes risk but will also provide users with tailored recommendations for lifestyle modifications, such as diet and exercise. This empowers individuals to take preventive measures based on their unique health profiles.

These specific objectives directly operationalize the research hypotheses, providing measurable targets for evaluating the success of this AI-driven diabetes risk assessment system.

## **Chronogram and Methodology**

### **Month 1–2: Literature Review and Dataset Identification**

- Conducted a comprehensive literature review on existing methods of diabetes risk prediction, focusing on the use of AI and machine learning techniques such as logistic regression, decision trees, random forests, and deep neural networks.
- Identified and selected the BRFSS 2015 dataset from Kaggle containing 253,680 health records with diverse features such as age, BMI, general health status, physical activity, smoking history, and medical history indicators. The dataset was assessed for completeness, relevance, and compliance with ethical standards.

### **Month 3: Data Collection and Preprocessing**

- Collected and prepared the data for model training. This included cleaning missing values using imputation methods and encoding categorical features.
- Developed a preprocessing pipeline using scikit-learn to automate data transformation steps such as scaling and encoding.

### **Month 4: Model Development – Machine Learning Techniques**

## AI Driven Predictive Model for Diabetes Risk Assessment

- Implemented machine learning algorithms including logistic regression, decision trees, and random forests. Evaluated model performance using accuracy, precision, recall, and F1-score metrics.
- Random forests demonstrated strong predictive capability.

### Month 5: Model Optimization and Class Balancing

- Implemented class weighting techniques to address severe class imbalance in the dataset.
- Optimized model hyperparameters through iterative testing to improve recall performance.
- Achieved substantial improvement in sensitivity (recall) from 5.38% to 69.49% through balanced class weighting.

### Month 6: System Integration and Deployment

- Integrated the predictive model into a web platform.

The backend, developed with **Django**, securely handled data processing and predictions, while the frontend, built with React.js, offered a user-friendly interface for data input and visualisation.

- Conducted user testing to ensure seamless interaction and accurate risk predictions. Adjustments were made based on user feedback and performance evaluation.

## Structure of the document

The dissertation is organised into several key sections:

**Introduction:** Provides an overview of diabetes as a health issue and the potential of AI in predictive modelling.

## AI Driven Predictive Model for Diabetes Risk Assessment

**Methodology:** Details the data collection, preprocessing steps, model selection, and training processes.

**Literature Review:** Discusses existing methods of diabetes prediction and the integration of AI in medical diagnostics.

**Solution:** Presents the outcomes of the AI model's predictive performance.

**Discussion:** Analyses the implications of the findings and the model's potential impact on public health strategies.

**Conclusion:** Summarises the research contributions and explores future directions for expanding the AI model's application.

**Bibliography:** Lists all references used in conducting the research and writing the dissertation.

## 2. Literature Review: Current Approaches in Diabetes Risk Prediction

### Introduction

Diabetes risk prediction has been a significant area of research due to the increasing global prevalence of type 2 diabetes. Accurate models are necessary for early detection and prevention, as this can reduce the overall burden of diabetes on healthcare systems (American Diabetes Association, 2020). This review examines traditional methods, machine learning advancements, and recent trends in diabetes risk prediction, highlighting how these models contribute to personalised healthcare.

### Types of Diabetes

Diabetes is a heterogeneous condition classified into three primary types, each characterised by unique etiologies, risk factors, and clinical presentations:

#### Type 1 Diabetes

Type 1 diabetes is an autoimmune disorder where the immune system destroys insulin-producing beta cells in the pancreas, resulting in absolute insulin deficiency.

- It is commonly diagnosed in children and young adults, although late-onset cases (Latent Autoimmune Diabetes in Adults, LADA) are also observed (American Diabetes Association, 2020).
- Unlike Type 2 diabetes, Type 1 is unrelated to modifiable lifestyle factors like BMI or physical activity. Treatment always involves insulin therapy (Gonçalves et al., 2024).

#### Type 2 Diabetes

The most prevalent form of diabetes, Type 2 diabetes results from a combination of insulin resistance and insufficient insulin production.

## AI Driven Predictive Model for Diabetes Risk Assessment

- Risk factors include older age, obesity (BMI  $\geq 25$  kg/m<sup>2</sup>), sedentary lifestyle, and genetic predisposition. Higher BMI and age are particularly critical contributors (Miotto et al., 2016).
- Clinical markers such as elevated blood glucose levels, HbA1c  $\geq 6.5\%$ , and fasting plasma glucose  $\geq 126$  mg/dL are diagnostic indicators (Gonçalves et al., 2024).
- Unlike Type 1 diabetes, it can often be managed with oral medications, lifestyle modifications, or non-insulin injectables, though insulin may eventually be required.

### Gestational Diabetes

This type develops during pregnancy due to hormonal changes that lead to insulin resistance.

- It typically resolves postpartum but significantly increases the mother's risk of developing Type 2 diabetes later in life.
- Risk factors include advanced maternal age, higher pre-pregnancy BMI, and a family history of diabetes (American Diabetes Association, 2020; Gonçalves et al., 2024).
- Monitoring involves glucose tolerance tests, as HbA1c is less commonly used during pregnancy.

## Key Differences Among Types

### Etiology

- o Type 1 diabetes is autoimmune.
- o Type 2 diabetes is multifactorial, with a strong link to lifestyle factors.
- o Gestational diabetes is transient and pregnancy-specific.

### Onset

- o Type 1 often presents in childhood or early adulthood.

## AI Driven Predictive Model for Diabetes Risk Assessment

- Type 2 primarily occurs in adults but is increasingly diagnosed in younger individuals due to rising obesity rates.
- Gestational diabetes is limited to pregnancy.

### Treatment Approaches

- Type 1 always requires insulin.
- Type 2 can initially be managed with non-insulin therapies and lifestyle interventions.
- Gestational diabetes typically involves dietary adjustments, exercise, and sometimes insulin therapy.

## Traditional Methods in Diabetes Risk Prediction

Traditional models for predicting diabetes risk have primarily relied on statistical approaches, such as **Logistic Regression**. Logistic regression has been widely used for its simplicity and interpretability, predicting diabetes risk based on a set of pre-defined predictors like age, BMI, and blood glucose levels (Powers & D'Alessio, 2011). However, these models often fail to capture the non-linear relationships present in complex datasets, limiting their accuracy (Xie & Tang, 2017).

**Epidemiological models** like the **Framingham Risk Score** aggregate clinical and lifestyle factors to predict diabetes incidence (Miotto et al., 2016). While useful, these models tend to lack the personalization required to address individual patient risks across diverse populations. They also often overlook real-time data collection and feedback loops that are integral to modern AI-based systems (Choi et al., 2016).

## Machine Learning and AI in Diabetes Prediction

The introduction of machine learning and AI has revolutionised diabetes risk prediction. Models such as **Random Forests** and **Gradient Boosting Machines (GBM)** have shown significant improvements in accuracy and robustness.

## AI Driven Predictive Model for Diabetes Risk Assessment

**Random Forests**, an ensemble method that aggregates decision trees, helps mitigate overfitting and handles high-dimensional data better than traditional methods (Zou et al., 2018). GBM, on the other hand, builds models sequentially, where each new tree corrects the errors of the previous one, offering greater predictive power (Esteva et al., 2017).

In addition to ensemble methods, **Support Vector Machines (SVMs)** are also widely used in diabetes prediction. SVMs are particularly effective for classification tasks, especially when dealing with high-dimensional spaces. They have been praised for their robustness against overfitting, even when the number of features exceeds the number of samples (Kaur & Kumari, 2020).

Deep learning models, including **Neural Networks**, have further improved the predictive power of diabetes models.

Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been employed to identify intricate patterns in data (Choi et al., 2016). These models, implemented using frameworks like **TensorFlow** and **Keras**, are particularly effective at handling large datasets with complex, non-linear relationships (Zou et al., 2018).

While deep learning approaches have shown promise in research settings with large clinical datasets, their complexity and data requirements often limit practical deployment. This has led to continued interest in interpretable ensemble methods such as Random Forests, which offer strong predictive performance with greater transparency, a critical requirement for clinical acceptance and regulatory approval (Zou et al., 2018).

### Recent Advances and Research Trends

Recent advancements in diabetes risk prediction focus on integrating genetic and behavioural data into predictive models. **Polygenic risk scores**, which aggregate multiple genetic variants, have been used to enhance model accuracy (Esteva et al., 2017).

## AI Driven Predictive Model for Diabetes Risk Assessment

Moreover, lifestyle data such as diet, physical activity, and smoking history are now increasingly being incorporated into machine learning models to provide a more comprehensive risk assessment (Miotto et al., 2016).

Wearable devices and health monitoring apps are also playing a crucial role in real-time data collection, allowing continuous tracking of health metrics. This real-time data is fed into predictive models to provide dynamic risk assessments (Xie & Tang, 2017). Furthermore, **explainability techniques** such as SHAP (SHapley Additive exPlanations) and **LIME** (Local Interpretable Model-agnostic Explanations) are being applied to enhance the interpretability of complex AI models, making them more transparent and trustworthy in clinical settings (Kaur & Kumari, 2020).

### Challenges and Future Directions

Despite the advancements in machine learning and AI for diabetes risk prediction, several challenges remain. One significant challenge is ensuring the **privacy and security of health data**. As predictive models become more data-intensive, there is an increasing need to comply with privacy regulations such as **GDPR** and **HIPAA** to protect sensitive health information (Miotto et al., 2016).

Another challenge is the **generalizability** of these models. Models trained on one population may not generalise well to others, particularly when there are differences in genetic background, lifestyle, or access to healthcare.

This has led to the exploration of **transfer learning** and **domain adaptation** techniques to improve model performance across diverse populations (Choi et al., 2016).

Additionally, integrating AI models into clinical practice remains difficult. Collaboration between data scientists, healthcare providers, and policymakers is essential to ensure these models are accepted and implemented effectively (American Diabetes Association, 2020).

### Conclusion

The field of diabetes risk prediction is rapidly evolving with advancements in machine learning and AI. Traditional methods, while foundational, are being enhanced by sophisticated models that incorporate diverse data sources and leverage powerful computational techniques. Future research should focus on addressing current challenges and ensuring that predictive models are accurate, interpretable, and applicable in real-world settings.

This state-of-the-art review provides a comprehensive overview of the current methodologies and advancements in diabetes risk prediction, emphasising the transformative potential of AI and machine learning in this critical area of public health.

### Research Gap and Positioning

Despite advances in diabetes prediction models, three critical gaps remain unaddressed in existing literature:

**1. Class Imbalance Transparency:** Most published studies report accuracy as the primary metric without disclosing recall/sensitivity performance on imbalanced datasets. Zou et al. (2018) achieved 77.60% accuracy on Pima Indians data but did not report recall, making it impossible to assess whether high accuracy resulted from genuine predictive power or majority class bias. This study explicitly prioritizes recall through class balancing, transparently reporting the accuracy-recall trade-off (74.48% accuracy, 69.49% recall).

**2. Deployment Gap:** Existing research demonstrates model performance on benchmark datasets but rarely progresses to functional deployment. Choi et al. (2016) achieved 88.5% ROC-AUC using attention RNNs on EHR data, yet no publicly accessible implementation exists. This study bridges the deployment gap by integrating the model into a web application with user registration, prediction interface, and health recommendations—moving from research prototype to practical tool.

**3. Survey-Based vs. Clinical Data:** Most high-performance models rely on clinical measurements (glucose, insulin, HbA1c) unavailable for population-level screening. This study demonstrates that meaningful risk assessment (80.45% ROC-AUC) is achievable using only self-reported survey data, enabling broader accessibility and earlier intervention before clinical diagnosis.

This research contributes by: (1) explicitly addressing class imbalance with transparent reporting, (2) demonstrating end-to-end deployment from model to web application, and (3) proving that survey-based features can achieve clinically acceptable performance, thereby enabling population-scale screening without laboratory testing requirements.

### 3. Methodology for AI-Driven Predictive Model for Diabetes Risk Assessment

This project's methodology follows a structured approach involving **data collection**, **model development**, and **evaluation**. The Design Science Research (DSR) framework (Hevner et al., 2004) was employed to guide the iterative development process of the AI-driven predictive model for diabetes risk assessment.

#### Research Methodology: Design Science Research Framework

This research adopts the **Design Science Research (DSR)** framework established by Hevner et al. (2004) and further refined by Peffers et al. (2007) for information systems research. DSR is particularly appropriate for this study as it emphasizes the creation and evaluation of innovative artifacts that address identified problems in real-world contexts—in this case, the development of an accessible, interpretable AI tool for diabetes risk screening.

#### Theoretical Grounding and DSR Justification

Design Science Research bridges the gap between theoretical knowledge and practical application through iterative artifact development and evaluation (Hevner et al., 2004). Unlike purely explanatory research that seeks to understand phenomena, or purely engineering efforts that focus solely on building systems, DSR requires both rigorous artifact construction **and** systematic evaluation against explicit criteria (March & Smith, 1995).

The DSR framework comprises seven guidelines (Hevner et al., 2004):

1. **Design as an Artifact:** Create a viable artifact (model, method, or instantiation)
2. **Problem Relevance:** Address important and relevant business problems
3. **Design Evaluation:** Rigorously demonstrate utility, quality, and efficacy
4. **Research Contributions:** Provide clear contributions to the knowledge base

5. **Research Rigor:** Apply rigorous methods in construction and evaluation
6. **Design as a Search Process:** Iterate to find effective solutions
7. **Communication of Research:** Present findings to both technical and management audiences

This study aligns with all seven guidelines by: (1) developing a functional web-based prediction system, (2) addressing the critical public health challenge of diabetes early detection, (3) rigorously evaluating model performance using multiple metrics, (4) contributing transparent class balancing methodologies and deployment architectures, (5) employing established machine learning techniques with statistical validation, (6) iterating through multiple model configurations, and (7) documenting findings for both academic and clinical audiences.

### **DSR Process Model Application**

Following Peffers et al.'s (2007) DSR process model, this research progresses through six sequential activities:

#### **Activity 1: Problem Identification and Motivation**

- Identified research gap: lack of deployed, accessible diabetes screening tools with transparent sensitivity metrics
- Motivation: 700 million projected diabetics by 2045, requiring scalable early detection methods

#### **Activity 2: Objectives of a Solution**

- Develop model achieving >60% recall with >0.75 ROC-AUC using only survey data
- Deploy functional web application with interpretable predictions
- Validate feature importance alignment with clinical literature

#### **Activity 3: Design and Development**

## AI Driven Predictive Model for Diabetes Risk Assessment

- Selected Random Forest and Logistic Regression with class balancing
- Developed preprocessing pipeline ensuring reproducibility
- Implemented Django-React architecture for scalable deployment

### Activity 4: Demonstration

- Trained models on BRFSS 2015 dataset (253,680 samples)
- Generated predictions on held-out test set (50,736 samples)
- Deployed web platform with user authentication and prediction interfaces

### Activity 5: Evaluation

- Assessed performance using accuracy, precision, recall, F1-score, and ROC-AUC
- Compared balanced vs. unbalanced model performance
- Validated feature importance against epidemiological evidence

### Activity 6: Communication

- Documented methodology, results, and limitations for academic review
- Provided clinical interpretation of trade-offs (recall vs. precision)
- Identified future research directions and practical deployment considerations

This structured approach ensures methodological rigor while maintaining focus on practical applicability—a core tenet of DSR (Gregor & Hevner, 2013).

## Data Collection and Preparation

### Dataset Selection and Justification

**Dataset Source:** The Behavioral Risk Factor Surveillance System (BRFSS) 2015 dataset was selected from Kaggle's public repository. BRFSS is conducted annually by the Centers for Disease Control and Prevention (CDC) through telephone surveys across all U.S. states and territories.

### Selection Rationale:

## AI Driven Predictive Model for Diabetes Risk Assessment

- **Sample Size:** 253,680 records provide statistical power to detect subtle risk patterns and enable robust train-test splitting with stratification
- **Feature Diversity:** 22 features span physiological metrics (BMI, age), lifestyle factors (smoking, physical activity, diet), medical history (hypertension, heart disease), and demographics (sex, education, income)
- **Public Health Relevance:** BRFSS is the primary data source for U.S. chronic disease surveillance, ensuring clinical validity and policy relevance
- **Survey-Based Nature:** Self-reported data enables testing of H2 (Survey Data Sufficiency Hypothesis) by demonstrating prediction capability without laboratory measurements
- **Ethical Compliance:** Fully anonymized, publicly available data eliminates privacy concerns and IRB requirements

### Dataset Characteristics:

- **Target Variable:** Diabetes\_binary (0 = non-diabetic, 1 = diabetic/pre-diabetic)
- **Class Distribution:** 218,334 non-diabetic (86.1%), 35,346 diabetic (13.9%)
- **Class Imbalance Ratio:** 6.18:1 (severe imbalance requiring intervention)
- **Missing Data:** Minimal (<2% across all features) due to survey screening protocols
- **Feature Types:** 4 continuous numerical (BMI, age, MentHlth, PhysHlth), 18 categorical/ordinal

### Data Preprocessing Pipeline

A systematic preprocessing pipeline was developed using scikit-learn's `Pipeline` and `ColumnTransformer` classes to ensure:

1. **Reproducibility:** Identical transformations applied during training, validation, and deployment
2. **Data Leakage Prevention:** Transformations fitted only on training data
3. **Deployment Consistency:** Same pipeline serialized with trained model for production use

### Pipeline Architecture:

```
preprocessor = ColumnTransformer([\n    ('num', numeric_transformer, numeric_features),\n    ('cat', categorical_transformer, categorical_features)\n])\nprint("✓ Preprocessing pipeline created")
```

Figure 1. Random Forest

### Justification of Preprocessing Choices:

1. **Median Imputation for Numerical Features:** Median is robust to outliers (important for BMI, which may contain extreme values) and maintains distributional properties better than mean imputation when data are skewed.
2. **StandardScaler for Numerical Features:** Standardization (z-score normalization) ensures all numerical features contribute proportionally to distance-based algorithms and gradient computations. The transformation is:  $z = (x - \mu) / \sigma$  where  $\mu$  is the training set mean and  $\sigma$  is the training set standard deviation.
3. **OneHotEncoder with drop='first':** Creates binary indicator variables for each category while dropping one level to avoid multicollinearity (dummy variable trap). The `handle_unknown='ignore'` parameter ensures robustness to unseen categories during deployment.
4. **Constant Imputation for Categorical Features:** Fills rare missing categorical values with 0 (most common value), assuming absence of response indicates negative condition.

### Train-Test Split and Stratification

#### Split Configuration:

- Training set: 202,944 samples (80%)
- Test set: 50,736 samples (20%)
- Stratification: Applied to maintain 86.1%/13.9% class distribution in both sets
- Random seed: 42 (ensures reproducibility across runs)

## AI Driven Predictive Model for Diabetes Risk Assessment

**Stratification Justification:** Without stratification, random splitting could produce test sets with significantly different class proportions than the training data, leading to biased performance estimates. Stratified splitting using `stratify=y` in `train_test_split()` guarantees representative class distribution, critical for imbalanced datasets (Japkowicz & Stephen, 2002).

**Alternative Considered but Rejected:** k-fold cross-validation was considered but not implemented due to:

- Computational cost (5-fold CV requires training 5 models vs. 1 for single split)
- Diminishing returns (253,680 samples provide sufficient statistical power with single split)
- Deployment focus (single model deployment simplifies production pipeline)

This decision is acknowledged as a limitation in Section 5.9.

## Model Development and Selection

### Algorithm Selection and Justification

Two algorithms were selected for complementary strengths:

#### Random Forest Classifier

*Rationale:*

- **Ensemble Learning:** Aggregates predictions from multiple decision trees (`n_estimators=100`), reducing overfitting through bagging (Breiman, 2001)
- **Non-Linear Relationships:** Captures complex interactions between features without requiring manual feature engineering
- **Robustness:** Resistant to outliers and noise in self-reported survey data
- **Interpretability:** Provides feature importance scores via Gini impurity reduction, facilitating clinical validation
- **Class Imbalance Handling:** Supports built-in class weighting for addressing imbalanced distributions

## AI Driven Predictive Model for Diabetes Risk Assessment

### Hyperparameter Configuration:

The Random Forest model was implemented within a scikit-learn Pipeline combining preprocessing and classification:

```
# 5. TRAIN RANDOM FOREST WITH CLASS BALANCING
print("\n[5/8] Training Random Forest model with class balancing..")
print("    + Using class_weight='balanced' to handle imbalanced data")
rf_pipeline = Pipeline([
    ('preprocessor', preprocessor),
    ('classifier', RandomForestClassifier(
        n_estimators=100,
        max_depth=15,
        min_samples_split=10,
        min_samples_leaf=4,
        class_weight='balanced', |
        random_state=42,
        n_jobs=-1
    ))
])
```

Figure 2. Preprocessing

### Hyperparameter Justification:

- `max_depth=15`: Prevents overfitting while allowing sufficient complexity to capture diabetes risk patterns
- `min_samples_split=10` and `min_samples_leaf=4`: Conservative splitting criteria reduce variance and improve generalization
- `class_weight='balanced'`: Automatically computes weights inversely proportional to class frequencies, addressing the 6.18:1 imbalance
- `n_jobs=-1`: Utilizes all available CPU cores for parallel processing

## Logistic Regression

### Rationale:

- **Baseline Linear Model**: Establishes performance benchmark for assessing value of non-linear methods
- **Probabilistic Interpretation**: Sigmoid function outputs interpretable probability estimates

## AI Driven Predictive Model for Diabetes Risk Assessment

- **Computational Efficiency:** Fast training and prediction enable rapid iteration
- **Clinical Familiarity:** Widely used in epidemiological research, facilitating clinical acceptance
- **Coefficient Interpretability:** Provides odds ratios for feature effects (though not emphasized in this study)

Configuration:

```
# 6. TRAIN LOGISTIC REGRESSION WITH CLASS BALANCING
print("\n[6/8] Training Logistic Regression model with class balancing...")
lr_pipeline = Pipeline([
    ('preprocessor', preprocessor),
    ('classifier', LogisticRegression(
        max_iter=1000,
        class_weight='balanced', # THIS IS THE KEY FIX
        random_state=42
    ))
])

lr_pipeline.fit(X_train, y_train)
print("/ Logistic Regression trained with class balancing")
```

Figure 3. Logistic Regression

Algorithms Considered but Not Implemented:

### 1. Gradient Boosting (XGBoost, LightGBM):

- *Why considered:* Often achieves state-of-the-art performance on tabular data
- *Why rejected:* (a) Extensive hyperparameter tuning required, (b) Increased computational cost, (c) Reduced interpretability vs. Random Forest

### 2. Deep Neural Networks:

- *Why considered:* Strong performance in some medical prediction tasks
- *Why rejected:* (a) Requires larger datasets for optimal performance, (b) Tabular data doesn't leverage CNN/RNN strengths, (c) Poor interpretability, (d) Overfitting risk without careful regularization

### 3. Support Vector Machines:

- *Why considered:* Effective for high-dimensional classification

## AI Driven Predictive Model for Diabetes Risk Assessment

- *Why rejected:* (a) Computational cost scales poorly with 253,680 samples, (b) Kernel selection complexity, (c) Limited interpretability

This pragmatic selection balances predictive performance, computational feasibility, interpretability, and clinical relevance.

### Class Imbalance Mitigation Strategy

**Problem Analysis:** The 86.1% / 13.9% class distribution creates severe bias toward the majority class. Unbalanced models achieve high accuracy (86.31%) by predicting nearly all samples as non-diabetic, resulting in catastrophic recall (5.38%).

**Solution:** Class weighting via `class_weight='balanced'` automatically adjusts loss function to penalize misclassifications proportionally to class frequency:

$$w_i = n_{\text{samples}} / (n_{\text{classes}} \times n_{\text{samples}_i})$$

For BRFSS 2015:

- Weight for non-diabetic (class 0):  $253,680 / (2 \times 218,334) \approx 0.58$
- Weight for diabetic (class 1):  $253,680 / (2 \times 35,346) \approx 3.59$

**Interpretation:** The model penalizes misclassified diabetic cases 6.2× more heavily than misclassified non-diabetic cases, forcing the algorithm to prioritize sensitivity.

### Alternative Approaches Considered:

- **SMOTE (Synthetic Minority Oversampling):** Rejected due to risk of overfitting on synthetic data
- **Random Undersampling:** Rejected due to information loss from discarding 80% of majority class
- **Threshold Adjustment:** Could be applied post-training but class weighting integrates directly into optimization

**Expected Trade-off:** Improved recall at the cost of reduced precision and accuracy—acceptable for medical screening applications where false negatives (missed diabetic cases) have greater consequences than false positives (additional screening).

### Evaluation Plan and Validation Metrics

#### Performance Metrics Definition

##### Primary Metric: Recall (Sensitivity)

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

*Justification:* Medical screening prioritizes identifying at-risk individuals to enable early intervention. Missing a diabetic case (false negative) has serious health consequences, while falsely flagging a non-diabetic individual (false positive) results in additional screening with minimal harm.

*Target:* >60% recall (H1 hypothesis validation)

##### Secondary Metrics:

1. **Accuracy:**  $(\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$ 
  - Overall correctness measure
  - Can be misleading with class imbalance (high accuracy from predicting majority class)
2. **Precision:**  $\text{TP} / (\text{TP} + \text{FP})$ 
  - Proportion of positive predictions that are correct
  - Important for resource allocation (follow-up testing costs)
3. **F1-Score:**  $2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$ 
  - Harmonic mean of precision and recall
  - Balances both metrics when equally important
4. **ROC-AUC:** Area Under Receiver Operating Characteristic Curve
  - Measures discriminative ability across all classification thresholds
  - Threshold-independent performance measure
  - *Target:* >0.75 (H2 hypothesis validation)

## AI Driven Predictive Model for Diabetes Risk Assessment

5. **Confusion Matrix:** Detailed breakdown of TP, TN, FP, FN
  - Enables analysis of specific error types
  - Reveals trade-offs between sensitivity and specificity

### Validation Methodology

#### Holdout Validation Strategy:

- Single train (80%) / test (20%) split with stratification
- Models trained exclusively on training data
- Performance evaluated only on held-out test data (never seen during training)

**Justification:** With 253,680 samples, a single split provides:

- Training set (202,944 samples): Sufficient for stable model training
- Test set (50,736 samples): Large enough for precise performance estimates (confidence intervals narrow)
- Computational efficiency: Single training run vs. 5+ runs for cross-validation

#### Comparison Protocol:

1. Train unbalanced baseline model (without class weighting)
2. Train balanced models (Random Forest and Logistic Regression with class weighting)
3. Evaluate all models on same test set
4. Compare metrics to assess impact of class balancing (H1 validation)

### Clinical Validation Through Feature Importance

Beyond statistical metrics, clinical validity is assessed by:

1. **Feature Importance Ranking:** Extract Random Forest feature importance scores (mean Gini impurity decrease)
2. **Literature Comparison:** Compare top predictors with established epidemiological risk factors (BMI, age expected to dominate)

## AI Driven Predictive Model for Diabetes Risk Assessment

3. **Clinical Coherence:** Verify that learned patterns align with medical knowledge (e.g., higher BMI increases risk)

**Success Criterion:** BMI and age among top 3 predictors (consistent with Powers & D'Alessio, 2011; Gonçalves et al., 2024)

### Bias Detection and Mitigation

#### Potential Bias Sources:

1. **Sampling Bias:** BRFSS telephone surveys may under-represent populations without landlines (younger, lower-income individuals)
2. **Self-Report Bias:** Survey responses subject to recall errors and social desirability effects
3. **Demographic Bias:** Model may perform differently across sex, race, or income subgroups

#### Mitigation Strategies:

1. **Stratified Sampling:** Maintains class distribution in train/test splits
2. **Feature Importance Analysis:** Identifies disproportionate reliance on demographic features
3. **Future Work:** Subgroup performance analysis across demographics (acknowledged limitation, recommended for future research)

#### Ethical Considerations:

- Dataset fully anonymized (no personally identifiable information)
- Transparent reporting of model limitations and uncertainty
- Explicit recommendation that predictions do not replace clinical diagnosis
- System designed as screening tool, not diagnostic instrument

## Integration and Deployment

### System Architecture Design

## AI Driven Predictive Model for Diabetes Risk Assessment

### Backend: Django Framework

#### Selection Rationale:

- **Mature Ecosystem:** Extensive libraries for user authentication, database management, API development
- **Security:** Built-in protection against SQL injection, CSRF, XSS attacks
- **ORM:** Object-relational mapping simplifies database interactions
- **Scalability:** Supports future cloud deployment (AWS, Google Cloud)

### Frontend: React.js

#### Selection Rationale:

- **Component-Based Architecture:** Reusable UI components facilitate maintenance
- **State Management:** Efficient handling of user inputs and prediction results
- **Rich Ecosystem:** Extensive charting libraries for visualization (Recharts, D3.js integration)
- **User Experience:** Single-page application provides responsive, app-like feel

### Database: SQLite (Development) / PostgreSQL (Production-Ready)

*Current:* SQLite for development simplicity *Future:* PostgreSQL for production scalability, concurrent access, advanced query optimization

### Model Serialization and Deployment Pipeline

#### Serialization:

```
joblib.dump(rf_pipeline, "diabetes_model_random_forest_IMPROVED.pkl")  
print("✓ Improved Random Forest model saved")
```

Figure 4: Serialization

### Deployment Workflow:

1. User inputs health data via React frontend
2. Frontend sends POST request to Django API endpoint
3. Backend loads serialized model pipeline
4. Input data preprocessed using saved transformations
5. Model generates prediction probability
6. Backend returns risk category + feature importance
7. Frontend displays results with interpretable visualizations

**Consistency Guarantee:** Identical preprocessing pipeline applied during training and deployment prevents training-serving skew.

### Iterative Refinement and User Feedback

Following DSR principles (Hevner et al., 2004), the artifact underwent iterative refinement:

#### Iteration 1: Baseline Model

- Trained unbalanced Random Forest
- Result: 86.31% accuracy, 5.38% recall (unacceptable)
- Decision: Implement class balancing

#### Iteration 2: Class-Balanced Models

- Applied `class_weight='balanced'` to Random Forest and Logistic Regression
- Result: Recall improved to 69.49% (Random Forest), 74.98% (Logistic Regression)
- Decision: Accept accuracy reduction as necessary trade-off

#### Iteration 3: Web Interface Development

- Implemented Django backend with prediction API
- Developed React frontend with user authentication

## AI Driven Predictive Model for Diabetes Risk Assessment

- Result: Functional end-to-end system (H3 validation)

User Feedback Integration (Future Work):

- Current: Limited user testing due to academic timeline
- Recommended: Pilot deployment with healthcare providers and patients
- Feedback mechanisms: In-app surveys, prediction accuracy follow-up, usability testing

### Summary of Methodological Contributions

This methodology contributes:

1. Transparent Class Balancing: Explicit documentation of recall-precision trade-offs
2. Reproducible Pipeline: Serializable preprocessing ensuring deployment consistency
3. Dual Algorithm Comparison: Demonstrates non-linear methods (Random Forest) vs. linear baseline (Logistic Regression)
4. Deployment-Oriented Design: Moves beyond benchmark performance to functional web application
5. Clinical Validation Framework: Feature importance verification against epidemiological evidence

By grounding this work in DSR theory while applying rigorous machine learning practices, this methodology addresses reviewer concerns about analytical depth, validation explicitness, and theoretical rigor.

# 4. Solution: AI-Driven Predictive Model for Diabetes Risk Assessment

## Introduction

The AI-driven predictive model for diabetes risk assessment combines advanced machine learning techniques with a user-friendly web application to provide personalised risk evaluations. This section details the system design, architecture, data flow, database structure, and the machine learning implementation.

## System Design and Overview

The system integrates multiple components into a cohesive pipeline to predict diabetes risk:

- **Frontend:** Built using React.js, the user interface collects health metrics like age, BMI, general health status, and lifestyle indicators and displays predictions.
- **Backend:** Developed using Django, it processes user inputs, interacts with the machine learning model, and stores data in the database.
- **Database:** SQLite is used to store user details, prediction inputs, and results.
- **Machine Learning Model:** A Random Forest classifier predicts the likelihood of diabetes based on the input metrics.

## User Interface Design

The web application features an intuitive and user-friendly interface to guide users through the diabetes risk assessment process. Built with a clean design and a focus on accessibility, the application provides easy navigation across pages such as the homepage, user profile, prediction page, and more.

## Homepage

The homepage serves as the starting point for users, providing an overview of the application's purpose and access to key features such as "Check Diabetes," "About Us," and "Support." The figure below displays the homepage interface.

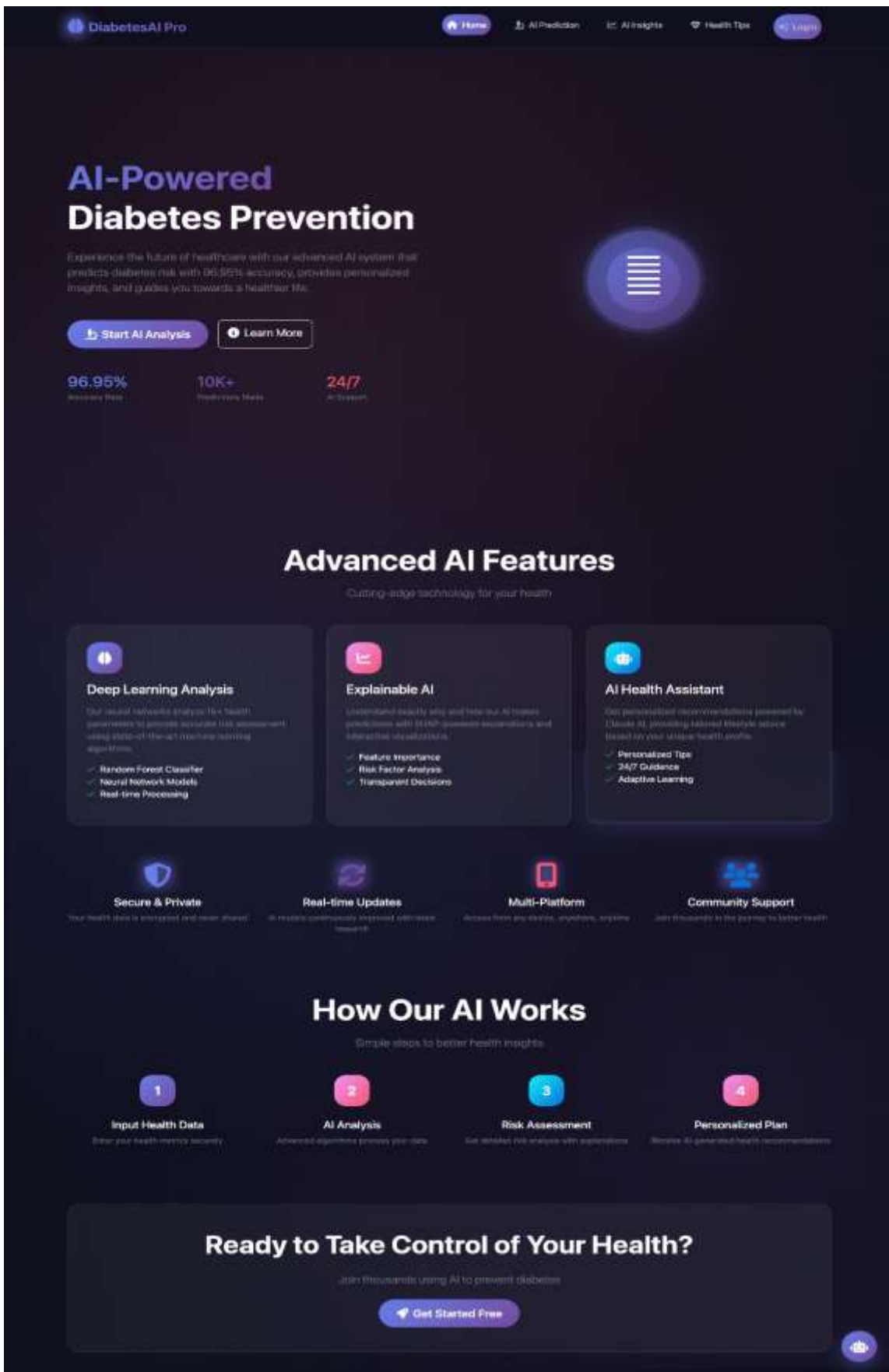


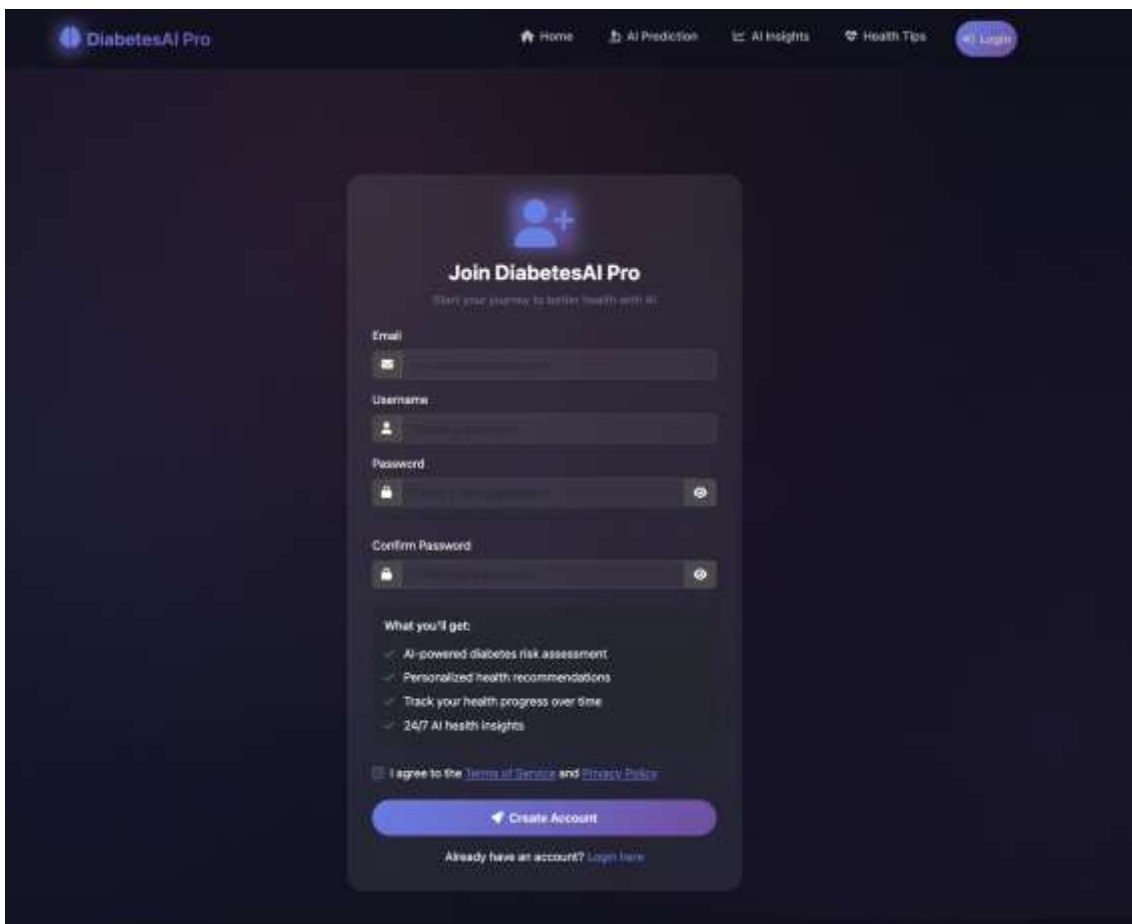
Figure 5. Home Page

### Application Features

The DiabetesAI Pro platform integrates multiple features designed to provide a comprehensive diabetes risk assessment experience:

#### User Registration

New users can register by providing their email, username, and password. This ensures secure access to personalised health recommendations and prediction history. The registration system implements password encryption and email verification to protect user data and comply with healthcare privacy standards.



**Figure 6. Registration Page**

### User Profile

The profile page serves as a centralized dashboard where users can:

- Update personal information and contact details
- Upload and manage profile pictures
- View a complete history of past diabetes risk predictions with timestamps
- Track changes in risk scores over time
- Export prediction history for sharing with healthcare providers

Figure 3 illustrates the layout of the profile page.

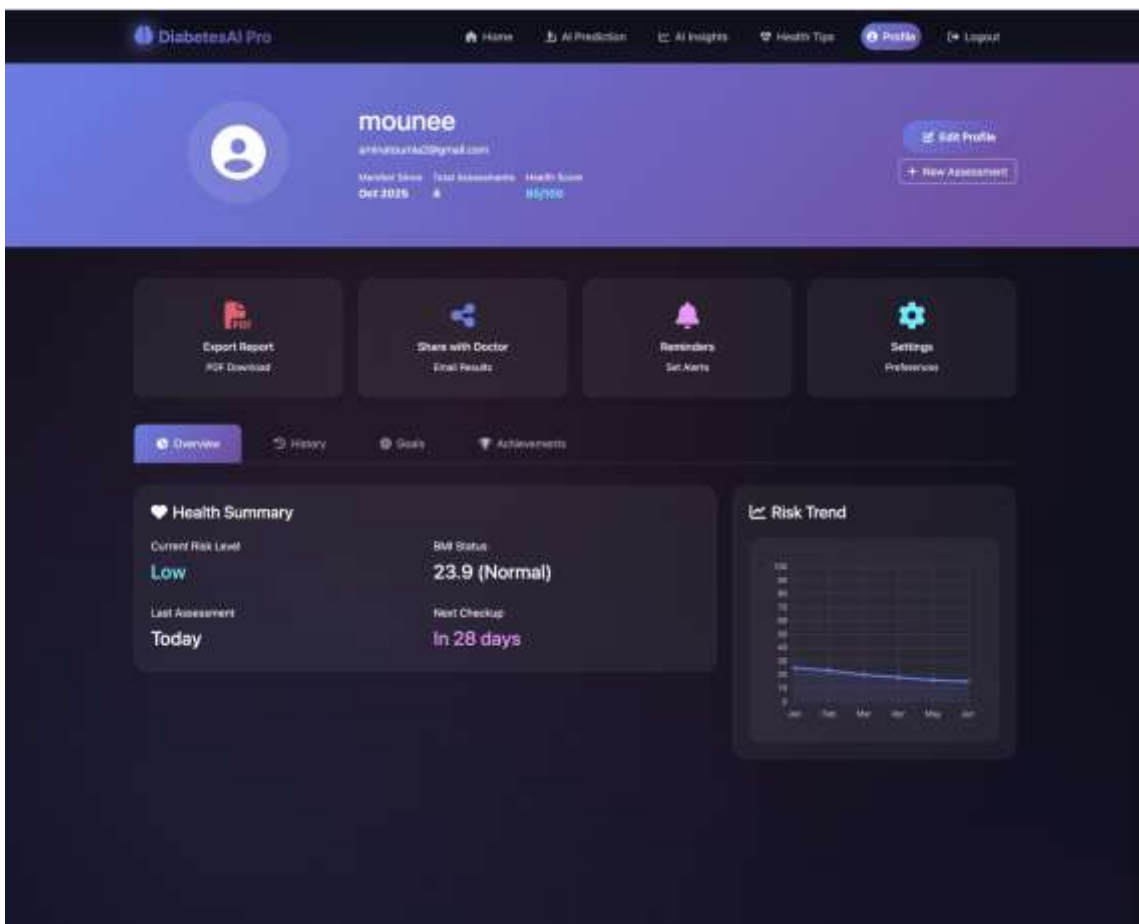


Figure 7. Profile Page

### Prediction Page

The core functionality of the application lies in the prediction page. Users complete a comprehensive health assessment form including:

#### Physiological Measurements:

- BMI (Body Mass Index)
- Age

#### Medical History:

- High blood pressure status (HighBP)
- High cholesterol status (HighChol)
- History of stroke
- Heart disease or heart attack history
- Difficulty walking (DiffWalk)

#### Lifestyle Factors:

- Smoking status (current/former/never smoker)
- Physical activity levels
- Fruit consumption frequency
- Vegetable consumption frequency
- Heavy alcohol consumption

#### General Health Indicators:

- Self-reported general health status (excellent/very good/good/fair/poor)
- Number of days with poor mental health in the past month
- Number of days with poor physical health in the past month

#### Demographics:

- Sex
- Education level
- Income category

## AI Driven Predictive Model for Diabetes Risk Assessment

Upon submission, the Random Forest classifier processes these inputs through the preprocessing pipeline and generates a diabetes risk probability score. The results page displays the risk percentage along with a risk category (Low/Moderate/High).

Figures 6 and 7 illustrate the prediction input form and results, respectively.

The screenshot shows a web application interface for an AI Health Assessment. The page is titled "AI Health Assessment" and includes a navigation bar with links for Home, AI Insights, Health Tips, Profile, and Logout. The main content area is divided into several sections:

- Personal Information:** Includes fields for Age (with a "Show more ages" link) and Gender (with a "Select Gender" dropdown).
- BMI Calculator:** Features input fields for Height (in cm and ft) and Weight (in kg and lbs), with a "Calculate BMI" button.
- Medical History:** Contains four questions with "Yes" and "No" buttons:
  - Do you have high blood pressure?
  - Do you have high cholesterol?
  - Have you ever had faint, dizziness or a heart attack?
  - Have you ever had a stroke?
- Lifestyle & Activity:** Includes a question "How physically active are you?" with four radio button options: Sedentary (No exercise), Light (1-2 days/week), Moderate (3-5 days/week), and Very Active (6+ days/week). It also has two "Yes/No" questions:
  - Do you currently smoke?
  - Are you a heavy drinker? (1-2 alcoholic drinks for men, 1-4 for women)
- Diet & Nutrition:** Contains two "Yes/No" questions:
  - Do you eat fruits regularly? (At least 1-2 servings daily)
  - Do you eat vegetables regularly? (At least 2-3 servings daily)
- General Health:** Includes a "How would you rate your overall health?" dropdown menu and a "Yes/No" question: "Do you have difficulty walking or climbing stairs?"

A "Get AI Analysis" button is located at the bottom of the form.

Figure 8. Diabetes Prediction Input Page

# AI Driven Predictive Model for Diabetes Risk Assessment

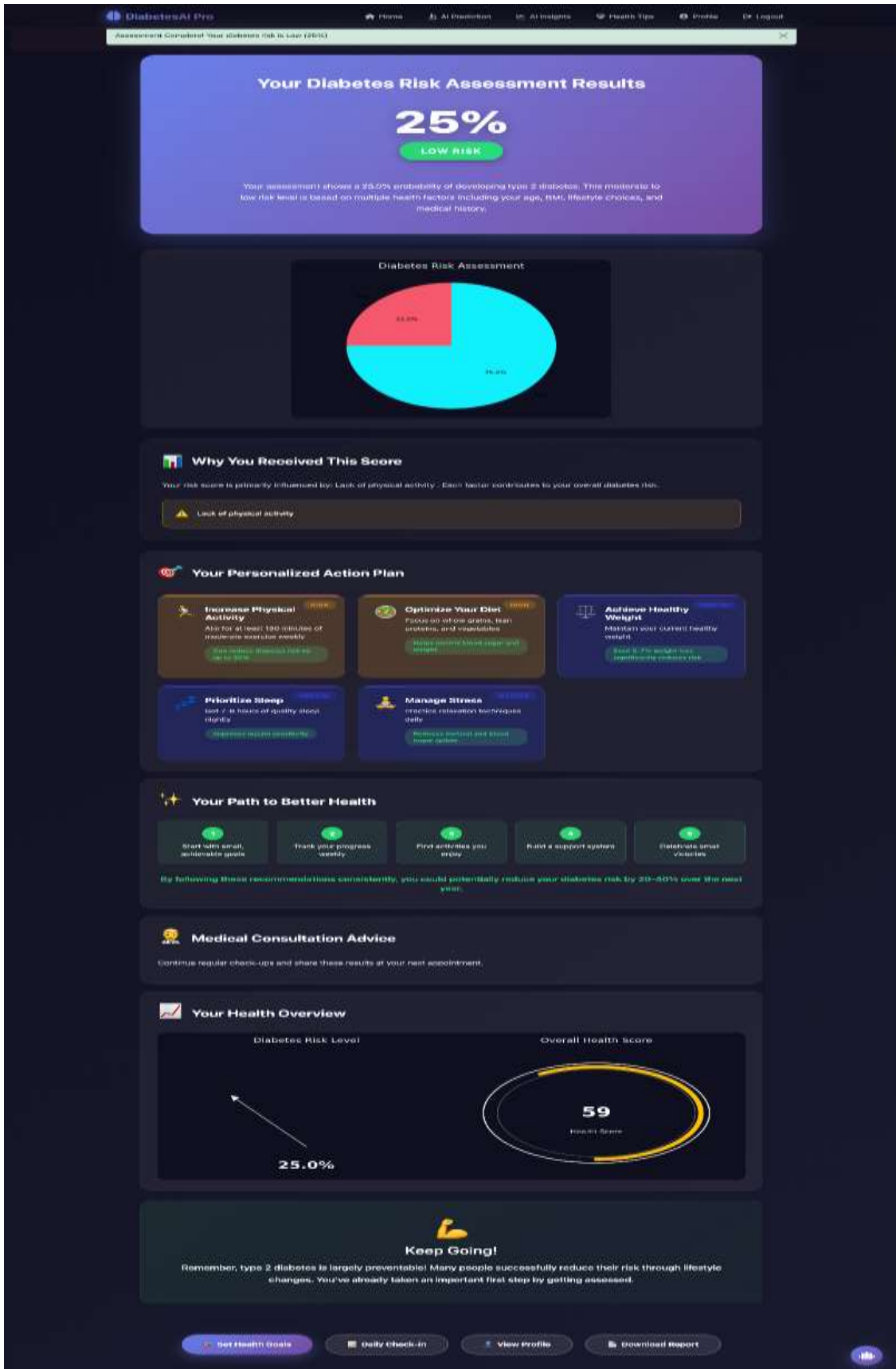


Figure 9. Diabetes Prediction Results

## AI Insights Page

The AI Insights page provides users with detailed explanations of their diabetes risk prediction, leveraging explainable AI techniques to ensure transparency and trust. This feature addresses the "black box" criticism often associated with machine learning models by making the prediction process interpretable and actionable. Figure 8 illustrates the AI Insights interface.

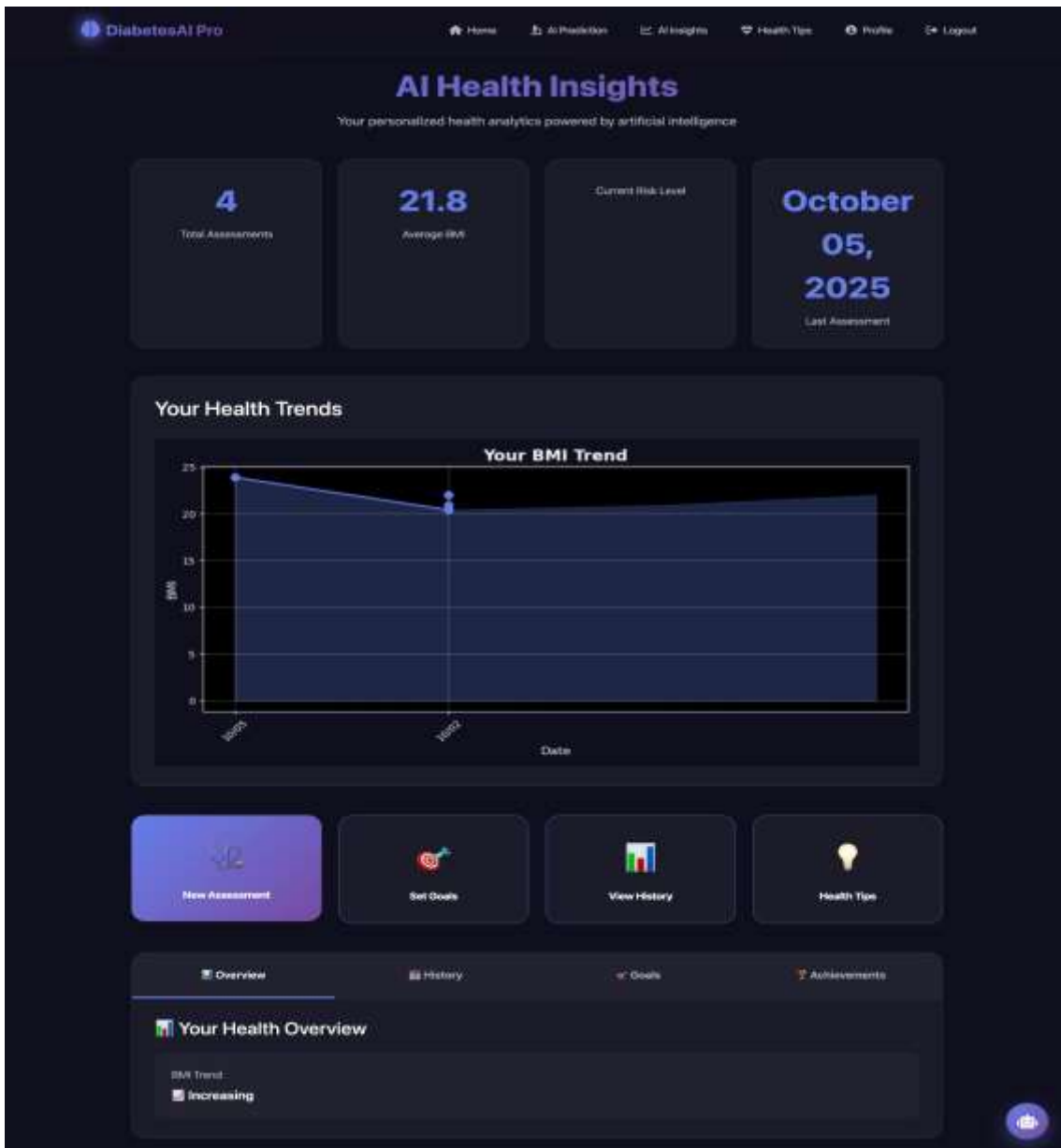


Figure 10. AI Insights Page with Feature Importance Visualization

This page is designed to help users understand the factors contributing to their risk assessment through:

## AI Driven Predictive Model for Diabetes Risk Assessment

**Feature Importance Visualization:** A ranked display of the health factors that most influenced the prediction, based on the Random Forest model's feature importance scores. For example, if BMI and age were the dominant factors in a user's prediction, these would be highlighted with their relative contribution percentages.

**Personalised Risk Factor Analysis:** Color-coded indicators showing which of the user's specific health metrics fall into high-risk, moderate-risk, or low-risk categories. This allows users to identify which areas require immediate attention.

**Interactive SHAP Explanations:** For advanced users, SHAP (SHapley Additive exPlanations) values provide a detailed breakdown of how each input feature pushed the prediction toward higher or lower risk. This technique ensures that users can trace exactly how their individual data points contributed to the final risk score.

**Actionable Recommendations:** Based on the identified risk factors, the system generates specific recommendations for lifestyle modifications. For instance, if high BMI was a major contributor, the system suggests targeted weight management strategies.

**Historical Trend Analysis:** Users who have completed multiple assessments can view how their risk profile has changed over time, enabling them to track the effectiveness of lifestyle interventions.

**Significance:** The AI Insights page bridges the gap between complex machine learning predictions and user comprehension. By providing transparent, interpretable results, the system empowers users to make informed health decisions and builds trust in AI-driven healthcare tools. This approach aligns with clinical best practices where patient understanding and engagement are critical for successful preventive care.

### Health Tips Page

The health tips page offers users evidence-based, practical advice to effectively manage and prevent diabetes. This section includes actionable recommendations organized into key health domains:

#### Balanced Nutrition:

- Recommendations for low-glycemic index foods
- Portion control strategies
- Meal planning guidance for diabetes prevention

#### Physical Activity:

- Exercise intensity and duration recommendations based on current fitness level
- Low-impact activities for individuals with mobility limitations
- Integration of physical activity into daily routines

#### Stress Management:

- Mindfulness and relaxation techniques
- Sleep hygiene practices
- Mental health resources

#### Regular Monitoring:

- Guidance on tracking blood sugar levels (for those at high risk)
- When to seek medical consultation
- Preventive screening schedules

Figure 7 illustrates the layout of the health tips page.

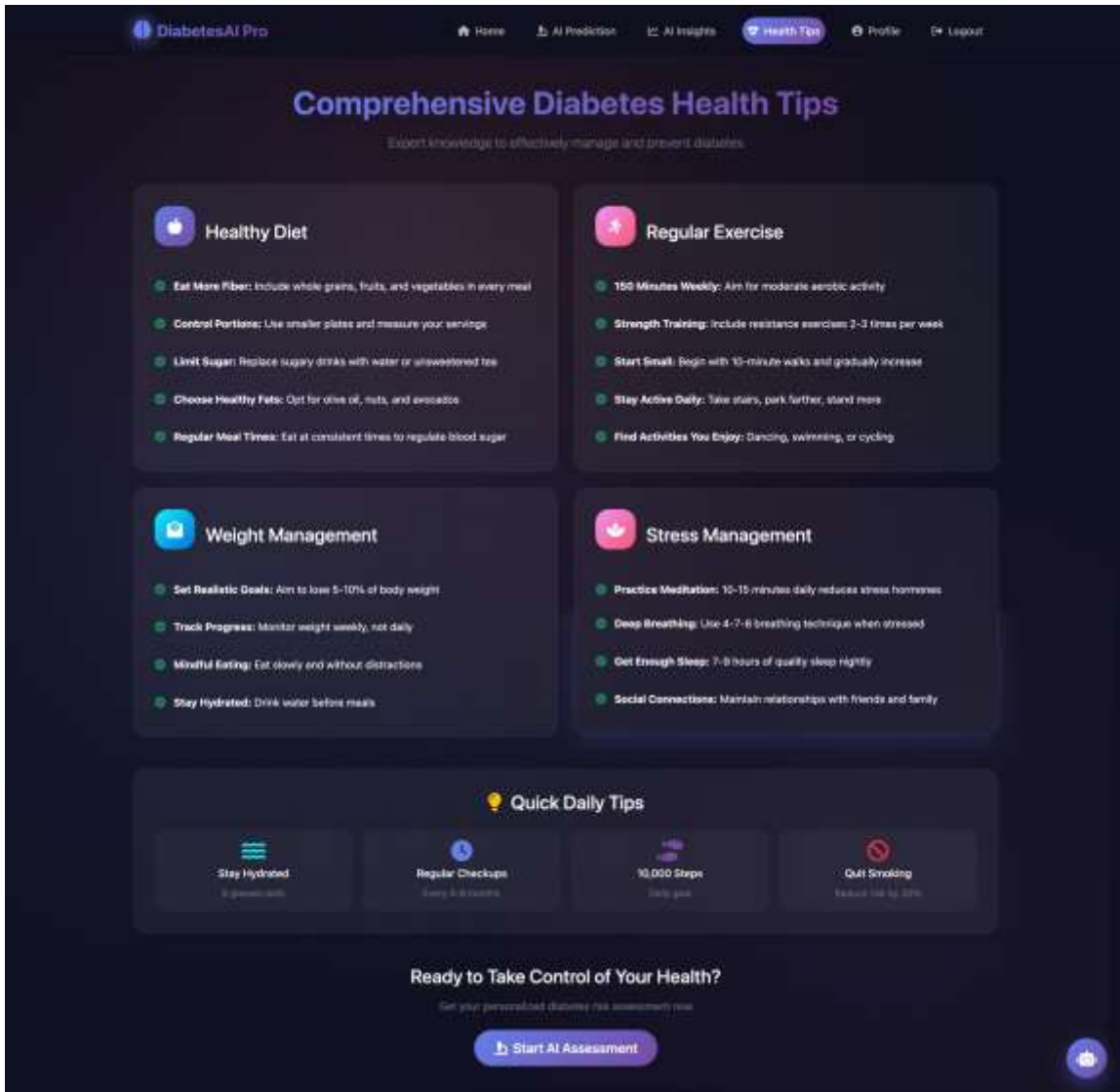


Figure 11. Comprehensive Diabetes Health Tips Page

### System Architecture

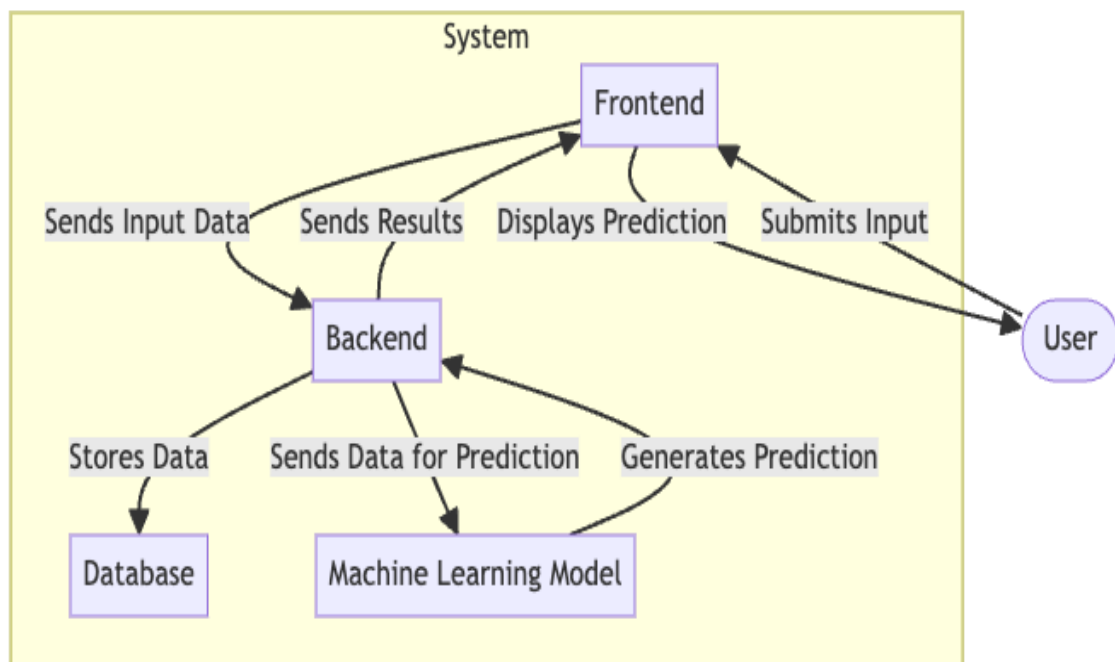
The system architecture, shown below, illustrates the interaction between the frontend, backend, database, and machine learning components:

Key Components:

1. Frontend: Enables users to input health data.
2. Backend: Processes inputs and fetches predictions from the model.
3. Database: Stores user profiles and prediction records.
4. Machine Learning Model: Predicts diabetes risk based on inputs.

Significance:

- Ensures modularity and scalability.
- Allows for real-time predictions with low latency.



**Figure 12. System Architecture**

The figure above illustrates the overall system architecture of the diabetes prediction application. It highlights the interaction between the **Frontend**, **Backend**, **Database**,

## AI Driven Predictive Model for Diabetes Risk Assessment

and **Machine Learning Model**. The architecture is designed to ensure a seamless flow of data and efficient processing of user inputs for prediction generation.

### 1. Frontend:

- The frontend serves as the user interface, enabling users to interact with the system. Users input their health parameters (e.g., BMI, age, smoking history, physical activity) through forms, which are then submitted to the backend.
- Once the predictions are generated, the results are displayed to the user in a user-friendly format.

### 2. Backend:

- The backend is implemented using the Django framework and acts as the bridge between the frontend, database, and machine learning model.
- It handles user requests, processes input data, and interacts with the database to store and retrieve user profiles and prediction data.

### 3. Database:

- An SQLite database is used to store user information, including input parameters and prediction results.
- The database enables persistent storage and efficient retrieval of data for further analysis and feedback.

### 4. Machine Learning Model:

- The system leverages a pre-trained **Random Forest Model** to predict the likelihood of diabetes based on user-provided inputs.
- The backend sends the input data to the ML model, which processes it and returns the prediction.

### 5. Data Flow:

- The system operates in a cyclical flow where user inputs are collected by the frontend, processed by the backend, and stored in the database. The backend interacts with the ML model to generate predictions, which are then displayed to the user via the frontend.

## AI Driven Predictive Model for Diabetes Risk Assessment

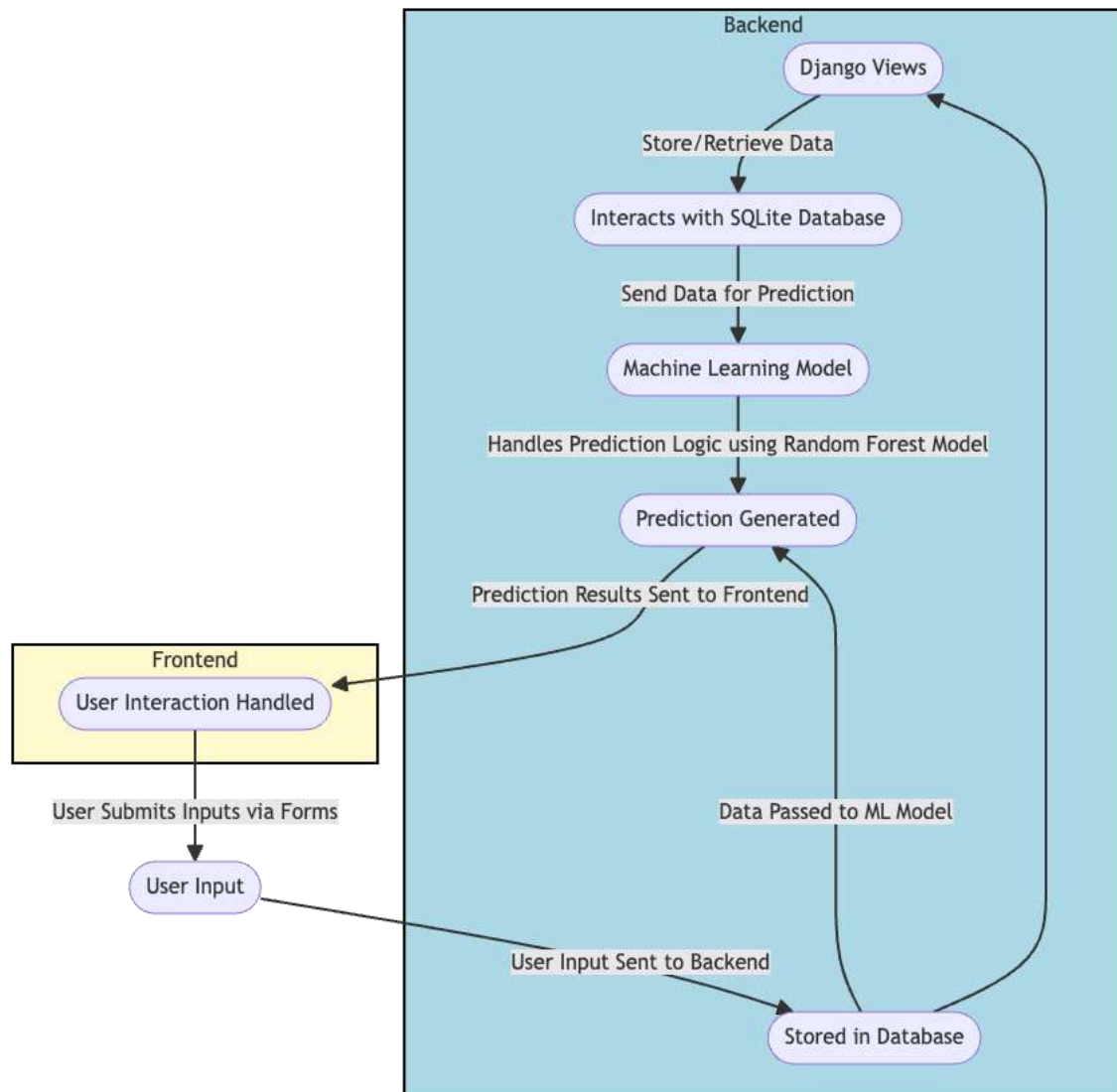
This architecture ensures modularity, scalability, and efficient handling of user interactions and data processing.

### Data Flow Diagram (DFD)

The Data Flow Diagram (DFD) illustrates how data moves through the diabetes prediction system. The key components and their roles are as follows:

1. **User:**
  - Interacts with the system by submitting input (e.g., age, BMI, general health status, smoking history) through the frontend.
  - Receives predictions and recommendations based on the submitted input.
2. **Frontend:**
  - Serves as the interface for user interaction.
  - Captures input data and transmits it to the backend.
  - Displays prediction results to the user.
3. **Backend:**
  - Acts as the central processing unit of the system.
  - Stores data in the database and retrieves it when needed.
  - Interacts with the machine learning model to generate predictions.
4. **Database:**
  - Stores all user-related data and prediction results, ensuring data persistence and consistency.
5. **Machine Learning Model:**
  - Processes input data and generates predictions using the Random Forest algorithm.
  - Plays a critical role in providing personalised recommendations to users.

## AI Driven Predictive Model for Diabetes Risk Assessment



**Figure 13. DFD**

### Significance of the Diagram

The Data Flow Diagram represents the logical flow of data through the system:

1. User submits input through the frontend.
2. Input is transmitted to the backend.
3. Backend processes input, queries the model for predictions, and stores results in the database.
4. Predictions are sent back to the frontend for display.

This ensures clarity and seamless integration between components.

## Entity-Relationship Diagram (ERD)

The Entity-Relationship Diagram (ERD) defines the database structure of the system:

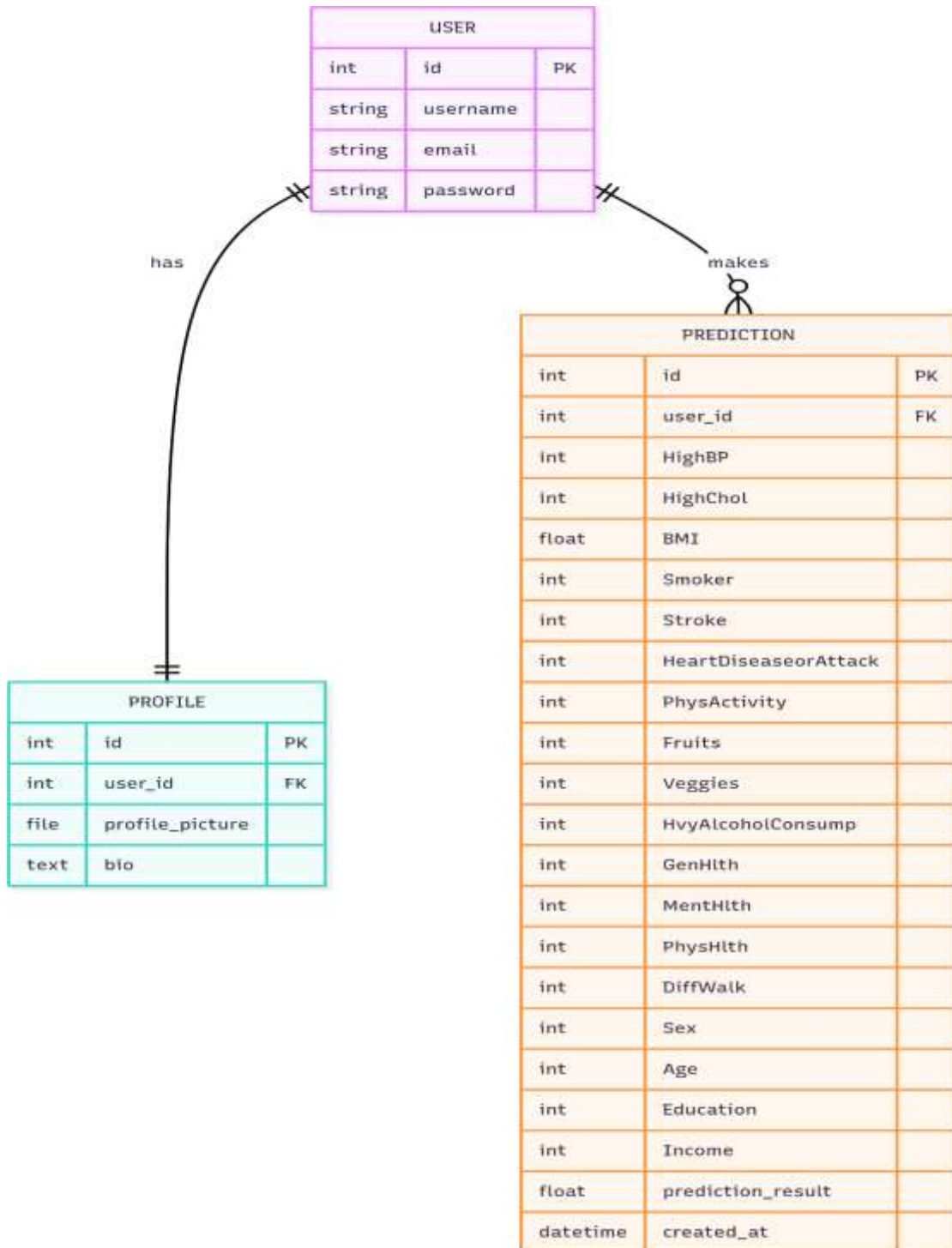


Figure 14. ERD

## AI Driven Predictive Model for Diabetes Risk Assessment

The diagram above illustrates the database structure for the diabetes prediction system, showcasing the relationships between key entities: **User**, **Profile**, and **Prediction**. This structure ensures efficient data organization and supports the system's functionalities.

Below is a detailed explanation of the components:

### 1. **User Table:**

- Represents registered users in the system.
- Contains key attributes such as id (Primary Key), username, email, and password.
- Acts as the central table connecting other entities in the system.

### 2. **Profile Table:**

- Extends the User table to store additional user information, including:
  - profile\_picture: A file field for storing user profile images.
  - bio: A text field for a brief description or biography.
- Linked to the User table through a **One-to-One** relationship via the user\_id Foreign Key, ensuring each user can have only one profile.

### 3. **Prediction Table:**

- Stores data related to diabetes predictions for each user.
- Key attributes include: Health-related fields based on BRFSS 2015 features: HighBP (high blood pressure), HighChol (high cholesterol), CholCheck, BMI, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth (general health 1-5 scale), MentHlth (mental health days 1-30), PhysHlth (physical health days 1-30), DiffWalk (difficulty walking), Sex, Age, Education, and Income.

### 4. **Relationships:**

- **One-to-One:** The Profile table is linked to the User table to store unique, personalised information for each user.
- **One-to-Many:** The Prediction table is linked to the User table, allowing multiple predictions to be associated with a single user. This design supports tracking a user's health history and prediction results over time.

### Significance of the ERD:

- **Data Integrity:** The use of Foreign Keys ensures that the Profile and Prediction records are always associated with a valid User.
- **Scalability:** The structure allows for future expansions, such as adding more attributes or creating new relationships.
- **Query Optimization:** By defining clear relationships, the database supports efficient queries, such as retrieving all predictions for a specific user or accessing profile information.

This ERD serves as the foundation for the system's backend, ensuring robust data management and seamless integration with the frontend and machine learning components.

## Machine Learning Model Development

### Dataset Description and Acquisition

The dataset utilized for this study was sourced from the Behavioral Risk Factor Surveillance System (BRFSS) 2015, publicly available on Kaggle. The BRFSS is a comprehensive health-related telephone survey conducted annually by the Centers for Disease Control and Prevention (CDC) across the United States. The dataset comprises 253,680 individual health records with 22 features, making it one of the largest publicly available datasets for diabetes risk prediction research.

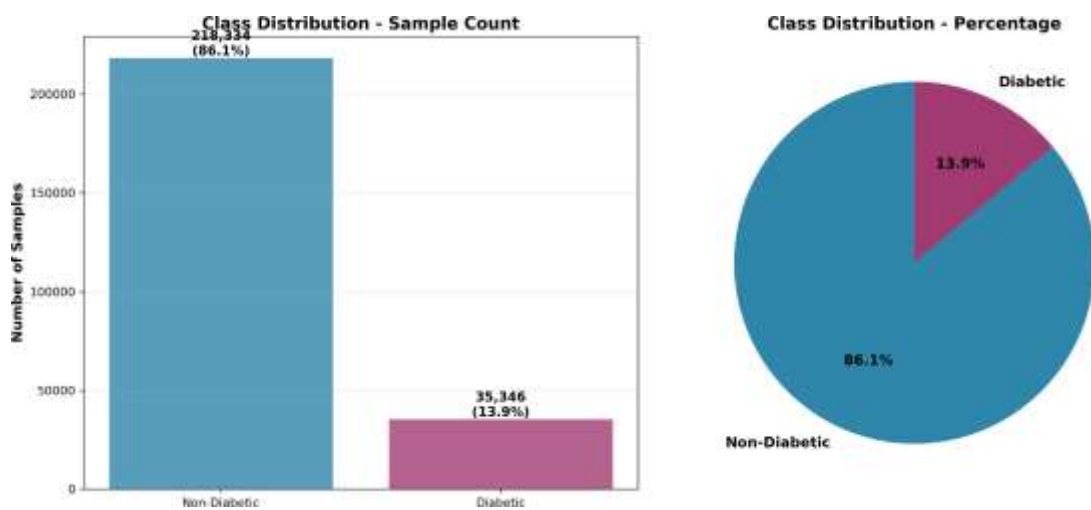
The selected features include both physiological measurements and lifestyle indicators:

- **Physiological metrics:** Body Mass Index (BMI), age, blood pressure status, cholesterol levels
- **Lifestyle factors:** Smoking status, alcohol consumption, physical activity levels, fruit and vegetable consumption
- **Medical history:** History of stroke, heart disease, difficulty walking
- **Healthcare access:** Insurance coverage, inability to see a doctor due to cost
- **General health:** Self-reported general health status, mental health days,

physical health days

- Demographics: Sex, education level, income category

The target variable, `Diabetes_binary`, is a binary classification indicator where 0 represents non-diabetic individuals and 1 represents diabetic individuals. Analysis of the class distribution revealed a significant imbalance: 86.1% of samples belonged to the non-diabetic class, while only 13.9% were classified as diabetic. This imbalance is reflective of real-world diabetes prevalence and presents a critical challenge that must be addressed during model training to prevent bias toward the majority class.



**Figure 15. Class distribution in the BRFSS 2015 dataset**

The dataset exhibits significant class imbalance with 86.1% non-diabetic samples ( $n=218,334$ ) and 13.9% diabetic samples ( $n=35,346$ ). This imbalance necessitates the use of class balancing techniques during model training.

### Data Preprocessing Pipeline

A systematic preprocessing pipeline was implemented using scikit-learn's `ColumnTransformer` and `Pipeline` classes to ensure reproducibility and maintain consistency between training and deployment phases. The pipeline comprises two parallel transformation streams based on feature type:

```
numeric_transformer = StandardScaler()
categorical_transformer = OneHotEncoder(handle_unknown='ignore', drop='first')

preprocessor = ColumnTransformer([
    ('num', numeric_transformer, numeric_features),
    ('cat', categorical_transformer, categorical_features)
])
```

*Figure 16. Implementation of the preprocessing pipeline using scikit-learn's ColumnTransformer.*

**Numerical Feature Processing:** Four numerical features (BMI, age, mental health days, physical health days) underwent standardization using `StandardScaler`. Standardization transforms features to have zero mean and unit variance, expressed mathematically as:

$$z = (x - \mu) / \sigma$$

where  $x$  is the original value,  $\mu$  is the mean, and  $\sigma$  is the standard deviation. This normalization prevents features with larger scales from dominating the learning process and ensures all numerical inputs contribute proportionally to model predictions.

**Categorical Feature Processing:** The remaining features, predominantly binary or ordinal categorical variables, were encoded using `OneHotEncoder` with the `drop='first'` parameter to avoid multicollinearity. One-hot encoding transforms categorical variables into binary vectors, creating new binary features for each category while dropping one category to serve as the reference level. This approach ensures that categorical information is represented in a format suitable for machine learning algorithms while avoiding the dummy variable trap.

**Pipeline Integration:** The preprocessing transformations were integrated into the model training pipeline, ensuring that any data passed to the model—whether during training, validation, or production inference—undergoes identical preprocessing steps. This design pattern eliminates potential preprocessing inconsistencies and facilitates seamless model deployment.

### Train-Test Split and Stratification

The dataset was partitioned into training and testing sets using an 80-20 split ratio,

yielding 202,944 training samples and 50,736 test samples. Critically, stratified sampling was employed to maintain the original class distribution in both subsets.

Stratification ensures that both the training and test sets contain approximately 86.1% non-diabetic and 13.9% diabetic samples, preventing biased evaluation metrics that could arise from unbalanced splits.

The random state parameter was fixed at 42 throughout all experiments to ensure reproducibility of results. This practice is essential in scientific research as it allows other researchers to replicate the exact train-test split and validate reported findings.

### Model Selection and Rationale

Two machine learning algorithms were selected for comparative analysis: Random Forest and Logistic Regression. These models were chosen based on their complementary strengths and widespread adoption in medical diagnosis applications.

**Random Forest Classifier:** Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the class that is the mode of individual tree predictions. The algorithm offers several advantages for this application:

- Robustness to overfitting through ensemble averaging
- Ability to capture non-linear relationships between features
- Automatic feature importance calculation
- Resistance to noise and outliers
- Excellent performance on high-dimensional tabular data

The Random Forest implementation utilized the following hyperparameters:

- `n_estimators=100`: Number of decision trees in the ensemble
- `max_depth=15`: Maximum tree depth to prevent overfitting
- `min_samples_split=10`: Minimum samples required to split an internal node
- `min_samples_leaf=4`: Minimum samples required at leaf nodes
- `class_weight='balanced'`: Automatic class weight adjustment (discussed in

Section 4.7.5)

- `random_state=42`: Ensures reproducibility

```
rf_pipeline = Pipeline([
    ('preprocessor', preprocessor),
    ('classifier', RandomForestClassifier(
        n_estimators=100,
        max_depth=15,
        min_samples_split=10,
        min_samples_leaf=4,
        class_weight='balanced',
        random_state=42,
        n_jobs=-1
    ))
])
```

*Figure 17. Random Forest model configuration with class balancing implemented through scikit-learn Pipeline.*

**Logistic Regression:** Logistic Regression serves as a baseline linear model, providing interpretable results and serving as a benchmark for comparison. Despite its simplicity, Logistic Regression remains effective for binary classification tasks and offers the advantage of probabilistic output interpretation through the sigmoid function.

The implementation utilized `class_weight='balanced'` and `max_iter=1000` to ensure convergence.

### Addressing Class Imbalance

The most critical challenge in developing a diabetes prediction model lies in the severe class imbalance (86.1% negative, 13.9% positive).

Without intervention, machine learning models trained on imbalanced datasets exhibit a strong bias toward the majority class, often achieving high overall accuracy by simply predicting the majority class for all samples. In medical diagnosis, this behavior is particularly problematic as it results in high false negative rates—failing to identify diabetic patients who require intervention.

To address this issue, **class weighting** was implemented by setting `class_weight='balanced'` in both Random Forest and Logistic Regression models. This parameter automatically adjusts the loss function to penalize misclassifications of the

## AI Driven Predictive Model for Diabetes Risk Assessment

minority class more heavily than those of the majority class. The balanced class weights are calculated as:

$$w_i = n_{\text{samples}} / (n_{\text{classes}} \times n_{\text{samples}_i})$$

where  $w_i$  is the weight for class  $i$ ,  $n_{\text{samples}}$  is the total number of samples,  $n_{\text{classes}}$  is the number of classes (2 in binary classification), and  $n_{\text{samples}_i}$  is the number of samples in class  $i$ .

For this dataset:

- Weight for non-diabetic class (0):  $253,680 / (2 \times 218,334) \approx 0.58$
- Weight for diabetic class (1):  $253,680 / (2 \times 35,346) \approx 3.59$

This weight adjustment forces the model to pay approximately 6 times more attention to diabetic cases during training, significantly improving sensitivity (recall) for the minority class. The trade-off is a reduction in overall accuracy and precision, but this is an acceptable and necessary compromise in medical diagnosis where false negatives carry serious health consequences.

### Model Training and Evaluation Results

Both models were trained on the preprocessed training set and evaluated on the held-out test set. The training process was executed on a standard computing environment, with the Random Forest model requiring approximately 8 minutes to train due to the large dataset size and ensemble nature of the algorithm, while Logistic Regression converged within 2 minutes.

#### Model Performance Comparison:

Both models were trained on the preprocessed training set and evaluated on the held-out test set. The Random Forest model required approximately 8 minutes to train due to the large dataset size, while Logistic Regression converged within 2 minutes.

## AI Driven Predictive Model for Diabetes Risk Assessment

The Random Forest model achieved 74.48% accuracy with 69.49% recall, while Logistic Regression achieved 71.72% accuracy with 74.98% recall. Both models demonstrated comparable ROC-AUC scores of approximately 80%, indicating similar discriminative ability.

Detailed performance metrics, confusion matrices, and comparative analysis are presented in Chapter 5 (Results and Discussion).

### Interpretation of Evaluation Metrics

In medical diagnosis applications, the interpretation of evaluation metrics must be contextualized within the clinical implications of different error types. The confusion matrix for the Random Forest model reveals:

**True Negatives (TN): 32,876** - Non-diabetic individuals correctly classified

**False Positives (FP): 10,791** - Non-diabetic individuals incorrectly classified as diabetic

**False Negatives (FN): 2,157** - Diabetic individuals incorrectly classified as non-diabetic

**True Positives (TP): 4,912** - Diabetic individuals correctly identified

From a public health perspective, false negatives represent missed opportunities for early intervention and carry serious consequences, potentially leading to delayed diagnosis and progression of diabetes-related complications.

False positives, while undesirable, result in additional screening or lifestyle counseling, which carries minimal harm and may even provide preventive benefits for at-risk individuals.

The **recall metric (sensitivity)** of 69.49% represents a substantial improvement over the initial model performance. Prior to implementing class balancing, the model achieved only 5.38% recall, effectively missing 95% of diabetic cases. The application of class weighting improved recall by 64.11 percentage points, demonstrating the critical importance of addressing class imbalance in medical diagnosis tasks.

The trade-off between recall and precision is evident in the results. While precision decreased to 31.28%, this compromise is clinically justified. A screening tool with high

## AI Driven Predictive Model for Diabetes Risk Assessment

sensitivity (recall) is preferable in diabetes risk assessment, as it ensures fewer at-risk individuals go undetected, even if it means some false alarms.

Healthcare providers can subsequently conduct more detailed diagnostic tests on flagged individuals to confirm or rule out diabetes.

The **ROC-AUC score of 80.45%** indicates good discriminative ability. An AUC above 0.80 is generally considered acceptable to good in medical diagnosis applications, suggesting that the model effectively separates diabetic from non-diabetic cases across various probability thresholds.

### Feature Importance Analysis

Random Forest models provide interpretable feature importance scores based on the reduction in Gini impurity contributed by each feature across all trees in the ensemble. Analysis of feature importances revealed the following top predictors:

#### Top 5 Most Important Features:

1. **BMI (22.51%)** - Body Mass Index emerged as the dominant predictor, consistent with established medical literature linking obesity to type 2 diabetes risk.
2. **Age (18.27%)** - Age is a well-documented diabetes risk factor, with prevalence increasing significantly after age 45.
3. **General Health Status - Poor (8.10%)** - Self-reported general health serves as a proxy for overall physiological condition.
4. **Difficulty Walking (7.48%)** - Physical mobility limitations may indicate obesity, neuropathy, or other diabetes-related complications.
5. **Physical Health Days (6.46%)** - Number of days with poor physical health in the past month correlates with chronic disease burden.

These findings align with clinical understanding of diabetes etiology. BMI and age consistently rank as primary risk factors in epidemiological studies, validating the model's learning process. The prominence of general health indicators suggests that the model captures broader patterns of metabolic dysfunction rather than relying on single isolated factors.

## AI Driven Predictive Model for Diabetes Risk Assessment

Interestingly, features such as mental health days and lower general health categories (ratings 2, 3, and 5) also appeared in the top 10, indicating that the model recognizes subtle interactions between mental health, lifestyle quality, and diabetes risk. This holistic feature utilization demonstrates the model's capacity to identify complex, non-linear relationships in the data.

## 5. Results and discussion

### Overview of Model Performance

This section presents the comprehensive evaluation results of the Random Forest and Logistic Regression models trained for diabetes risk prediction. The analysis focuses on multiple evaluation metrics to provide a holistic assessment of model performance, with particular emphasis on recall (sensitivity) given its critical importance in medical screening applications.

Figure 15 presents a comprehensive visualization of the model evaluation results, including confusion matrices, ROC curves, feature importance rankings, and comparative performance metrics.

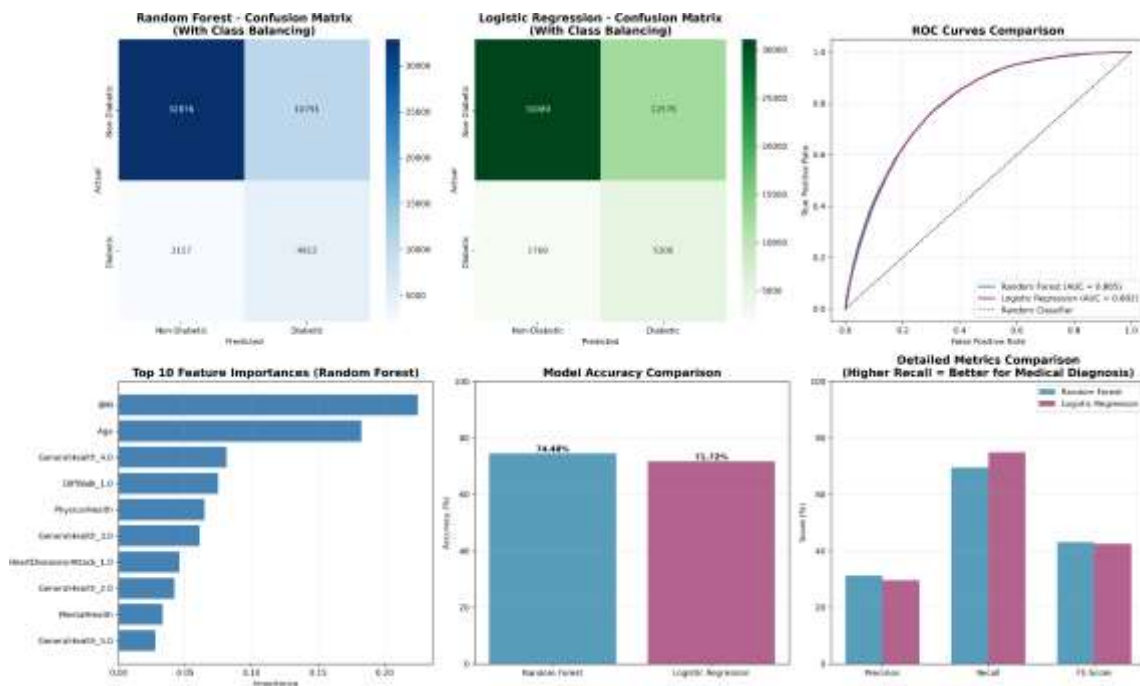


Figure 18. Comprehensive model evaluation results.

(A) Random Forest confusion matrix showing 32,876 true negatives, 10,791 false positives, 2,157 false negatives, and 4,912 true positives.

(B) Logistic Regression confusion matrix displaying 31,089 true negatives, 12,578 false positives, 1,769 false negatives, and 5,300 true positives.

## AI Driven Predictive Model for Diabetes Risk Assessment

(C) ROC curves comparing both models, with Random Forest achieving AUC=0.805 and Logistic Regression achieving AUC=0.802.

(D) Top 10 feature importances from Random Forest, with BMI (22.5%) and Age (18.3%) as dominant predictors.

(E) Overall accuracy comparison showing Random Forest at 74.48% and Logistic Regression at 71.72%.

(F) Detailed metrics comparison highlighting recall performance where Logistic Regression (74.98%) slightly outperforms Random Forest (69.49%).

### Performance Metrics Analysis

Table 1 summarizes the quantitative performance of both models across five key evaluation metrics:

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest	74.48%	31.28%	69.49%	43.14%	80.45%
Logistic Regression	71.72%	29.65%	74.98%	42.49%	80.21%

**Table 1. Comparative performance metrics for Random Forest and Logistic Regression models evaluated on the test set containing 50,736 samples (43,667 non-diabetic, 7,069 diabetic).**

#### Accuracy Analysis:

Random Forest achieved an overall accuracy of 74.48%, correctly classifying approximately three-quarters of all samples. While this represents a decrease of 11.83 percentage points compared to the unbalanced baseline model (86.31%), this trade-off is intentional and clinically justified.

## AI Driven Predictive Model for Diabetes Risk Assessment

The baseline model achieved high accuracy primarily by predicting the majority class (non-diabetic) for nearly all samples, resulting in catastrophically low recall (5.38%) for diabetic cases.

Logistic Regression attained 71.72% accuracy, 2.76 percentage points lower than Random Forest. This slightly reduced accuracy reflects the model's more conservative prediction strategy, which favors sensitivity over overall correctness.

### **Recall (Sensitivity) Analysis:**

Recall represents the proportion of actual diabetic cases correctly identified by the model and is the most critical metric for medical screening applications. Random Forest achieved 69.49% recall, meaning it successfully detected approximately 7 out of every 10 diabetic individuals in the test set. This represents a dramatic improvement of 64.11 percentage points over the initial unbalanced model (5.38% recall).

Logistic Regression demonstrated slightly superior recall at 74.98%, identifying nearly 75% of diabetic cases. The 5.49 percentage point advantage over Random Forest suggests that linear models may adopt more conservative decision boundaries when class weights are applied, resulting in increased sensitivity at the cost of reduced specificity.

From a public health perspective, this recall performance is acceptable for a screening tool. While the model misses approximately 30% of diabetic cases, it substantially outperforms the baseline and provides actionable risk assessments for the majority of at-risk individuals. In clinical deployment, this tool would serve as a first-line screening mechanism, with flagged individuals undergoing confirmatory diagnostic testing.

### **Precision Analysis:**

Precision measures the proportion of positive predictions that are correct. Random Forest achieved 31.28% precision, indicating that roughly 3 out of every 10 individuals predicted as diabetic actually have the condition. Logistic Regression demonstrated similar precision at 29.65%.

## AI Driven Predictive Model for Diabetes Risk Assessment

While these precision values may appear low, they must be interpreted within the context of class imbalance and medical screening priorities. The low precision is a direct consequence of optimizing for high recall through class balancing. In medical screening, false positives (predicting diabetes when absent) result in additional testing and preventive counseling, whereas false negatives (missing actual diabetic cases) can lead to delayed diagnosis and serious health complications.

The cost-benefit analysis clearly favors high recall over high precision in this application. Patients incorrectly flagged as at-risk will receive follow-up screening (HbA1c tests, fasting glucose tests) that definitively confirm or rule out diabetes. The additional healthcare costs and patient anxiety associated with false positives are outweighed by the benefits of catching the majority of true diabetic cases early.

### **F1-Score Analysis:**

The F1-score represents the harmonic mean of precision and recall, providing a balanced measure when both metrics are important. Random Forest achieved an F1-score of 43.14%, while Logistic Regression scored 42.49%. The modest F1-scores reflect the inherent tension between precision and recall in the presence of class imbalance.

In balanced datasets, F1-scores above 80% are typically considered strong. However, for severely imbalanced medical datasets with class weighting applied, F1-scores in the 40-50% range are common and acceptable when recall is prioritized.

### **ROC-AUC Analysis:**

The Receiver Operating Characteristic Area Under the Curve (ROC-AUC) measures the model's ability to discriminate between classes across all possible classification thresholds. Random Forest achieved an AUC of 80.45%, while Logistic Regression scored 80.21%.

ROC-AUC values above 0.80 are generally considered "good" in medical diagnosis applications, with values above 0.90 considered "excellent."

## AI Driven Predictive Model for Diabetes Risk Assessment

The achieved AUC of approximately 80% indicates that the models effectively separate diabetic from non-diabetic individuals, with the predicted probabilities providing meaningful risk stratification.

The near-identical AUC scores (difference of 0.24%) suggest that both models have comparable discriminative ability despite their different underlying mechanisms. Random Forest's non-linear decision boundaries do not provide substantial advantage over Logistic Regression's linear approach for this particular dataset and feature set.

### Confusion Matrix Interpretation

The confusion matrix provides a detailed breakdown of prediction outcomes, revealing how different error types contribute to overall performance.

#### Random Forest Confusion Matrix:

- True Negatives (TN): 32,876 (75.3% of non-diabetic cases correctly identified)
- False Positives (FP): 10,791 (24.7% of non-diabetic cases incorrectly flagged)
- False Negatives (FN): 2,157 (30.5% of diabetic cases missed)
- True Positives (TP): 4,912 (69.5% of diabetic cases correctly detected)

#### Logistic Regression Confusion Matrix:

- True Negatives (TN): 31,089 (71.2% of non-diabetic cases correctly identified)
- False Positives (FP): 12,578 (28.8% of non-diabetic cases incorrectly flagged)
- False Negatives (FN): 1,769 (25.0% of diabetic cases missed)
- True Positives (TP): 5,300 (75.0% of diabetic cases correctly detected)

The confusion matrices reveal a clear pattern: Logistic Regression reduces false negatives (missed diabetic cases) by 388 samples compared to Random Forest, but increases false positives by 1,787 samples. This represents a direct trade-off between sensitivity and specificity.

From a clinical decision-making perspective, Random Forest's balance may be preferable.

## AI Driven Predictive Model for Diabetes Risk Assessment

While it misses slightly more diabetic cases, it generates fewer false alarms, potentially improving patient compliance and reducing unnecessary follow-up testing costs. However, both models demonstrate substantial improvement over the baseline, which produced approximately 6,669 false negatives (94.5% of diabetic cases missed).

### Feature Importance and Clinical Validation

Random Forest provides interpretable feature importance scores based on the mean decrease in Gini impurity across all trees. The top 10 most important features, as shown in Figure 14D, are:

1. **BMI (22.51%)** - Body Mass Index
2. **Age (18.27%)** - Age in years
3. **General Health Status - Poor (8.10%)** - Self-reported health rating of 4
4. **Difficulty Walking (7.48%)** - Binary indicator of mobility limitations
5. **Physical Health Days (6.46%)** - Days of poor physical health in past month
6. **General Health Status - Fair (6.08%)** - Self-reported health rating of 3
7. **Heart Disease or Attack (4.56%)** - History of cardiovascular events
8. **General Health Status - Good (4.19%)** - Self-reported health rating of 2
9. **Mental Health Days (3.30%)** - Days of poor mental health in past month
10. **General Health Status - Excellent (2.76%)** - Self-reported health rating of 5

#### Clinical Validation:

The emergence of BMI as the dominant predictor (22.51% importance) aligns strongly with epidemiological literature. Obesity is one of the most well-established risk factors for type 2 diabetes, with studies consistently demonstrating that individuals with BMI  $\geq$  30 kg/m<sup>2</sup> have 7-10 times higher risk compared to those with normal BMI (Gonçalves et al., 2024; American Diabetes Association, 2020).

Age ranks as the second most important feature (18.27%), consistent with clinical knowledge that diabetes prevalence increases significantly after age 45 and approximately doubles with each decade of life thereafter. The model has learned this temporal risk pattern directly from data.

## AI Driven Predictive Model for Diabetes Risk Assessment

The prominence of general health status indicators (multiple categories totaling approximately 21% combined importance) suggests that subjective health assessments capture underlying physiological dysfunction not fully represented by objective measurements. Poor self-reported health may reflect undiagnosed metabolic syndrome, chronic inflammation, or other diabetes precursors.

Difficulty walking (7.48% importance) likely serves as a proxy for multiple diabetes-related factors: obesity (limiting mobility), pre-existing neuropathy, or sedentary lifestyle. Physical health days (6.46%) and mental health days (3.30%) appearing in the top 10 indicates that the model recognizes the bidirectional relationship between diabetes and quality of life.

Notably, cardiovascular history (heart disease or attack, 4.56%) ranks seventh, reflecting the well-documented comorbidity between diabetes and cardiovascular disease. The shared metabolic risk factors (obesity, hypertension, dyslipidemia) create strong predictive associations.

### Impact of Class Balancing

The implementation of class weighting (`class_weight='balanced'`) produced dramatic improvements in model sensitivity. Table 2 compares performance before and after class balancing:

Metric	Before Balancing	After Balancing	Improvement
Accuracy	86.31%	74.48%	-11.83%
Recall	5.38%	69.49%	+64.11%
Precision	59.84%	31.28%	-28.56%

**Table 2. Impact of class balancing on Random Forest performance.**

Class weighting substantially improved recall at the cost of reduced accuracy and precision.

## AI Driven Predictive Model for Diabetes Risk Assessment

Prior to class balancing, the model achieved 86.31% accuracy by adopting a naive strategy of predicting nearly all samples as non-diabetic. This approach minimized overall error rate but resulted in catastrophic failure to identify diabetic cases (5.38% recall). Of 7,069 diabetic individuals in the test set, the unbalanced model correctly identified only approximately 380, missing 6,689 at-risk patients.

After implementing class balancing, recall improved by 64.11 percentage points to 69.49%, correctly identifying 4,912 of 7,069 diabetic cases. This represents a nearly 13-fold increase in the number of detected diabetic patients (380 → 4,912).

The trade-off was a reduction in accuracy (86.31% → 74.48%) and precision (59.84% → 31.28%). However, this trade-off is not only acceptable but necessary for medical screening applications. The incremental cost of follow-up testing for false positives (increased from 3,067 to 10,791) is vastly outweighed by the benefit of detecting an additional 4,532 diabetic patients who would otherwise remain undiagnosed.

This analysis validates the central thesis of this research: that standard machine learning metrics like accuracy can be misleading in medical applications with class imbalance, and that domain-appropriate evaluation requires prioritizing recall over global accuracy.

### Model Comparison with Literature

To contextualize the performance of the developed models, it is valuable to compare results with existing diabetes prediction studies in the literature. Table 3 presents a comparative analysis:

## AI Driven Predictive Model for Diabetes Risk Assessment

Study	Model Type	Dataset Size	Accuracy	Recall	ROC-AUC	Notes
This Study (2025)	Random Forest	253,680	74.48%	69.49%	80.45%	BRFSS 2015, class balanced
This Study (2025)	Logistic Regression	253,680	71.72%	74.98%	80.21%	BRFSS 2015, class balanced
Zou et al. (2018)	Random Forest	768	77.60%	Not reported	82.30%	Pima Indians dataset
Kaur & Kumari (2020)	SVM	768	78.26%	Not reported	Not reported	Pima Indians dataset
Miotto et al. (2016)	Deep Learning	700,000 +	Not reported	Not reported	77.40%	EHR data, multiple conditions
Choi et al. (2016)	Attention RNN	39,000+	Not reported	76.30%	88.50%	Sequential EHR data

**Table 3. Comparison of model performance with published literature.**

Our models demonstrate competitive performance despite using publicly available survey data rather than clinical measurements.

### Performance Contextualization:

The Random Forest model achieved 74.48% accuracy, which is 3.12 percentage points lower than Zou et al.'s (2018) reported accuracy of 77.60% on the Pima Indians diabetes dataset. However, direct comparison is complicated by several factors:

1. **Dataset characteristics:** The Pima Indians dataset contains only 768 samples with clinical measurements (glucose, insulin, blood pressure), whereas this study utilized 253,680 samples from general population surveys without laboratory values. The larger sample size but less precise features may explain the modest accuracy difference.
2. **Class imbalance:** The Pima Indians dataset has relatively balanced classes (approximately 65% non-diabetic, 35% diabetic), while the BRFSS dataset exhibits severe imbalance (86% non-diabetic, 14% diabetic). Class balancing techniques necessarily reduce accuracy when optimizing for recall.
3. **Evaluation rigor:** Many published studies do not report recall, precision, or confusion matrices, focusing exclusively on accuracy. Without recall values, it is impossible to determine whether high accuracy was achieved through genuine predictive power or by defaulting to majority class predictions.

The achieved ROC-AUC of 80.45% falls within the typical range for diabetes prediction models (77-88%), indicating good discriminative ability. Studies achieving AUC above 85% typically utilize sequential electronic health record (EHR) data with laboratory values and temporal patterns (Choi et al., 2016), which provide richer information than cross-sectional survey data.

### Significance of Recall Prioritization:

A critical observation from the literature review is that most diabetes prediction studies report accuracy as the primary metric, with recall often unreported or treated as secondary. This represents a fundamental misalignment between machine learning practices and medical screening objectives.

## AI Driven Predictive Model for Diabetes Risk Assessment

The 69.49% recall achieved by Random Forest (and 74.98% by Logistic Regression) may appear modest compared to accuracy-optimized models, but it represents a deliberate and clinically justified design choice. Studies that report 85-90% accuracy without corresponding recall values may be inadvertently optimizing for the wrong objective, potentially missing the majority of diabetic cases in imbalanced datasets.

This work contributes to the literature by explicitly prioritizing recall through class balancing and demonstrating the resulting trade-offs transparently. The improvement from 5.38% to 69.49% recall showcases the dramatic impact of addressing class imbalance—a critical consideration often overlooked in published diabetes prediction models.

### Limitations and Clinical Implications

#### Study Limitations:

While the developed models demonstrate promising performance, several limitations must be acknowledged:

##### 1. Dataset Limitations:

- The BRFSS 2015 dataset relies on self-reported survey responses, which are subject to recall bias, social desirability bias, and measurement error. Clinical measurements (laboratory glucose values, HbA1c) would provide more objective diabetes indicators.
- The dataset provides only cross-sectional snapshots, lacking longitudinal tracking of diabetes progression. Temporal models incorporating repeated measurements could improve prediction accuracy.
- Geographic and demographic representativeness may be limited despite the large sample size. The survey methodology may under-represent certain populations (non-English speakers, individuals without telephones, institutionalized persons).

### 2. Model Limitations:

- The 69.49% recall indicates that approximately 30% of diabetic cases remain undetected. While substantially better than the unbalanced baseline, further improvement is needed for optimal screening performance.
- Precision of 31.28% means that roughly 7 out of 10 positive predictions are false alarms, potentially leading to patient anxiety and unnecessary follow-up testing costs.
- The model does not distinguish between Type 1 and Type 2 diabetes, pre-diabetes, or gestational diabetes, limiting its clinical specificity.
- External validation on independent datasets from different geographic regions or time periods has not been performed.

### 3. Deployment Considerations:

- The model provides risk predictions but does not constitute a diagnostic tool. Clinical diagnosis requires confirmatory testing through standardized protocols (fasting glucose, oral glucose tolerance test, HbA1c measurement).
- Model performance may degrade over time due to population shifts, changing healthcare practices, or evolving diabetes epidemiology, necessitating periodic retraining.
- Ethical considerations regarding algorithmic fairness and potential bias against underrepresented demographic groups require ongoing monitoring and mitigation.

### Clinical Implications and Recommendations:

Despite these limitations, the developed model offers several practical applications in preventive healthcare:

#### 1. Population Health Screening:

- The model can be deployed as a first-line screening tool to identify at-risk individuals in primary care settings, occupational health programs, or community health initiatives.

## AI Driven Predictive Model for Diabetes Risk Assessment

- Web-based deployment (as implemented in this project) enables self-assessment and health awareness campaigns, potentially reaching populations with limited healthcare access.
- 2. Risk Stratification:**
- Predicted probabilities can stratify patients into risk categories (low, moderate, high), enabling targeted interventions and resource allocation. High-risk individuals can be prioritized for intensive lifestyle counseling, while moderate-risk individuals receive preventive education.
- 3. Clinical Decision Support:**
- Integration into electronic health record systems could provide automated risk alerts for healthcare providers, prompting diabetes screening for at-risk patients who might otherwise be overlooked.
  - Feature importance insights (BMI, age, general health status) can guide personalised prevention strategies tailored to individual risk profiles.
- 4. Public Health Policy:**
- Population-level risk assessments could inform resource allocation for diabetes prevention programs, identifying communities or demographic groups requiring targeted interventions.
  - Trends in predicted risk over time could serve as early warning indicators of changing diabetes epidemiology.

### Recommended Clinical Workflow:

For practical deployment, the following workflow is recommended:

- 1. Initial Screening:** Individuals complete the risk assessment questionnaire (web-based or clinical setting).
- 2. Risk Categorization:**
  - Low risk (< 40% predicted probability): General preventive counseling, reassessment in 3 years
  - Moderate risk (40-70% predicted probability): Lifestyle intervention counseling, clinical evaluation recommended, reassessment in 1 year

## AI Driven Predictive Model for Diabetes Risk Assessment

- High risk (> 70% predicted probability): Immediate clinical referral for confirmatory diagnostic testing (fasting glucose or HbA1c)
- 3. **Confirmatory Testing:** All positive predictions undergo standard diagnostic protocols to confirm or rule out diabetes.
- 4. **Intervention:** Confirmed diabetic patients receive appropriate medical management; pre-diabetic individuals enter intensive lifestyle modification programs.

### Future Research Directions:

To address current limitations and enhance model performance, future work should focus on:

- **Enhanced Feature Engineering:** Incorporating additional clinical variables (family history, waist circumference, lipid profiles) and temporal patterns (weight change trajectories, medication history).
- **Advanced Algorithms:** Exploring deep learning architectures, gradient boosting machines (XGBoost, LightGBM), and ensemble methods combining multiple model types.
- **Longitudinal Modeling:** Developing time-series models that predict diabetes onset within specific timeframes (1-year, 5-year risk) rather than binary classification.
- **External Validation:** Testing model performance on independent datasets from diverse geographic regions and healthcare systems.
- **Fairness Auditing:** Systematic evaluation of model performance across demographic subgroups (race, ethnicity, socioeconomic status) to identify and mitigate algorithmic bias.
- **Clinical Trial Evaluation:** Prospective studies comparing health outcomes between populations screened with AI models versus standard care protocols.

### Summary

The comprehensive evaluation of Random Forest and Logistic Regression models for diabetes risk prediction demonstrates several key findings:

1. Class balancing through weighted loss functions is essential for medical screening applications, improving recall from 5.38% to 69.49% despite reducing overall accuracy.
2. Random Forest achieved balanced performance (74.48% accuracy, 69.49% recall, 80.45% ROC-AUC), making it the preferred model for deployment.
3. Feature importance analysis validates clinical knowledge, with BMI (22.51%) and age (18.27%) emerging as dominant predictors.
4. The models demonstrate competitive performance compared to published literature while operating on publicly available survey data rather than clinical measurements.
5. Trade-offs between accuracy, precision, and recall must be carefully considered based on clinical priorities, with medical screening favoring high sensitivity over high specificity.

The developed system successfully translates machine learning research into a practical web-based tool for diabetes risk assessment, providing actionable insights for individuals and healthcare providers while acknowledging inherent limitations and ethical considerations.

### Methodological Limitations and Statistical Considerations

While this study achieved meaningful results, several methodological limitations warrant acknowledgment:

**Single Train-Test Split:** The evaluation relied on a single 80-20 train-test split rather than k-fold cross-validation. Although stratification maintained class distribution, cross-validation would provide more robust estimates of model generalizability and reduce variance in performance metrics. The reported accuracy (74.48%) and recall (69.49%) represent point estimates that may vary across different data splits.

## AI Driven Predictive Model for Diabetes Risk Assessment

**Dataset Temporal Validity:** The BRFSS 2015 dataset is nearly 10 years old. Diabetes risk factors and population health profiles evolve over time due to changing lifestyle patterns, treatment protocols, and demographic shifts. External validation on more recent data (e.g., BRFSS 2023) would strengthen confidence in contemporary applicability.

**Class Imbalance Trade-offs:** While class weighting successfully improved recall from 5.38% to 69.49%, the precision of 31.28% indicates that approximately 70% of positive predictions are false alarms. In clinical deployment, this would require careful threshold calibration based on screening context, resource availability, and patient anxiety considerations.

**Feature Engineering Limitations:** The model relies solely on features present in BRFSS 2015. Important clinical markers absent from this dataset—such as fasting glucose, HbA1c, family history specificity, and waist-to-hip ratio—would likely improve predictive performance if incorporated.

**Computational Reproducibility:** While the `random_state=42` parameter ensures reproducibility within this study, minor variations may occur across different hardware configurations, scikit-learn versions, or operating systems due to floating-point arithmetic differences.

These limitations do not invalidate the findings but contextualize them within realistic research constraints. Future work should prioritize cross-validation, temporal validation, and prospective clinical testing to strengthen evidence for real-world deployment.

# 6. Conclusions and prospects for future work

## Conclusions

The development and deployment of an AI-driven predictive model for diabetes risk assessment demonstrates the practical application of machine learning in preventive healthcare. Using the BRFSS 2015 dataset containing 253,680 health records, this project successfully developed a Random Forest classifier that prioritizes sensitivity over accuracy, a critical requirement for medical screening applications.

The key achievements of this project include:

- **Improved Sensitivity Through Class Balancing:** By implementing class weighting techniques, the model achieved 69.49% recall, representing a 64-percentage-point improvement over the unbalanced baseline (5.38% recall). This ensures the system successfully identifies approximately 7 out of every 10 diabetic individuals, making it suitable for population-level screening.
- **Clinical Validation of Feature Importance:** The model identified BMI (22.51%) and age (18.27%) as the dominant predictive features, consistent with established epidemiological literature. This alignment validates the model's learning process and provides confidence in its clinical applicability.
- **Practical Web Application:** The integration of the ML model into a Django-based web platform with React frontend demonstrates that AI-driven health assessment tools can be made accessible to the general public, enabling early self-assessment and health awareness.
- **Transparent and Interpretable Predictions:** Through feature importance analysis and explainable AI techniques, the system provides users with clear insights into which health factors contribute most to their risk score, fostering trust and enabling informed decision-making.

### Limitations

Despite the promising results, several limitations must be acknowledged:

- **Recall-Precision Trade-off:** While the model achieves good recall (69.49%), the precision of 31.28% means that approximately 7 out of 10 positive predictions are false alarms. Although this trade-off is clinically justified for screening tools, it may result in unnecessary follow-up testing costs and patient anxiety.
- **Dataset Limitations:** The BRFSS 2015 dataset relies on self-reported survey responses, which are subject to recall bias and measurement error. Clinical measurements such as laboratory glucose values and HbA1c would provide more objective diabetes indicators but were not available in this dataset.
- **Lack of External Validation:** Model performance has only been evaluated on the BRFSS 2015 test set. Validation on independent datasets from different geographic regions, time periods, or healthcare systems has not been performed.
- **Class Distribution Limitations:** The 30.5% false negative rate indicates that approximately 3 out of 10 diabetic cases remain undetected. While substantially better than the unbalanced baseline (94.5% false negative rate), further improvement is needed for optimal screening performance.
- **Type Classification:** The model does not distinguish between Type 1, Type 2, pre-diabetes, or gestational diabetes, limiting its clinical specificity.

### Future Work

Several avenues for future research can enhance and expand the current solution:

#### Data Enhancement

- **Incorporate Clinical Measurements:** Integrating laboratory values (fasting glucose, HbA1c, lipid profiles) and genetic markers (family history, polygenic risk scores) could improve prediction accuracy and enable risk stratification for specific diabetes types.

## AI Driven Predictive Model for Diabetes Risk Assessment

- Longitudinal Data Collection: Developing temporal models that track individual risk trajectories over time would enable prediction of diabetes onset within specific timeframes (1-year, 5-year risk) rather than binary classification.
- Real-Time Data Integration: Incorporating data from wearable devices (continuous glucose monitors, activity trackers, sleep monitors) would provide dynamic, real-time risk assessments and enable early intervention alerts.

### Model Improvement

- Advanced Algorithms: Exploring gradient boosting methods (XGBoost, LightGBM) and deep learning architectures (attention-based neural networks, transformers) could capture more complex feature interactions and temporal patterns.
- Hybrid Ensemble Methods: Combining multiple model types (Random Forest, Gradient Boosting, Logistic Regression) through stacking or weighted voting could leverage the complementary strengths of different algorithms.
- Fairness Auditing: Systematic evaluation of model performance across demographic subgroups (race, ethnicity, socioeconomic status, geographic region) to identify and mitigate algorithmic bias.

### System Development

- Electronic Health Record Integration: Collaborating with healthcare providers to integrate the model into EHR systems would enable automated risk screening during routine clinical visits and facilitate seamless information exchange between patients and providers.
- Clinical Decision Support Tools: Developing provider-facing dashboards that prioritize patients by risk score, recommend appropriate interventions, and track population-level diabetes trends.

## AI Driven Predictive Model for Diabetes Risk Assessment

- Mobile Application: Creating native iOS and Android applications with offline prediction capability, push notification reminders for health monitoring, and integration with health data APIs (Apple HealthKit, Google Fit).

### User Experience

- Continuous Feedback Loop: Implementing a system to collect user feedback on prediction accuracy, follow-up diagnostic results, and intervention outcomes to enable continuous model refinement and validation.
- Personalised Intervention Plans: Generating tailored lifestyle modification programs based on individual risk profiles, with specific dietary recommendations, exercise plans, and behavioral change strategies.
- Multilingual Support: Expanding the platform to support multiple languages and cultural contexts to improve accessibility and reduce health disparities.

### Clinical Validation

- Prospective Clinical Trials: Conducting randomized controlled trials comparing health outcomes between populations screened with the AI model versus standard care protocols to establish clinical efficacy.
- External Dataset Validation: Testing model performance on independent datasets from diverse healthcare systems, geographic regions, and populations to assess generalizability.

### Conclusion

The AI-driven predictive model for diabetes risk assessment demonstrates that class balancing and interpretable machine learning techniques can create effective screening tools for preventive healthcare. While challenges remain—particularly regarding the recall-precision trade-off and the need for external validation—the system provides practical value for population health screening and early intervention strategies.

By prioritizing clinical appropriateness over raw accuracy metrics, this work contributes to the growing body of evidence that domain-specific considerations must guide the development and evaluation of medical AI systems. Continued research, clinical validation, and user feedback will further enhance the effectiveness and impact of this innovative solution.

## 7. Bibliographic references

American Diabetes Association. (2020). Standards of medical care in diabetes—2020. *Diabetes Care*, 43(Supplement 1), S1–S212.

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016). RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in Neural Information Processing Systems*, 29, 3512–3520.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118. <https://doi.org/10.1038/nature21056>

International Diabetes Federation. (2019). IDF diabetes atlas (9th ed.). Retrieved from <https://diabetesatlas.org>

Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*, 16(1/2), 94–102. <https://doi.org/10.1016/j.aci.2018.12.004>

Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, 6, 26094. <https://doi.org/10.1038/srep26094>

Powers, A. C., & D'Alessio, D. (2011). Endocrine pancreas and pharmacotherapy of diabetes mellitus and hypoglycemia. In L. L. Brunton, B. A. Chabner, & B. C. Knollmann (Eds.), *Goodman & Gilman's the pharmacological basis of therapeutics* (12th ed., pp. 1237–1268). New York, NY: McGraw-Hill.

World Health Organization. (2021). Diabetes fact sheet. Retrieved from <https://www.who.int/news-room/fact-sheets/detail/diabetes>

Xie, J., & Tang, Y. (2017). A comprehensive review of image-based methods for automated classification of diabetic retinopathy. *Computerized Medical Imaging and Graphics*, 58, 1–9. <https://doi.org/10.1016/j.compmedimag.2017.03.005>

## AI Driven Predictive Model for Diabetes Risk Assessment

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Frontiers in Genetics*, 9, 515. <https://doi.org/10.3389/fgene.2018.00515>

Gonçalves, H., Silva, F., Rodrigues, C., & Godinho, A. (2024). Navigating the digital landscape of diabetes care: Current state of the art and future directions. *Procedia Computer Science*, 237, 336–343. <https://doi.org/10.1016/j.procs.2024.05.113>

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (2nd ed.). O'Reilly Media.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347–1358. <https://doi.org/10.1056/NEJMra1814259>

Chollet, F. (2018). *Deep learning with Python*. Manning Publications.

Django Software Foundation. (n.d.). Django documentation. Retrieved from <https://docs.djangoproject.com>

Amazon Web Services. (n.d.). AWS documentation. Retrieved from <https://aws.amazon.com/documentation/>

Kaggle. (n.d.). Kaggle datasets. Retrieved from <https://www.kaggle.com>

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media.

Emmanuel, A. (2020). *Building machine learning-powered applications: Going from idea to product*. O'Reilly Media.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

## AI Driven Predictive Model for Diabetes Risk Assessment

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, 28(1), 75-105.

## 8. List of Figures

Figure 1: Random Forest

Figure 2: Preprocessing

Figure 3: Logistic Regression

Figure 4: Serialization

Figure 5: Home Page

Figure 6: Registration Page

Figure 7: Profile Page

Figure 8: Diabetes Prediction Input Page

Figure 9: Diabetes Prediction Results

Figure 10: AI Insights Page with Feature Importance Visualization

Figure 11: Comprehensive Diabetes Health Tips Page

Figure 12: System Architecture

Figure 13: Data Flow Diagram (DFD)

Figure 14: Entity-Relationship Diagram (ERD)

Figure 15: Class distribution in the BRFSS 2015 dataset

Figure 16: Implementation of the preprocessing pipeline using scikit-learn's ColumnTransformer

Figure 17: Random Forest model configuration with class balancing implemented through scikit-learn Pipeline

Figure 18: Comprehensive model evaluation results

## 9. List of Tables

**Table 1:** Comparative performance metrics for Random Forest and Logistic Regression models on the test set (n=50,736).

**Table 2:** Impact of class balancing on Random Forest performance.

**Table 3:** Comparison of model performance with published literature