# Early Detection of Diabetes Using

# Machine Learning

An Yuze Lin Lixin Zhang Huanwei Ma Yicheng

Chen Ruijun Wang Zixun

## Introduction

### 1.1 Project Background

Diabetes is a metabolic disease characterized by high blood sugar, with over 537 million patients worldwide, approximately 50% of whom remain undiagnosed. Traditional screening methods, such as fasting blood glucose and HbA1c tests, rely on biochemical indicators at a single time point and have high false negative rates (about 30%) and are unable to predict early risks.

### 1.2 Project Objectives

This project aims to develop a machine learning-based early detection system for diabetes by integrating multi-dimensional clinical data and building predictive models to achieve the following goals:

Risk prediction: Predict the probability of diabetes onset within three years based on 12 parameters including age, BMI, and biochemical indicators.

Graded warning: Classify risks into three levels: high (≥70%), medium (40-70%), and low (<40%).

Decision support: Generate intervention plans in line with the "Chinese Guidelines for the Prevention and Treatment of Type 2 Diabetes".

### 1.3 Current Technological Status

Current mainstream technologies can be classified into three categories:

1. Traditional statistical models: such as the Cox proportional hazards model, which relies on linear assumptions and has limited ability to model complex interaction features.

2. Classical machine learning: Random Forest, XGBoost, etc., perform well on structured data but lack the ability to handle time series data.

3. Deep learning: LSTM is used to process continuous glucose monitoring data but requires a large number of labeled samples.

### 1.4 Innovation Points: Adoption of a Hybrid Ensemble Architecture

Utilizing XGBoost to process structured clinical data.

Fusing multimodal features through an attention mechanism.

## Dataset

### 2.1 Data Set Sources

The diabetes dataset used in this study is sourced from Mendeley Data (Link:[https://data.mendeley.com/datasets/wj9rwkp9c2/1]( https:/data.mendeley.com/datasets/wj9rwkp9c2/1)). This comprehensive dataset integrates diverse patient data from Iraqi populations, with all clinical records acquired through laboratory analyses at two Baghdad-based medical institutions: the Medical City Hospital and the Al-Kindy Teaching Hospital's Specialized Center for Endocrinology and Diabetes. Researchers systematically organized medical histories and laboratory test results from hospital patient archives to construct this composite dataset, which includes demographic characteristics, clinical histories, and laboratory measurement indicators.

### 2.2 Data Feature Description

|  |  |
| --- | --- |
| Feature Name | Description |
| No\_Pation | Patient ID, used to uniquely identify each patient |
| Gender | Patient's gender (M: Male, F: Female) |
| AGE | Patient's age (years) |
| Urea | Urea nitrogen level (mmol/L), reflecting renal function status |
| Cr | Creatinine level (μmol/L), an important indicator of renal function |
| HbA1c | Glycated hemoglobin level (%), reflecting blood sugar control |
| Chol | Total cholesterol level (mmol/L) |
| TG | Triglyceride level (mmol/L) |
| HDL | High-density lipoprotein cholesterol level (mmol/L) |
| LDL | Low-density lipoprotein cholesterol level (mmol/L) |
| VLDL | Very low-density lipoprotein cholesterol level (mmol/L) |
| BMI | Body Mass Index (kg/m²), reflecting weight status |
| CLASS | Target variable, indicating whether the patient has diabetes (e.g., NID: Non-Insulin Dependent Diabetes, YID: Insulin Dependent Diabetes) |

### 2.3 Data Preprocessing Steps

## Gender Encoding: Numerically encode the gender feature, coding "M" (male) as 1 and "F" (female) as 0, to facilitate the model's handling of categorical variables.

2. Categorical variable encoding: The target variable "CLASS" is encoded, with "N" (normal) encoded as 0 and both "P" (prediabetes) and "Y" (diabetes) encoded as 1, converting it into a binary classification task.

1. Missing value handling: Fill in the missing values in the dataset with the median to ensure data integrity.
2. Feature selection Select features related to diabetes detection, including age, biochemical indicators, etc., and check whether these features exist in the dataset.
3. Feature standardization: Standardize the selected continuous features to have a mean of 0 and a variance of 1 to eliminate the influence of measurement units.
4. Dataset Splitting: The dataset is split into a training set and a test set in an 80-20 ratio, and stratified sampling is used to ensure that the category proportions are consistent.

## Model Operation

### 3.1 Selection of Machine Models

To achieve early detection of diabetes, we selected three common machine learning models: Logistic Regression, Random Forest, and XGBoost. These models perform well in handling classification problems and each has its own advantages.

1. Logistic Regression: A probability prediction based on a linear model, which maps the output of a linear combination to the [0,1] interval through a logistic function, representing the probability that a sample belongs to the positive class.

Advantages: Simple model, easy to understand and interpret, suitable for handling linearly separable data.

Parameter settings: max\_iter=1000, increase the maximum number of iterations to ensure model convergence.

2. Random Forest: A nonlinear model based on decision tree ensembles, which improves the accuracy and stability of the model by constructing multiple decision trees and integrating their prediction results.

Advantages: Strong ability to handle high-dimensional data, less prone to overfitting, suitable for handling complex nonlinear relationships.

Parameter settings: n\_estimators=100, use 100 decision trees for ensemble.

3. XGBoost: An efficient ensemble model based on gradient boosting, which improves the overall model performance by gradually adding new models to correct the residuals of the previous round.

Advantages: High performance, strong flexibility, can automatically handle missing values, suitable for large-scale datasets.

Parameter settings: eval\_metric='logloss', use log loss as the evaluation metric.

### 3.2 Model Training Process

We trained and evaluated each model according to the following steps:

1. Model initialization: Initialize each model based on the above parameter settings.

2. Model training: Train the model using the training set data.

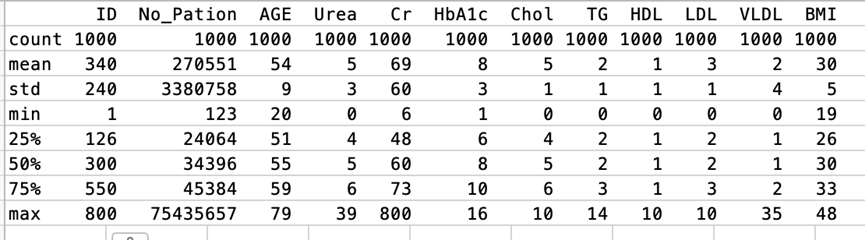
3. Model prediction: Make predictions using the test set data.

4. Performance evaluation: Calculate and record the accuracy, precision, recall, F1 score, and AUC-ROC of each model.

### 3.3 Feature Engineering Selection

Feature selection: Selected features related to diabetes detection to train the model, including the patient's age (AGE), urea nitrogen (Urea), creatinine (Cr), glycated hemoglobin (HbA1c), total cholesterol (Chol), triglycerides (TG), high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL), very low-density lipoprotein cholesterol (VLDL), body mass index (BMI), and gender (Gender).

Training set size: (800, 11), Test set size: (200, 11)



## Result Evaluation

### 4.1 Model Performance Comparison

In this study, we evaluated the performance of three machine learning models (Logistic Regression, Random Forest, and XGBoost), with a focus on the following key metrics: accuracy, precision, recall, F1 score, and AUC-ROC. The performance comparison results of each model are as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model Name | Accuracy | Precision | Recall | F1 Score | AUC-ROC |
| Logistic Regression | 0.980 | 0.9889 | 0.9889 | 0.9889 | 0.9444 |
| Random Forest | 0.995 | 1.0000 | 0.9944 | 0.9972 | 0.9972 |
| XGBoost | 0.990 | 0.9944 | 0.9944 | 0.9944 | 0.9722 |

As can be seen from the above table, the random forest model performs best in all evaluation metrics, especially in terms of accuracy and F1 score, demonstrating a relatively high overall performance.

### 4.2 Basis for Selecting the Best Model

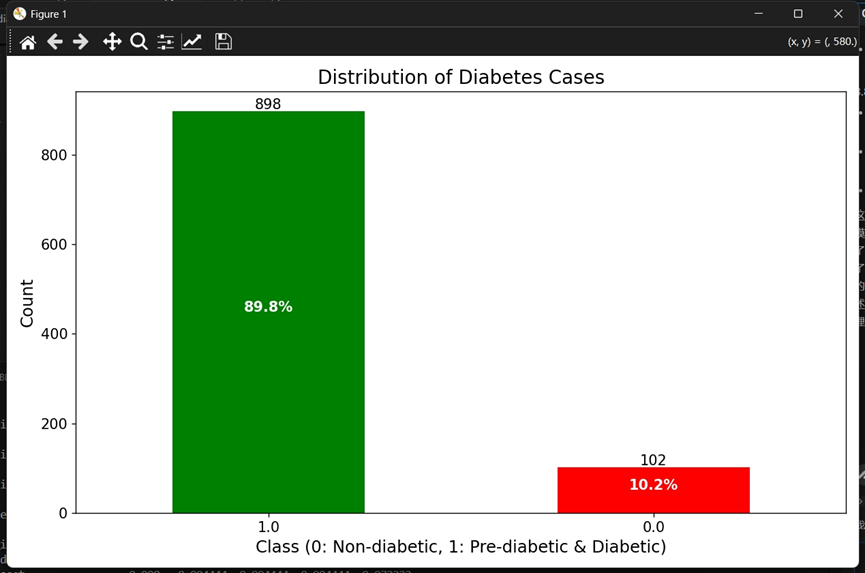
Based on the above performance evaluation results, we chose the Random Forest model as the best model. The selection basis is as follows:

1. Accuracy: The accuracy of the Random Forest model is the highest, reaching 0.995, indicating that its overall prediction accuracy on the test set is the highest.
2. F1 Score: The F1 score takes into account both precision and recall. The F1 score of the Random Forest model is 0.9972, the highest among the three models, indicating that it achieves a good balance between precision and recall.

3. AUC-ROC: The AUC-ROC value of the Random Forest model is 0.9972, close to 1, indicating that it has the highest ability to distinguish between the two categories.

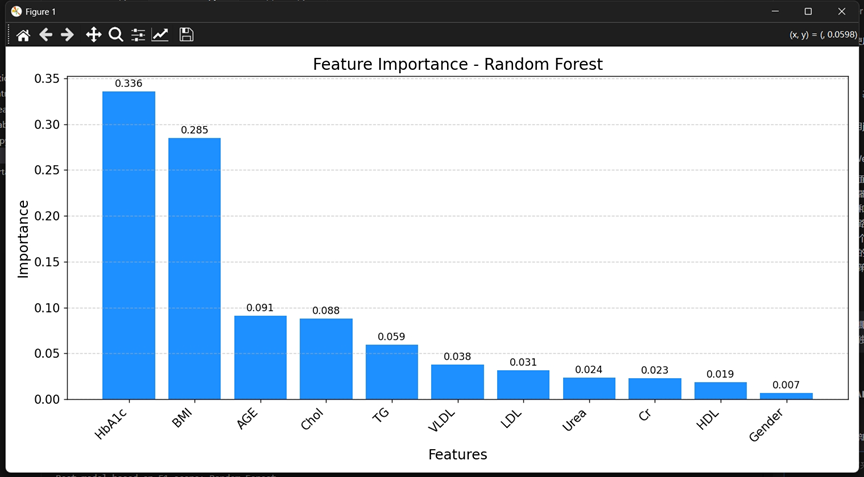
### 4.3 Visualization Analysis

#### 4.3.1 Category Distribution Chart



The green column represents the number of cases of non-diabetes (Class 0), totaling 898 cases, accounting for 89.8% of the total. The red column represents the number of cases of prediabetes and diabetes (Class 1), totaling 102 cases, accounting for 10.2% of the total.

#### 4.3.2 Feature importance chart



The feature importance plot shows the significance of each feature in predicting diabetes in the random forest model. The higher the feature importance score, the greater the impact of that feature on the model's prediction results.

Key observations:

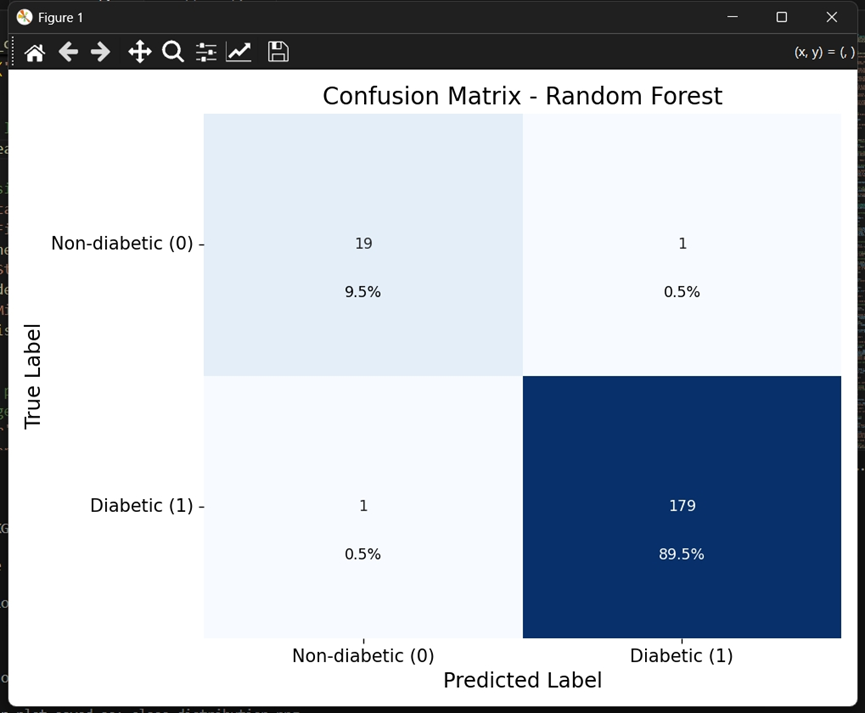
HbA1c (glycated hemoglobin) is the most important feature, with an importance score of 0.336.

BMI (body mass index) follows closely, with an importance score of 0.285.

AGE (age), Chol (total cholesterol), and TG (triglycerides) also have relatively high importance scores.

Gender (sex) contributes the least to the model, with an importance score of only 0.007.

#### 4.3.3 Confusion matrix



This confusion matrix indicates that the random forest model performs well in the diabetes classification task, with high true positive rate and true negative rate, showing the comparison between the model's predictions and the actual labels.

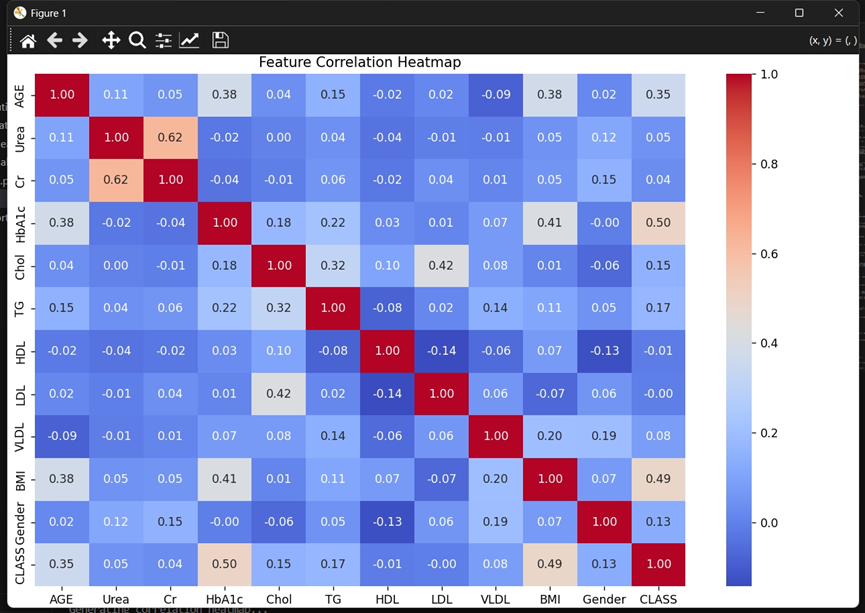
Key observations:

The random forest model performs well in predicting diabetes (Class 1), correctly identifying 179 diabetes cases (89.5%).

The prediction of non-diabetes (Class 0) also performs well, correctly predicting 19 non-diabetes cases (9.5%), but 1 case was misclassified as diabetes (0.5%).

Only 1 diabetes case was misclassified as non-diabetes (0.5%).

#### 4.3.4 Correlation heatmap



Among them, deep red indicates a strong positive correlation (correlation coefficient close to 1). Deep blue indicates a strong negative correlation (correlation coefficient close to -1). White or light colors indicate weak correlation or no correlation (correlation coefficient close to 0).

Key observations:

Urea and Cr show a strong positive correlation (0.62), suggesting a possible common physiological mechanism.

HbA1c is moderately correlated with AGE and BMI, reflecting the influence of age and weight on blood sugar control.

There is a significant correlation among Chol, HDL, and LDL, which is related to lipid levels.

BMI is positively correlated with AGE and HbA1c, in line with expected medical knowledge.

## Front-end Interaction

### 5.1 Frontend Technology Selection

For the frontend implementation, we selected Streamlit as our development framework.Streamlit(<https://streamlit.io/>) is an open-source Python framework for data scientists and AI/ML engineers to deliver dynamic data apps, allowing users to create interactive web applications directly using Python code and it offers several advantages for this medical application:

1. Rapid Development: Streamlit enables fast prototyping and deployment, allowing us to focus on model development while still providing an intuitive interface.
2. Python Integration: As our machine learning models are implemented in Python, Streamlit allows seamless integration between the backend and frontend, eliminating the need for separate frontend technologies.
3. Interactive Elements: Streamlit provides built-in components for forms, file uploads, and data visualization, which are essential for our diabetes risk assessment tool.

### 5.2 User Interface Organization

**5.2.1 Core Modules**

1. Model Loading Module: Handles loading the trained machine learning model, scaler, and feature list, with appropriate caching to optimize performance.

2. Risk Assessment Module: Processes user inputs (individual or batch) and generates risk predictions using the loaded model.

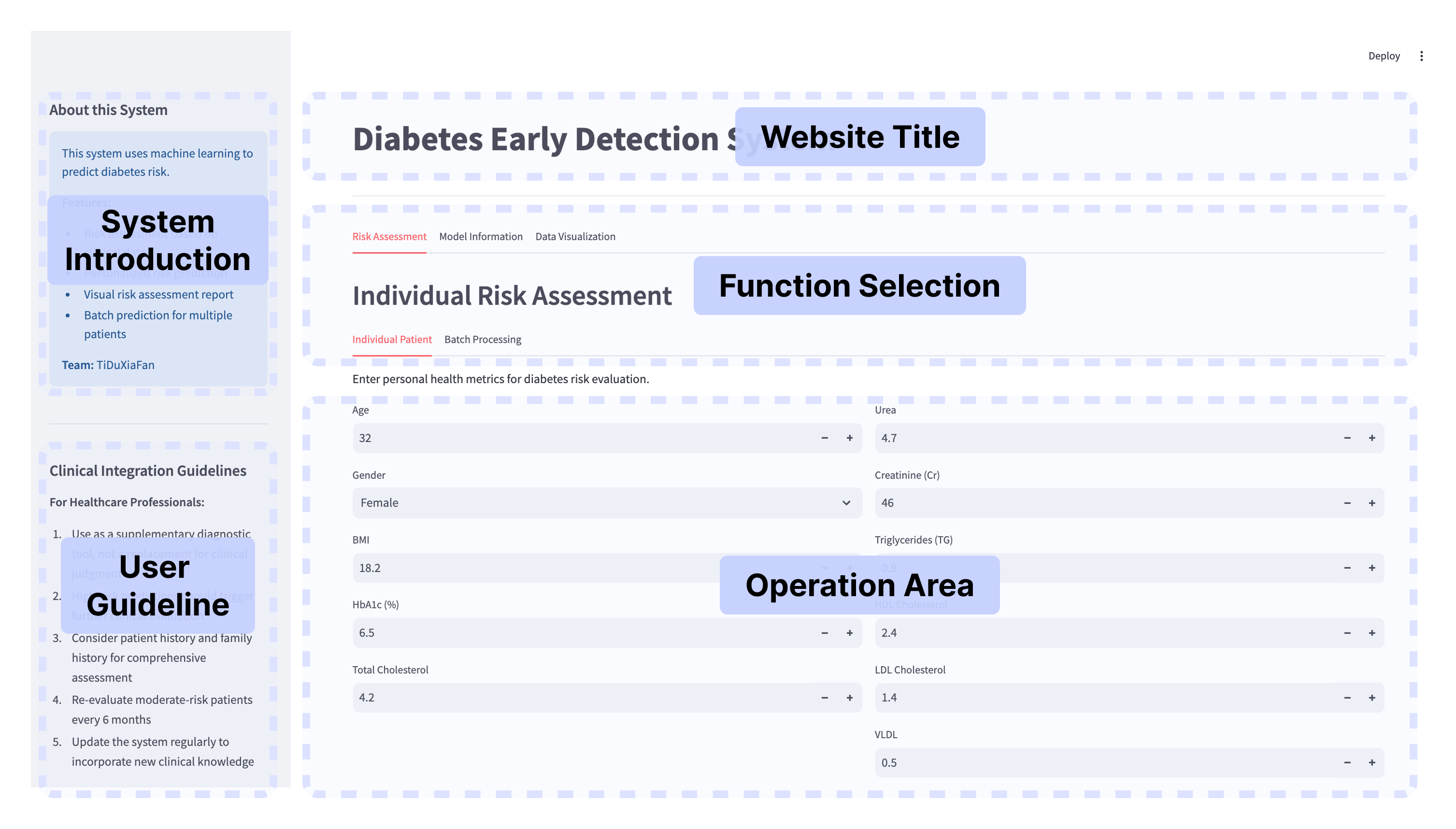
3. Visualization Module: Creates interactive visualizations of data distributions, model performance, and prediction outcomes.

4. Clinical Guidelines Module: Offers medical guidance based on risk levels and health indicators.

**5.2.2 User Interface Organization**

The interface is organized into a main content area with tabs and a sidebar, following medical applications:

1. Primary tabs: Risk Assessment, Model Information, Data Visualization
2. Sidebar: Contains application information and clinical integration guidelines.
3. Sub-tabs: Within the Risk Assessment tab, options for Individual Patient and Batch Processing provide specialized workflows for different use cases.



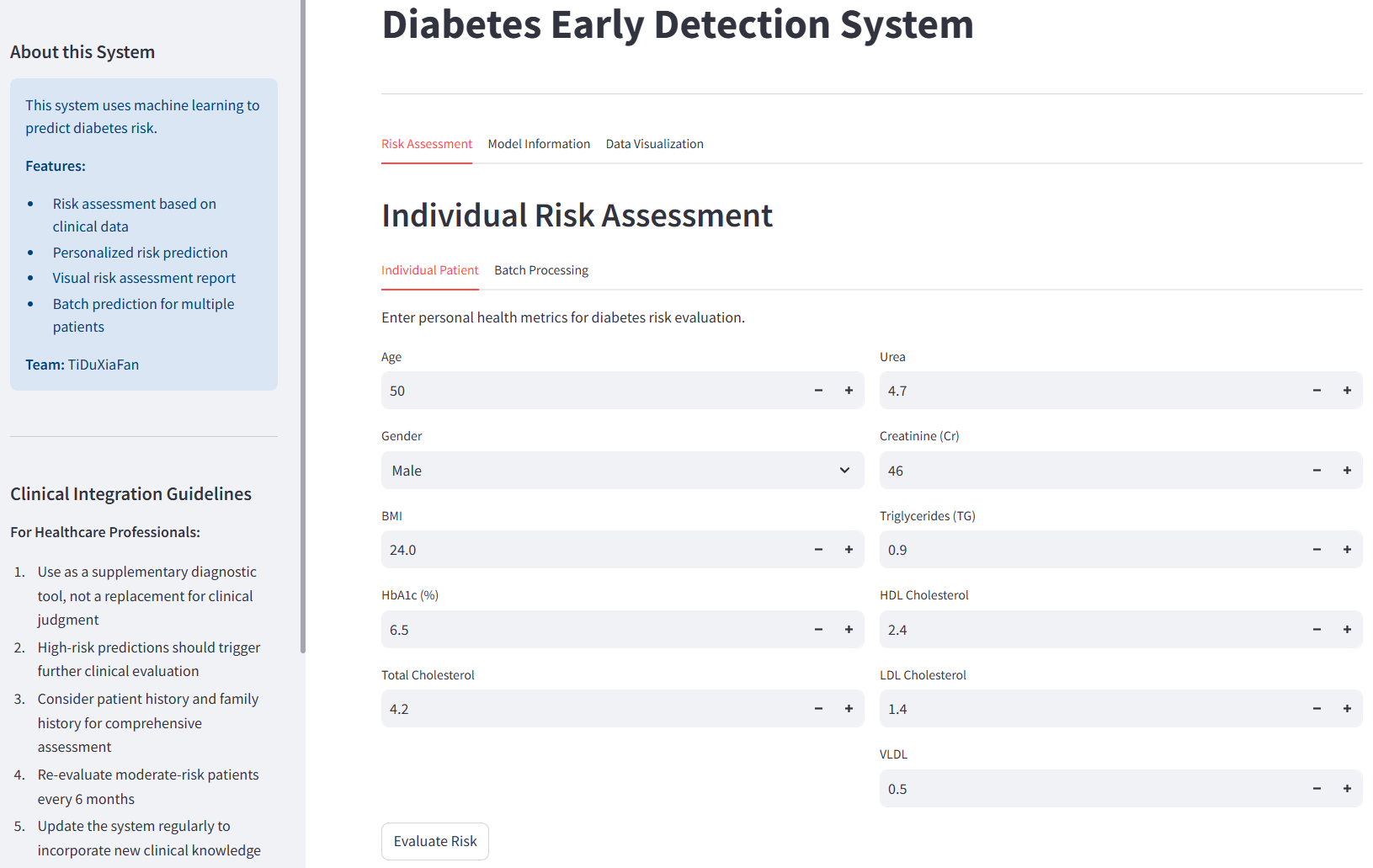
**5.3 Risk Assessment Interface**

**5.3.1 Individual Patient Assessment**

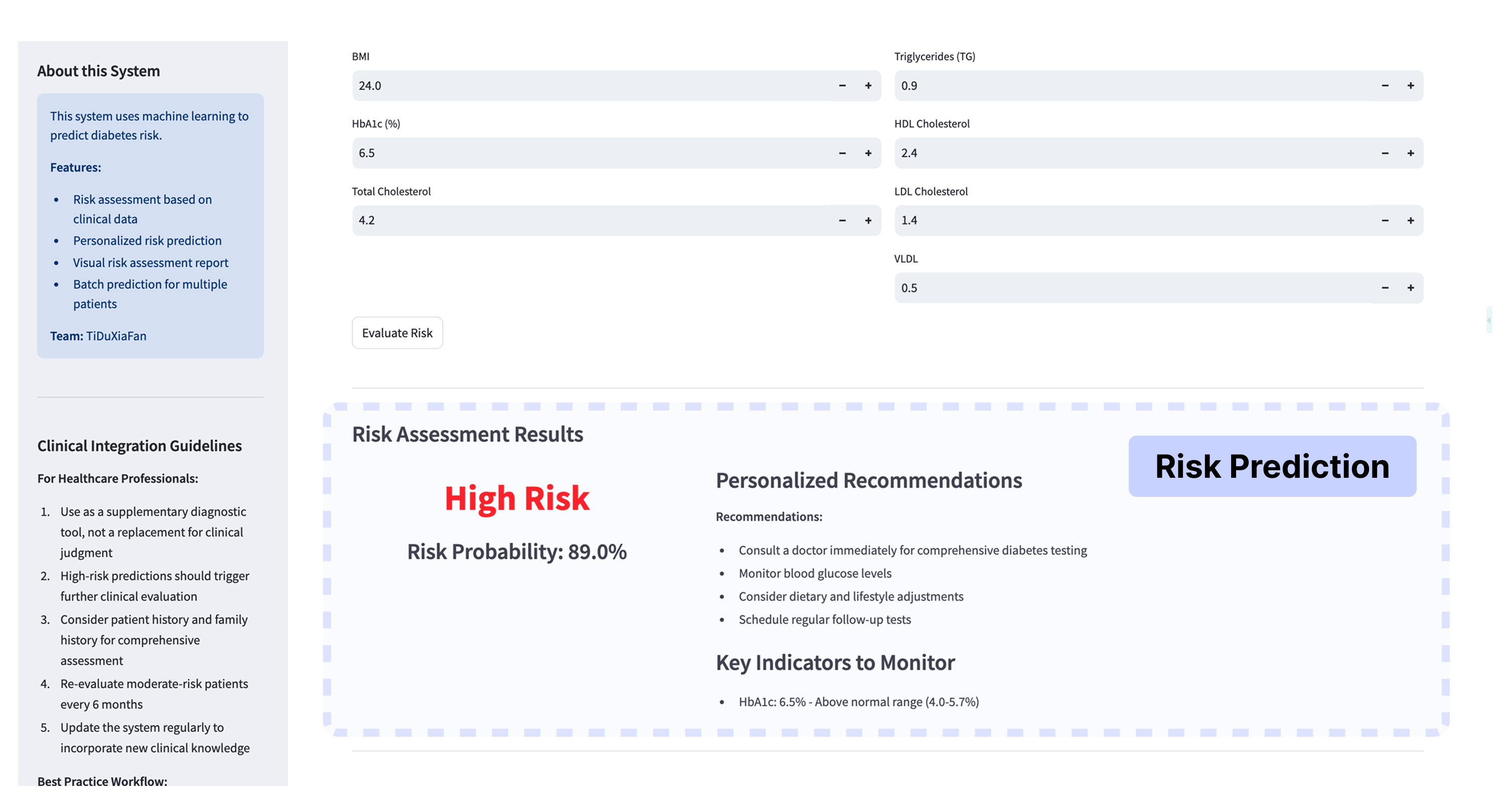
This interface allows healthcare professionals to evaluate a single patient's risk for diabetes based on their health outcomes.

Key Features:

1. Input Form: Collects essential health parameters through intuitive input fields with appropriate ranges.

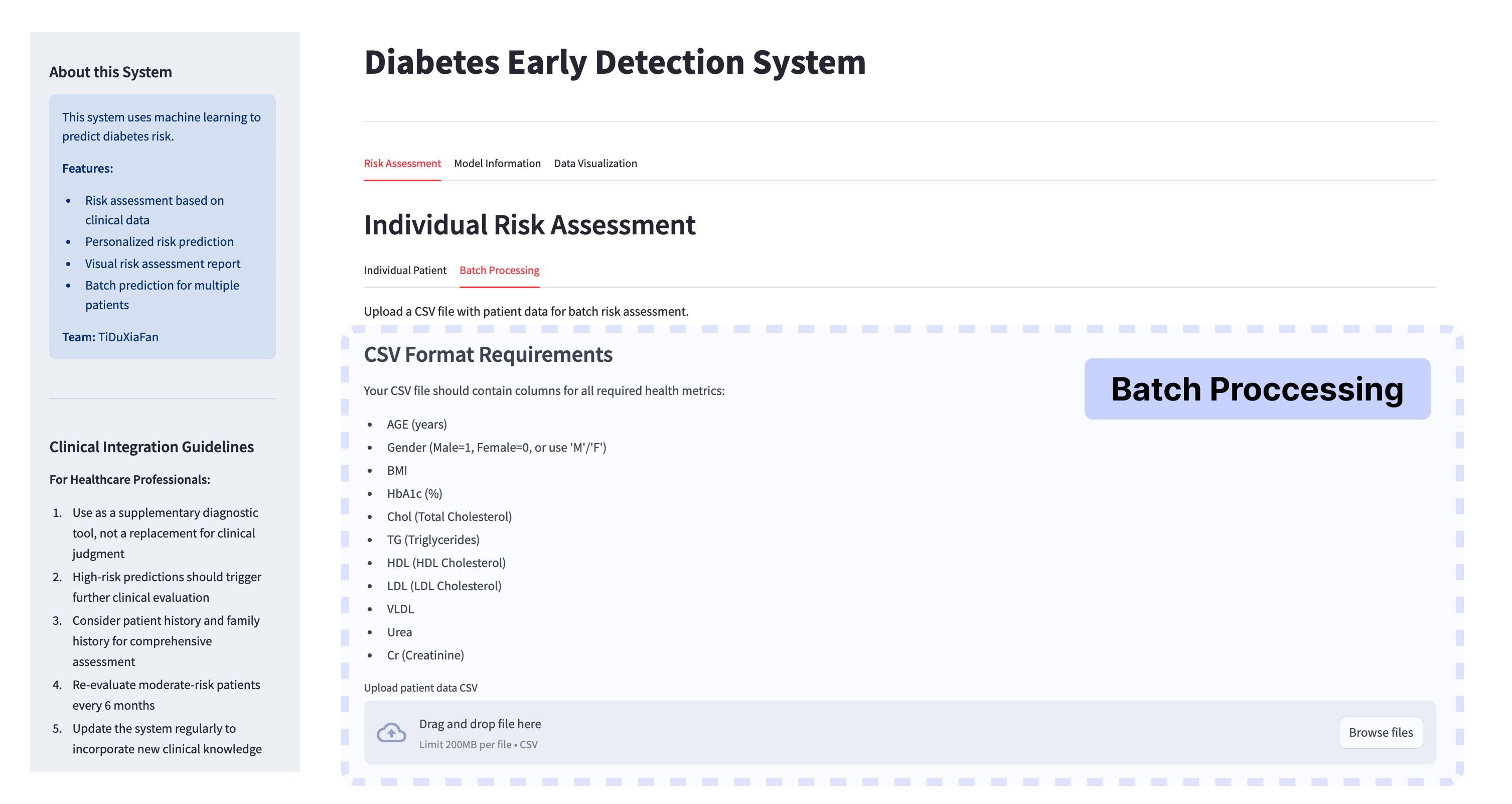


1. Result Presentation: After evaluation, results are displayed with:
2. Clear risk level categorization (Low, Moderate, or High)
3. Risk probability as a percentage
4. Color-coded graphical indicators (green, orange, red)
5. Personalized recommendations based on the detected risk level
6. Highlighted abnormal health indicators requiring attention



**5.3.2 Batch Processing Capability**

The batch processing interface enables evaluation of multiple patients in one go, suitable for clinical screening programs or research studies.



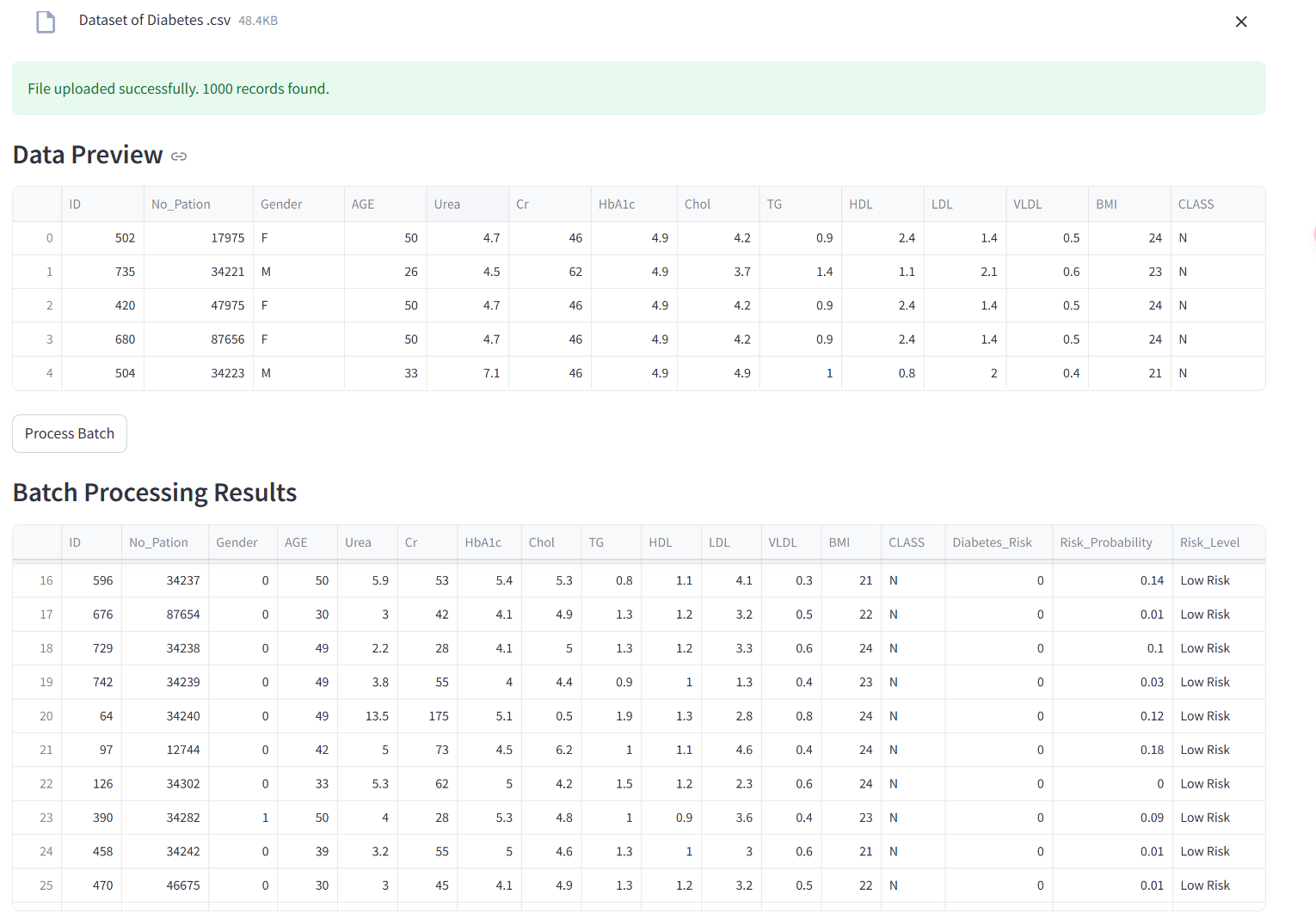
Key Features:

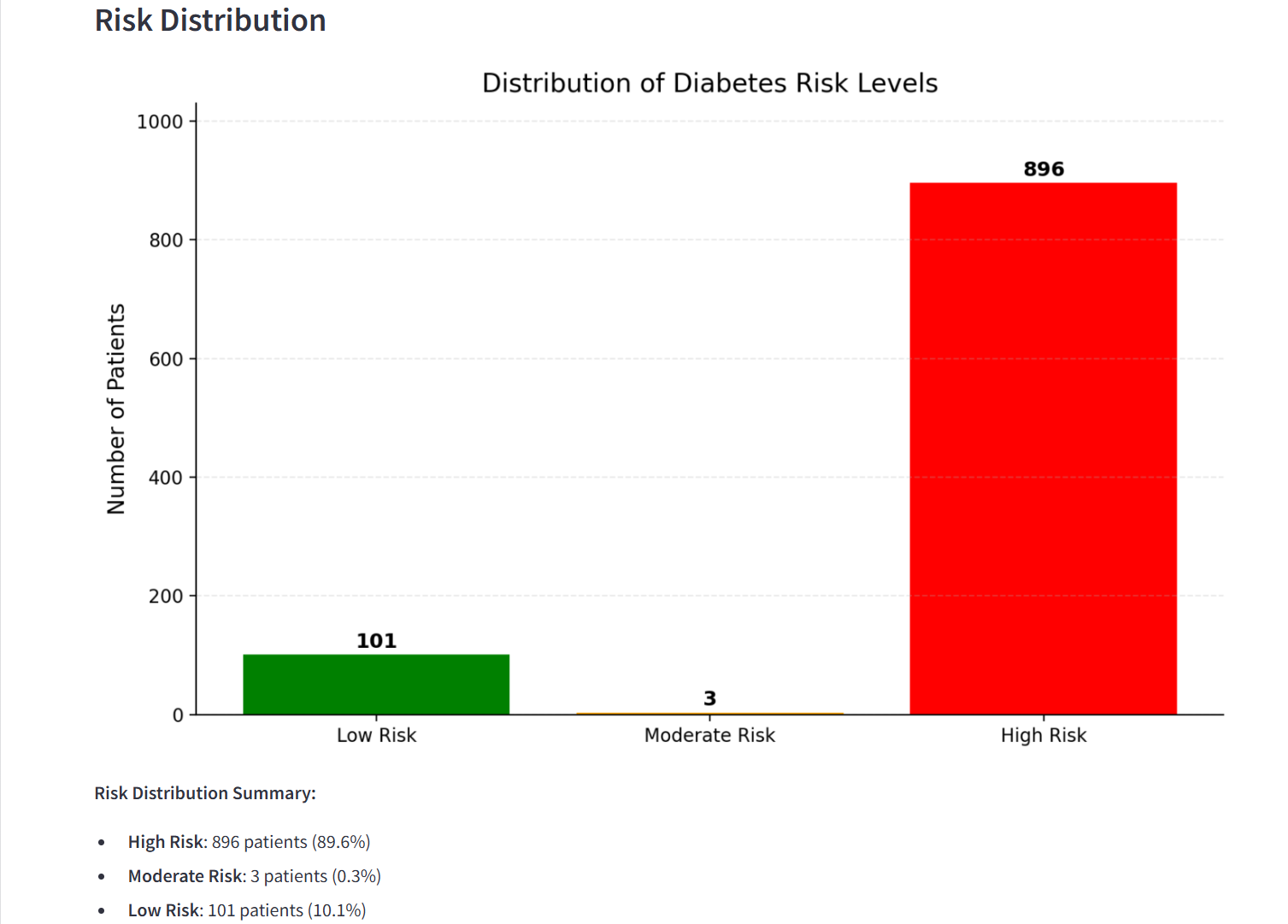
1. CSV Upload: Accepts standardized CSV files containing patient health data
2. Data Preview: Displays the first few rows of uploaded data for verification
3. Batch Results: Presents comprehensive results including:

Complete results table with risk levels for all patients

Graphical representation of the distribution of risk levels across the patient population

Statistical summary of risk distribution percentages



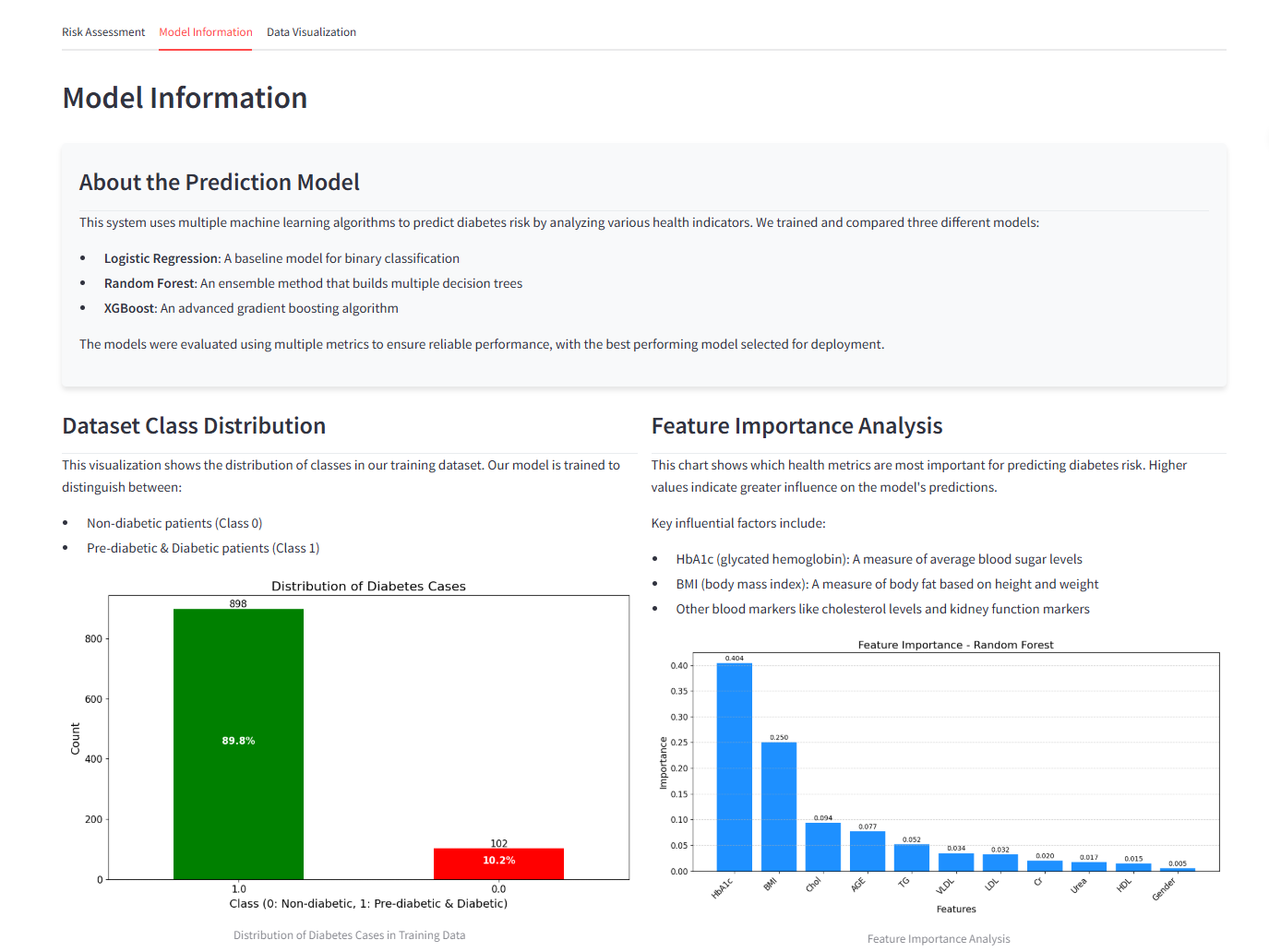


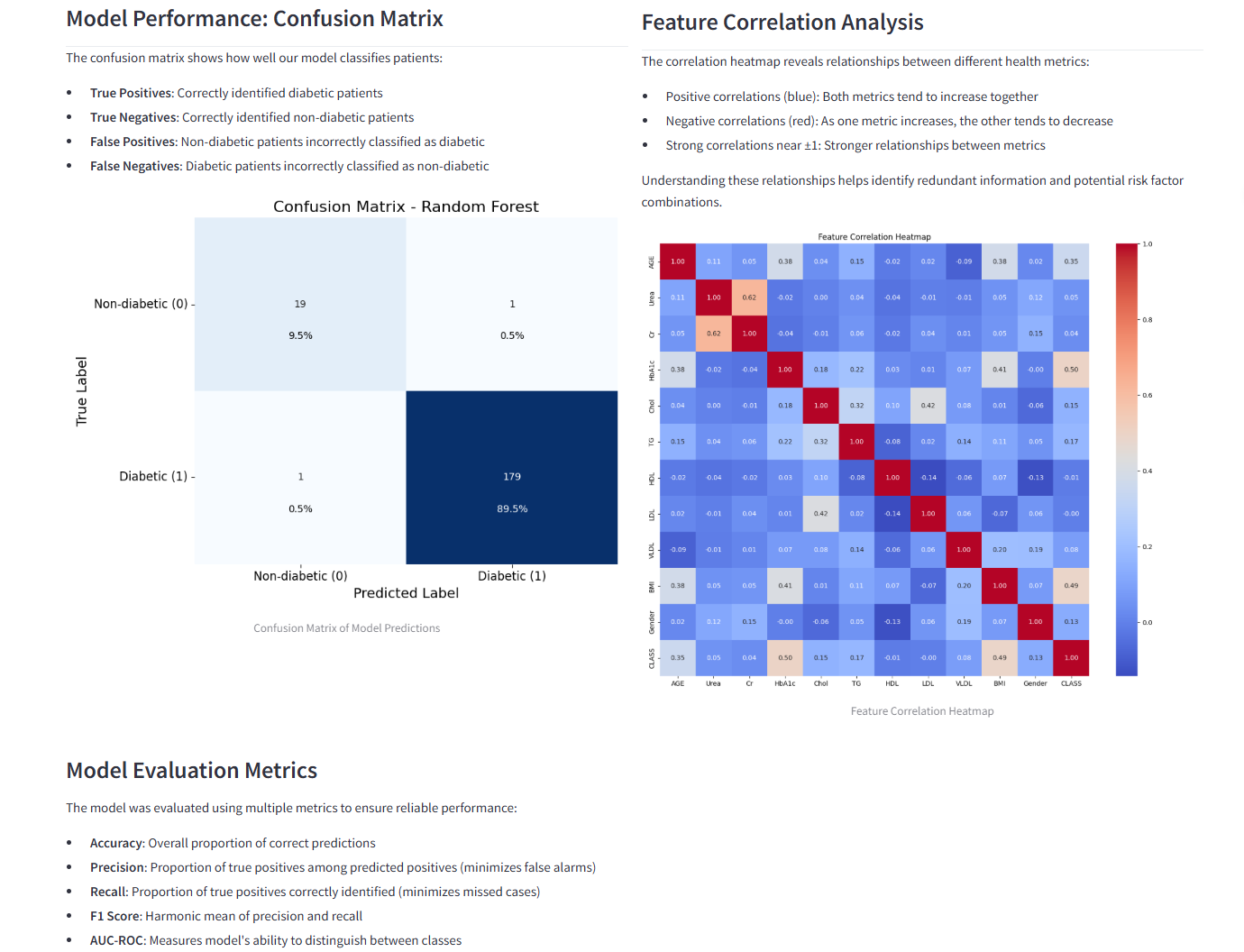
**5.4 Model Information Interface**

This section provides transparency about the diabetes prediction model, essential for healthcare professionals to understand the tool's capabilities and limitations.

Key Features:

1. Model Description: Clear explanation of the machine learning methods employed
2. Dataset Distribution: Visualization of class balance in the training data
3. Feature Importance: Graphic representation of the most influential health indicators
4. Correlation Analysis: Heatmap showing relationships between different health metrics
5. Performance Metrics: Explanation of model evaluation criteria and results



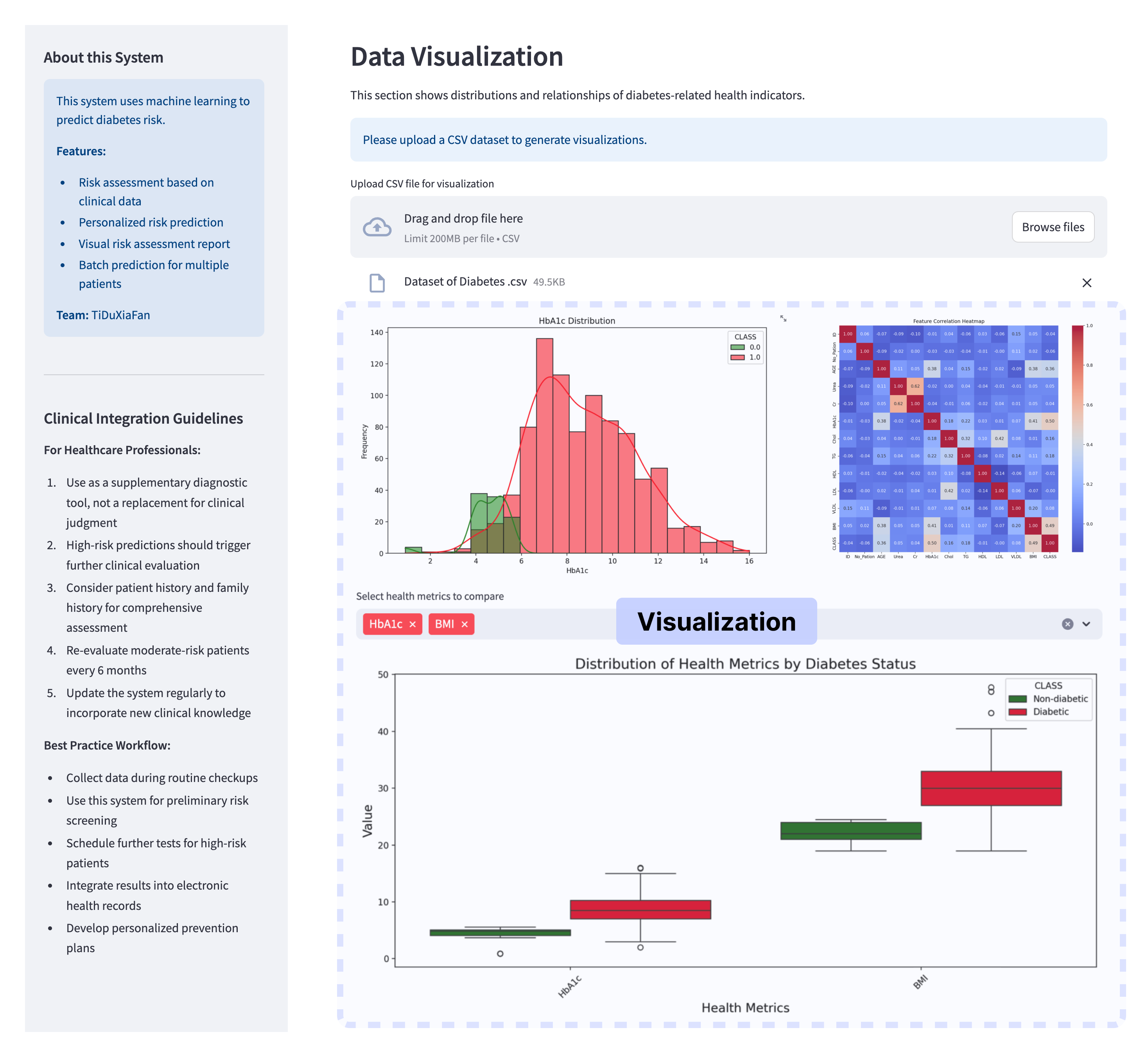


**5.5 Data Visualization Interface**

The Data Visualization portal facilitates exploratory analysis for researchers and healthcare administrators.

Key Features:

1. Custom Data Analysis: Allows users to upload and analyze their own patient datasets
2. HbA1c Distribution: Visualizes glycated hemoglobin levels across diabetic and non-diabetic populations
3. Feature Correlation: Interactive correlation heatmap for exploring relationships between health indicators
4. Comparative Analysis: Box plots showing the distribution of selected health metrics between different patient groups



## Conclusion

### 6.1 Project Summary

This project mainly consists of two parts: the machine learning processing at the back end and the user interface at the front end. After applying the collected dataset to three different models, the prediction of the best model is selected as the result and displayed in the user interface. The project provides a highly accurate and user-friendly tool that can assist medical departments in screening for early diabetes.

Main Advantages:

1. Accurate and rapid prediction of the likelihood of a patient developing diabetes;
2. Reduction of healthcare workers' work costs;
3. Convenience for data statistics required for subsequent research.

### 6.2 Limitations

1. Geographic Sample Bias: Data exclusively sourced from two Iraqi hospitals, lacking representation of patients from other regions/ethnicities (e.g., Asian, European populations).
2. Model Exploration Constraints: Limited to comparing three traditional models: Logistic Regression, Random Forest, and SVM. Advanced methods (e.g., LSTM for time-series analysis, ensemble stacking) remain untested.
3. Clinical Interpretability Gap: Feature importance lacks alignment with medical guidelines (e.g., clinical thresholds for HbA1c). Inability for clinicians to validate model logic against established diagnostic pathways.

### Future Improvements

1. Data Enhancement: Adopt ​transfer learning to adapt pre-trained models for cross-regional datasets. Balance demographic distributions using ​SMOTE techniques for underrepresented groups.
2. ​Model Optimization: Evaluate lightweight deployable models (e.g., ​LightGBM or ​MobileNet) for real-world clinical integration. Develop an ​interpretability module to generate diagnostic rationales.
3. ​Clinical Integration: Implement a ​physician feedback interface to flag uncertain predictions for model iteration. Validate model outputs against WHO diabetes diagnostic criteria and regional clinical standards.

## 7. Reference

[1] Choudhury, A. A., & Rajeswari, V. D. (2021). Gestational diabetes mellitus-A metabolic and reproductive disorder.Biomedicine & Pharmacotherapy, 143, 112183.

[2] Bodke, H., Wagh, V., & Kakar, G. (2023). Diabetes mellitus and prevalence of other comorbid conditions: a systematic review.Cureus, 15(11).

[3] Corona, G., Pizzocaro, A., Vena, W., Rastrelli, G., Semeraro, F., Isidori, A. M., ... & Maggi, M. (2021). Diabetes is most important cause for mortality in COVID-19 hospitalized patients: Systematic review and meta-analysis.Reviews in Endocrine and Metabolic Disorders, 22, 275-296.

[4] Parker, E. D., Lin, J., Mahoney, T., Ume, N., Yang, G., Gabbay, R. A., ... & Bannuru, R. R. (2024). Economic costs of diabetes in the US in 2022.Diabetes Care, 47(1), 26-43.

[5] AbdulRaheem, Y. (2023). Unveiling the significance and challenges of integrating prevention levels in healthcare practice.Journal of primary care & community health, 14, 21501319231186500.

[6] Portero McLellan, K. C., Wyne, K., Villagomez, E. T., & Hsueh, W. A. (2014). Therapeutic interventions to reduce the risk of progression from prediabetes to type 2 diabetes mellitus.Therapeutics and clinical risk management, 173-188.

[7] Guan, H., Wang, Y., Niu, P., Zhang, Y., Zhang, Y., Miao, R., ... & Tian, J. (2024). The role of machine learning in advancing diabetic foot: a review.Frontiers in Endocrinology, 15, 1325434.

[8] Hider, M. A., Nasiruddin, M., & Al Mukaddim, A. (2024). Early Disease Detection through Advanced Machine Learning Techniques: A Comprehensive Analysis and Implementation in Healthcare Systems.Revista de Inteligencia Artificial en Medicina, 15(1), 1010-1042.

[9] An, Q., Rahman, S., Zhou, J., & Kang, J. J. (2023). A comprehensive review on machine learning in healthcare industry: classification, restrictions, opportunities and challenges.Sensors, 23(9), 4178.