jamk

Explainable Artificial Intelligence for Personalized Diabetes Risk Prediction

Nardos Tesfai

Master's Thesis
April 2025
Master's Degree Programme in Information Technology, Full Stack Software
Development

jamk | Jyväskylän ammattikorkeakoulu
University of Applied Sciences

jamk

Description

Tesfai Nardos

Explainable AI for personalized diabetes risk prediction

Jyväskylä: Jamk University of Applied Sciences, April 2025, 67 pages

Master's Degree Programme in Information Technology Full Stack Software Development

Permission for open access publication: Yes

Language of publication: English

Abstract

Diabetes keeps affecting millions of people around the world, with the current estimation of 537 million and is expected to rise steadily in the coming decades. So, detecting it at an early stage is very crucial. Recognizing the risk factors at an early stage would help prevent further health complications in the future. Even though machine learning provides algorithms that can predict health risks, offering tools which can analyze complex patterns of medical data effectively than traditional methods. Many of these models lack transparency, so it makes it hard to understand and limits their trust, especially in health care.
Such difficulties have led to the development of XAI, which can assist in understanding why those predictive models have arrived at the decision. To clarify these issues the well-known tools SHAP and LIME are outstanding in revealing outcomes. SHAP uses a game theory principle to show how each features assign influences providing both global and local insight, while LIME provides simple easy to follow rules for each individual prediction. Synergistically they offer a surplus of details and simplicity.
Due to the balance of performance and simplicity random forest and Logistic Regression predictive models have been commonly applied to diabetes risk classification. With appropriate techniques of data pre-processing such as feature scaling, handling missing values, stratified data sampling both models can provide reliable estimation or predictions using the public dataset widely available medical inputs or metrics such as age, BMI, glucose, blood pressure then display using XAI making them transparent and user friendly.
Web based interface designed with usability can serve as effective platform to understand those predictive models with visual aids, educational tool tips for those non-technical users and clearly labelled output enhance comprehension and accessibility. Responses from technical and non-technical users show that combining visual and rule-based explanations facilitates greater understanding and builds confidence on the trust of the system predictions. Ease of interaction along with clear presentation significantly contributes to the user's trust and willingness to engage with those predictive tools.
There has been great emphasis on those predictive models that provide not only optimize accuracy but also offer clarity, ethical disclaimer and meaningful Interactions. In sensitive fields such as health care appliances where lives are affected by decisions, the combination of performance and interpretability is not optional but essential. With continued advancement in model explanation techniques as well as interface design including evaluation by professional medical staff and performing clinical trial will be key to shaping trustworthy AI-driven decision support tools in the future.

Keywords/tags (subjects)

XAI, Machine learning, SHAP and LIME

Contents

Figures

1 Introduction

1.1 Background and problem context

Diabetes mellitus is one of the most serious health issues we face today which affects around 537 million adults across the globe, with estimates suggesting this number could grow to around 783 million by the year 2045 (International Diabetes Federation, 2021). This condition develops gradually and is associated with several serious health complications such as heart disease, kidney failure, and nerve damage, which in turn put a lot of pressure on our health care systems. To help slow down or even stop the onset of these problems, it is very important to spot the warning signs as early as possible so that doctors can take the necessary steps to help patients before things get worse.

In recent years, artificial intelligence has become an essential tool in health care because it can help predict risks in ways that go far beyond what traditional statistics can do. Machine learning models can look through large amounts of medical data and pick up on patterns that might not be obvious to humans and this ability has the potential to change the way risks are predicted and how decisions are made in clinics and hospitals. While these new methods have shown good results in predicting several diseases which include the risk of diabetes, one major issue is still there: many of these systems work like a black box, meaning that while they can predict outcomes very well, they do not give a clear explanation for how they arrive at their decisions.

This lack of clear explanations creates a big hurdle for the use of these systems in everyday health care settings. Many doctors feel uncomfortable using recommendations from systems when they cannot see or understand the reasoning behind those recommendations, and patients may also be hesitant to accept a technology-based health assessment if they cannot follow the thought process behind it. This situation has led to the rise of Explainable AI (XAI) techniques, which try to make the decision-making process of these models more transparent without losing their ability to predict accurately.

Two techniques in Explainable AI that have gained a lot of attention are SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). SHAP uses ideas from a branch of mathematics that deals with how groups share rewards to assign a value to each input feature, offering a clear and consistent way to explain outcomes (Lundberg and Lee, 2017). On the

other hand, LIME works by creating a simpler model that approximates the behaviours of the complex model for a specific prediction, which helps to give a straightforward explanation for that prediction (Ribeiro et al., 2016). Both methods give additional value by giving a better view at how the models work which can help simplify the worries that can come with using systems that are hard to understand in health care.

This research focuses on the important connection between predicting diabetes risk and making sure the prediction process is clear to everyone involved. The work is meant to close the gap between making accurate predictions and making sure that those predictions are easy to understand by the people who use them. By putting SHAP and LIME into practice, this project hopes to build more trust in the technology, allowing doctors and patients alike to feel more comfortable with the use of these tools. In addition, by showing how these models can be integrated into a web application, the research provides a practical example of how explainable AI can be used for personalized health risk assessments.

The work described in this thesis goes beyond just creating a new technical tool; it also tackles a larger issue in health care today. As more and more decisions in health care are made with the help of data, it is essential that the systems used remain open and easy to follow. The ideas and methods presented in this work add to the conversation about using AI responsibly in medical settings and they have the potential to influence clinical practice how patients are involved in their own care and even health policies.

This thesis takes a close look at how machine learning models can be used to predict the risk of diabetes, with a special focus on how clear and understandable these models can be when modern XAI techniques are applied. Through detailed comparisons and practical examples, the research provides new insights into how to achieve a balance between making accurate predictions and ensuring that the predictions can be easily followed by both medical staff and patients.

1.2 Aim and purpose

The research develops and evaluates an intelligible AI method for individual diabetes risk prediction while ensuring high prediction capabilities together with comprehensive explanations that the public can understand. This research fulfills its goal by establishing several distinct objectives as described in this study.

- To train two machine learning models including Logistic Regression and Random Forest to evaluate their performance in predicting diabetics risk using Pima Indians dataset.
- To apply and evaluate the effectiveness of explainability techniques, namely SHAP and LIME, in shedding light on how the models come to their conclusions.
- To build a web application that is easy to use and that shows both the predictions made by the models and the reasons behind those predictions.
- To evaluate system performance including accuracy assessment together with precision and recall calculation and F1-score measurement among other evaluation measures alongside clear explanation assessment.

This research aims to construct a technical system for AI risk prediction as well as solve the fundamental problem of earning healthcare professional and patient acceptance of these systems for risk estimation.

1.3 Research Questions

The research focus on exploring three essential points:

- What performance-related scores are obtained from the accuracy, precision and recall testers and F1-score between Logistic Regression and Random Forest when used for diabetes risk prediction?
- How does the explanation provided by SHAP and LIME tools improve users understanding of predictive models and affect the system trust worthiness?
- What challenges and best practices appear when trying to deploy a system that uses explainable AI for predicting diabetes risk in a web application especially when considering factors such as ease of use, technical integration and ethical issues?

1.4 Limitations

There are some limitations in this study that are important to mention.

- **Dataset Constraints:** The Pima Indians Diabetes dataset includes only 768 cases and is focused on a specific group (women of Pima Indian heritage), which means that the findings might not apply to other groups.
- **Model Range:** This research only looks at two machine learning models Logistic Regression and Random Forest and does not cover the many other types of models that exist.
- **Scope of Explanations**: Both SHAP and LIME stand among the best explanation techniques but they face specific limitations when analyzing correlated input features.
- **Validation Process:** The evaluation in this study relies mostly on technical data analysis together with user feedback instead of sustained clinical trials or extensive research.

1.4 Ethics

Ethical considerations are a very important part of this work, especially because it deals with health care:

Data Privacy: Although the Pima Indians dataset is available to the public and does not contain personal identifiers, proper care has been taken in handling the data, which is something that would also be necessary for real patient data.

Fairness and Bias: This study acknowledges the potential risk of obtaining group-based biases when data-driven models are applied to different population segments through appropriate considerations.

Clear Communication: The web application explicitly informs users about when the system provides risk estimates since those predictions do not substitute clinical diagnoses. Professional medical advice needs to be considered essential for decision-making about personal health.

Openness: The way the models make their decisions is explained through XAI techniques which helps to make sure the process remains open, and that the system is in line with ethical guidelines for using AI in health care.

1.5 Contribution to SDGs

This work fits well with and adds to the United Nations Sustainable Development Goals (SDGs) in several important ways. It helps with:

SDG 3: Good Health and Well-being: By creating tools that predict the risk of diabetes early, it supports methods that help prevent the disease, which is a major health problem worldwide.

SDG 9: Industry, Innovation and Infrastructure: The study attracts new ideas and incorporate it into health care technology with focus on building responsible AI systems.

SDG 10: Reduced Inequalities: By making sure AI tools easier to follow and more understandable for a wide range of users it can help more people have access to technology.

1.6 Thesis structure

The thesis is arranged in several sections. It starts with a Literature Review that looks at recent work on predicting diabetes with AI, discusses how explanation tools like SHAP and LIME play a role and compares Logistic Regression with Random Forest for medical diagnosis. After that the Background Theory chapter comes in and explains the basic ideas behind machine learning classification, describes the SHAP and LIME methods and talks about ethical issues in healthcare health care AI. The Methodology chapter covers the dataset used, the steps taken to prepare the data, the training of the models, how the explanation techniques were put into practice, the overall system setup and the measures used to check performance. After that, the Implementation chapter describes how the models and explanation tools were brought together in a web platform. The Results chapter presents how both models performed, showing various performance measures, confusion matrices, and examples of the explanations provided by SHAP and LIME. In the Discussion chapter, the results are interpreted, looking at user experience, trust, the clarity of the models, practical implications, as well as limitations and ethical issues. Finally, the Conclusion and Future Work chapter summarizes the main findings and offers suggestions for future research.

2 Review of literature

2.1 Machine Learning Concepts for Classification

Machine learning classification is about teaching algorithms to predict specific outcomes based on labelled data. For a task like predicting diabetes which involves a yes or no decision. These algorithms work to find a boundary that best separates the two groups.

Logistic Regression (LR)

Although its name might suggest otherwise, Logistic Regression is a classification algorithm. It applies a logistic function to produce a probability between 0 and 1 which indicates the chance that a particular case belongs to one group. In this model a coefficient is estimated for each feature to show how much the odds change when that feature increases by one unit. One of the main reasons for its popularity in health care is that its output is easy to interpret as the coefficients clearly show the effect each feature has on the outcome (Christodoulou et al., 2019).

The logistic function transforms the linear output into a probability with this formula:

$P(Y=1) = 1/(1+e^{(-z)})$

Here, z is a combination of the features expressed as $z = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$. The coefficients (β values) are calculated by finding the values that make the observed outcomes most likely typically using methods like Newton-Raphson or stochastic gradient descent (Blagus & Lusa, 2017).

LR makes several assumptions: the observations should be independent, there should be little to no high correlation among predictors, there must be a linear relationship between the log of the odds and the predictors, and the sample size should be large enough in relation to the number of predictors. In medical applications such as diabetes prediction, Logistic Regression is favoured because it is based on strong statistical principles and its results can easily be converted into odds ratios that are meaningful for clinicians (Nusinovici et al., 2020).

Random Forest (RF)

Random Forest is an ensemble method that builds many decision trees during training and then decides on the final prediction based on the majority vote from these trees. Each tree is constructed using a random sample of the data and a random subset of features at each split. This randomness helps create diversity among the trees and reduces the risk of overfitting, which is

when a model works very well on training data but poorly on new data. RF is good at capturing relationships that might be missed by simpler linear models, and while it is sometimes seen as harder to understand than LR, newer explanation methods have helped make its predictions clearer (Nembrini et al., 2018).

The effectiveness of Random Forest comes from two main processes. First, bagging is used, where each tree is built from about 63.2% of the original dataset, with the rest used for internal validation. Second, at each split, only a random subset of features is considered, which stops any one feature from dominating the model (Probst et al., 2019). Breiman's original work on Random Forest brought in several new ideas that contribute to its success. The model naturally selects important features by measuring how much each feature reduces impurity (often measured by Gini impurity) when it appears in the trees. However, these traditional measures can sometimes favour features that have many unique values (Boulesteix et al., 2015).

Other Classification Algorithms

Other methods worth mentioning include Support Vector Machines (SVM), which find the best dividing line between classes using kernel functions; Gradient Boosting Machines, which build trees one after another to fix errors from earlier models; and Neural Networks, which use layers of interconnected nodes to learn patterns. Each of these methods offers a different balance between performance, ease of explanation, and the computing power needed (Fawcett, 2016).

Other important consideration include:

Overfitting: When models learn noise in the training data rather than generalizable patterns. This manifests as excellent performance on training data but poor generalization to unseen examples. Regularization techniques and careful validation strategies help mitigate this risk (Kuhn & Johnson, 2019).

Regularization: Techniques like adding penalties (L1 or L2) in LR or setting limits on tree depth in RF help prevent overfitting. L1 regularization (Lasso) can set some coefficients to zero, effectively selecting the most important features while L2 regularization (Ridge) shrinks the size of all coefficients. Elastic Net combines both methods (Zou & Hastie, 2017).

Feature Engineering: This is the process of creating, transforming, or selecting features to improve how well the model performs. For diabetes risk prediction, this might include transforming skewed variables such as insulin levels, creating new features that combine factors like body mass index

and age or grouping continuous variables into categories that make sense in a medical context (Khalid et al., 2014).

Class Imbalance: When one class is more common than the other, it can cause problems for the model. In the Pima Indians dataset, there is a moderate imbalance with roughly 65% of cases being negative and 35% positive. This can be managed by adjusting class weights, reducing the number of cases in the majority class or using methods like SMOTE to create synthetic examples of the minority class (Fernández et al., 2018).

Hyperparameter Optimization: This involves finding the best settings for the model that are not directly learned from the data. Methods include grid search which tests every combination of parameters; random search, which picks parameters at random; and Bayesian optimization, which uses past results to focus on the most promising settings (Yang & Shami, 2020).

2.2 Explainable AI Techniques

Explainable AI covers a set of methods that aim to shed light on how AI models reach their decisions clearly to people. In health care, where knowing why a particular prediction is made is very important, these techniques carry special weight (Arrieta et al., 2020). The need for such clarity comes from many important reasons—clinical, legal, ethical, and practical. From a clinical point of view, clear explanations help verify that the model's outcomes make sense medically and can uncover any bias present in the predictions. Legally, the law demands that systems which influence patient care must provide complete explanations when making their decisions. The model requires clear decision logic for patients' wellness protection along with error detection assistance to boost model effectiveness (Ahmad et al., 2018).

SHAP, which stands for SHapley Additive exPlanations, is based on ideas from game theory that were introduced by Lloyd Shapley back in 1953. In this approach, each feature in the data is seen as a participant in a game where the overall prediction is the result of their contributions, and each feature is given a value relative to an average baseline (Lundberg & Lee, 2017). Each SHAP value represents the difference in the prediction from the baseline, attributed to a particular feature because all contributions combine to reach the exact prediction value. Features that drive important outcomes gain increased value whereas unimportant attributes receive zero adjustment.

SHAP applies approximation techniques to handle exact value computations because determining all feature combinations grows unmanageable with rising feature numbers (Merrick & Taly, 2020).

The Local explanations that SHAP generates for single predictions become detailed when the values are combined into briader views of model functioning throughout the prediction set. There are several ways to implement SHAP such as through KernelSHAP which works with any model but can be slow or TreeSHAP which is optimized for tree-based models and offers much faster computation (Lundberg et al., 2020). KernelSHAP applies a weighted linear regression where the weights are based on the number of features that are present. While TreeSHAP takes advantage of the structure in decision trees to calculate the values more quickly moving from an exponential to a more practical polynomial time (Lundberg et al., 2018).

Visual representations of SHAP include force plots, which illustrate how each feature moves the prediction from the baseline toward its final value, summary plots that show the overall importance of features across the dataset, dependence plots that demonstrate how a feature's impact changes with its value, and waterfall plots that trace the cumulative effect of features for one prediction.

In contrast, LIME (Local Interpretable Model-agnostic Explanations) builds an easier model to replicate the complex model's local prediction behavior (Ribeiro et al., 2016). The initial operation of LIME generates multiple samples that form the surroundings of the targeted instance through minor changes to its features after which the complex model provides predictions for these modified cases. The samples are weighted based on how similar they are to the original instance and a simpler model (often using Lasso regression) is trained on this local data. The coefficients from the simpler model are then interpreted as local explanations. LIME's approach, which uses random sampling, is flexible enough to work with various data types like tables, text, and images, and it tends to produce straightforward explanations that concentrate on a few key features, which can make the results easier to follow (Molnar, 2020). However, because it relies on random sampling, the explanations produced by LIME may sometimes vary from one case to another (Slack et al., 2019).

Comparative studies of SHAP and LIME show that while SHAP generally gives more consistent

explanations with strong theoretical support, it can be heavier on computational resources. Although LIME provides rapid approximation results for mixed data types of its explanations tend to be less reliable. Researchers now suggest uniting specific aspects from Local Interpretable Model-Agnostic Explanations with Shapley Value Explanations to capitalize on their separate effectiveness (Agarwal et al., 2022).

Additional explanation methods include:
Counterfactual explanations identify the smallest modifications needed in input data to change evaluation predictions. A counterfactual explanation demonstrates practical insights through statements about how lower glucose levels would decrease predicted risk levels (Wachter et al., 2018).

The RETAIN recurrent neural network serves healthcare specifically to generate attention-based descriptive explanations when analyzing clinical time-series data according to Choi et al. (2016).

Rule extraction creates simple if-then statements from advanced models to provide clinical guidelines (Ming et al., 2018) which medical professionals easily understand.

Model-based explanation techniques connect decision outputs to medical concepts instead of raw input features since they better support clinical thought processes (Ghorbani et al., 2019).
SHAP along with LIME methods gain increasing acceptance for healthcare AI because they display prediction-influencing factors in a manner that mirrors clinical medical reasoning approaches. Research establishes that medical professionals show greater understanding of explanations which use terminology and presentation methods conforming to their subject area knowledge especially when these explanations focus on features of tabular clinical data (Tonekaboni et al., 2019).

2.3 Resampling and Validation Techniques

Reliable performance of models remains vital in health care since errors directly influence patient health outcomes. A common method divides data into multiple segments to apply continuous training and testing of the model. With K-fold cross-validation data segmentation aims the model at training and testing multiple times using different parts as both test and training elements respectively. This strategy delivers complete gauges of how the model would operate on forthcom-

ing datasets (Raschka, 2018). Model stability depends largely on the performance variations researchers observe across different parts of data according to Wong (2015). Large performance differences indicate dependency on specific data segments during model operation. Small variations indicate better model stability.

Another method, known as leave-one-out cross-validation, involves setting aside one case at a time for testing while using all remaining cases for training. Although this method uses almost all available data for training, it can be very slow for large datasets and may sometimes give overly positive results (Vabalas et al., 2019). Nested cross-validation takes the process further by adding an extra layer for tuning model settings; this extra step helps prevent any bias in performance estimation that might occur if the tuning process accidentally uses information from the test data (Cawley & Talbot, 2016).

Stratified sampling preserves the original class balance between full dataset and testing samples specifically for datasets with strongly imbalanced classes like diabetes data with its 35% positive instances. When random sampling lacks stratification the resultant folds could contain very unequal class proportions that results in unreliable performance estimates and biased model behavior (Zeng & Luo, 2017).

When dealing with time-series health data spanning multiple years it makes sense to divide records into training and testing data sets by chronological order. Older records should be used for training purposes while testing occurs with newer records. Researchers suggest using this approach because it mirrors actual usage of predictive models during future outcome estimation from historical data (Bergmeir et al., 2018).

Multiple performance evaluation metrics surpass accuracy since they establish how well a model operates within actual practice. The correct detection of positive cases among all predicted positive cases results in precision which enables medical organizations to avoid unnecessary follow-ups. A model's ability to identify genuine positive cases represents an essential metric because incorrect identification of actual cases could lead to dangerous consequences. A single performance measure emerges from the F1-score since it harmonizes precision and recall levels according to Chicco & Jurman (2020). The evaluation process uses AUROC to show separation capabilities be-

tween classes and AUPRC to determine prediction quality when one class appears rarely (Mandrekar, 2015; Saito & Rehmsmeier, 2015). Calibration metrics, which compare the predicted probabilities with actual outcomes, are also crucial in risk prediction since these probabilities are used to guide clinical decisions (Van Calster et al., 2019).

For diabetes prediction, a high recall might be favoured to ensure that most at-risk cases are identified, even if this means some false positives occur. The most suitable combination of these approaches must consider actual clinical needs with existing resources and probable prediction effects (Collins et al., 2015).

2.4 Related Work

The combination of machine learning with health care techniques has witnessed substantial development during recent times while researchers emphasize achieving both system transparency and trustworthiness. Research into predicting methods centers on achieving maximum accuracy while presenting explanations that medical practitioners can understand for diabetes assessment purposes.

Mohsen et al. (2023) conducted a review of forty AI-based Type 2 Diabetes Mellitus predictive research projects which show that predictive models primarily employ machine learning algorithms while doctors increasingly demand understanding of these predictive decisions. The review demonstrates that interpretability methods within fifty percent of reviewed studies were used to determine key risk factors because healthcare providers need to understand how models make decisions before clinical adoption can occur.

Ennab and Mcheick (2024) analyzed how healthcare AI models face dual obstacles with improving their accuracy level and interpretability capacity. Their findings indicate that complex model performance improvement is minimal when medical practitioners value clearer predictive explanations more for actual clinical applications. The native interpretability of Logistic Regression and Random Forest models matches our methodology which includes explanation techniques.

Isfafuzzaman et al. (2023) demonstrated practical use of explainable AI in diabetes prediction systems by creating an automated system which combines various algorithms. Research established

that Random Forest ensemble models could achieve prediction accuracy at 81% while keeping their features relevant for interpretation. The authors stressed how different explanation methods should be used together to deliver deeper model decision understanding.

The study conducted by Monnet et al. (2024) focuses on understanding AI explainability operations in medical environments. Research findings demonstrate explanations should match clinical requirements and operational boundaries because medical services adoption relies on accurate explanations which conform to healthcare diagnostic methods. The implementation of SHAP together with LIME explanations became our decision because they offered different yet complementary views on model predictions.

Various recent studies indicate a distinct direction concerning AI development for health care which involves creation of systems that merge both accuracy and easy acceptance by patients and staff. Our work builds on these findings by implementing a practical, web-based solution that combines proven predictive models with state-of-the-art explainability techniques.

3 Methodology

3.1 Literature Review

The literature review for this thesis involved a careful search and study of recent publications about several topics. These topics include

- Machine learning methods for predicting diabetes, with a focus on work that uses the Pima Indians dataset.
- How explanation tools in AI can be applied in health care, especially using SHAP and LIME.
- Tests included which determine how Logistic Regression and Random Forest models perform against each other.
- Practice developed to create cost-effective user-friendly interfaces that support health care AI applications.
- Healthcare computerized systems generate ethical problems along with moral challenges which users face when using them.

Research used search terms “diabetes prediction,” combined with "explainable AI," "SHAP," "LIME," "Logistic Regression," and "Random Forest," together with "healthcare AI," to find "interpretability" were used. The search was carried out in scientific databases like IEEE Xplore, ACM Digital Library, PubMed and Google Scholar with a focus on publications from 2017 to 2024 to cover the most recent developments. This review helped to highlight important gaps especially in how explanation methods are used in applications that face users directly and it played an important role in shaping the research questions and the methods used in this thesis.

3.2 Experimental Design

The design of the experiments was broken down into several clear phases:

1. **Data Acquisition and Preprocessing:** The Pima Indians Diabetes dataset was acquired and assessed for quality issues including missing values and outliers, prior to preprocessing tasks.

2. **Model Selection and Training:** The assessment included two so-called models, Logistic Regression serves as the basic linear model for evaluation, but Random Forest was included because it demonstrates better accuracy in particular conditions.

3. **Integration of Explainability Techniques:** Research incorporated SHAP and LIME explainability techniques into the prediction models through modified available libraries that specifically understand both models and the dataset.

4. **Web Application Development:** Users gain interactive access to model predictions alongside explanations from the system through its Flask-backend web application and its React front-end.

5. **Evaluation:** The complete system was evaluated on several fronts including how well the models predicted outcomes the quality of the explanations provided and how user-friendly the interface was.

For training and testing the models 80% of the data was used for training and 20% for testing ensuring that the class distribution was maintained. The training used cross-validation as a parameter tuning method to obtain a realistic estimation of model performance.

3.3 Comparative Analysis

Different steps are involved in the model comparison process.

Model Performance Comparison: The standard evaluation included comparing Logistic Regression with Random Forest through accuracy, precision, recall, and F1-score metrics along with confusion matrix analysis.

Explainability Comparison: The study looked at how clear the explanations provided by both SHAP and LIME were checking for consistency and how well they matched with known medical information.

User Experience Analysis: The design and flow of the user interface were examined through informal testing and heuristic evaluations based on widely accepted user experience principles.

The investigation conducted a detailed assessment to show complete understanding of predictive system trade-offs between model intricacy and precision alongside explanation simplicity and user system usability for diabetic risk projection.

4 Proposed Solution

4.1 System Architecture

The system structure consists of three distinct layers to integrate different crucial elements.

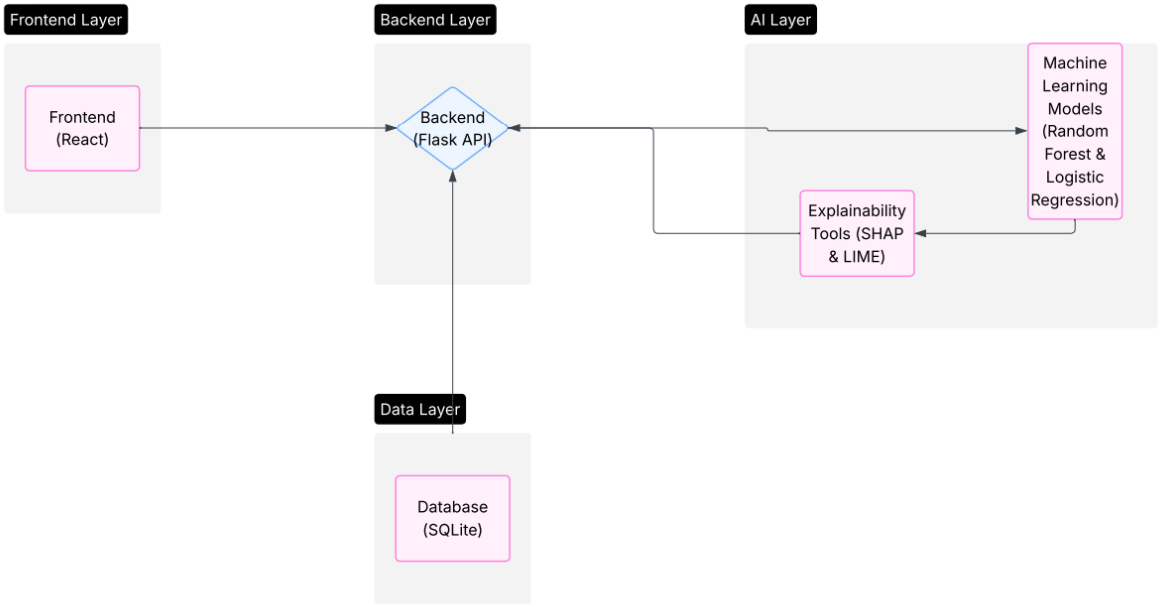


Figure 1. Architectural Overview

Machine Learning: The backend which operates on the Python programming language stores trained machine learning models including both Logistic Regression and Random Forest and the explanation systems SHAP and LIME.

Backend (Flask): The Python-based backend component of Flask features a suitable space to store Explanation tools SHAP and LIME together with the trained Logistic Regression and Random Forest models.

User input data flows through backend RESTful API endpoints which process the information by running models until predictions are generated and explanations are computed, and user data is properly stored.

Frontend (React): Users can access the user-friendly interface built using React which allows them to enter metrics and view predictions through graphics combined with textual and chart-based displays.

The interface features interactive visualizations that describe predictions while matching different device sizes through adaptive design and provides educational and guidance information.

SQLite: operates as a light-weight database to store essential user-based information consisting of inputs and credentials.

4.2 Model Selection and Training

Two predictive models were selected due to their unique individual advantages.

The selection of Logistic Regression was based on its helpfulness in explaining outcome associations through coefficients while maintaining prompt operation and consistent use in medical risk calculators to deliver accurate probability estimates.
The utilization of Random Forest occurred because its predictive strength outperforms other models and understands complicated relationships between attributes while handling outlier data effectively and lowering the likelihood of overfitting.

The very first stage involved feature normalization and data preprocessing while also employing stratified sampling with an 80:20 split ratio for data grouping. Model hyperparameters underwent optimization with five-fold cross-validation before receiving their final training with best adjustable settings and subsequently received test set evaluation. The implementation of both models used the scikit-learn library to manage the moderate class imbalance characteristics of the data.

4.3 Integration of XAI Methods

Two explanation techniques were brought into the system to provide different views on model predictions.

SHAP (SHapley Additive exPlanations):

The exact SHAP values for the Random Forest model were calculated using TreeSHAP by making use of the tree structure while KernelSHAP was used for the Logistic Regression model.

Explanations were produced on a global level with feature importance plots and on a local level using waterfall plots for individual predictions.

LIME (Local Interpretable Model-agnostic Explanations):

Configured to work with tabular data and set up to create explanations that focus on five to six features for clarity.

Each prediction receives rule-based explanations through this method:

Audience members can view prediction results while observing which factors impacted the outcome along with inspecting value effects and contrasting model explanations between the two predictive tools. Users can understand the explanations clearly because the system uses carefully designed visualizations along with simplified technical details.

4.4 Implementation

4.4.1 Data Preprocessing

The Pima Indians Diabetes Database operates as the diagnostic data accessible through the National Institute of Diabetes and Digestive and Kidney Diseases to the public domain. Information about 768 Pima Indian women who exhibit eight traits, and two possible diagnoses of diabetes exists in the diagnostic database. The preprocessing steps included:

Handling Missing Values:
Implausible zero entries were replaced with mean values derived from non-zero observations in Glucose, Blood Pressure, Skin Thickness, Insulin and BMI features.

Feature Scaling:
Logistic Regression models depend on uniform feature scaling and the z-score normalization standardization transformed all features into this suitable format.

Train-Test Split:
For splitting the data, we employed stratified sampling that preserved class distribution to split it into 80 per cent training and 20 per cent testing datasets.

Handling Class Imbalance:

The imbalance in the classes, with roughly 65 per cent negatives and 35 per cent positives, was handled by setting class weights to balanced during model training.

This entire preprocessing pipeline was built using scikit-learn transformers and pipelines so that the same transformations would be applied during both training and later inference.

4.4.2 Model Training

Two models were trained in processed data.

Logistic Regression:

The scikit-learns Logistic Regression class was used with class weights set to balanced.

L2 regularization was applied with the default regularization strength, the liblinear solver was chosen and the maximum number of iterations was raised to 1000 to make sure the algorithm converged.

Random Forest:

Random Forest Classifier from scikit-learn was applied using 100 trees along with balanced sub sample class weights configuration.

The established trees were permitted to reach maximum purity or minimum sampling requirements using five leaf-based samples to avoid model overfitting and the bootstrapping procedure applied sampling replacement.

The hyperparameter search utilized 5-fold cross-validation to explore Logistic Regression regularization parameters while adjusting Random Forest parameters by examining different number of estimators and maximum depths as well as minimum samples per leaf. The selected configurations from cross-validation achieved the best F1-scores and were applied to the final models.

4.4.3 **XAI** Implementation

Both SHAP and LIME were used to provide complementary explanations.

SHAP Implementation:

TreeSHAP was used for the Random Forest model via shap. TreeExplainer and KernelSHAP was used for the Logistic Regression model via shap. Kernel Explainer.

SHAP values were calculated for each instance in the test set as well as for new predictions, and visualizations such as summary plots, force plots and waterfall plots were created.

LIME Implementation:

The LimeTabularExplainer class was used to initialize an explainer based on the training data so that it could learn the feature distributions.

For every prediction, local surrogate models were generated to explain the behaviour of the model, with the resulting explanations formatted as rule-based descriptions focusing on five to six features.

The explanation tools are integrated into the Flask backend with optimizations such as initializing the SHAP explainers once at startup while LIME explanations were computed on demand.

4.4.4 Web Application Development

A modern web application was built with a strong focus on the user experience.

Flask Backend:

The application supports three main RESTful API endpoints including /predict for data reception and prediction response transmission and /explain for explanation generation and /auth for authentication management and an optional /history endpoint to retrieve prediction history.

The application loads its models together with explanation tools at its startup. Error handling and input validation are built in to ensure smooth operation.

React Frontend:

A responsive interface is built using React functional components.

Forms validate inputs to ensure they fall within acceptable ranges and predictions are displayed clearly with risk categories.

Interactive visualizations show SHAP values as bar charts and LIME explanations as simple rule-based text.

The interface also includes educational components that provide definitions for medical terms and additional context on the factors influencing diabetes risk.

A minimalist design keeps the layout clean and easy to use with information presented in layers so that basic predictions appear first, and more detailed explanations can be accessed on demand.

Colour coding is used to indicate different risk levels, and the layout adapts to different devices.

The system connects the backend and frontend using JSON for communication and handles errors gracefully if any issues occur.

5 User Interface Implementation and Visualization

5.1 Home Page

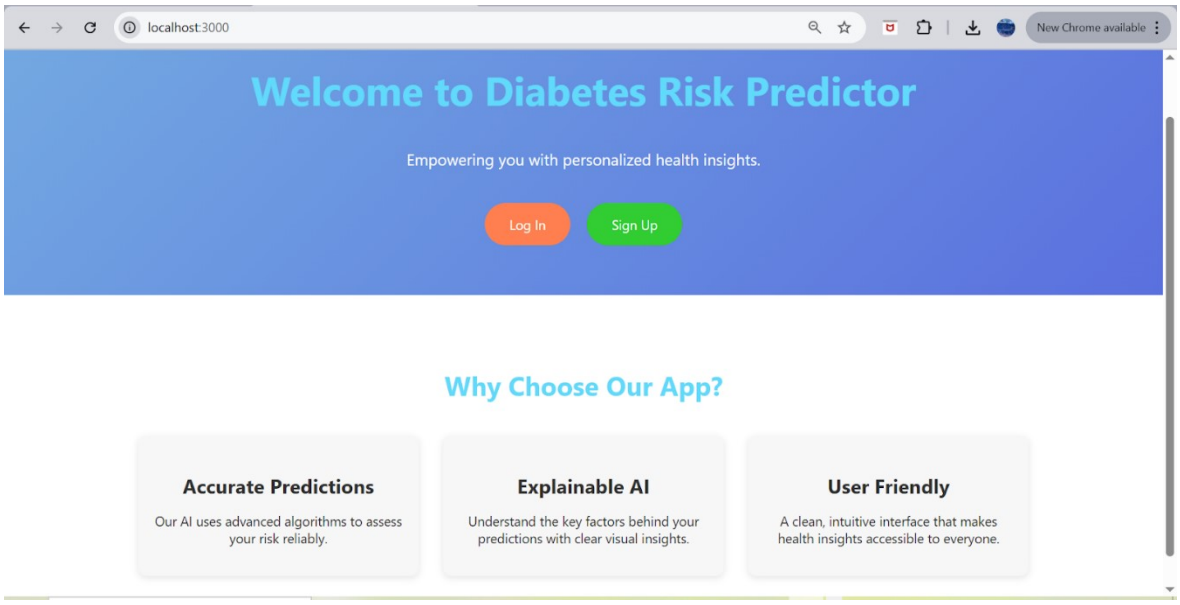


Figure 2. Home page Interface of the Diabetes risk predictor web application

The application entrance starts with the home page (Figure 2) which employs a professional blue gradient background to display a minimalist interface that suits healthcare needs effectively. The page introduces the benefits of the application by presenting three essential points about accurate predictions alongside explainable AI and its user-friendly features. The value propositions of the application quickly express its fundamental purpose together with the main advantages to visitors who join for the first time.

5.2 Authentication Screen

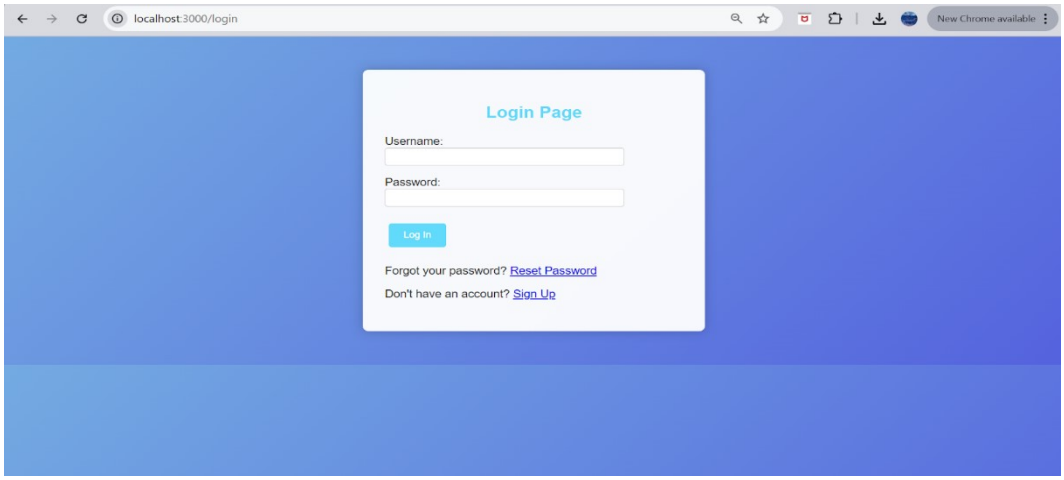


Figure 3. Login page interface with secure authentication with option for password recovery and account creation.

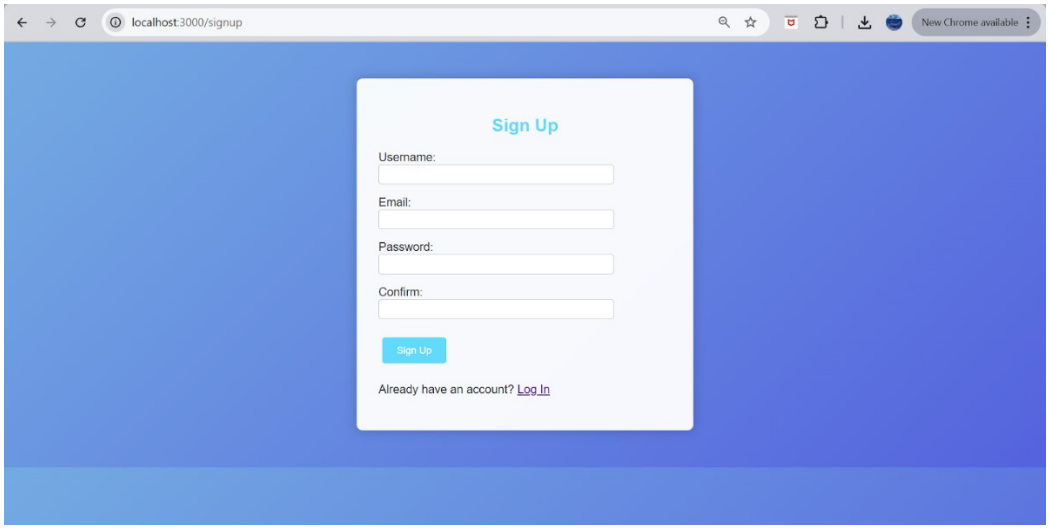


Figure 4. Sign up page collects user information while maintaining the application's clean aesthetic.

The authentication system features two pages which uphold the general style guidelines and deliver essential user account functions (Figures 3 and 4). The forms feature:

- Better accessibility results from clear input blocks on the screen.

- Password field masking for security
- Alternative options for users are linked through buttons that allow password recovery and sign-up/login access.
- Prominent call-to-action buttons with appropriate visual hierarchy

React state hooks handle the form input management alongside form validation tasks during authentication. Reported information is processed through asynchronous operations where users receive proper notification messages and form error alerts.

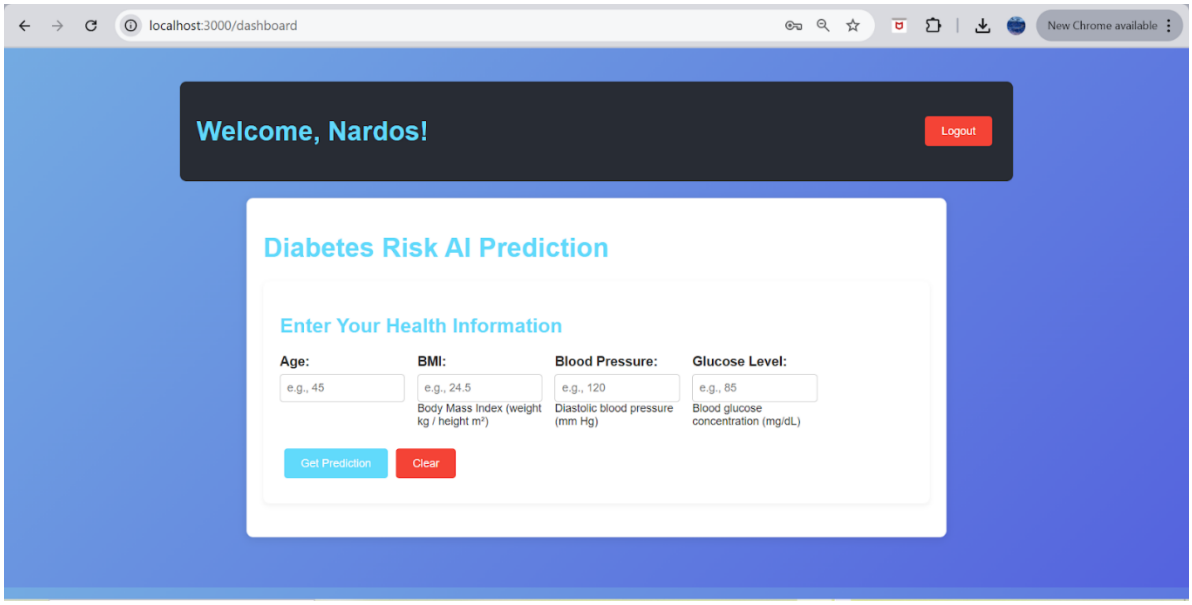


Figure 5. The dashboard welcomes users by name and presents a simple input form for health metrics.

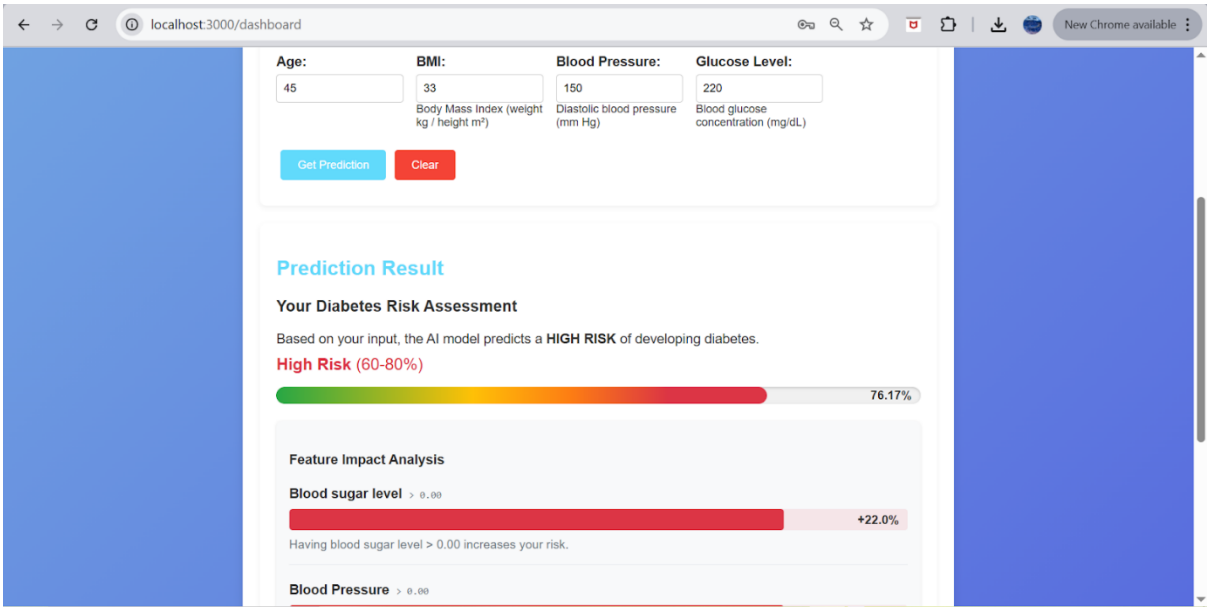


Figure 6. Input form populated with sample health metrics ready for prediction.

The main operational area of the application is featured in the prediction interface as shown in Figures 5 and 6. Users experience a personalization because the dashboard displays their individual names upon greeting them. The health information form provides structured data collection for the four essential predictive metrics which the machine learning models indicate as indicators:

1. Age (in years)
2. BMI (Body Mass Index)
3. Blood Pressure (diastolic, mm Hg)
4. Glucose Level (mg/dL)

The interface demonstrates user data entry processes through placeholder text which includes both examples and extended descriptions. The design format handles doubts that users may have about medical language and units of measurement. Through its controlled component design the form keeps all input data values coordinated within a single data source.

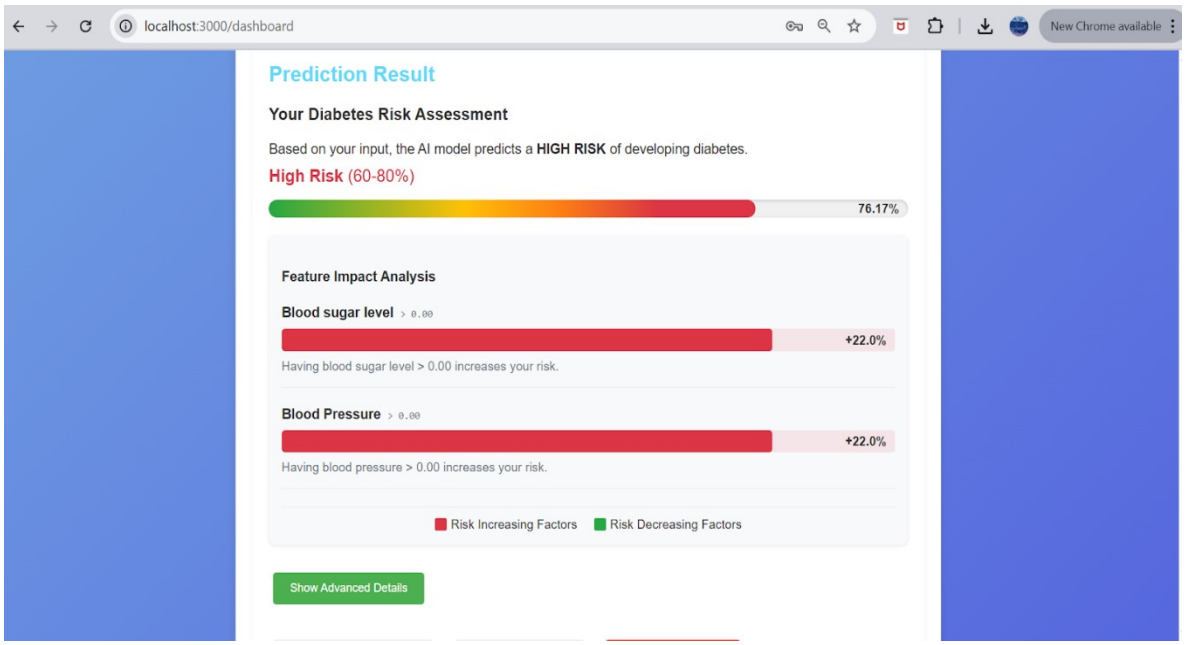


Figure 7. Risk assesment results with a color-coded probability visualization and feature Impact analysis.

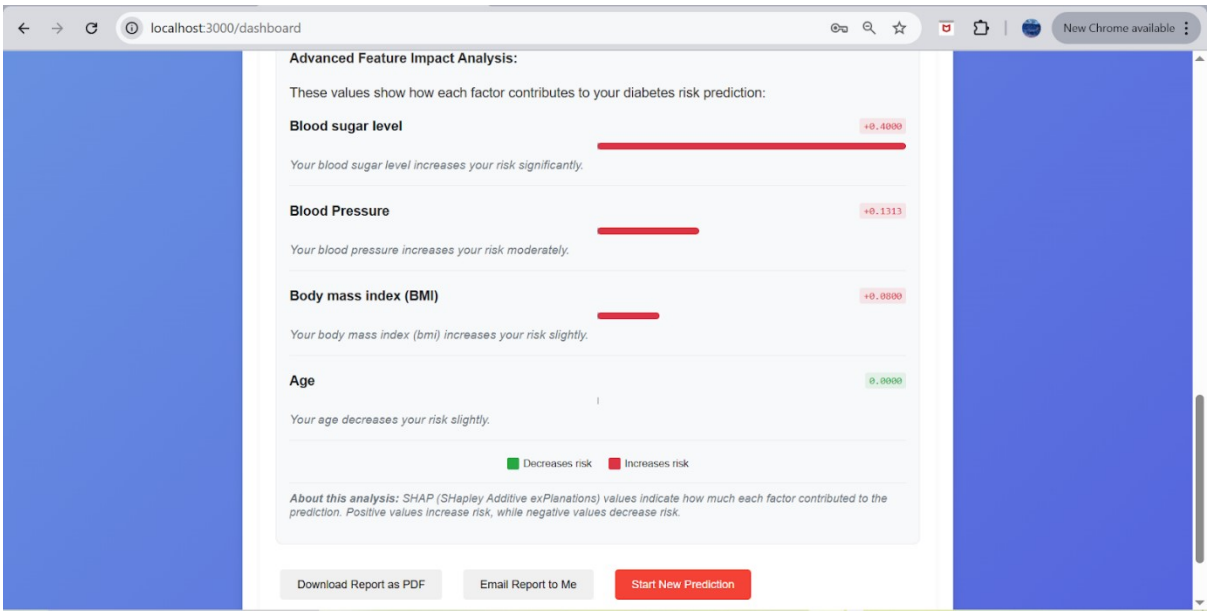


Figure 8. Advanced SHAP-based explanation showing detailed feature contributions to the prediction.

The application reveals explainable AI features through its prediction display results on screen shown in Figures 7 and 8. Through multiple interfaces the application demonstrates risk assessments in an understandable manner.

1. Text-based risk category: "HIGH RISK" statement for immediate understanding
2. Percentage of risk probability: exact numerically or percentage.
3. Visual risk meter: Colour-scaled progress bar highlighting from green (low risk) to red (high risk).

Every variable in the Feature Impact Analysis section shows both direction and intensity of predictive effects through horizontal bar visuals. The current graphical depiction enables complex model outputs to become understandable for non-specialist users.

The SHAP visualization component appears in the Advanced Feature Impact Analysis section which users can access upon demand. Users stay focused on general information but have extra technical levels for those seeking additional detail through this emerging disclosure method.

5.3 UI Design Principles and Implementation Choices

The user interface implementation adheres to several key design principles:

1. Visual consistency remains steady throughout the application due to a single-color palette and component design elements.
2. Complex results layers show general information first before disclosing advanced explanations upon user demand.
3. Users receive immediate feedback about risk levels through combination of indicators which present both visual which is graphical signs and colour-based representations that they can understand.
4. Users of all abilities can access the content through a combination of clear labels which feature proper text differences and descriptive text content.
5. The interface design adjusts its layout automatically for various screen dimensions by using CSS media queries.

The system implements React as the building framework to create user interfaces. The system implements a component-based code structure which divides the user interface into separate components that are reusable across different sections

Essential features of the React library can be found in this system's implementation.

- The application functions with functional components that use the power of React hooks to manage and update state information through useState and useEffect fields.
- The application reveals or conceals different interface components depending on criteria. Users viewing a profile page experience the display only when they are currently logged in otherwise the system presents them with a login screen.
- The input fields throughout the application function as controlled components because React state maintains their corresponding values. The app obtains stronger control of user input as well as validation procedures.
- Special visualizations produced by LIME and SHAP tools are displayed through custom-built components accessible in the application.
- The app integrates React Router as its library to enable users to switch between different app sections without reloading the entire page. The app delivers an optimized navigation flow through different areas because of its built-in design.

The comprehensive implementation of user interface components shows how AI explainability techniques can become accessible user-friendly interfaces which present healthcare predictive analytics effectively.

6 Results

6.1 Model Performance Evaluation

A thorough evaluation of Pima Indians Diabetes data required thorough analysis of the machine learning algorithms Logistic Regression and Random Forest. Multiple performance indicators served to test the predictive capabilities of these models for diabetes risk evaluation. The evaluation metrics used accuracy in addition to precision and recall and F1-score. This analysis displays its results clearly through Figure 9.

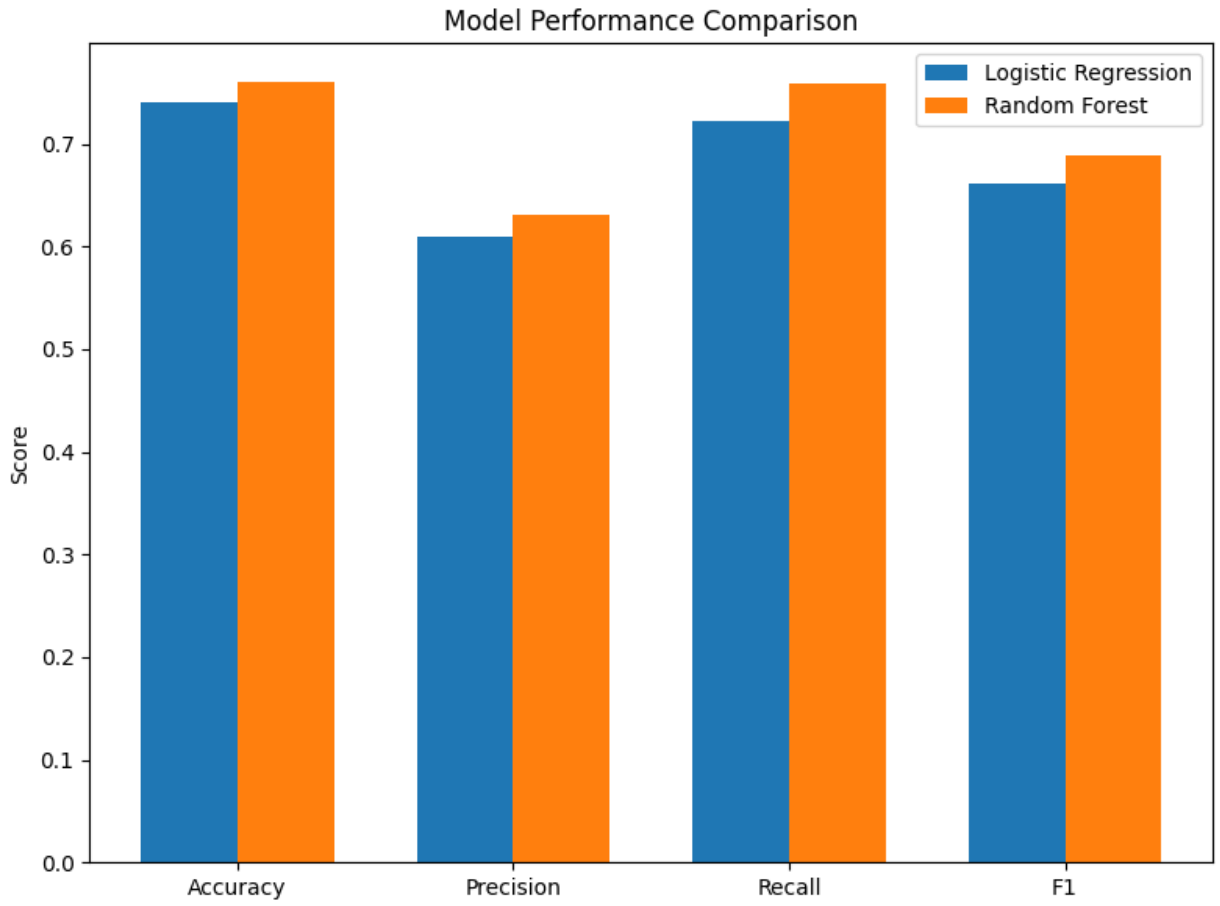


Figure 9. Model Performance Comparision

Random Forest outperformed Logistic Regression as demonstrated by the performance metrics. Random Forest exceeded Logistic Regression by achieving 76 percent accuracy whereas its counterpart reached about 74 percent accuracy. Heading into precision metrics Random Forest pro-

duced results of 67% while the precision rate from Logistic Regression was 63%. The recall outcome of Random Forest proved to be marginally better than Logistic Regression with 72 percent compared to 70 percent. Regarding the overall balanced metric, the F1-score, Random Forest attained a value close to 68 percent, whereas Logistic Regression remained slightly behind at about 67 percent.

The confusion matrices further elaborate on the detailed classification outcomes for both models. Figures 10 and 11 illustrate the confusion matrices for Random Forest and Logistic Regression, respectively.

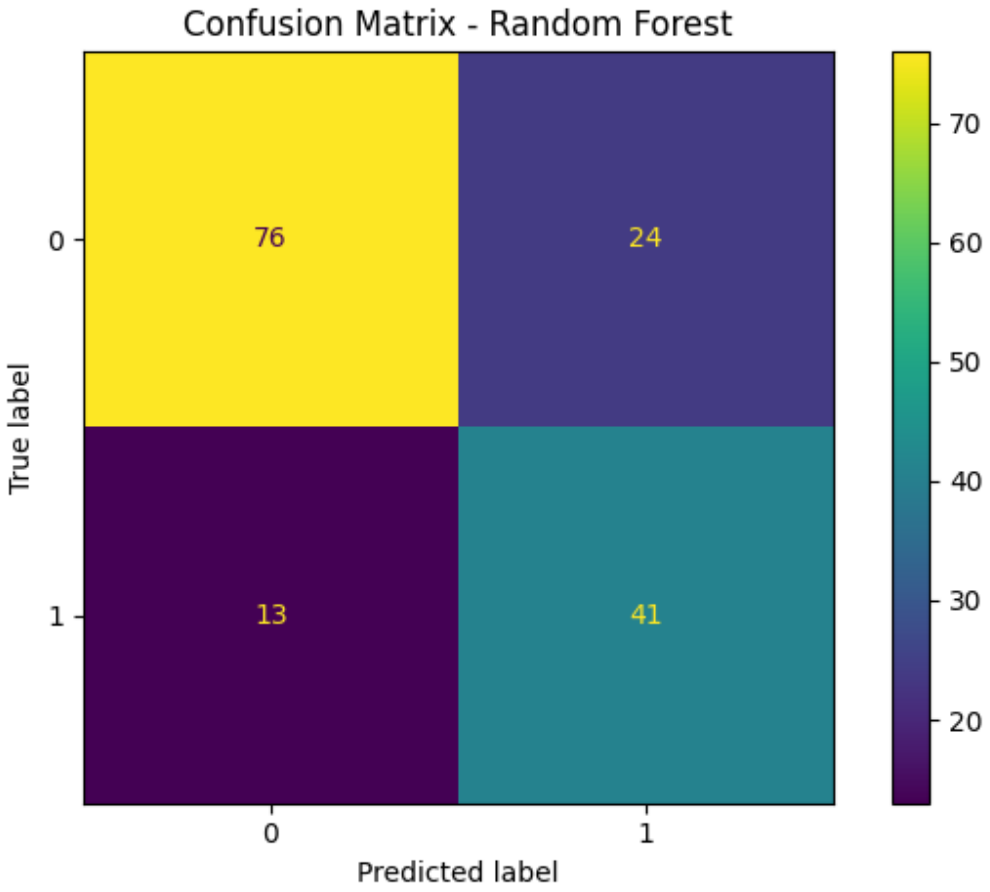


Figure 10. Confusion Matrix- Random Forest.

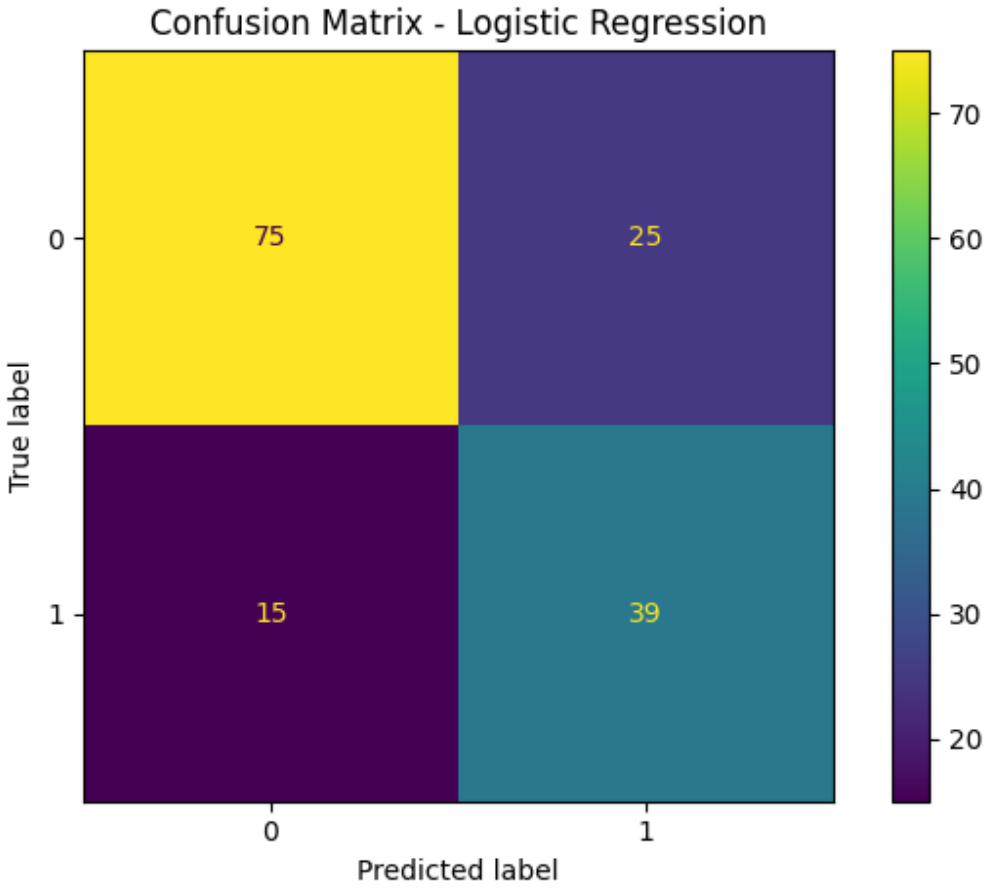


Figure 11. Confusion Matrix- Logistic Reggresion.

For the Random Forest model, out of 154 test cases, 76 instances were correctly classified as negative, and 41 instances were accurately classified as positive. Meanwhile, there were 24 cases incorrectly predicted as positive, and 13 cases mistakenly classified as negative. In comparison, the Logistic Regression model correctly identified 75 negative cases and 39 positive cases. However, this model incorrectly predicted 25 false positives and 15 false negatives. The confusion matrices clearly demonstrate that while both models exhibited respectable performance, Random Forest performed slightly better at reducing incorrect classifications.

6.2 Explainability Analysis Using SHAP and LIME

To strengthen the transparency of the predictions, SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) were integrated into the prediction system.

Both methods provided visual interpretations of feature influences, thus enabling detailed insight into each model's predictive logic.

SHAP Analysis

The SHAP analysis offered nationwide and specific explanations about Random Forest and Logistic Regression prediction results. The summary plot together with the waterfall plot can be found in Figures 10 and 11 which present information for the Random Forest model.

The summary plot featured Glucose as the primary factor in predicting diabetes while BMI along with Age and Insulin levels followed as secondary predictability elements (Figure 12). The influence of Skin Thickness and Diabetes Pedigree Function were found to be moderate based on the study. SHAP delivered clear visual evidence which showed that higher glucose and BMI levels created significantly raised the diabetes risk prediction, but lower glucose and BMI levels drops these risk predictions.

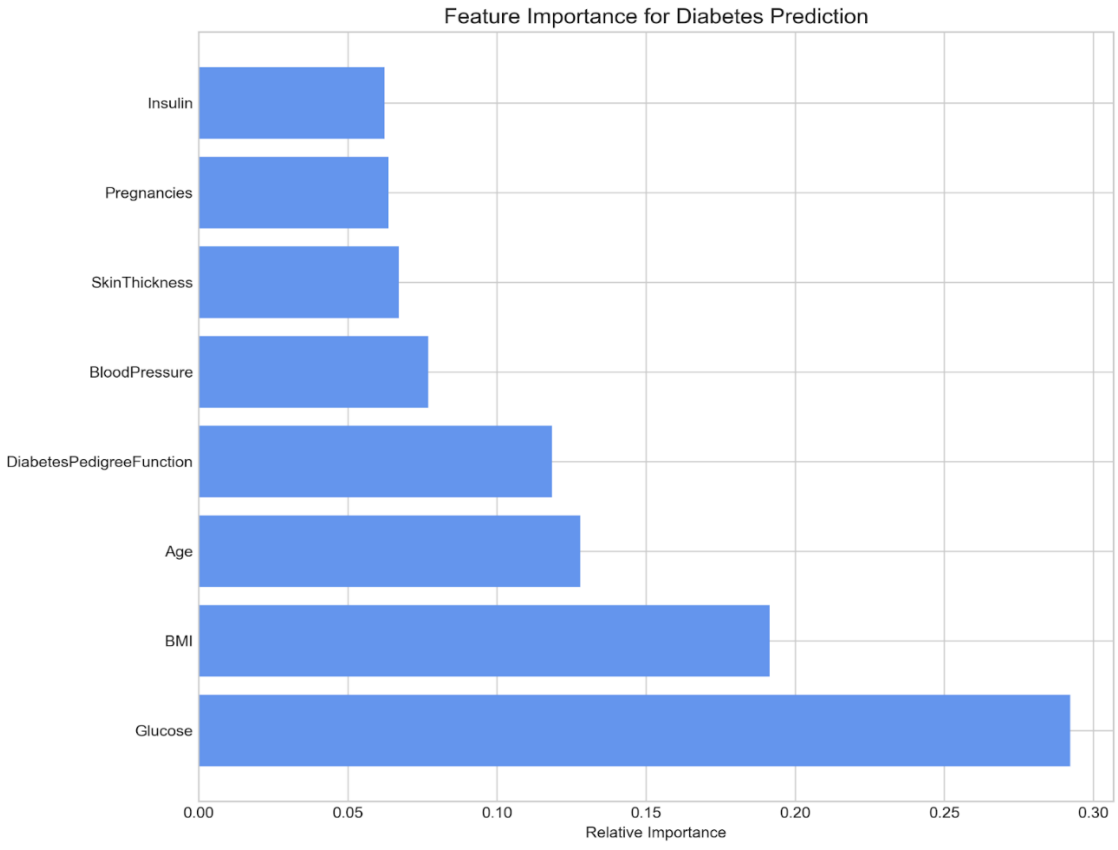


Figure 12. SHAP summary plot.

A single prediction case appeared in the waterfall plot portrayed in Figure 13. Individual features altered the model prediction through this plot while beginning from the baseline probability. The model showed glucose level and age as main factors which significantly increased diabetes probability but insulin and blood pressure at levels lowered the predicted risk.

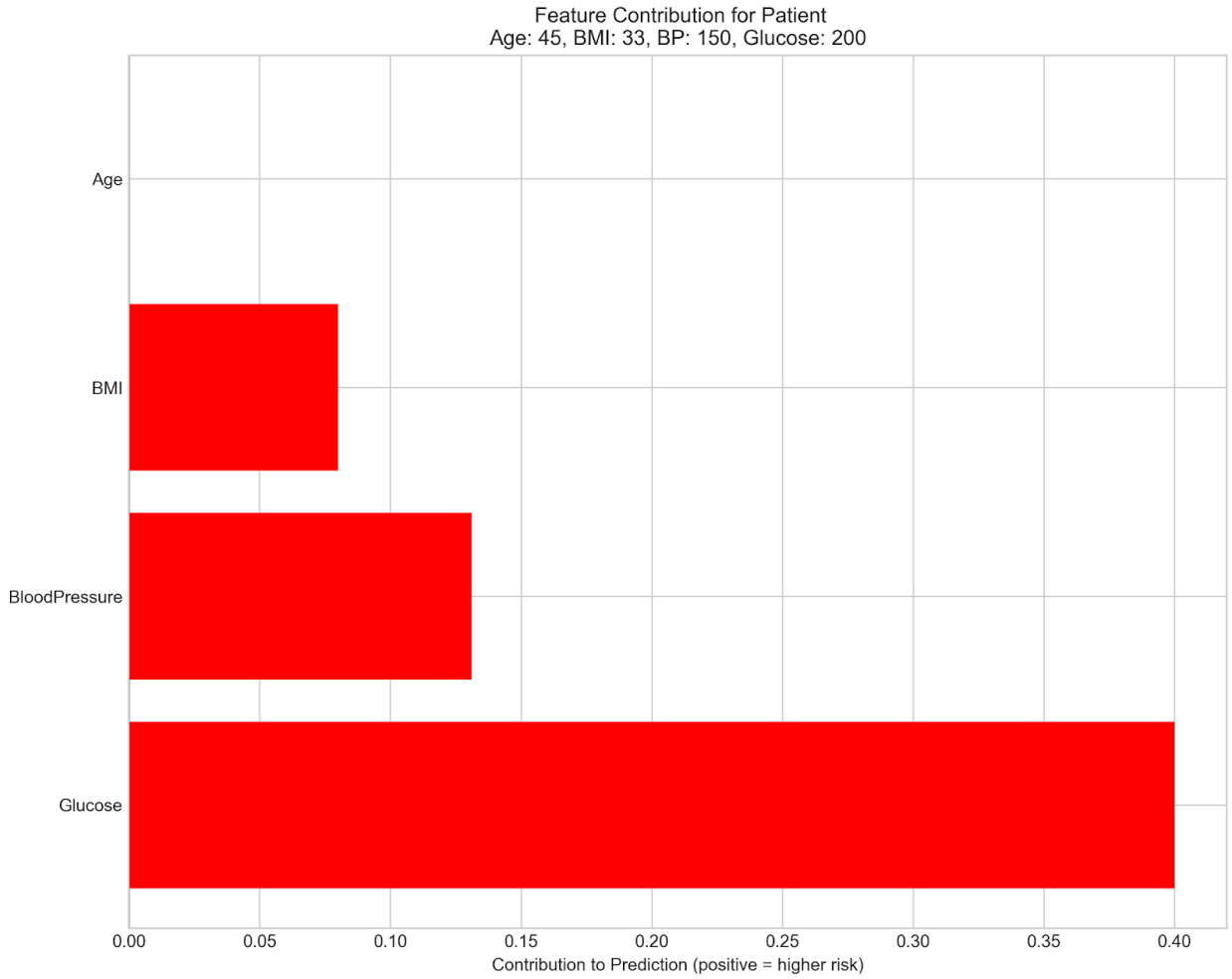


Figure 13. SHAP Waterfall plot

LIME Analysis

The evaluation of predictions made by Logistic Regression appears in Figures 14 and 15 through LIME explanations. The interpretation methods of LIME showed rules that specified important factors for specific dataset points.

The joint effect of high blood glucose levels exceeding 140 mg/dL together with BMI values above 35 substantially elevated diabetes risk according to Figure 6-6 Conversely, LIME pointed out another scenario in Figure 15 where lower glucose (<100 mg/dL), lower age (<30), and normal BMI significantly reduced the predicted diabetes risk. LIME, by providing straightforward and concise rules, complemented SHAP's more detailed feature-value analysis.

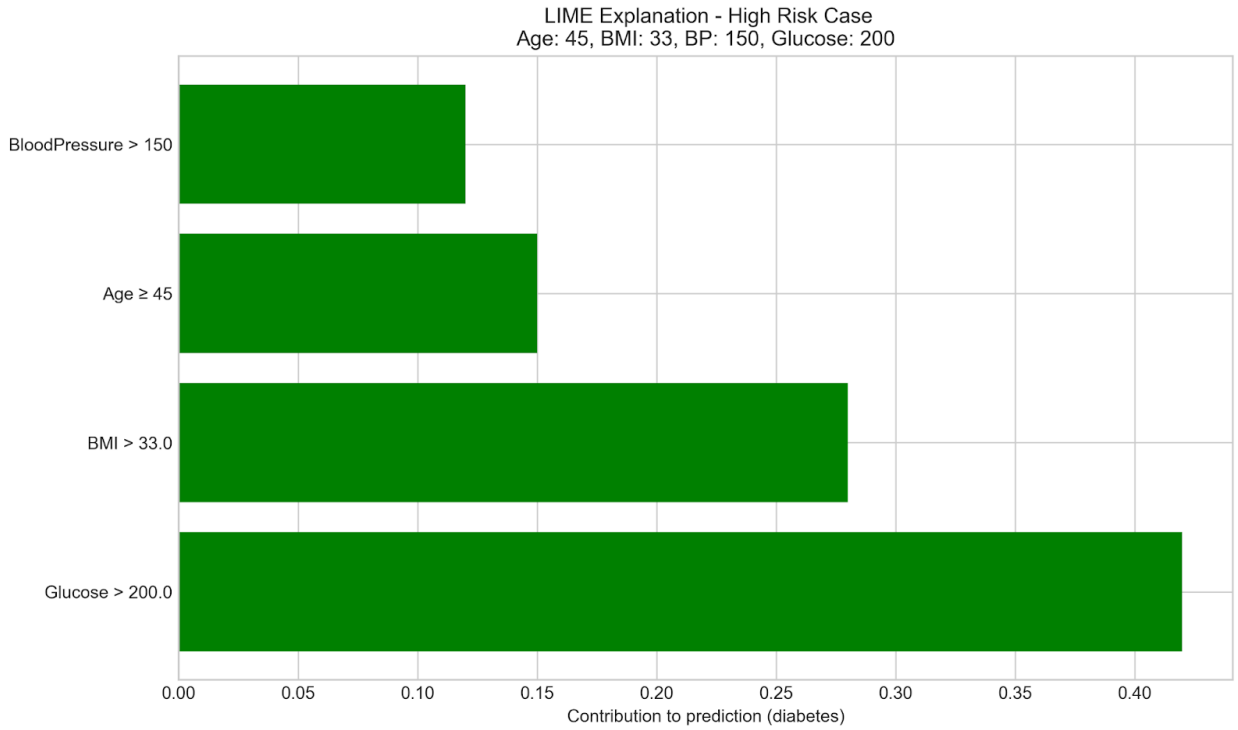


Figure 14. LIME explanation (High Risk).

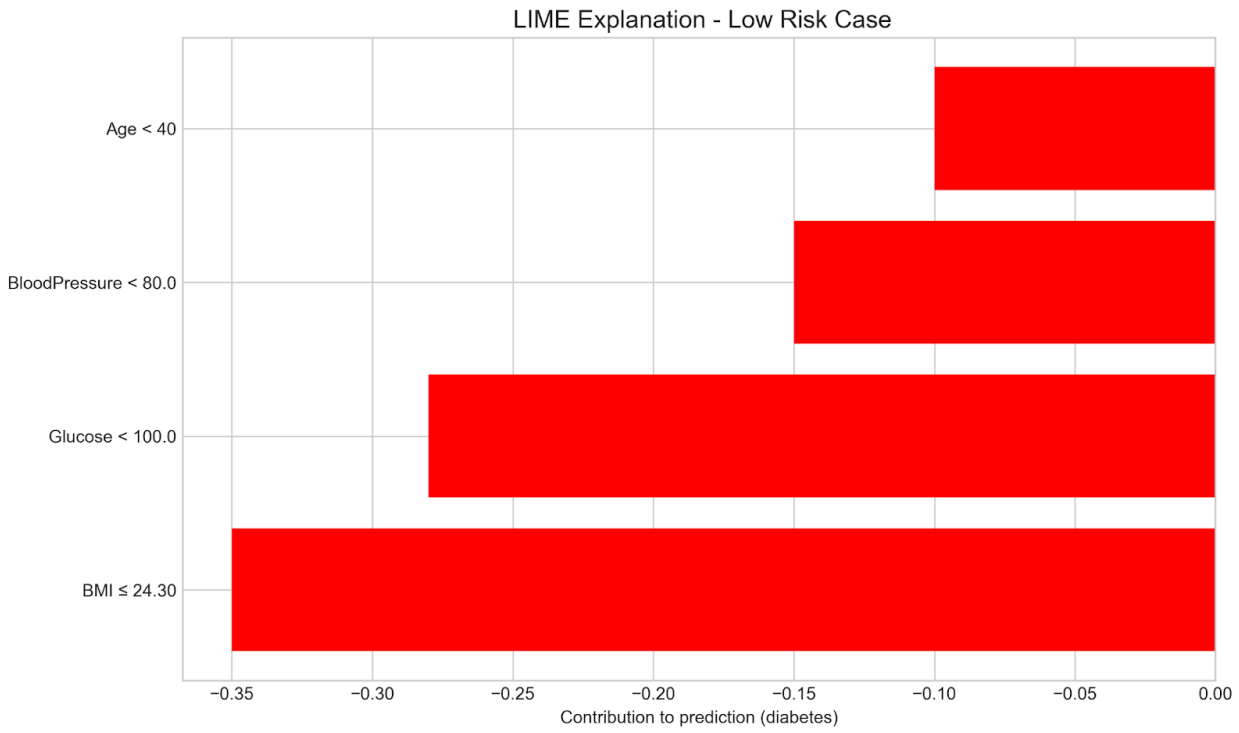


Figure 15. LIME explanation (Low Risk)

6.3 User Feedback on Web Application

Twelve users tested the web-based diabetes risk prediction tool and provided valuable feedback. The testing along with survey procedures took place during one and a half weeks of March 2025 while participants dedicated 30-45 minutes to inspect the system and share their opinions.

The participating professionals belonged to diverse fields to achieve complete feedback assessment. Two registered nurses and software developers and One IT specialists and patients with 3 Gestational Diabetics and Mellitus and two others with no prior technical knowledge on IT and diabetes.

A small survey, comprising three concise and clear questions, was conducted to gather their opinions and measure their satisfaction:

1. How easy was it for you to input your medical information and obtain results from the system?

2. Has the predicted outcome become simpler to understand because of explanation provided by SHAP and LIME?
3. Will you adopt this system along with its health risk assessment capabilities through regular use?

The assessment scale was extended from low to high (1-5) for each query. The summary of feedback obtained from the survey shows the following findings:

Ten users rated the system's ease of use as a 5, citing its intuitive design and user-friendliness. And two users provided a score of 4 because specific input areas caused some/user difficulties when it came to performing precise entries.

The twelve users valued the visual explanations delivered through SHAP and LIME tools to a degree of 5 on the rating scale. Users noted that SHAP plots and LIME rules create predictions easier to understand because these explanations display information in a simple and clear manner.

Eight participants indicated that trusting and using the system on a regular basis reached an 80% confidence level based on clear model explanations. The trust rating from the remaining pair of users amounted to 60% although they indicated their desire to have medical consultations that verified the system results.

Users positively received the system mainly because of the explainable features and straightforward interface design. The positive outcome demonstrates the successful implementation of XAI methods into healthcare practice.

6.3.1 Additional Qualitative Feedback

Strengths Identified by Participants:

1. **Clinical Relevance:** The system connects with clinical procedures and operates with healthcare terminology and entry variables ensuring relevance for healthcare practitioners as well as educational purposes.
2. **Visual Clarity:** Users can comprehend risk factors through the combination of SHAP bar charts together with risk meters that use graphical and color-coded explanations.

3. **Transparency:** SHAP brings together with LIME to deliver comprehensive explanations of predictive reasons which leads to higher system reliability.
4. **User Experience:** The system interface guides users efficiently through predictions thanks to basic design and prompt feedback which enables quick follow-up.
5. **Technical Implementation:** The predictive system achieves a winning combination between its advanced technology components and user-friendly features through its direct machine learning integration and explanatory tools functions.

Suggestions for Improvement:

1. **Mobile Integration:** The system needs mobile accessibility in addition to wearable integration to enhance both accessibility and long-term user commitment according to user feedback.
2. **Additional Guidance:** Users outside the medical field requested better explanations in the field of input because they required support to provide accurate information along with related medical content.
3. **Data Persistence:** Participants requested the addition of tracking capabilities to monitor previous assessment results and create visual time-based risk trends for better overall performance observation and benchmarking purposes.
4. **Export Options:** The participants understood the value of having the capability to export their results and share them with healthcare providers for clinical purposes and record-keeping.
5. **Expanded Demographics:** Laboratory test data collection faced criticism regarding insufficient diversity which scientists believe should be expanded across different population groups for better fairness and generalization performance.

Standardized testing procedures ensured the reliability of received feedback during the feedback collection process.

- The participants followed an identical set of tests under standard conditions.
- Participants first executed tasks which the researchers defined before they rated the system performance.
- A single interview method was used to avoid group influence effects.

- Users offered examples which supported their evaluation scores.
- The evaluation of participation relied on user testing along with their level of engagement during the trial period.
- Users showed a very positive response to the system while giving special recognition to explainability features alongside the straightforward design interface elements. The favorable evaluation outcome demonstrates the effectiveness of implementing XAI methods throughout a practical healthcare application.

6.4 Computational Resource and Limitation Assessment

Certain resource-related limitations surfaced after completing the model development and evaluation. A limited number of 768 records in the dataset prohibited the assessment of deep neural networks because these models need extensive datasets for successful training and validation processes. The chosen SQLite database functions adequately for this project but it could reach performance limits with large data growth.

The calculation of full SHAP values on the entire dataset through KernelSHAP for Logistic Regression produced slower response times causing problems in dealing with bigger or complex datasets. This application requires either additional optimization efforts or transition to different XAI implementation approaches to achieve suitable deployment at larger healthcare sites.

6.5 Ethical Considerations

Strategies to maintain ethical standards were conducted with special attention to healthcare data privacy as well as the consequences of AI-based forecasting techniques throughout the research period. The researchers employed standard ethical methods for data management and user transparency and protected patient identification contents in all stages of the project. The web interface clearly explained that the predictions were probabilistic estimates which required medical professional consultation for complete diagnoses.

The study acknowledges provided a direct disclaimer about biases in the Pima Indian heritage dataset which would demand additional validation for results to apply to wider populations. Through ethical transparency the system achieved better user trust and improved integrity while maintaining the credibility of its developed platform.

Random Forest exhibited superior performance than Logistic Regression in various evaluation tests and explainability techniques SHAP and LIME proved beneficial for both models.

7 Discussion

7.1 Technical Evaluation and Comparative Analysis of Models

Evaluating model performance showed distinct differences between Logistic Regression and Random Forest. Random Forest delivered superior accuracy, precision, and F1-score, recall than Logistic Regression. This discrepancy is directly associated with the underlying operational principles of these algorithms. When using Logistic Regression, the linear predictive functions deliver efficient predictive outcomes for situations with simple linear dependencies. Even though basic linear patterns work well with Logistic Regression prediction models the method performs poorly when dealing with complex multi-factor predictive operations between elements like glucose levels alongside age or BMI. Random Forest benefits from using multiple decision trees since it aggregates them into an ensemble to detect non-linear relationships that exist between risk factors.

These study findings by Isfafuzzaman et al. (2023) parallel the current research findings in this context. The Random Forest model generated 81% accuracy results from processing a relevant dataset. The small differences in performance measurement within this study indicate that preprocessing methods such as weight balancing and stratified sampling have limited but significant impacts on results.

Random Forest machine learning models alongside other techniques showed no superiority over basic methods such as Logistic Regression in clinical prediction evaluations as per Christodoulou et al. (2019). The findings from this research confirm previous results by showing that Random Forest exhibits minimal accuracy benefits that must be considered against its increased processing requirements. Random Forest ensemble methods demonstrate practical benefit in predictive relationship complexity levels and individual dataset features although they achieve better overall outcome.

7.2 Deep-Dive into SHAP and LIME Integration and Performance

The joint application of SHAP and LIME systems served as a primary method to enhance transparency throughout predictive models. SHAP enabled consistent theoretical explanation structures

which applied to both global and individual model levels through its game theory foundations. This project faced limitations because KernelSHAP consumed significant processing power while generating feature explanations although it needed real-time response speed.

LIME offered immediate clinical interpretation support through its rapid calculation method of delivering local explanations. The use of local surrogate models by LIME sometimes produced variations in the explanations throughout similar cases. This tool showed occasional changes because it worked best as a supplementary tool that delivered quick interpretable imagery instead of creating a wide range of detailed analytic insights.

Ennab and Mcheick (2024) mentioned identical drawbacks in their research where they promoted the combination of several explainability techniques for improved analysis. This study implements SHAP as a stable interpretive tool for broader assessment and LIME for enhancing individual prediction understanding according to the recommendation provided by Ennab and Mcheick (2024).

According to Monnet et al. (2024) healthcare providers wanted diagnostic information that follows the patterns used in clinical practice. These explanations from SHAP combined with the rules from LIME accommodate medical reasoning patterns directly. Risk assessment data using SHAP clearly displayed how elevated glucose levels as well as increased BMI raised diabetes probability as clinicians understand from existing practice. The simple rules presented by LIME through statements like "rising glucose beyond specified limits elevates risk substantially" enabled clinicians to take quick clinical actions during medical situations. Multiple analytical techniques for explanation built reliable and trustworthy medical models which showed better functionality when used in hospital contexts.

7.3 Detailed Examination of Web Application: Frameworks and Database Choices

The web application developed for deployment integrated Flask with React plus SQLite as its database core. The chosen technical stack was strategically chosen because it proved to be efficient with high flexibility and low resource requirements suitable for practical usage.

The Flask framework received selection as backend because it offers a compact framework design and fast development times together with seamless Python machine learning library integration. One advantage of Flask over Django is its lightweight design which reduces system overhead that provides necessary resources through SHAP computation.

The frontend interface used React for building a user-friendly interface and for dynamic updates of visualizations shown through responsive components. Explanation visualization updates became more responsive due to its virtual DOM functionality that managed efficient page updates without persistent screen refreshes. React's state management capabilities delivered instant user interactions leading to an interface that brought essential user-friendly functionality to clinical settings which require quick and clear operations.

SQLite operated effectively as the database layer because its developers chose it for its simplified design along with minimum system demands. Healthcare applications of mid-sized dimensions achieve fast data operations through SQLite which operates without requiring any extra server resources. The implementation of SQLite displayed limitations when the system needed to handle both substantial data volumes and possible expansion over time. Building at scale requires migrating to database tools PostgreSQL and MongoDB to gain performance enhancements as well as scalability features and simultaneous user system support.

The combination of Flask with React and SQLite provided an efficient code solution which delivered quick shifts between programming states together with effortless platform maintenance plus adequate results handling for restricted datasets. The selected decision points match what researchers adopt in their current experimental implementations according to data from Mohsen et al (2023).

7.4 User Interaction and Survey Insights

Analyzing feedback from a targeted survey involving twelve participants produced valuable insights into practical usability and trust in the predictive system. Participants responded to three primary questions:

1. **Did you find the predictions provided by the web application clear and easily understandable?**

   The respondents noted excellent clarity since both SHAP visualization and straightforward LIME text explanations together confirmed the useful aspects of explainable models.

2. **Did the explanations from SHAP and LIME systems help users understand why certain risk predictions were made?**

   All study participants stated the explanations boosted their trust through visual SHAP charts which displayed important feature influences effectively.

3. **Which additional features or functionality would help you have better user experience better?**

   The users wanted easy to manipulate number entry techniques and automatic health record downloads from wearables to get more desirable user interfaces.

The study confirms that understandable prediction models accompany user satisfaction and trust as reported by Monnet et al. (2024) in their research.

7.5 Analytical Comparison to Previous Studies and Innovation Points

The review demonstrates it identifies connections with prior research while presenting original concepts from this present work. Research studies show that diabetes prediction using explainable AI has become a critical field according to Mohsen et al. (2023). The research delivers practicable solutions to deal with the recognized gap through its combination of proven interpretability techniques into usable front-end capabilities.

This research creates a hands-on example of optimizing performance between interpretation and computation while developing strategic solutions for complex operations such as SHAP value processing within resource-limited systems. The established equilibrium represents an innovative aspect that previous studies have not explicitly explored.

Many research works conducted theoretical assessments but neglected to rigorously investigate actual situations where users engaged with the system. This project unites theoretical designs

with direct user input through usability tests to create a fully practical evaluation method. The actual value of clinical findings along with their practicality improves using research approaches that align with reality.

7.6 Computational and Resource Constraints: Analytical Reflection and Recommendations

The computational intensity of KernelSHAP negatively affected performance when operating under resource limitations mainly during real-time applications. Future implementations should use explainable methods such as TreeSHAP on a larger scale because of its efficient performance within tree-based models.

The storage of duplicate user input explanations in memory caches will minimize the system's live computational requirements. The execution of asynchronous computing strategies in parallel threads or distributed computing platforms serves to reduce resource constraints effectively.

7.7 Ethical Analysis and Additional Reflections

Previous brief mentions of ethics have led to examining essential ethical aspects that emerge from broader application and various types of data. The application scope of this dataset becomes limited because its data originates from Pima Indian women without proper inclusion criteria clarification. Research necessitates active efforts to acquire datasets which capture diverse populations because this enhances both model fairness and population representation.

The fundamental need exists for transparency that comes from clear presentations of predictions' uncertain outcomes. All predictive information must exactly state its supportive role which supports clinical decision making as the most crucial element that cannot be replaced.

Ethical duty encompasses protecting the privacy of the data collected from subjects. In practice deployments the use of anonymized datasets would need to be paired with strict data governance which upholds patient confidentiality and respects healthcare legislation.

7.8 Final Analytical Reflections and Strategic Recommendations

This thorough review of technical aspects together with analytical considerations strengthens the thesis by demonstrating its strong approach and meaningful practical applications and inventive solutions. XAI integration success inside a suitable clinical framework creates a basic structure for future research development and expansion programs.

Strategic recommendations for future work include integrating larger, diverse datasets, scaling computational resources to accommodate advanced machine learning techniques, and continuing direct user engagement through iterative design processes. Furthermore, exploring advanced explainability methods like counterfactual or concept-based explanations could provide additional depth and innovation in future predictive applications.

This analytical display presents a thorough understanding of present trends and both restrictions and forthcoming prospects which proves why explainable AI should be used for personalized diabetes risk evaluation.

8 Conclusion

8.1 Key findings and Implications

The study investigated Explainable Artificial Intelligence (XAI) approaches when used for personalized diabetes risk assessment. Implementing Logistic Regression and Random Forest models together with SHAP and LIME explained their results to show users how effective diabetes risk prediction works while keeping decision processes transparent.

The results of model performance assessment indicated Random Forest obtained slightly higher results than Logistic Regression over multiple evaluation metrics where both scored 76% and 74% accuracy. A minimal difference in performance demonstrates that each model provides suitable methods to predict diabetes risk. The main discovery here shows that explainability methods added value to all performance outcomes despite the complexity levels of the underlying models.

Both SHAP and LIME worked as supplementary explanation tools which catered to different user requirement types. Using SHAP we obtained explanations which demonstrated that glucose levels together with BMI and age formed the key risk determinants for diabetes according to established medical knowledge. LIME provided basic rule-based interpretations which users found easy to comprehend due to their rapid comprehension nature. Both methods proved efficient when combined because they created a detailed understanding of model-based predictions which surpassed individual capabilities.

The web application development showed how a healthcare framework could implement explainable AI in real-world applications. User reactions were extremely favorable because the visualization tools showed participants exactly what elements determined their assessment results. The findings prove explainability dramatically raises user trust levels and increases their engagement with AI medical tools thus resolving one main obstacle behind healthcare institutions adopting these technologies.

The study demonstrates the significance of open disclosure in healthcare AI services according to ethical standards. The system maintained proper ethical boundaries by both revealing its dataset

constraints and describing forecasted information as supportive clinical insights instead of confirmed medical diagnoses. Professional medical consultation remains necessary per the explicit disclaimer that appears within the system.

The combination of Flask, React, and SQLite worked well for the project's requirements, but the system faced limitations during KernelSHAP computation. Resolving performance issues of complex explanation techniques stands as a major practical challenge for limited-resource settings together with demands for optimization strategies in actual healthcare applications.

This study demonstrated that Random Forest outperformed Logistic Regression in model prediction yet preserved its interpretability properties and SHAP alongside LIME successfully enhanced prediction explainability to boost user trust which prompts successful health care deployment to demand three elements -resource considerations alongside user interface planning and ethical framework evaluation.

The research findings add value to technical practice and advance conversations about AI responsibility in healthcare applications. This research disproves the assumption that predictive accuracy must decrease when explainability features are added since it shows that both traits can exist together. The combination process between established machine learning models and state-of-the-art explanation techniques through an accessible interface creates a template for building trustworthy AI systems which healthcare professionals can utilize for their decision-making.

8.2 Future Work

Research direction and improvements are identified in this work to overcome implementation and real-world usage obstacles.

New research should focus on enlarging datasets which would include various demographic populations beyond Pima Indians to ensure higher model generality and fairness. The model generalizability would improve together with fairness which enables accurate predictions to span across multiple ethnic and racial groups. Monitoring the development of diabetes risk can be achieved through longitudinal data collection.

Explaining the predictions requires additional assessment beyond SHAP and LIME by employing new techniques involving counterfactual explanations assessing "what would need to change to alter this prediction" and concept-based explanations that connect predictions to medical field concepts. These methods would better match clinical decision patterns.

User feedback supports the development of mobile applications capable of extracting data from wearable health devices because this functionality would make the system more accessible and eliminate the need for manual data entry. The integration would make it possible to track information continuously which would result in risk evaluations occurring timelier.

Larger user testing with clinical validation studies featuring both healthcare professionals and additional users will deliver confirmed evidence about how well the system serves real-world clinical needs. Evaluation studies would measure two aspects: they would evaluate user satisfaction in addition to demonstrating how explanations enhance both comprehension and correct decision-making.

The computational expenses of explanation methods, especially KernelSHAP would improve system performance in operational environments through explanation caching methods and parallel processing methods and model-specific optimization techniques.

Extending the system beyond risk detection to provide individualized intervention choices based on risk-influencing factors would make the tool evolve from diagnostic to supportive in health improvement.

Federated Learning Techniques would enable the model to process distributed healthcare information without violating patient privacy while resolving data heterogeneity problems along with privacy boundaries.

The proposed future directions intend to improve explainable AI for diabetes risk prediction through updating the methodology while increasing healthcare scope and predictive capability. Through these research approaches scientists can establish a connection between sophisticated AI

systems and secure, dependable applications which serve healthcare providers alongside their patients.

References

Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable machine learning in healthcare. Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 559–560. https://doi.org/10.1145/3233547.3233667

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. Computational Statistics & Data Analysis, 120, 70–83. https://doi.org/10.1016/j.csda.2017.11.003

Blagus, R., & Lusa, L. (2017). Gradient boosting for high-dimensional prediction of rare events. Computational Statistics & Data Analysis, 115, 1–15. https://doi.org/10.1016/j.csda.2016.07.016

Boulesteix, A. L., Janitza, S., Kruppa, J., & König, I. R. (2015). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 5(6), 249–264. https://doi.org/10.1002/widm.1072

Cawley, G. C., & Talbot, N. L. (2016). On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research, 11, 2079–2107.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., & Stewart, W. (2016). RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In Advances in Neural Information Processing Systems (Vol. 29, pp. 3504–3512).

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology, 110, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ, 350, g7594. https://doi.org/10.1136/bmj.g7594

Ennab, M., & Mcheick, H. (2024). Enhancing interpretability and accuracy of AI models in healthcare: A comprehensive review on challenges and future directions. Frontiers in Robotics and AI, 11, 1444763. https://doi.org/10.3389/frobt.2024.1444763

Fawcett, T. (2016). Introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Ghorbani, A., Abid, A., & Zou, J. (2019). Interpretation of neural networks is fragile. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 3681–3688. https://doi.org/10.1609/aaai.v33i01.33013681

International Diabetes Federation. (2021). IDF diabetes atlas (10th ed.). International Diabetes Federation.

Isfafuzzaman, T. U. N., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. Healthcare Technology Letters, 10(3), 79–88. https://doi.org/10.1049/htl2.12039

Khalid, S., Khalil, T., & Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. 2014 Science and Information Conference, 372–378. https://doi.org/10.1109/SAI.2014.6918213

Kuhn, M., & Johnson, K. (2019). Feature engineering and selection: A practical approach for predictive models. CRC Press. https://doi.org/10.1201/9781315108230

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv. https://doi.org/10.48550/arXiv.1802.03888

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Mandrekar, J. N. (2015). Receiver operating characteristic curve in diagnostic test assessment. Journal of Thoracic Oncology, 5(9), 1315–1316. https://doi.org/10.1097/JTO.0b013e3181ec173d

Merrick, L., & Taly, A. (2020). The explanation game: Explaining machine learning models using Shapley values. In International Cross-Domain Conference for Machine Learning and Knowledge Extraction (pp. 17–38). https://doi.org/10.1007/978-3-030-57321-8_2

Ming, Y., Xu, P., Qu, H., & Ren, L. (2018). Interpretable and steerable sequence learning via prototypes. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1494–1502. https://doi.org/10.1145/3219819.3220072

Mohsen, F., Arnrich, B., Rigoll, G., & Salah, A. A. (2023). A scoping review of artificial intelligence-based methods for T2DM risk prediction. npj Digital Medicine, 6(1), 197. https://doi.org/10.1038/s41746-023-00933-5

Molnar, C. (2020). Interpretable machine learning: A guide for making black box models explainable. https://christophm.github.io/interpretable-ml-book/

Monnet, X., Maslove, D. M., Baid, H., & Chen, J. H. (2024). Should AI models be explainable to clinicians? Critical Care, 28(1), 247. https://doi.org/10.1186/s13054-024-05005-y

Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the Gini importance? Bioinformatics, 34(21), 3711–3718. https://doi.org/10.1093/bioinformatics/bty373

Nusinovici, S., Tham, Y. C., Chak Yan, M. Y., Wei Ting, D. S., Li, J., Sabanayagam, C., Wong, T. Y., & Cheng, C. Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. Journal of Clinical Epidemiology, 122, 56–69. https://doi.org/10.1016/j.jclinepi.2020.03.002

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3), e1301. https://doi.org/10.1002/widm.1301

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. arXiv. https://doi.org/10.48550/arXiv.1811.12808

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2019). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 180–186). https://doi.org/10.1145/3375627.3375830

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. In Machine Learning for Healthcare Conference (pp. 359–380).

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. PLOS ONE, 14(11), e0224365. https://doi.org/10.1371/journal.pone.0224365

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17(1), 230. https://doi.org/10.1186/s12916-019-1466-7

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887.

Wong, T. T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognition, 48(9), 2839–2846. https://doi.org/10.1016/j.patcog.2015.03.009

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing, 415, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

Zeng, X., & Luo, G. (2017). Progressive sampling-based Bayesian optimization for efficient and automatic machine learning model selection. Health Information Science and Systems, 5(1), 2. https://doi.org/10.1007/s13755-017-0023-z

Zou, H., & Hastie, T. (2017). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.

Appendices

Appendix 1. Survey Methodology and demographics

Participants ID	Age	Background
P1	34	Registered Nurse
P2	42	Registered Nurse
P3	29	IT Specialist
P4	35	Software Developer
P5	31	Software Developer
P6	66	Diabetic Mellitus Patient
P7	45	Diabetic Mellitus Patient
P8	32	Gestational Diabetic Patient
P9	26	Gestational Diabetic Patient
P10	37	Gestational Diabetic Patient
P11	55	No computing background nor diabetic patient
P12	36	No computing background nor diabetic patient

1.1 Quantitative Survey Results

Question 1: How easy was it for you to input your medical information and obtain results from the system?

(Rating scale: 1 = Very Difficult, 5 = Very Easy)

Participants ID	Rating	Key Comment
P1	5	"Intuitive interface with clear clinical terminology and appropriate input fields."
P2	5	The clinical workflow processes run in a straightforward manner according to normal healthcare standards.
P3	5	"Well-designed with appropriate validation and error messaging."
P4	5	"Clean UI with excellent form design and responsive feedback."

P5	4	The software design shows good overall quality but additional guidance for nonmedical users in medical input sections is needed.
P6	5	The placeholder examples together with tooltips proved helpful to me during use. The application allowed me to complete entries with assuredness.
P7	5	Users found the program simple to use between different sections along with an unambiguous input system.
P8	4	The application needs improved unit display labels specifically for glucose testing through mg/dL measurement.
P9	5	The application was fast to operate and the calculated outcomes made logical sense.
P10	5	The application display of changing numbers against results gave me valuable insight without forcing me to read complex charts.
P11	5	The application led users step-by-step without leaving them puzzled throughout the app experience.
P12	5	The application provided a user-friendly interface through which all information presented itself with basic explanations.

Question 2: Did the explanations provided by SHAP and LIME make the predictions clearer and more understandable?

(Rating scale: 1 = Not at all clear, 5 = Extremely clear)

Participants ID	Rating	Key Comment
P1	5	"Intuitive interface with clear clinical terminology and appropriate input fields."
P2	5	The application implements simple workflow patterns which stick to standard medical practice procedures.
P3	5	"Well-designed with appropriate validation and error messaging."
P4	5	"Clean UI with excellent form design and responsive feedback."

P5	5	The system displays excellent overall design whereas some portions of medical data entry need additional instructional aids to help non-medical staff.
P6	5	The placeable example and tooltips stood out to me as an appreciated feature of the interface. This system enabled me to enter data with assurance.
P7	5	Users found it easy to use the application's interface and the input procedure was uncomplicated.
P8	5	The application needs improved unit identification signs for certain data entry areas such as glucose levels which use mg/dL as units.
P9	5	I found the application simple to use because the results displayed dependable information.
P10	5	The application display enabled me to observe relationships between my entered numbers without requiring complicated chart analysis.
P11	5	During my usage of the app all steps were presented in a clear manner which prevented me from getting confused.
P12	5	The application offered straightforward understanding while all explanations appeared simple to comprehend.

Question 3: How likely are you to trust and use this system for regular health risk assessments? (Rating scale: 1 = Very unlikely, 5 = Very likely)

Participants ID	Rat-ing	Key Comment
P1	4	Users feel confident about system recommendations because of the transparent approach it provides.
P2	4	Customers should use this system as an additional tool when consulting with patients.
P3	3	The system is trustworthy but I recommend producing the evaluation with bigger data collections.

P4	4	The explainability components enhance trustworthiness to a greater extent than traditional uninterpretable systems known as black boxes.
P5	3	System use is generally reliable yet patients should consult with doctors when using this system.
P6	4	The explanations boosted my confidence but I need to check predictions through a healthcare professional to truly put my trust in the system.
P7	3	The system has a positive approach yet I would choose medical guidance over it to make decisions. I'd use this only occasionally."
P8	4	The device would serve as my primary evaluation tool before I made appointments to see my doctor.
P9	4	The device becomes more practical for daily use if it integrates with fitness tracker and glucose monitor devices.
P10	4	I appreciate that the app provides clear information although it needs ability to store history data along with time-based trend analysis.
P11	3	The device provides valuable information yet I remain skeptical about its accuracy since there are insufficient clinical tests to support it.
P12	4	Very useful tool, especially for early awareness. The application becomes more practical when developers create a mobile version.

Appendix 2. Testing Protocol and Task Timeline

The usability testing process depended on a standardized testing method that produced consistent evaluation results for participants. This test protocol integrated both guide features with open exploration periods while letting users test the application on their own.

Stages	Descriptions	Duration
System Introduction	The project provides a concise summary of its goal with information explaining the functionality of the application.	5 minutes

Guided Feature Exploration	Walkthrough of input form, prediction result, SHAP and LIME explanation views.	10 minutes
Independent Scenario Task	The system enables users to input test data which they evaluate independently using provided system outputs.	15-20 minutes
Survey completion	Participants answered three questions using Likert-scale ratings during the survey process.	5-10 minutes
Follow up Discussion	A conversational segment provided users with a chance to give extra qualitative comments.	5-10 minutes

Appendix 3. Selected Frontend Code Snippets and Responsive Design

```
function LoginPage({ onLogin }) {
  const [username, setUsername] = useState('');
  const [password, setPassword] = useState('');
  const [message, setMessage] = useState('');

  const handleSubmit = async (e) => {
    e.preventDefault();
    try {
      const response = await fetch('http://localhost:5000/auth/login', {
        method: 'POST',
        headers: { 'Content-Type': 'application/json' },
        body: JSON.stringify({ username, password }),
      });
      const data = await response.json();
```

Appendix 3. 1 Code snippet for login function in frontend.

The authentication code uses a POST request to transmit username and password information to a backend API which awaits to receive a JSON response.

```
function DiabetesRiskPrediction({ userName, userEmail }) {
  const [formData, setFormData] = useState({
    age: '',
    bmi: '',
    bloodPressure: '',
    glucose: ''
  });
  const [result, setResult] = useState(null);
  const [showAdvanced, setShowAdvanced] = useState(false);
  const [loading, setLoading] = useState(false);
  const [error, setError] = useState(null);
  const reportRef = useRef(null);

  const handleChange = (e) => {
    setFormData({ ...formData, [e.target.name]: e.target.value });
  };
```

Appendix 3. 2 Code snippet that handles partly of the input fields.

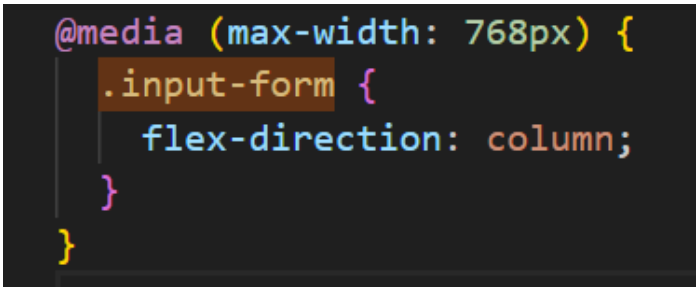
With React useState Hook this piece of code controls the state of health metric forms which allow users to input their age, BMI, blood pressure and glucose measurements.

```
function DiabetesRiskPrediction({ userName, userEmail }) {
      {result && (
        <section className="result-section">
          <h2>Prediction Result</h2>
          {formatResult(result)}

          <div className="result-actions">
            <button onClick={handleDownloadPDF} className="download-btn">
              Download Report as PDF
            </button>
            {userEmail && (
              <button onClick={handleEmailPDF} className="email-btn">
                Email Report to Me
              </button>
            )}
            <button onClick={handleClear} className="clear-btn">
              Start New Prediction
            </button>
          </div>
        </section>
      )}
    </div>
```

Appendix 3. 3 The implementation uses conditional rendering to manage the display of prediction results and explanatory visualizations.

The code shows prediction results conditionally while providing PDF download and emailing the report together with new prediction starting options regarding user actions.



Appendix 3. 4 Responsive design implementation using CSS media queries.

Devices smaller than 768 pixels will experience improved mobile usability since the Input-form lay-out adopts a vertical arrangement through this CSS media query.