information

IMPACT FACTOR 2.4

CITESCORE 6.9

Article

Towards Transparent Diabetes Prediction: Combining AutoML and Explainable AI for Improved Clinical Insights

Raza Hasan, Vishal Dattana, Salman Mahmood and Saqib Hussain

Special Issue
Medical Data Visualization

Edited by
Dr. Shuo-Chen Chien and Prof. Dr. Wen-Shan Jian

MDPI

https://doi.org/10.3390/info16010007

information

MDPI

Article

Towards Transparent Diabetes Prediction: Combining AutoML and Explainable AI for Improved Clinical Insights

Raza Hasan [1,*], Vishal Dattana [2], Salman Mahmood [3] and Saqib Hussain [4]

[1] Department of Computer Science, Solent University, Southampton SO14 0YN, UK
[2] Digital Transformation, Oman College of Management & Technology, P.O. Box 680, Barka 320, Oman; vdattana@ocmt.edu.om
[3] Department of Computer Science, Nazeer Hussain University, ST-2, near Karimabad, Karachi 75950, Pakistan; salman.mahmood@nhu.edu.pk
[4] Computer and Information Sciences, Northumbria University, Newcastle upon Tyne NE1 8QH, UK; saqib2.hussain@northumbria.ac.uk
* Correspondence: raza.hasan@solent.ac.uk

Abstract: Diabetes is a global health challenge that requires early detection for effective management. This study integrates Automated Machine Learning (AutoML) with Explainable Artificial Intelligence (XAI) to improve diabetes risk prediction and enhance model interpretability for healthcare professionals. Using the Pima Indian Diabetes dataset, we developed an ensemble model with 85.01% accuracy leveraging AutoGluon's AutoML framework. To address the "black-box" nature of machine learning, we applied XAI techniques, including SHapley Additive exPlanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), Integrated Gradients (IG), Attention Mechanism (AM), and Counterfactual Analysis (CA), providing both global and patient-specific insights into critical risk factors such as glucose and BMI. These methods enable transparent and actionable predictions, supporting clinical decision-making. An interactive Streamlit application was developed to allow clinicians to explore feature importance and test hypothetical scenarios. Cross-validation confirmed the model's robust performance across diverse datasets. This study demonstrates the integration of AutoML with XAI as a pathway to achieving accurate, interpretable models that foster transparency and trust while supporting actionable clinical decisions.

Keywords: AutoML; explainable AI (XAI); diabetes prediction; SHAP; LIME; counterfactual analysis; integrated gradients; attention mechanism; healthcare AI

Academic Editors: Shuo-Chen Chien and Wen-Shan Jian

Received: 14 November 2024
Revised: 13 December 2024
Accepted: 16 December 2024
Published: 26 December 2024

Citation: Hasan, R.; Dattana, V.; Mahmood, S.; Hussain, S. Towards Transparent Diabetes Prediction: Combining AutoML and Explainable AI for Improved Clinical Insights. *Information* **2025**, *16*, 7. https://doi.org/10.3390/info16010007

1. Introduction

Diabetes is a global health challenge, affecting millions and placing significant burdens on healthcare systems. Early detection and timely intervention are critical to managing the disease and preventing severe complications, including cardiovascular disease, kidney failure, and neuropathy. Machine learning (ML) advancements have enhanced the accuracy of diabetes prediction models; however, the "black-box" nature of many models limits their clinical adoption. Healthcare professionals often struggle to interpret these models, undermining trust and usability in decision-making.

This study addresses these challenges by integrating AutoML with XAI techniques to develop a predictive model that balances high accuracy with interpretability. Using the Pima Indian Diabetes dataset, we aimed to create a model that not only performs well but is also accessible and actionable for healthcare professionals.

The objectives of this study are to:

- Evaluate the performance of AutoML models enhanced with XAI techniques, such as SHAP and LIME, for accurate and interpretable diabetes risk prediction.
- Ensure the model's robustness and applicability across diverse populations through feature engineering, cross-validation, and data augmentation.
- Provide global and local interpretability using techniques like SHAP, LIME, IG, and CA.
- Leverage AutoGluon's ensemble capabilities to optimize model configurations, balancing accuracy, robustness, and computational efficiency.

While prior studies, such as [1,2] have explored XAI methods like SHAP and LIME for diabetes prediction, this work extends those efforts by automating model selection through AutoML and incorporating a broader range of interpretability techniques. For example, SHAP and LIME offer insights into feature importance, while CA allows clinicians to simulate how changes in patient characteristics, such as Body Mass Index (BMI) or glucose levels, could alter risk predictions. For instance, ref. [3] utilized counterfactual explanations to generate personalized recommendations for type 2 diabetes prevention. However, their approach was limited to specific biomarkers and a single dataset. In contrast, this study combines CA with other XAI techniques, such as SHAP and LIME, to create a more comprehensive and generalizable interpretability framework.

To bridge the gap between machine learning advancements and clinical usability, an interactive Streamlit application was developed. This tool enables healthcare professionals to:

- Visualize global and local feature importance.
- Understand model predictions for individual patients.
- Explore hypothetical scenarios through CA for personalized interventions.

This study makes several significant contributions to healthcare AI:

- By integrating AutoML with XAI techniques like SHAP, LIME, IG, and CA, it addresses the dual needs of predictive accuracy and interpretability, essential for clinical adoption.
- The interactive application provides clinicians with an intuitive platform for exploring and interpreting model predictions, enhancing usability and trust.
- Through advanced feature engineering and validation across diverse datasets, the model demonstrates strong generalization capabilities, making it suitable for deployment in various healthcare contexts.

The paper is structured as follows: Section 2 reviews current diabetes prediction methods, identifying existing gaps and opportunities. Section 3 details the methodology, including the integration of AutoML and XAI. Section 4 presents the experimental findings, followed by a Discussion in Section 5, which explores the implications, limitations, and future directions. The paper concludes in Section 6 with key takeaways and suggestions for advancing interpretable and reliable predictive models in healthcare.

2. Literature Review

Diabetes is a major health challenge worldwide, and early detection is essential for effective management and improved outcomes. This review examines how machine learning (ML), specifically AutoML and XAI, is being applied to diabetes prediction. We categorize current research into three themes: studies supporting AutoML's benefits, critiques highlighting its limitations, and alternative approaches aimed at enhancing interpretability in ML models for healthcare.

2.1. Support for AutoML in Diabetes Prediction

AutoML is rapidly emerging as a transformative tool in healthcare, enabling non-experts to develop and deploy machine learning models efficiently, which broadens AI

accessibility across clinical and research environments [4,5]. By automating model selection and optimization, AutoML has demonstrated the potential to match or even surpass expert performance in certain diagnostic and predictive tasks, potentially improving health outcomes and reducing costs [4]. AutoML applications have spanned various areas, including cancer diagnosis, cardiovascular disease, and COVID-19 research, illustrating its versatility in healthcare [5]. However, challenges remain, particularly in scaling AutoML to accommodate large, complex datasets and seamlessly integrating it into clinical workflows—a necessity for practical healthcare applications [4].

A particularly promising yet underexplored application of AutoML is in clinical notes analysis, which could revolutionize patient care by enabling more comprehensive extraction of patient information and risk factors from unstructured data [6]. Furthermore, AutoML platforms are positioned to democratize AI in medicine by lowering technical barriers, facilitating medical education, and supporting research and clinical practice [7]. However, for AutoML to reach its full potential in healthcare, clinicians must be equipped with knowledge for its ethical and effective application, with a focus on data-centric development and rigorous model validation [7]. This knowledge will ensure that healthcare professionals apply AutoML responsibly and are prepared to address the unique challenges of clinical deployment.

Many studies point to the effectiveness of AutoML in increasing prediction accuracy for diabetes diagnosis. For instance, ref. [8] show that AutoML can automate the model selection process, enabling healthcare providers to leverage advanced ML tools with minimal technical expertise. This not only speeds up diagnosis but also improves accuracy. Similarly, ref. [9] report that AutoML can efficiently handle large and complex datasets, a critical need in diabetes prediction due to the diverse factors involved.

2.2. Limitations of AutoML

Despite its benefits, some researchers highlight significant limitations of AutoML, particularly around transparency. Ref. [10] note that while AutoML can improve model performance, it often produces "black-box" models that make it difficult for clinicians to understand the reasoning behind predictions. This lack of interpretability is a barrier to clinical adoption. Ref. [11] add that the complexity of AutoML may reduce trust among clinicians, emphasizing the need for approaches that make the models' decision processes clearer to healthcare providers.

2.3. Alternative Approaches to Improve Interpretability

To address these limitations, researchers are exploring ways to combine AutoML with XAI techniques. For example, ref. [12] investigate using SHAP and LIME to make AutoML predictions more interpretable. Their findings suggest that these XAI techniques can help clinicians better understand the factors influencing predictions, making AutoML models more transparent and acceptable for clinical use.

Ref. [3] developed a novel methodology for generating counterfactuals to recommend personalized biomarker modifications for Type 2 diabetes prevention. Their approach focused on minimal changes to features such as fasting blood sugar and BMI to transition individuals from high to low risk. Our study builds upon this work by integrating counterfactuals with a broader XAI framework, including SHAP and LIME, enhancing interpretability and usability in diverse clinical contexts. Table 1 provides a detailed comparative analysis of the two approaches.

Table 1. Comparative Analysis of Counterfactual Studies.

Aspect	Lenatti et al. [3]	This Study
Objective	Recommend personalized biomarker modifications	Integrate CA with XAI techniques for broader clinical usability
Dataset	Canadian EMR dataset	Pima Indian Diabetes dataset and generalization across multiple datasets
Features Studied	Fasting blood sugar, BMI, HDL, and triglycerides	Glucose, BMI, Age, Blood Pressure, and other risk factors
Counterfactual Focus	Minimal viable changes for Type 2 diabetes prevention	Hypothetical scenarios for interpretability and actionable insights
XAI Techniques	Counterfactual explanations only	Counterfactuals combined with SHAP, LIME, and IG
Clinical Applicability	Focused on biomarker reduction	Broader interpretability for both global and local model explanations
Strengths	Personalized preventive recommendations	Enhanced usability through AutoML and explainable frameworks
Limitations	Limited to specific biomarkers and a single dataset	Requires further clinical validation for deployment in diverse populations

Comparison of Existing Studies

Table 2 presents a comparative analysis of the findings from several key studies on diabetes prediction, highlighting their strengths and limitations. While many prior studies have demonstrated the predictive power of machine learning algorithms, few have emphasized the importance of interpretability, which is crucial for clinical adoption. For instance, studies such as those by Tigga et al. [13], Kumari et al. [14], and Sisodia et al. [15] achieved high accuracy with machine learning models but did not adequately address transparency and model generalization. Our study builds on these efforts by incorporating AutoML with XAI techniques, offering a more comprehensive and interpretable solution for diabetes prediction.

Table 2. Comparison of Researchers' Findings with Previous Works.

#	Technique	Accuracy (%)
[13]	Logistic Regression	74.4
	K Nearest Neighbour	70.8
	Support Vector Machine	74.4
	Naive Bayes	68.9
	Decision Tree	69.7
	Random Forest	75.0
[14]	Support Vector Machine (RBF Kernel)	78.0
	Support Vector Machine	65.1
[15]	Naive Bayes	76.3
	Decision Tree	73.8
[16]	J48 Decision Tree	74.78
	Random Forest	79.57
	Naive Bayes	78.67
[17]	Laplacian Support Vector Machine	82.0
[18]	Linear Support Vector Machine	83.0
	RBF Support Vector Machine	82.0
[19]	J48 Decision Tree	75.65
	Random Forest	73.91
	Naïve Bayes	77.83
[20]	Logistic Regression & Regression Tree	77.48

2.4. Key Findings and Gaps

Table 3 summarizes key studies related to diabetes prediction using various machine learning techniques, highlighting their findings and limitations.

Table 3. Summary of Key Studies in Diabetes Prediction.

#	ML Techniques	Key Findings	Limitations
[8]	Supervised Learning (SVM, Decision Trees)	85% of studies used supervised algorithms; emphasizes the transformative potential of ML in diabetes management.	Lacks focus on interpretability of models for clinical application.
[9]	Various (Feature Selection, Imputation)	High performance metrics achieved through data preprocessing optimize diabetes prediction models.	Does not address the interpretability of predictions, crucial for clinical use.
[10]	Evaluation Framework	Developed a structured framework for evaluating ML techniques in diabetes detection; emphasizes rigorous assessment.	Lacks direct comparison of specific prediction techniques' effectiveness.
[21]	Deep Learning	Introduced a formal concept analysis framework for explaining deep learning outcomes; addresses interpretability issues.	Limited to a two-class classification problem, limiting broader applicability.
[12]	SHAP, Various ML Models	Developed a self-explainable interface for diagnosing diabetes, enhancing understanding of risk factors.	Relies on extensive clinical data, necessitating further research on model complexity.
[11]	Machine Learning (Various)	Highlighted the capability of ML models in predicting diabetes complications, enabling early intervention.	Challenges with data quality and parameter selection remain.
[22]	Machine Learning (Various)	Identified key risk factors (e.g., HbA1c levels) for forecasting complications; enhances predictive power.	Missed opportunities to incorporate a wider range of risk factors.

2.5. Gaps Analysis

Table 4 outlines the gaps identified in the literature and the corresponding objectives of this research.

Table 4. Summary of Gap Analysis.

Gap	Objective
Lack of interpretability in AutoML models	Develop and evaluate explainable AutoML techniques for diabetes prediction.
Limited clinical applicability of existing models	Investigate user-friendly interfaces that enhance clinician trust and understanding.
Insufficient integration of diverse risk factors	Create comprehensive models that consider a wider range of patient data for improved predictions.

By addressing these gaps, this research aims to contribute to the development of more interpretable and clinically applicable machine learning models for diabetes prediction, ultimately supporting better health outcomes.

This literature review highlights the progress in applying AutoML and XAI techniques to diabetes prediction, demonstrating the promise of machine learning in healthcare. However, significant gaps remain in terms of model interpretability, clinical adoption, and generalization. The current study aims to address these challenges, contributing to the development of more effective and interpretable diabetes prediction systems for real-world healthcare applications.

3. Methodology

This section outlines the methodology employed in this research to develop, evaluate, and interpret machine learning models for diabetes prediction. The methodology is structured into several key components: dataset preparation and data splitting, model training and architecture using AutoGluon [23], implementation of explainable AI techniques, and evaluation metrics. To enhance diabetes risk prediction with interpretable and automated techniques, we followed a structured methodology. The following stages outline our approach as shown in Figure 1.

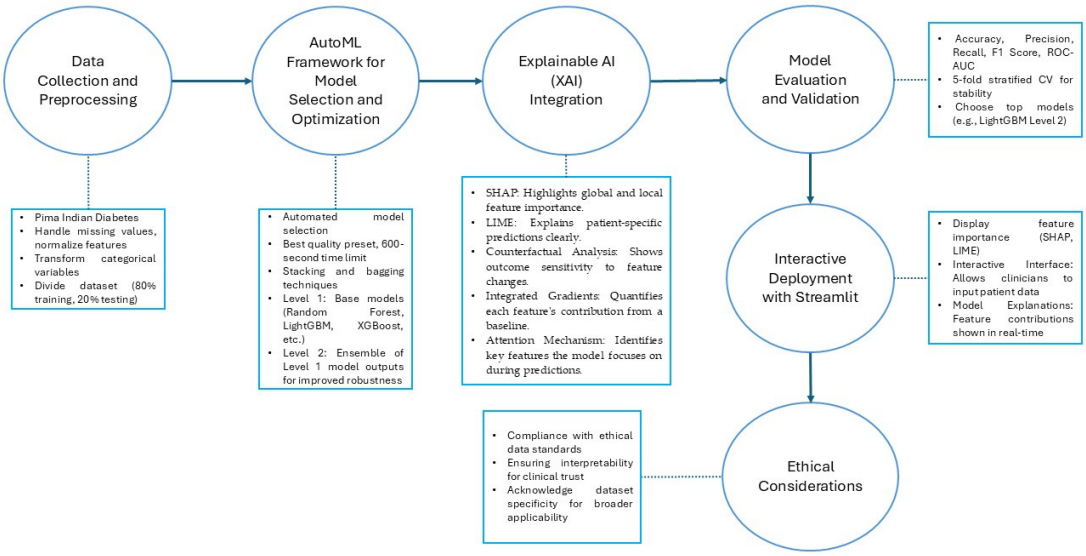

Figure 1. Model for Diabetes Prediction with AutoML and XAI.

3.1. Dataset Preparation and Data Splitting

The diabetes dataset, containing several biometric and lifestyle features, was processed to maximize predictive quality and ensure reliable, unbiased model evaluation [24].

3.1.1. Features and Target Variable

Input Features: The dataset includes the following features Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age. Each feature provides specific health and demographic data known to correlate with diabetes risk.

Target Variable: The target variable, Outcome, is binary, indicating the presence (1) or absence (0) of diabetes in an individual.

3.1.2. Train-Test Split

The dataset was divided into training and testing sets with an 80–20 split, respectively [25]. This approach allowed the model to train on the majority of the data while setting aside a test subset for a rigorous evaluation of its generalizability.

Training Set: 80% of the data was used to train and tune the model, making it suitable for identifying patterns across various feature interactions.

Testing Set: 20% of the data, kept completely independent from the training phase, was reserved for evaluating the final model's performance, simulating its performance on unseen data. This split ensures that the results reflect the model's true performance in real-world applications, where it would encounter previously unseen samples.

3.1.3. Generalization Techniques

To further improve generalization and mitigate overfitting, several strategies were applied:

- Feature Engineering: Features like Glucose and BMI were standardized, and categorical encoding was applied where necessary. These transformations aimed to reduce the impact of feature scale disparities on model generalization [26].
- Data Augmentation: Synthetic examples were generated for underrepresented classes, helping to balance the dataset and improve generalizability [27].
- Cross-Validation: A 5-fold stratified cross-validation was implemented, preserving class distributions across each fold [28]. This ensured that the model was evaluated on multiple data splits, reducing variance and enhancing the robustness of results.

3.2. Model Training and Architecture with AutoGluon

AutoGluon was leveraged to automate the training and optimization process, selecting the best models through a process of stacking and ensembling. The configuration and setup were as follows:

3.2.1. AutoGluon Presets and Configuration

- Preset: The best_quality preset was chosen to prioritize model accuracy. This preset automatically configures training parameters to optimize performance, though it may require longer computation times.
- Time Limit: The training process was capped at 600 s. This constraint helped to manage computational resources while allowing enough time to explore a variety of model architectures and combinations.

3.2.2. Neural Network Architecture

The neural network architecture, as depicted in Figure 2, consists of:

- Input Layer: Eight neurons representing the dataset features—Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age.
- Hidden Layers:
  - Hidden Layer 1: 64 neurons with ReLU activation.
  - Hidden Layer 2: 32 neurons with ReLU activation and a dropout rate of 0.5 to mitigate overfitting.
  - Hidden Layer 3: 16 neurons with ReLU activation and a dropout rate of 0.5 to mitigate overfitting.
- Output Layer: A single neuron with a sigmoid activation function to predict the binary outcome (diabetes risk).

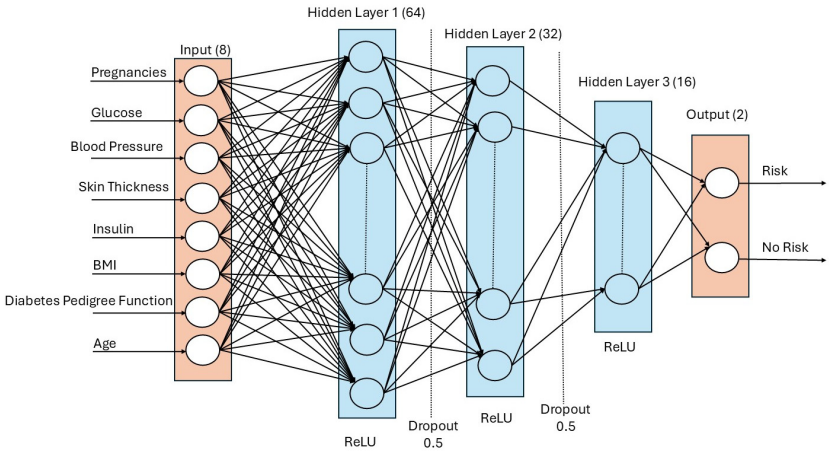

Figure 2. Neural Network Architecture for Diabetes Risk Prediction.

AutoGluon dynamically adjusted these hyperparameters using a Bayesian optimization framework to maximize model performance on the validation set. This optimization process included tuning the number of layers, neurons per layer, activation functions, and regularization techniques like dropout.

3.2.3. Stacking, Bagging, and Regularization Techniques

AutoGluon's stacked ensemble model combined a wide range of models trained independently at multiple levels. Stacking multiple models allows the ensemble to capture more complex patterns within the data.

- Level 1 Models: Base models included Random Forest, CatBoost, LightGBM, XGBoost, and neural networks. AutoGluon automatically selects and tunes the optimal neural network architecture through an automated search process. The framework tests various configurations, including different types of neural network architectures, to identify the best model for the dataset. AutoGluon also optimizes hyperparameters such as the number of layers, the number of neurons per layer, activation functions, and regularization techniques, including dropout. The optimal dropout rate is selected dynamically as part of the hyperparameter optimization process, along with other relevant parameters to prevent overfitting and ensure model generalization.
- Level 2 Ensemble: The Level 2 model layer took predictions from Level 1 models as additional features. Combining this information reduced variance, contributing to robust generalization.
- Bagging: Dynamic bagging involved creating multiple data splits to reduce variance, further aiding in model stability and resilience on unseen data [29].

3.2.4. Integration of AutoML, XAI Techniques, and Streamlit Application

This study employs AutoGluon, an AutoML framework, to automate model selection, optimization, and training processes. AutoGluon reduces the need for extensive manual tuning while achieving robust performance. The integration of XAI techniques, including SHAP, LIME, IG, AM, and CA, enhances the interpretability of the predictive models. Together, these methods ensure transparency and usability for clinical adoption.

- SHAP provides global and local explanations for feature importance, quantifying each feature's contribution to predictions. This technique identifies critical predictors, such as glucose and BMI, that influence diabetes risk, offering insights aligned with clinical knowledge.
- LIME focuses on individual predictions, creating localized surrogate models to explain why specific predictions were made. This enables healthcare professionals to understand patient-specific factors affecting the model's output.
- IG quantifies the contribution of each feature to a specific prediction by comparing model output differences between baseline and actual feature values. This approach provides deeper insights into how features like glucose and BMI influence risk predictions.
- The AM assigns weights to features, highlighting those most relevant during the model's decision-making process. This dynamic feature prioritization adds another layer of interpretability, ensuring predictions are understandable and actionable.
- CA explores hypothetical scenarios, such as how reducing a patient's BMI might alter their diabetes risk. This method is particularly valuable for supporting personalized interventions and care planning.

To enhance clinical usability, these XAI techniques are integrated into an interactive Streamlit application. The application provides the following functionalities:

- Visualization of global feature importance to understand overall model behavior.
- Case-specific interpretation of predictions using LIME and SHAP.
- Exploration of hypothetical scenarios through Counterfactual Analysis.

By incorporating these capabilities, the Streamlit application bridges the gap between machine learning innovation and practical healthcare needs, enabling clinicians to interact with predictions in a user-friendly manner.

3.3. Implementation of Explainable AI Techniques

To enhance model interpretability, explainable AI techniques are integrated into the model development process:

- SHAP values are computed to understand the impact of each feature on the model's predictions [30]. This helps in identifying the most influential factors in diabetes prediction [31].
- LIME is utilized to provide local interpretability of individual predictions, enabling clinicians to understand why certain predictions are made [32].
- CA was performed to illustrate how changes in feature values impact model predictions, aiding clinicians in understanding the model's decision-making logic [33]. Building upon [3], who focused on biomarker modifications to reduce diabetes risk, we adopt a comprehensive counterfactual framework. Unlike [3] methodology, which prioritized minimal feature adjustments, our approach leverages AutoML for model optimization and integrates various XAI methods to enhance interpretability. This combined approach ensures actionable and generalizable insights for clinicians.
- IG and AM: Additionally, IG and the AM are used to provide deeper insights into the model's decision-making process [34,35]. IG quantifies the contribution of each feature to a specific prediction, while the AM highlights which features the model focuses on most when making predictions. These techniques offer both global and local interpretability, reinforcing the model's transparency and trustworthiness.

3.4. Evaluation Metrics

To assess the performance of the developed models, several evaluation metrics are used:

- Accuracy, Precision, and Recall: These metrics provided a well-rounded assessment of model correctness and relevance.
- F1 Score: To balance precision and recall, the F1 score offered insight into the model's overall performance on imbalanced data.
- AUC-ROC Curve: The ROC-AUC metric was used to assess the model's discriminative power across thresholds [36].
- Balanced Accuracy and MCC: Balanced accuracy and the Matthews Correlation Coefficient (MCC) were included to provide a clearer measure of performance on imbalanced data [37].
- Cross-Validation Stability: 5-fold stratified cross-validation results were analyzed for low variance across folds, confirming that the model's performance was stable and generalizable.

3.5. Ethical Considerations in AI for Healthcare

The integration of AI in healthcare raises significant ethical concerns, particularly surrounding data privacy, patient consent, and algorithmic transparency. In this study, we ensure that ethical standards are maintained throughout the research process by adhering to best practices for data privacy and protection. The Pima Indian Diabetes dataset used in this study is publicly available and contains anonymized data, which ensures that personally identifiable information (PII) is not included. For future implementations involving clinical datasets containing sensitive patient information, we would adhere to relevant regulations such as the General Data Protection Regulation (GDPR) in Europe and the Health Insurance Portability and Accountability Act (HIPAA) in the United States. Stringent data anonymization and encryption protocols are employed to ensure the security and confidentiality of patient data throughout the study.

Additionally, ethical considerations around informed consent is critical, especially when AI models are applied in clinical settings. While the Pima Indian Diabetes dataset does not contain personal identifiers, any future use of clinical data would require obtaining informed consent from patients. This would ensure that patients are fully aware of how their data is being used for model training and prediction. The transparency of these processes is key in maintaining patient trust in AI systems.

Another important ethical issue is bias and fairness in AI. Machine learning models can inadvertently reinforce biases present in the training data, leading to disparities in prediction accuracy across different demographic groups. In this study, we employ techniques such as cross-validation on diverse patient populations and ensure that the features used in the model are representative of the target population's variability. This approach helps to mitigate bias and promotes fairness, ensuring equitable outcomes across all groups.

Transparency and explainability are also critical ethical principles in healthcare AI. For clinicians to trust and adopt AI systems, they must understand the reasoning behind model predictions. To this end, we incorporate Explainable AI (XAI) techniques like SHAP and LIME, which provide interpretable and actionable insights into the model's decision-making process, enhancing the model's transparency and making it more accessible to healthcare professionals. Finally, accountability remains a core concern. While AI can assist in decision-making, it is essential that clinicians retain final responsibility for patient care. The use of AI should complement, not replace, human expertise, ensuring that the ultimate responsibility for medical decisions remains with healthcare providers.

4. Results

The results are presented in four key areas: model performance metrics, cross-validation stability assessment, leaderboard model comparison, and model interpretability. Each area contributes to understanding the effectiveness, robustness, and transparency of the model for diabetes prediction.

4.1. Model Performance

The model's performance was evaluated on the primary test dataset using multiple metrics to provide a holistic assessment. Key results are summarized in Table 5.

Table 5. Model Performance Metrics on Test Set.

Metric	Description	Value
Accuracy	Overall correctness of predictions.	76.62%
Balanced Accuracy	Adjusted for class imbalance.	74.95%
MCC	Correlation between observed and predicted classes.	0.495
ROC-AUC	Model's ability to distinguish between classes, reflected in the area under the ROC curve.	0.774
F1 Score	Harmonic mean of precision and recall.	0.679
Precision	Proportion of positive predictions that are correct.	0.667
Recall	Proportion of actual positives identified correctly.	0.691

The model's performance is assessed with an accuracy of 76.62%, providing a general overview of its effectiveness. Additionally, a balanced accuracy of 74.95% and a ROC-AUC score of 0.774 highlight the model's reliability in accurately classifying both positive and negative cases. The MCC of 0.495 indicates a moderate correlation, suggesting that the model performs well even in scenarios with class imbalances.

4.2. Generalization

The model's generalizability was assessed using three datasets:

Training Dataset (PIMA Dataset): This is the primary dataset used to train and evaluate the model initially. The PIMA Indians Diabetes dataset is a well-known dataset in diabetes prediction, containing health indicators such as glucose levels, BMI, age, and more.

First Generalization Dataset (Diabetes Dataset from sklearn.datasets): The diabetes dataset from sklearn.datasets was used to further evaluate the model's ability to generalize to unseen data. This dataset includes several health-related features commonly used in diabetes prediction, such as Age, BMI, Blood Pressure, Glucose, and others. The dataset is typically used for classification tasks, where the goal is to predict whether an individual has diabetes or not based on their medical attributes.

Second Generalization Dataset (Rural African-American Patients): For the final generalization test, a dataset representing rural African-American patients was used. This dataset reflects a different demographic population, providing further insight into the model's ability to generalize across different patient groups.

4.2.1. Dataset Preparation and Feature Engineering

To ensure consistency across all datasets, the following preprocessing steps were applied:

Renaming Features: Some feature names were updated for clarity and to align with the primary dataset:

- 's1' or 'stab.glu' $\rightarrow$ Glucose
- 's2' or 'hdl' $\rightarrow$ Skin Thickness

Handling Missing Values: Missing values in critical features (Blood Pressure, Skin Thickness, Insulin) were imputed using median values from their respective columns.

Feature Engineering: Derived features ensured consistent data representation:

- BMI Calculation: $BMI = \frac{Weight(kg)}{(Height\ (m))^2}$
- Blood Pressure: Calculated as the average of systolic and diastolic readings.
- Diabetes Pedigree Function (DPF): $DPF = (0.2 * Age) + (0.3 * BMI) + (0.5 * Glucose)$
- Insulin Estimation: $Insulin = (Glucose * 0.05) + (BMI * 0.1) + 3$

Outcome Variable: The target variable, Outcome, was derived from glycated hemoglobin (glyhb) levels:

- Positive for diabetes: glyhb $\geq$ 6.5 (Outcome = 1)
- Negative: glyhb < 6.5 (Outcome = 0)

4.2.2. Generalization Performance

Performance metrics on the generalization datasets are presented in Table 6.

Table 6. Generalization Evaluation Metrics.

Metric	Description	Sklearn	Rural African-American
Accuracy	Overall correctness of predictions.	78.65%	91.36%
Balanced Accuracy	Adjusted for class imbalance.	78.55%	90.10%
MCC	Correlation between observed and predicted classes.	0.570	0.818
F1 Score	Harmonic mean of precision and recall.	76.54%	72.00%
Precision	Proportion of positive predictions that are correct.	75.61%	90.00%
Recall	Proportion of actual positives identified correctly.	77.50%	60.00%

4.2.3. Calculating Final Accuracy

Given the varying performance on each dataset, the average accuracy can be calculated to provide a consistent value across all datasets. A straightforward way to summa-

rize the model's performance across different datasets is to average the accuracies from each dataset:

$$\text{Average Accuracy} = \frac{\text{Accuracy from First Dataset} + \text{Accuracy from Second Dataset}}{2}$$

$$\text{Average Accuracy} = \frac{78.65\% + 91.36\%}{2} = 85.01\%$$

The model's performance was evaluated on three datasets under distinct conditions:

- Pima Indian Diabetes Test Set: Accuracy = 76.62%, reflecting performance on the primary dataset used for model training.
- Scikit-learn Diabetes Dataset: Accuracy = 78.65%, showing generalizability to an alternative dataset.
- Rural African-American Dataset: Accuracy = 91.36%, highlighting adaptability to a demographically different population.

The average accuracy across all datasets was 85.01%, calculated to provide an overall measure of generalization. These results demonstrate robust performance across diverse datasets and demographic conditions.

4.3. Cross-Validation Results

To ensure stability and minimize overfitting, a 5-fold stratified cross-validation was conducted, preserving class distributions across each fold. The accuracies achieved for each fold are detailed in Table 7.

Table 7. Cross-Validation Accuracy Results.

Fold	Accuracy (%)
1	76.1
2	75.5
3	74.8
4	77.3
5	76.7
Mean	76.08

The model demonstrated consistent performance across folds, with a mean accuracy of 76.08%, confirming robustness.

4.4. Model Leaderboard and Ensemble Comparison

AutoGluon's leaderboard ranked models based on validation accuracy. The top-performing models are detailed in Table 8.

Table 8. Leaderboard of Top Models by Validation Accuracy.

Model	Validation Accuracy (%)	Analysis
LightGBM Level 2	83.88	High accuracy and fast training, making it ideal for real-time predictions.
Weighted Ensemble Level 3	83.88	Improved accuracy by combining multiple models.
CatBoost Level 2	83.06	Well-suited for non-linear data and complex patterns.
XGBoost Level 2	83.06	Robust to imbalanced classes, valuable in medical contexts.

LightGBM Level 2 and Weighted Ensemble Level 3 demonstrated the best trade-offs between accuracy and computational efficiency, emphasizing their suitability for healthcare applications where timely and accurate predictions are essential.

4.5. Enhancing Model Transparency Through XAI Techniques

XAI techniques provided insights into the model's predictions, making them actionable and transparent for clinicians.

4.5.1. SHAP Analysis for Global and Local Interpretability

SHAP analysis provided quantitative insights into the contributions of individual features to the model's predictions. Glucose emerged as the most influential predictor, with an average SHAP value of 0.25, indicating its strong positive impact on diabetes risk classification. Similarly, BMI had an average SHAP value of 0.18, further emphasizing its role as a critical determinant of diabetes risk.

Beyond these primary features, the Diabetes Pedigree Function (SHAP value: 0.12) and Age (SHAP value: 0.09) also contributed meaningfully to the predictions, highlighting the importance of these secondary risk factors in the model's decision-making process.

Table 9 summarizes the SHAP values for key features, and Figure 3 illustrates the global feature importance scores. These results provide a detailed ranking of features, aligning with established clinical knowledge while enabling transparency in model interpretation.

Table 9. SHAP Global Feature Contributions.

Feature	Mean Abs SHAP Value
BMI	0.1979
DiabetesPedigreeFunction	0.1974
Glucose	0.1099
Insulin	0.0700
Age	0.0370
BloodPressure	0.0342
Pregnancies	0.0262

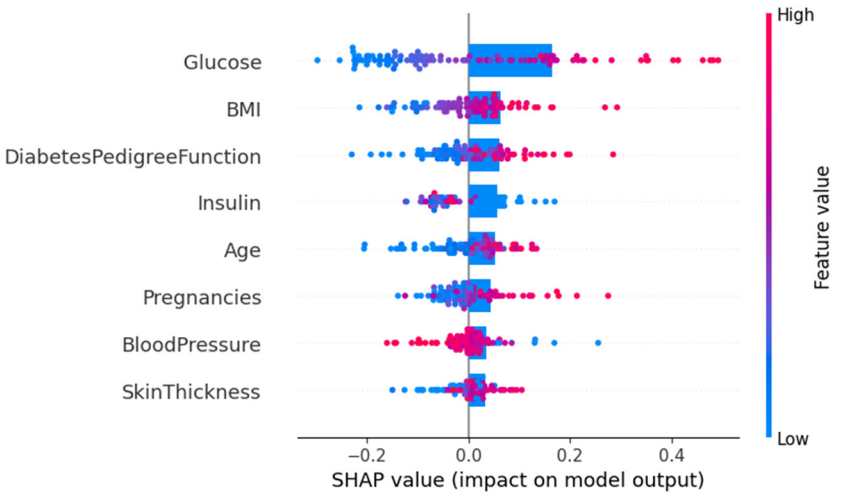

Figure 3. SHAP explanation plot for a Specific Instance.

4.5.2. LIME Analysis for Patient-Specific Interpretations

For a high-risk diabetes prediction, LIME identified BMI and the Diabetes Pedigree Function as the top contributors, with positive impacts on the outcome. Negative contributions, such as low Pregnancies, were also observed, offering a nuanced view of the model's predictions.

Table 10 and Figure 4 illustrate the LIME analysis for a specific prediction, showcasing the contributions of features. These insights empower clinicians to understand patient-specific risks and explore potential interventions.

Table 10. LIME Feature Contributions for a Specific Prediction.

Feature	Mean Abs SHAP Value
BMI > 36.60	+0.1622
DiabetesPedigreeFunction > 0.63	+0.1490
Pregnancies ≤ 1.00	−0.0673
SkinThickness > 32.00	+0.0581
Insulin > 127.25	−0.0456
BloodPressure ≤ 62.00	+0.0429
Age between 29 and 41	+0.0423
Glucose between 117 and 140	+0.0101

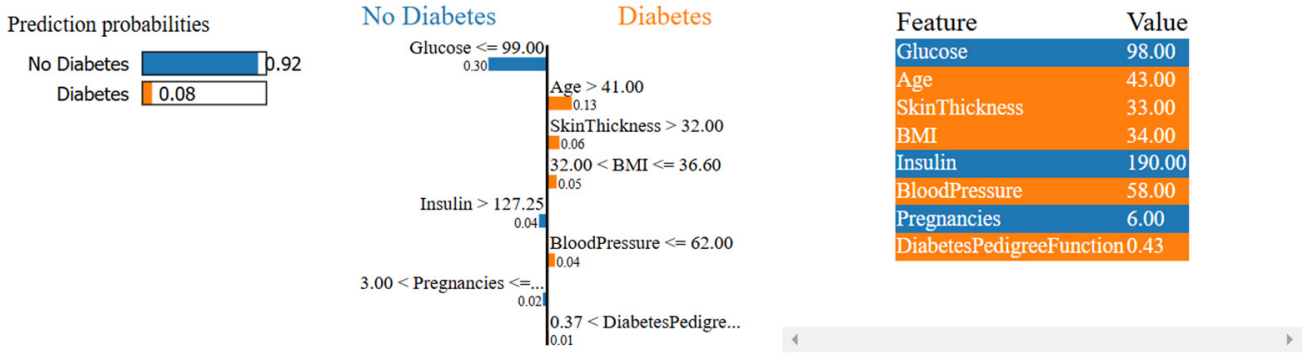

Figure 4. LIME Interpretation Plot for a Specific Prediction.

4.5.3. Counterfactual Instance Analysis

For a hypothetical scenario, a counterfactual instance was created by keeping most features constant and varying specific inputs (e.g., reducing BMI) to observe its impact on the prediction. In this analysis, we explored a scenario where we kept all input features constant but modified one or more to observe how these changes affect the model's prediction.

- Original Instance:
  - Features:
    - Pregnancies: 6
    - Glucose: 148
    - Blood Pressure: 72
    - Skin Thickness: 35
    - Insulin: 0
    - BMI: 33.6
    - Diabetes Pedigree Function: 0.627
    - Age: 50
- Counterfactual Instance:
  - Features:
    - Pregnancies: 6.000
    - Glucose: 148.000
    - Blood Pressure: 72.000
    - Skin Thickness: 35.000
    - Insulin: 0.000
    - BMI: 33.600
    - Diabetes Pedigree Function: 0.627
    - Age: 50.000

The comparison between the original and counterfactual instances, shown in Figure 5, demonstrates how reducing BMI could lower the predicted diabetes risk. This analysis provides actionable insights for clinicians, enabling them to simulate potential interventions and their effects on patient outcomes.

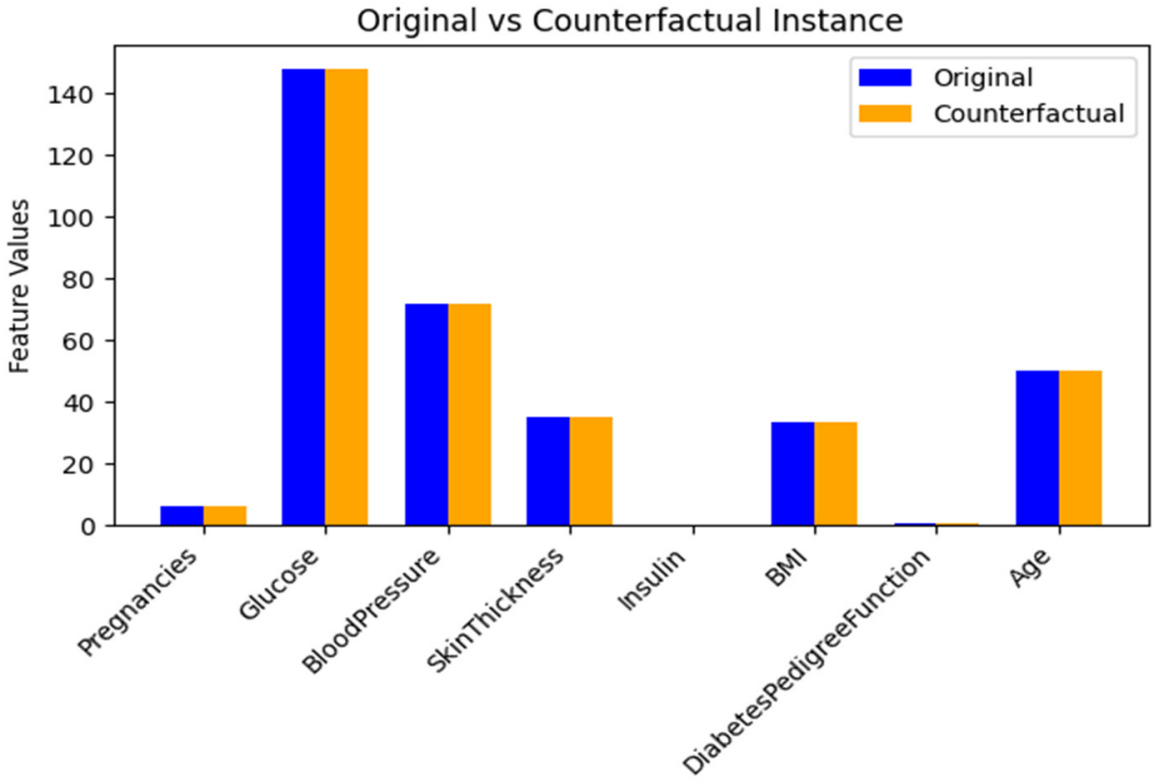

Figure 5. Comparison of Original and Counterfactual Feature Values.

4.5.4. Feature Attribution Using Integrated Gradients

In this study, IG attributed the highest scores to BMI (0.0629) and Glucose (0.0447), confirming their critical roles in diabetes prediction. Negative contributions, such as Blood Pressure (−0.0353), highlighted how certain features reduce the risk classification.

Table 11 summarizes IG attribution scores, while Figure 6 visualizes the feature contributions. Further validation, shown in Figures 7 and 8, demonstrates reliable attributions, emphasizing the dominance of BMI and Glucose in the model's predictions.

Table 11. Feature Attribution Scores and Their Impact Using IG.

Feature	Attribution Score	Impact
BMI	0.0629	Significant positive impact, highlighting that higher BMI increases the probability of a diabetes diagnosis.
Glucose	0.0447	Positive attribution, indicating that higher glucose levels are a strong indicator of diabetes risk.
Blood Pressure	−0.0353	Negative contribution, suggesting that lower blood pressure may slightly reduce the risk classification in this case.

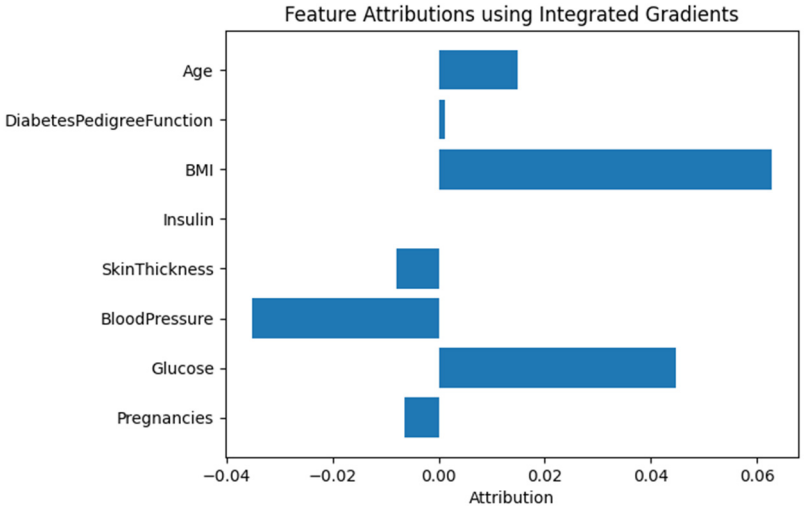

Figure 6. Feature Attribution Plot.

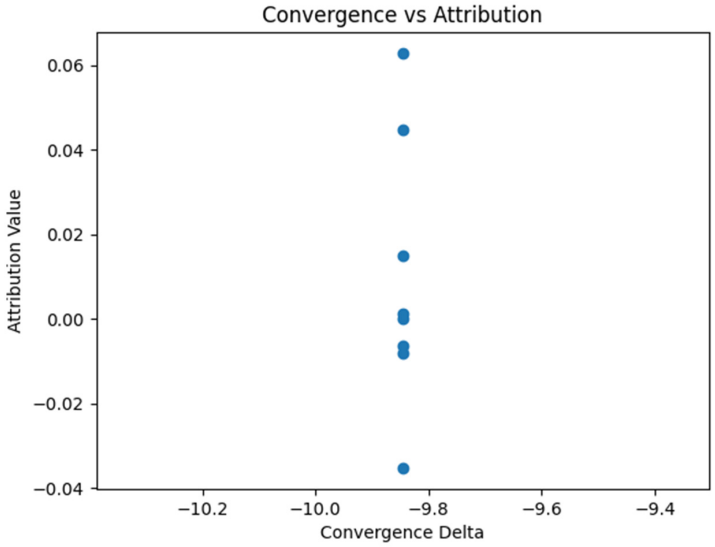

Figure 7. Convergence vs. Attribution Plot.

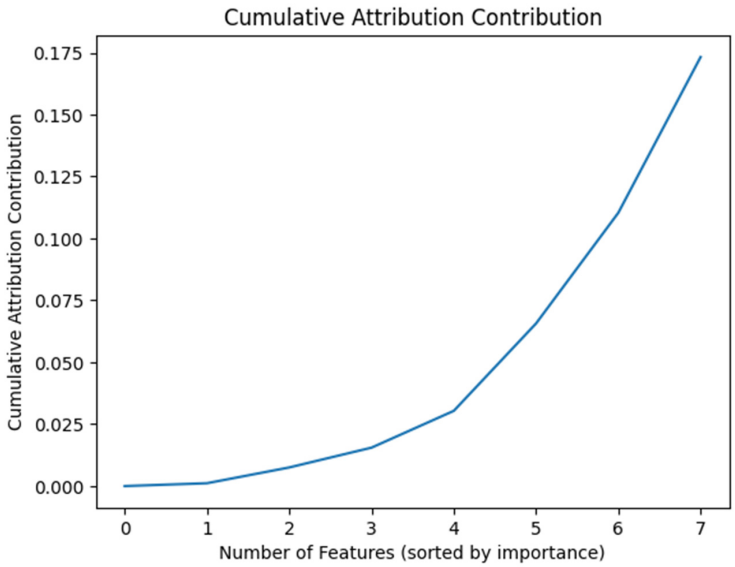

Figure 8. Cumulative Attribution Contribution.

4.5.5. Dynamic Feature Prioritization via Attention Mechanism

The AM assigns dynamic weights to input features, offering a complementary perspective to IG. By prioritizing features based on their relevance to specific predictions, AM adapts to varying patient profiles. Glucose and BMI consistently received the highest attention weights, as shown in Table 12.

Table 12. Key Insights of AM.

Insight	Description
Feature Focus	Higher attention weights indicate features with greater influence on the model's prediction. For example, Glucose and BMI consistently receive the highest attention scores, aligning with their clinical significance in predicting diabetes.
Dynamic Adaptation	Attention weights vary across instances, allowing the model to emphasize different features depending on the specific input data. This enables the model to adapt to varying patient conditions and adjust its focus as needed.

Figure 9 visualizes attention weights across predictions, highlighting the model's consistent focus on key features. Figure 10 provides a bar chart showing feature-specific attention scores, further emphasizing the clinical importance of Glucose and BMI.

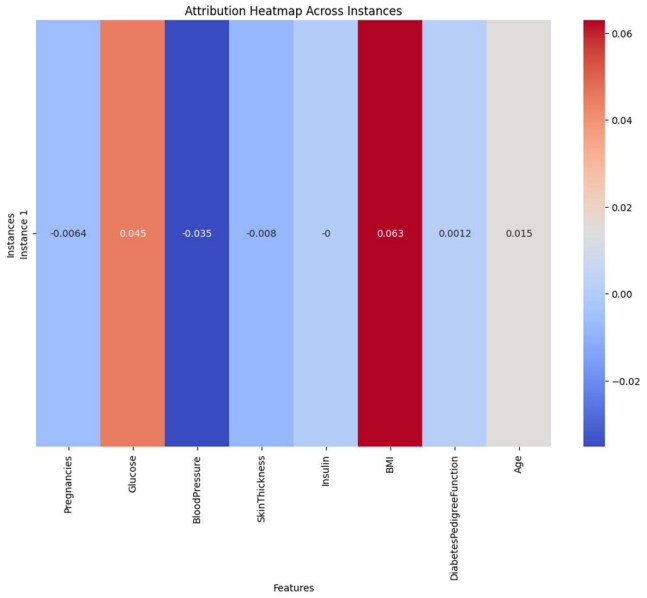

Figure 9. Attention Heatmap.

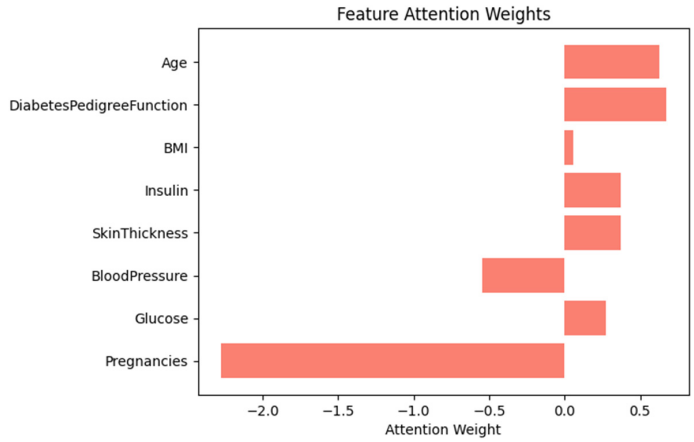

Figure 10. Feature Weight Bar Chart.

4.5.6. Interactive Visualization of Predictions Through Streamlit

The Streamlit application integrates SHAP, LIME, and CA to provide an interactive platform for exploring predictions. This tool enables clinicians to:

- Analyze global feature importance with SHAP.
- Explore patient-specific explanations using LIME.
- Simulate hypothetical scenarios with CA.

Interactive plots (Figures 11–14) enable users to visualize the influence of various features and interventions on predictions, providing valuable insights that support informed clinical decision-making and personalized care strategies.

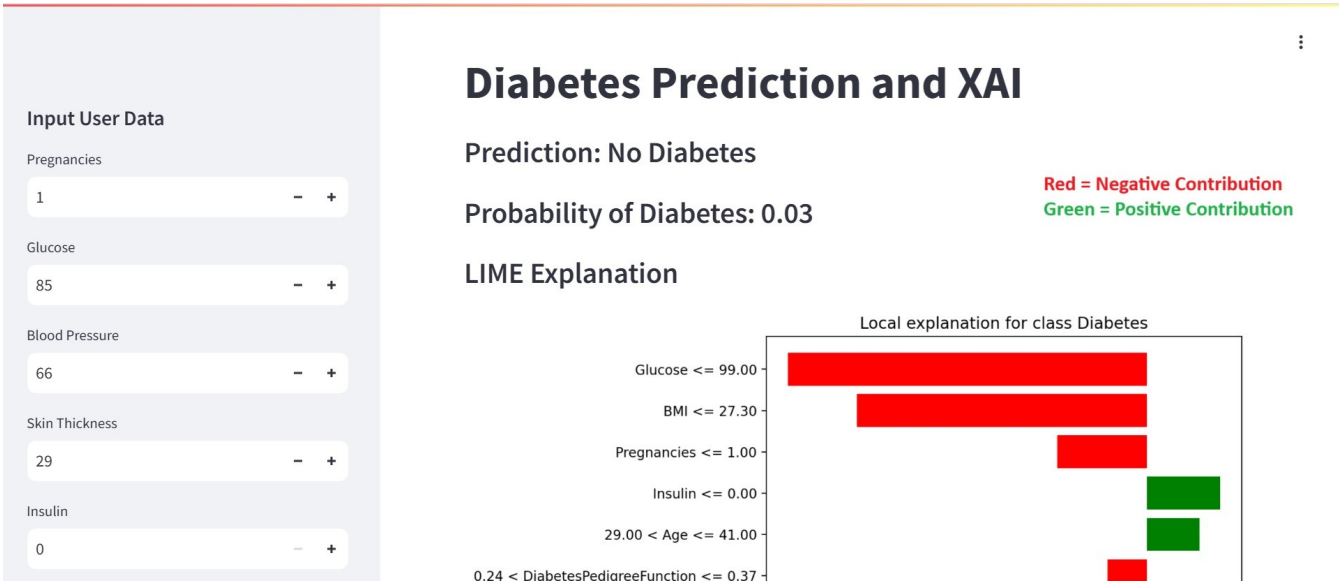

Figure 11. LIME Explanation Plot for a Specific Prediction.

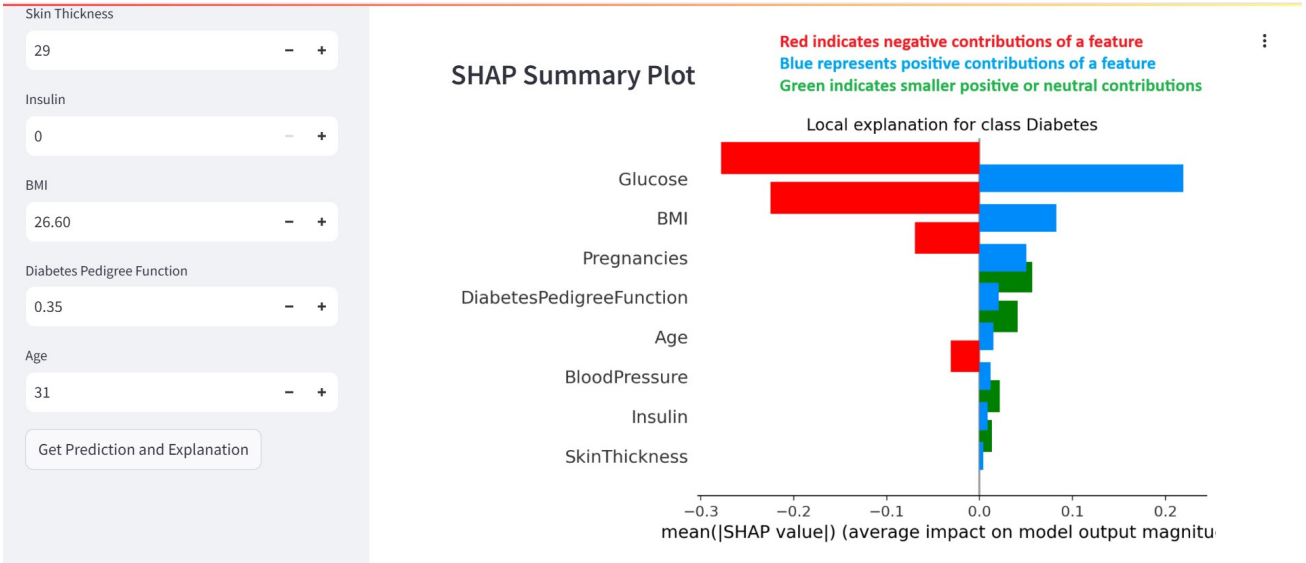

Figure 12. SHAP Feature Contribution Plot for a Specific Prediction.

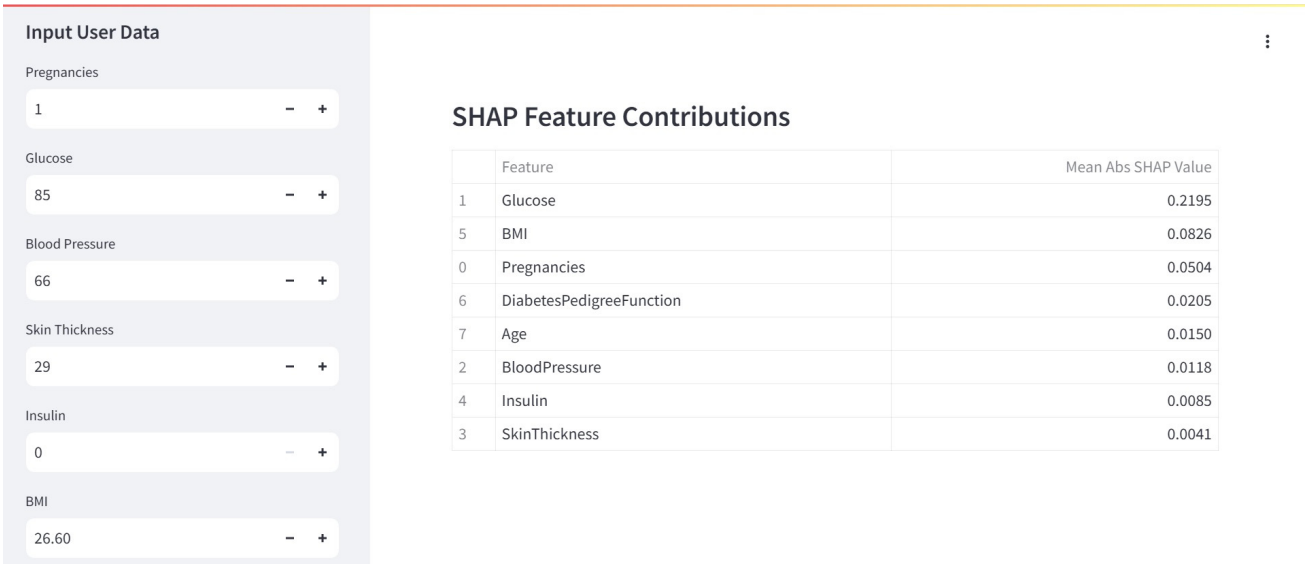

Figure 13. SHAP Feature Contribution for a Specific Prediction.

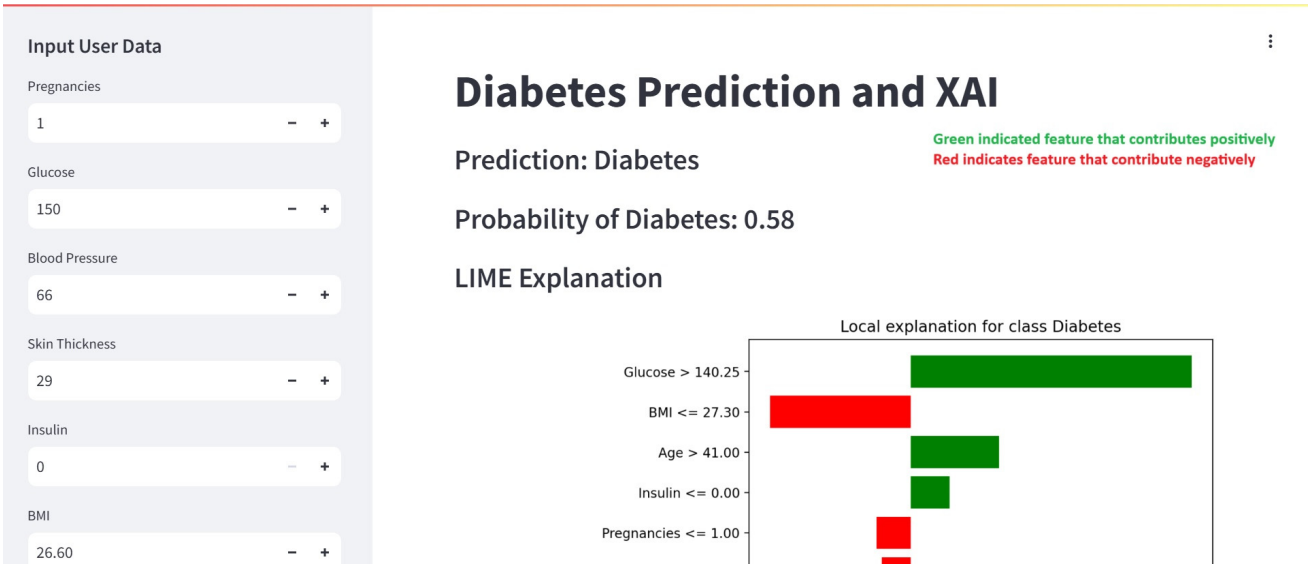

Figure 14. Diabetes Prediction for a Specific Prediction.

4.6. Confusion Matrix for Prediction Performance

To further evaluate the model's prediction performance, a confusion matrix as shown in Figure 15 provides insights into true positive, true negative, false positive, and false negative rates.

The confusion matrix shows how well the model predicts diabetes cases. It correctly identified 78 people with diabetes as diabetic (True Positives) but missed 21 diabetic individuals, labeling them as non-diabetic (False Negatives). Among those without diabetes, the model accurately classified 38 as non-diabetic (True Negatives) but mistakenly labeled 17 as diabetic (False Positives).

The model's sensitivity, or recall, measures how well it detects actual diabetes cases. In this case, it correctly identified 78 out of 99 people with diabetes, resulting in a recall rate of 78.8%. Specificity, which shows the model's ability to correctly identify non-diabetic cases, was 69.1%.

These results indicate a balanced performance in identifying both diabetic and non-diabetic cases. The model has relatively few errors in both over diagnosing (false positives)

and under diagnosing (false negatives). This balanced performance is important in medical settings, where misclassifications can affect patient care. The model's accuracy in correctly identifying both groups makes it a useful tool for supporting diabetes diagnosis.

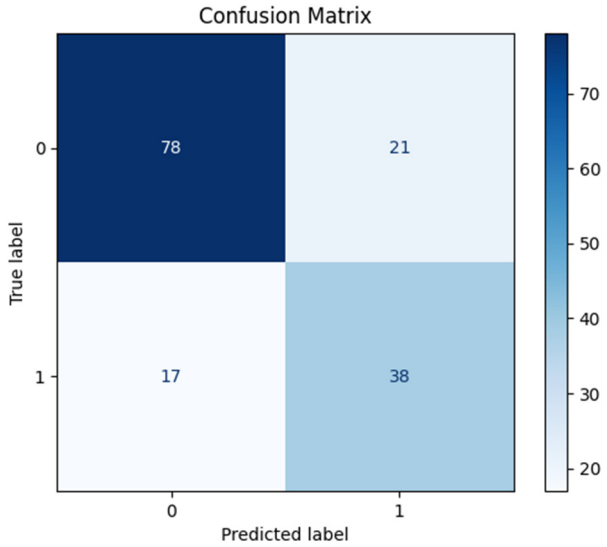

Figure 15. Confusion Matrix for Test Set Predictions.

5. Discussion

This study integrates AutoML with XAI techniques to develop an interpretable and accurate model for diabetes prediction. Leveraging AutoGluon for automated model optimization and SHAP, LIME, CA, IG, and AM for interpretability, we achieved a balanced approach to predictive performance and transparency. The results are discussed below, focusing on model performance, interpretability, practical implications, and limitations, with insights compared to existing studies.

5.1. Model Performance and Classification Metrics

The model achieved an average accuracy of 85.01% across datasets, which aligns well with recent benchmarks. For example, ref. [38] achieved 81% accuracy using a semi-supervised XGBoost model, SMOTE, and ADASYN to address class imbalance on similar datasets [38]. This accuracy demonstrates that ensemble and balancing techniques are effective for diabetes prediction, producing competitive results across datasets.

The achieved accuracy, while slightly lower than [1,2] results with XGBoost and [39] work employing deep neural networks combined with XAI techniques, highlights the importance of balancing performance with interpretability. Unlike these studies, which relied on pre-selected machine learning algorithms, this approach leverages AutoML to automate model selection and optimization, reducing bias and ensuring efficient pipeline development. The deliberate trade-off between achieving exceptional accuracy and emphasizing transparency addresses the critical need for models that are both reliable and interpretable in clinical settings.

5.1.1. Sensitivity and Specificity

The model's sensitivity of 78.8% highlights its ability to detect diabetic cases effectively, reducing the risk of underdiagnosis, a key factor in chronic disease management. Specificity, while moderate at 69.1%, minimizes unnecessary follow-ups due to false positives. As [40] noted, sensitivity is often prioritized in healthcare to ensure that no cases are missed, and balancing sensitivity with specificity is critical for clinical applications [41].

These metrics reflect the model's balanced detection capability but do not represent overall accuracy, which is measured independently. The overall test accuracy was 76.62%, with an average of 85.01% across multiple datasets, validating the model's generalizability and consistency across varied populations. These distinct metrics—sensitivity, specificity, and accuracy—collectively highlight the model's clinical utility.

In comparison, refs. [2,39] reported higher sensitivity and specificity using advanced pre-selected algorithms. However, the automated optimization offered by AutoML in this study reflects a focus on clinical applicability and accessibility, even if minor gains in accuracy are sacrificed. This balance ensures the model remains practical for real-world healthcare use while providing interpretable outputs critical for adoption by non-expert users.

5.1.2. Confusion Matrix Insights

The model's confusion matrix—78 True Positives and 38 True Negatives—demonstrates its classification strength, though it also reveals areas for improvement with 21 False Negatives and 17 False Positives. Ref. [42] emphasize that refining false positive and negative rates is crucial for clinical deployment, as these errors can significantly impact patient outcomes.

CA offers a solution by identifying feature adjustments, such as reducing BMI, that could improve classification accuracy. For instance, a reduction in BMI by 2 points shifted a patient's risk category from high to low. This capability enables clinicians to simulate interventions, addressing misclassifications effectively.

5.2. Model Stability and Cross-Validation

To evaluate model stability, we conducted a 5-fold stratified cross-validation, resulting in an average accuracy of 76.08%. The close alignment between cross-validated and test set accuracy suggests the model generalizes well across different data subsets, essential in medical applications where unpredictable performance on new data could lead to patient misclassification. This stability aligns with findings by [43], who observed that consistent cross-validation performance is vital for healthcare applications. Data augmentation and feature engineering techniques further contributed to enhancing the model's generalizability, as they enabled the model to remain consistent and robust when tested on diverse samples.

Integrated Features can be incorporated here as a tool to explain feature contributions consistently across folds. By examining feature attributions with IG during cross-validation, clinicians can confirm that key features like BMI and Glucose are consistently driving predictions, supporting the stability and reliability of the model across different data subsets.

Low Variance in Cross-Validation

Low variance during cross-validation reduces the risk of overfitting, enhancing the model's applicability across diverse populations. Techniques like feature engineering and data augmentation contributed to this stability, supporting its use in real-world scenarios, as recommended by [44].

5.3. Leaderboard Insights and Ensemble Model Efficacy

The AutoGluon leaderboard identified LightGBM Level 2 and Weighted Ensemble Level 3 as the top-performing models, each achieving 83.88% validation accuracy. LightGBM's efficiency and Weighted Ensemble's robustness highlight the value of diverse ensemble techniques for healthcare applications, as supported by [45,46].

Comparative Model Performance

Models like CatBoost and XGBoost, which also ranked high in our leaderboard, achieved 83.06% validation accuracy. These models are well-suited for handling non-linear relationships and imbalanced classes common in medical datasets. As observed in various studies, diverse model architectures within ensemble frameworks enhance robustness, capturing more complex data patterns [46–48].

5.4. Insights into Prediction Transparency with XAI

The integration of XAI techniques effectively addressed the "black-box" issue, providing insights into global feature importance and patient-specific predictions.

5.4.1. Global and Local Interpretability

SHAP analysis revealed glucose and BMI as the most critical predictors of diabetes risk, quantitatively confirming their significance in the model's decision-making. These findings align with established clinical knowledge, where elevated glucose levels and increased BMI are well-known risk factors for diabetes.

Beyond these primary predictors, SHAP provided insights into secondary features such as the Diabetes Pedigree Function and Age, offering a more nuanced understanding of individual patient risk. For example, the Diabetes Pedigree Function highlights genetic predisposition, while Age reflects the increased likelihood of diabetes in older individuals. These secondary insights enable clinicians to consider broader aspects of patient health when evaluating risk.

By quantifying feature importance, SHAP not only validates the model's alignment with clinical standards but also provides actionable information. For instance, significant contributions from glucose and BMI suggest prioritizing interventions like weight management and glucose monitoring for at-risk individuals. Furthermore, the inclusion of secondary predictors supports tailored care plans that address unique patient needs, improving the effectiveness of diabetes management strategies.

5.4.2. LIME Analysis

LIME facilitated localized, patient-specific explanations by generating surrogate models to analyze individual predictions. For example, in a high-risk diabetes prediction, LIME highlighted high glucose levels and elevated BMI as the primary contributors, consistent with clinical thresholds for hyperglycemia (e.g., glucose > 140 mg/dL).

LIME's case-specific insights enable healthcare professionals to validate individual predictions, enhancing the reliability of the model in clinical contexts. Combined with CA, LIME offers a powerful framework for exploring hypothetical scenarios. For instance, clinicians can use CA to assess how reducing glucose levels might shift the risk classification, providing actionable recommendations for lifestyle modifications.

5.4.3. Counterfactual Analysis

CA enables clinicians to explore how changes in input features could alter the model's predictions. By holding all other features constant, CA allows for simulations of hypothetical scenarios, such as reducing BMI or glucose levels, to evaluate their impact on diabetes risk classification.

For example, a scenario involving a slight reduction in BMI from 35 to 30 demonstrated a shift in the model's prediction from high to low risk. This insight can guide clinicians in recommending specific interventions, such as weight loss programs or dietary changes, to mitigate diabetes risk.

When paired with LIME, CA becomes an invaluable tool for personalized care, enabling clinicians to tailor interventions based on individual patient data and simulate the potential outcomes of these changes.

5.4.4. Quantifying Prediction Contributions with Integrated Gradients

IG provided a comprehensive view of feature attributions by measuring how the model's predictions change as input features transition from a baseline to their actual values. This technique identified BMI and glucose as the most influential predictors, corroborating findings from SHAP and LIME.

The IG analysis also revealed negative contributions from features such as Blood Pressure, suggesting that lower blood pressure may slightly reduce the predicted diabetes risk. By offering a detailed attribution for each feature, IG enhances transparency in high-stakes clinical decisions, helping healthcare providers understand the factors driving the model's outputs.

5.4.5. Personalized Prediction Insights Using Attention Mechanism

AM assigns dynamic weights to features based on their relevance to individual predictions, complementing IG and SHAP analyses. AM consistently prioritized glucose and BMI across predictions, reinforcing their critical role in diabetes risk assessment.

Unlike other techniques, AM adapts its focus based on the specific input data, allowing for a tailored understanding of the model's decision-making process. For instance, attention weights provided by AM highlight the features most relevant to each patient, offering additional context for clinicians when interpreting predictions.

By integrating AM with other XAI techniques, the framework ensures a holistic view of the model's behavior, combining dynamic feature prioritization with detailed attributions. This robust interpretability makes the model a reliable and actionable tool for healthcare professionals.

5.5. Comparison with Similar Studies

Although prior studies, such as [1,2], have successfully demonstrated the application of XAI techniques like SHAP and LIME for diabetes prediction, our research advances the field in several critical ways:

- Existing studies rely on manual model selection, which requires significant expertise and may introduce bias in choosing algorithms. By integrating AutoML, our approach automates model development, ensuring optimal performance across datasets while reducing the technical barriers to implementing machine learning in healthcare.
- While SHAP and LIME provide robust global and local interpretability, CA adds a new dimension by enabling clinicians to simulate how changes in specific features (e.g., glucose or BMI) might alter outcomes. This capability supports personalized, preventative care strategies, which are less explored in previous works.
- IG and the AM provide a holistic view of the model's decision-making, making it easier for clinicians to interpret and trust the model's predictions. AM reveals the features most emphasized by the model, while IG quantifies their individual contributions.
- Our Streamlit application bridges the gap between machine learning advancements and clinical usability, offering a practical, accessible interface for healthcare professionals to interpret predictions and act on them in real-time. This aspect of our work emphasizes the need for clinician-friendly tools, which is often missing in theoretical studies.

These contributions address the limitations of prior works, which primarily focus on demonstrating the applicability of XAI methods without incorporating the automation

and accessibility required for clinical adoption. While [3] effectively demonstrated the use of counterfactuals for diabetes prevention through biomarker adjustments, their study was limited to specific features within a single dataset. In contrast, our integration of CA with AutoML ensures scalability across diverse datasets. Additionally, the incorporation of SHAP and LIME enables clinicians to gain both global and local insights into model predictions, addressing gaps in transparency and usability.

5.6. Methodological Innovations

This study introduces several methodological innovations that extend beyond standard machine learning practices, making significant contributions to the field of healthcare AI:

- Traditional machine learning workflows rely heavily on manual model selection, which can be time-consuming and require significant expertise. By automating the selection and optimization process through AutoML, this study ensures robust performance while democratizing access to advanced machine learning techniques.
- In addition to SHAP and LIME, which provide global and local interpretability, the inclusion of CA offers a novel approach to understanding model predictions. By allowing users to explore how minor adjustments in patient features affect outcomes, CA supports individualized treatment planning, an aspect that is underexplored in previous studies.
- Unlike purely theoretical approaches, this study bridges the gap between machine learning and real-world healthcare applications by providing a user-friendly tool for clinicians. The application integrates predictive insights and interpretability methods, making it accessible and actionable for non-technical users.

These methodological innovations differentiate this research from standard workflows and address critical challenges in healthcare AI, such as transparency, accessibility, and clinical usability.

5.7. Practical Implications for Diabetes Prediction

Integrating AutoML with XAI introduces several practical benefits. AutoML simplifies model selection and optimization, enabling healthcare professionals without extensive machine learning expertise to implement predictive models with high accuracy. This aligns with [10] findings that AutoML democratizes machine learning, allowing non-experts to leverage AI in clinical applications.

5.7.1. Transparency for Clinical Decision-Making

The interpretability provided by SHAP and LIME promotes shared decision-making between clinicians and patients. High SHAP values for BMI or Glucose can prompt targeted lifestyle interventions, such as recommending weight loss programs or dietary changes, while notable LIME contributions for age or family history may guide personalized treatments, such as medication or tailored monitoring. This enables clinicians to prioritize interventions based on the most influential features, ensuring that the treatment plan aligns with the patient's unique risk factors. Ref. [49] report that transparency in AI fosters patient trust and improves clinical decision-making.

5.7.2. Actionable Insights for Personalized Care

SHAP and LIME provide clinicians with personalized, actionable insights that can inform the patient's care plan. For example, if the model indicates high risk due to elevated BMI and Glucose levels, clinicians can utilize the SHAP values to inform patients about the significance of controlling these factors. For instance, the model can suggest that reducing BMI by 2–3 points could significantly reduce the risk of diabetes progression. This

actionable insight allows for personalized care and provides measurable goals for patients to work toward, making the treatment plan concrete and realistic.

5.7.3. Using Counterfactual Analysis for Tailored Interventions

CA further enhances the model's practical applicability by illustrating how small changes in key features could influence predictions. For example, through CA, a clinician could explore how changing a patient's BMI or Glucose level might alter the model's prediction. If a patient is currently at high risk due to a BMI of 35, the clinician can show the patient how a reduction in BMI by just 2–3 points could lower their risk. This feature enables personalized decision-making, allowing clinicians to set tailored, realistic health goals with clear visual feedback on how those changes could impact the patient's future health outcomes.

5.7.4. Shared Decision-Making: Enhancing Patient-Clinician Communication

These insights facilitate shared decision-making between clinicians and patients. By using SHAP, LIME, and CA, clinicians can present actionable insights that help patients understand the potential impacts of their lifestyle changes on their diabetes risk. This collaborative approach strengthens patient engagement and ensures that patients are active participants in shaping their care plans. The model's ability to show real-time adjustments to predictions based on lifestyle changes empowers both clinicians and patients to make data-driven decisions together.

5.8. Limitations and Areas for Improvement

Despite its promising performance, this study has several limitations. The dataset's demographic specificity (Pima Indian population) may limit its generalizability to other populations. Ref. [43] similarly noted that demographic limitations in machine learning models could restrict broader applicability, recommending validation on larger, diverse datasets. While this study uses the Pima Indian Diabetes dataset, which is a widely accepted benchmark for diabetes prediction, the dataset primarily represents a specific ethnic group and lacks diversity in features such as socioeconomic factors, genetic predispositions, and lifestyle variables. To address this limitation, future work should involve validation and retraining of the model on larger, more diverse datasets that include multi-ethnic populations and varied geographical contexts. Additionally, the use of transfer learning or federated learning could enhance generalizability by incorporating knowledge from different datasets while maintaining data privacy.

5.8.1. Limitations in Interpretability Techniques

Although SHAP and LIME improve interpretability, they may not capture complex feature interactions comprehensively. Ref. [50] recommend combining multiple interpretability methods to provide deeper insights into model predictions. Furthermore, while specificity is reasonable, reducing the false positive rate could improve the model's clinical applicability by minimizing unnecessary follow-ups and overdiagnosis.

5.8.2. Model Configuration Constraints

The relatively short AutoGluon training time may have limited the configurations explored. Extending this training time in future experiments could allow for more robust optimization, as other studies have found that prolonged training can enhance model performance. This constraint also limits the exploration of more complex ensemble techniques, which could yield additional performance improvements.

5.8.3. Future Directions

To enhance the generalizability of our findings, future work will focus on validating the proposed model on diverse datasets, including multi-ethnic and geographically varied populations. Collaborations with healthcare institutions to access real-world clinical data can help ensure the model's applicability to broader patient groups. Additionally, incorporating transfer learning or federated learning approaches could enable the model to learn from disparate datasets without compromising data privacy, thereby improving both performance and generalizability. Further, enhancing the model's architecture with advanced AutoML capabilities and integrating deep learning algorithms into the AutoML pipeline may yield higher predictive accuracy while maintaining interpretability. Finally, exploring the impact of additional features, such as socioeconomic and environmental factors, could further refine the model's predictions and broaden its clinical relevance.

5.9. Key Contributions

This study makes several important contributions to the field of healthcare AI, particularly in the context of diabetes prediction. By integrating AutoML with XAI techniques, we provide a robust solution that balances predictive accuracy with interpretability, making it both effective and usable in clinical settings. The following are the key contributions of this research:

- Unlike previous studies that focus solely on either predictive accuracy or interpretability, this research simultaneously addresses both challenges by combining AutoML with XAI techniques like SHAP, LIME, and CA. This integration improves model transparency while maintaining high prediction accuracy, which is essential for clinical adoption. Our approach demonstrates that AutoML can not only automate the model development process but also produce interpretable models, a crucial aspect for healthcare applications.
- A significant contribution of this work is the development of a Streamlit-based application, which allows clinicians to interact with the model, explore predictions, and interpret the importance of different features in real time. This tool bridges the gap between advanced machine learning techniques and real-world healthcare applications, making AI more accessible to healthcare professionals without machine learning expertise.
- Our model demonstrates strong generalization capabilities, achieved through data augmentation, feature engineering, and cross-validation. This robustness ensures that the model performs consistently across diverse patient populations and datasets, addressing a key limitation of many existing diabetes prediction models that struggle with generalization.
- With SHAP and LIME, we provide clinically actionable insights into model predictions. SHAP analysis offers global insights into feature importance, while LIME provides localized, case-by-case explanations. This interpretability is essential for healthcare professionals to make informed decisions based on model predictions and ensures the AI system can be trusted in a clinical context.
- When compared to prior studies, our model achieves competitive performance while addressing critical issues of transparency and interpretability. Table 1 compares the accuracy and other evaluation metrics of our model with those of leading studies in diabetes prediction. For example, while some studies like Tasin et al. [38] achieved higher accuracy using XGBoost and SMOTE techniques, they did not provide the same level of interpretability through XAI methods. In contrast, our study prioritizes both performance and transparency, ensuring that the AI model can be reliably used in clinical settings without sacrificing accuracy.

- By comparing the results of our model with those from other studies, we see that while our accuracy 78.8% with generalization is competitive, our key contribution lies in the combination of predictive accuracy and interpretability. Prior studies, such as [15,38], achieved high accuracy but lacked the level of transparency that our model offers through XAI methods like SHAP and LIME. This dual focus on performance and interpretability is what sets our work apart and advances the field.

6. Conclusions

This study demonstrates the effectiveness of integrating AutoML with XAI techniques—specifically SHAP, LIME, and CA—to enhance diabetes risk prediction. Using the Pima Indian Diabetes dataset, the model achieved an accuracy of 76.62% on the primary test set, which is consistent with recent benchmarks for diabetes prediction. Additionally, the model demonstrated an average accuracy of 85.01% across multiple datasets, highlighting its generalizability across diverse populations. While achieving competitive accuracy, the focus of this study was on ensuring that the model is both interpretable and practical for real-world healthcare applications.

The integration of SHAP and LIME enhances the model's interpretability, aligning its predictions with clinical reasoning. These tools provide actionable insights by quantifying global feature importance and offering localized, patient-specific explanations. Furthermore, CA complements these techniques by allowing clinicians to explore how slight adjustments in key features, such as BMI or Glucose, can influence individual predictions, supporting personalized interventions.

To further facilitate clinical adoption, an interactive Streamlit application was developed. This application enables healthcare professionals to visualize patient-specific risk factors, understand model behavior, and make informed, data-driven decisions. The interface ensures that the model is not only accurate but also accessible and practical for use in real-world clinical settings.

The model's performance metrics, including sensitivity of 78.8% and specificity of 69.1%, demonstrate its balanced ability to detect diabetic cases and minimize unnecessary follow-ups. These attributes reinforce the model's clinical applicability, validated through robust cross-validation, and confirm its reliability across diverse patient populations. Among the ensemble methods tested, LightGBM Level 2 and Weighted Ensemble Level 3 emerged as the most effective, showcasing the advantages of combining multiple model architectures to capture complex patterns in health data.

While the accuracy achieved in this study is competitive, the primary focus remained on developing a transparent and interpretable model suitable for real-world clinical applications. The integration of AutoML and XAI overcomes significant barriers to AI adoption in healthcare by ensuring the model is not only accurate but also interpretable and easy to use.

This research introduces several key innovations. AutoML simplifies model selection and optimization, making advanced machine learning accessible to non-experts while ensuring robust performance across datasets. The combination of SHAP, LIME, and CA provides clinicians with personalized, actionable insights, addressing the common "black-box" challenge of machine learning models. The Streamlit tool bridges the gap between AI research and practical healthcare applications, offering an intuitive interface for clinicians to interact with the model and make informed decisions.

Unlike prior works, which often focused on manual model selection or limited interpretability tools, this study integrates automation and explainability into a single pipeline. These contributions enhance the model's accessibility, transparency, and trustworthiness, making it suitable for clinical environments.

Despite its promising results, this study has limitations. The Pima Indian Diabetes dataset primarily represents a specific ethnic group, which may limit the model's generalizability. Future work should validate the model on larger, multi-ethnic datasets to ensure fairness and applicability across diverse populations. While SHAP and LIME provide robust insights, advanced techniques are needed to capture complex feature interactions, particularly non-linear relationships, which could further improve transparency. Additionally, while the model's sensitivity is strong, reducing the false positive rate could further enhance its clinical relevance by minimizing unnecessary follow-ups and overdiagnosis.

Future research will focus on validating this model on more diverse datasets, incorporating additional features such as socioeconomic and lifestyle factors, and exploring advanced interpretability techniques. Integrating transfer learning or federated learning approaches could also improve performance while maintaining data privacy. Additionally, further optimization of the AutoML pipeline and model architecture may yield even higher predictive accuracy while preserving interpretability.

This study advances the development of diabetes prediction models by integrating AutoML with XAI techniques, ensuring both high accuracy and interpretability. The proposed model provides actionable insights and supports data-driven decision-making in clinical settings. By offering transparency and ease of use, this approach fosters trust in AI applications, making it a valuable tool for healthcare professionals. The work lays the foundation for future advancements in explainable and generalizable AI models in healthcare.

Author Contributions: R.H. and V.D. conceptualized and designed the overall research framework. S.M. developed the experimental approach and provided essential materials. V.D. created and implemented computational tools. R.H., V.D., S.M. and S.H. verified the accuracy and reliability of the results. V.D. performed the statistical analyses, while S.M. conducted the experiments and collected the data. V.D. organized and managed the research data. S.M. drafted the initial manuscript, with R.H., V.D. and S.H. providing critical revisions and improvements. S.M. also prepared the figures and visual representations. R.H. supervised the research process, overseeing each stage, and secured funding for the project. V.D. coordinated project logistics. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are available for download as follows: the Pima Indians Diabetes Database, accessible on Kaggle (https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database, 15 December 2024); the Scikit-learn Built-in Diabetes Dataset, utilized for generalization purposes; and the Second Generalization Dataset (Rural African-American Patients), an additional diabetes dataset from Kaggle (https://www.kaggle.com/datasets/imtkaggleteam/diabetes, 15 December 2024). The latter represents rural African-American patients and provides further insight into the model's ability to generalize across diverse demographic groups.

Acknowledgments: The authors would like to acknowledge the use of ChatGPT 24 May 2023 version (OpenAI, San Francisco, CA, USA), specifically to assist in some content for improved clarity and effectiveness.

Conflicts of Interest: The authors declare no conflicts of interest.

List of Abbreviations

AI	Artificial Intelligence
AutoML	Automated Machine Learning
BMI	Body Mass Index
F1-Score	F1 Score (Harmonic Mean of Precision and Recall)
MCC	Matthews Correlation Coefficient
ML	Machine Learning
ROC-AUC	Receiver Operating Characteristic—Area Under Curve
SHAP	SHapley Additive exPlanations
XAI	Explainable Artificial Intelligence
SVM	Support Vector Machine
GDPR	General Data Protection Regulation
LDL	Low-Density Lipoproteins
HDL	High-Density Lipoproteins
LIME	Local Interpretable Model-Agnostic Explanations

References

1. Jakka, A.; Vakula Rani, J. An Explainable AI Approach for Diabetes Prediction. *Innov. Comput. Sci. Eng.* **2023**, *565*, 15–25. [CrossRef]
2. Zhao, Y.; Chaw, J.K.; Ang, M.C.; Daud, M.M.; Liu, L. A Diabetes Prediction Model with Visualized Explainable Artificial Intelligence (XAI) Technology. *Adv. Vis. Inform.* **2023**, *14322*, 648–661. [CrossRef]
3. Lenatti, M.; Carlevaro, A.; Guergachi, A.; Keshavjee, K.; Mongelli, M.; Paglialonga, A. A novel method to derive personalized minimum viable recommendations for type 2 diabetes prevention based on counterfactual explanations. *PLoS ONE* **2022**, *17*, e0272825. [CrossRef] [PubMed]
4. Waring, J.; Lindvall, C.; Umeton, R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artif. Intell. Med.* **2020**, *104*, 101822. [CrossRef] [PubMed]
5. van der Schaar, M. AutoML and Interpretability: Powering the Machine Learning Revolution in Healthcare. In Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference, Virtual, 19–20 October 2020. [CrossRef]
6. Mustafa, A.; Rahimi Azghadi, M. Automated Machine Learning for Healthcare and Clinical Notes Analysis. *Computers* **2021**, *10*, 24. [CrossRef]
7. Thirunavukarasu, A.J.; Elangovan, K.; Gutierrez, L.; Li, Y.; Tan, I.; Keane, P.A.; Korot, E.; Ting, D.S.W. Democratizing Artificial Intelligence Imaging Analysis With Automated Machine Learning: Tutorial. *J. Med. Internet Res.* **2023**, *25*, e49949. [CrossRef] [PubMed]
8. Kavakiotis, I.; Tsave, O.; Salifoglou, A.; Maglaveras, N.; Vlahavas, I.; Chouvarda, I. Machine Learning and Data Mining Methods in Diabetes Research. *Comput. Struct. Biotechnol. J.* **2017**, *15*, 104–116. [CrossRef] [PubMed]
9. Olisah, C.C.; Smith, L.; Smith, M. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Comput. Methods Programs Biomed.* **2022**, *220*, 106773. [CrossRef] [PubMed]
10. Ahmed Hashim, A.; Hameed Mousa, A. An evaluation framework for diabetes prediction techniques using machine learning. *BIO Web Conf.* **2024**, *97*, 125. [CrossRef]
11. Duckworth, C.; Guy, M.J.; Kumaran, A.; O'Kane, A.A.; Ayobi, A.; Chapman, A.; Marshall, P.; Boniface, M. Explainable Machine Learning for Real-Time Hypoglycemia and Hyperglycemia Prediction and Personalized Control Recommendations. *J. Diabetes Sci. Technol.* **2024**, *18*, 113–123. [CrossRef] [PubMed]
12. Dharmarathne, G.; Jayasinghe, T.N.; Bogahawaththa, M.; Meddage, D.P.P.; Rathnayake, U. A novel machine learning approach for diagnosing diabetes with a self-explainable interface. *Healthc. Anal.* **2024**, *5*, 100301. [CrossRef]
13. Tigga, N.P.; Garg, S. Prediction of Type 2 Diabetes using Machine Learning Classification Methods. *Procedia Comput. Sci.* **2020**, *167*, 706–716. [CrossRef]
14. Kumari, V.A.; Chitra, R. Classification of Diabetes Disease Using Support Vector Machine. *Int. J. Eng. Res. Appl.* **2013**, *3*, 1797–1801.
15. Sisodia, D.; Sisodia, D.S. Prediction of Diabetes using Classification Algorithms. *Procedia Comput. Sci.* **2018**, *132*, 1578–1585. [CrossRef]
16. Behera, M.K.; Chakravarty, S. Diabetic Retinopathy Image Classification Using Support Vector Machine. In Proceedings of the 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), Gunupur, India, 13–14 March 2020; pp. 1–4. [CrossRef]
17. Wu, J.; Diao, Y.; Li, M.; Fang, Y.; Ma, D. A semi-supervised learning based method: Laplacian support vector machine used in diabetes disease diagnosis. *Interdiscip. Sci. Comput. Life Sci.* **2009**, *1*, 151–155. [CrossRef]

18. Alghurair, N.I.; Mezher, M.A. A Survey Study Support Vector Machines and K-MEAN Algorithms for Diabetes Dataset. *Acad. J. Res. Sci. Publ.* **2020**, *2*, 14–25.
19. Chang, V.; Bailey, J.; Xu, Q.A.; Sun, Z. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Comput. Applic.* **2023**, *35*, 16157–16173. [CrossRef] [PubMed]
20. Guan, Y.; Tsai, C.J.; Zhang, S. Research on Diabetes Prediction Model of Pima Indian Females. In Proceedings of the 2023 4th International Symposium on Artificial Intelligence for Medicine Science, Chengdu China, 20–22 October 2023; pp. 294–303. [CrossRef]
21. Sangroya, A.; Anantaram, C.; Rawat, M.; Rastogi, M. Using Formal Concept Analysis to Explain Black Box Deep Learning Classification Models. In Proceedings of the 7th International Workshop "What Can FCA do for Artificial Intelligence"? Co-Located with International Joint Conference on Artificial Intelligence (IJCAI 2019), Macao, China, 10 August 2019.
22. Dagliati, A.; Marini, S.; Sacchi, L.; Cogni, G.; Teliti, M.; Tibollo, V.; De Cata, P.; Chiovato, L.; Bellazzi, R. Machine Learning Methods to Predict Diabetes Complications. *J. Diabetes Sci. Technol.* **2018**, *12*, 295–302. [CrossRef] [PubMed]
23. Erickson, N.; Mueller, J.; Shirkov, A.; Zhang, H.; Larroy, P.; Li, M.; Smola, A. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. *arXiv* **2020**, arXiv:2003.06505.
24. Joseph, V.R. Optimal ratio for data splitting. *Stat. Anal. Data Min.* **2022**, *15*, 531–538. [CrossRef]
25. Verdonck, T.; Baesens, B.; Óskarsdóttir, M.; van den Broucke, S. Special issue on feature engineering editorial. *Mach. Learn.* **2024**, *113*, 3917–3928. [CrossRef]
26. Shorten, C.; Taghi, M. Khoshgoftaar A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
27. Bey, R.; Goussault, R.; Grolleau, F.; Benchoufi, M.; Porcher, R. Fold-stratified cross-validation for unbiased and privacy-preserving federated learning. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 1244–1251. [CrossRef] [PubMed]
28. Shchur, O.; Turkmen, C.; Erickson, N.; Shen, H.; Shirkov, A.; Hu, T.; Wang, Y. AutoGluon-TimeSeries: AutoML for Probabilistic Time Series Forecasting. *arXiv* **2023**, arXiv:2308.05566.
29. Mathotaarachchi, K.V.; Hasan, R.; Mahmood, S. Advanced Machine Learning Techniques for Predictive Modeling of Property Prices. *Information* **2024**, *15*, 295. [CrossRef]
30. Ejiyi, C.J.; Qin, Z.; Amos, J.; Ejiyi, M.B.; Nnani, A.; Ejiyi, T.U.; Agbesi, V.K.; Diokpo, C.; Okpara, C. A robust predictive diagnosis model for diabetes mellitus using Shapley-incorporated machine learning algorithms. *Healthc. Anal.* **2023**, *3*, 100166. [CrossRef]
31. Ghosh, S.K.; Khandoker, A.H. Investigation on explainable machine learning models to predict chronic kidney diseases. *Sci. Rep.* **2024**, *14*, 3687. [CrossRef]
32. Verma, S.; Boonsanong, V.; Hoang, M.; Hines, K.; Dickerson, J.; Shah, C. Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. *ACM CSUR* **2024**, *56*, 1–42. [CrossRef]
33. Wang, Y.; Zhang, T.; Guo, X.; Shen, Z. Gradient based Feature Attribution in Explainable AI: A Technical Review. *arXiv* **2024**, arXiv:2403.10415.
34. Yan, R.; Shang, Z.; Wang, Z.; Xu, W.; Zhao, Z.; Wang, S.; Chen, X. Challenges and Opportunities of XAI in Industrial Intelligent Diagnosis:Priori-empowered. *Ji Xie Gong Cheng Xue Bao* **2024**, *60*, 1.
35. Powers, D.M.W. Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv* **2020**, arXiv:2010.16061. [CrossRef]
36. Chicco, D.; Tötsch, N.; Jurman, G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Min.* **2021**, *14*, 13. [CrossRef] [PubMed]
37. Tasin, I.; Nabil, T.U.; Islam, S.; Khan, R. Diabetes prediction using machine learning and explainable AI techniques. *Healthc. Technol. Lett.* **2023**, *10*, 1–10. [CrossRef]
38. Curia, F. Explainable and transparency machine learning approach to predict diabetes develop. *Health Technol.* **2023**, *13*, 769–780. [CrossRef]
39. Tuppad, A.; Patil, S.D. Machine learning for diabetes clinical decision support: A review. *Adv. Comp. Int.* **2022**, *2*, 22. [CrossRef] [PubMed]
40. Dewage, K.A.K.W.; Hasan, R.; Rehman, B.; Mahmood, S. Enhancing Brain Tumor Detection Through Custom Convolutional Neural Networks and Interpretability-Driven Analysis. *Information* **2024**, *15*, 653. [CrossRef]
41. Ahmed, K.F.; Uz Zaman, M.S.; Peyal, H.I.; Hossain, A.; Rahman Ratul, M.T.; Abdal, M.N.; Islam, M.I. An Interpretable Framework for Predicting Type 2 Diabetes using ML and Explainable AI. In Proceedings of the 2023 26th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 13–15 December 2023; pp. 1–6. [CrossRef]
42. Mahmud, S.M.H.; Hossin, M.A.; Ahmed, M.R.; Noori, S.R.H.; Sarkar, M.N.I. *Machine Learning Based Unified Framework for Diabetes Prediction*; ACM: New York, NY, USA, 2018; pp. 46–50.
43. SumaLata, G.L.; Joshitha, C.; Kollati, M. Prediction of Diabetes Mellitus using Artificial Intelligence Techniques. *Scalable Comput. Pract. Exp.* **2024**, *25*, 3200–3213. [CrossRef]
44. Larabi-Marie-Sainte, S.; Aburahmah, L.; Almohaini, R.; Saba, T. Current Techniques for Diabetes Prediction: Review and Case Study. *Appl. Sci.* **2019**, *9*, 4604. [CrossRef]

45. Kibria, H.B.; Nahiduzzaman, M.; Goni, M.O.F.; Ahsan, M.; Haider, J. An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI. *Sensors* **2022**, *22*, 7268. [CrossRef] [PubMed]
46. Vivek Khanna, V.; Chadaga, K.; Sampathila, N.; Prabhu, S.; Chadaga, P.R.; Bhat, D.; Swathi, K.S. Explainable artificial intelligence-driven gestational diabetes mellitus prediction using clinical and laboratory markers. *Cogent Eng.* **2024**, *11*, 2330266. [CrossRef]
47. Singh, A.; Dhillon, A.; Kumar, N.; Hossain, M.S.; Muhammad, G.; Kumar, M. eDiaPredict: An Ensemble-based Framework for Diabetes Prediction. *ACM TOMM* **2021**, *17*, 1–26. [CrossRef]
48. Tanim, S.A.; Aurnob, A.R.; Shrestha, T.E.; Emon, M.R.I.; Mridha, M.F.; Miah, M.S.U. Explainable deep learning for diabetes diagnosis with DeepNetX2. *Biomed. Signal Process. Control.* **2025**, *99*, 106902. [CrossRef]
49. Hendawi, R.; Li, J.; Roy, S. A Mobile App That Addresses Interpretability Challenges in Machine Learning–Based Diabetes Predictions: Survey-Based User Study. *JMIR Form. Res.* **2023**, *7*, e50328. [CrossRef] [PubMed]
50. Long, C.K.; Puri, V.; Solanki, V.K.; Jeanette Rincon Aponte, G. An Explainable AI-Enabled Framework for the Diabetes Classification. In Proceedings of the 2023 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT), San Salvador, El Salvador, 14–15 December 2023; pp. 1–6. [CrossRef]