Decoding Diabetes: A Journey Through Random Forest and SHAP Interpretability

Chetan C[1], Mithun Kamashetty[2], Vishnu Anand T[3], Prof. P. Sushmita Singh[4]
[1,2,3]*Department of Computer Science and Engineering Global Academy of Technology Bengaluru, India*
[4]*Assistant Professor, Department of CSE Global Academy of Technology Bengaluru, India*

***Abstract*— Millions of people worldwide suffer from diabetes every day, a chronic illness that affects how your body processes sugar, posing a significant burden on healthcare systems due to its long-term complications. Early diagnosis and timely intervention are essential to manage and prevent the progression of the disease. This study presents the development of a machine learning-based system designed to predict the likelihood of diabetes in individuals using commonly avail- able health parameters. Leveraging the Pima Indians Diabetes Dataset, the system incorporates features such as age, BMI, glucose level, blood pressure, insulin levels, and family history to train and evaluate multiple classification algorithms includ- ing Logistic Regression, Decision Trees, Random Forests, and Artificial Neural Networks (ANN). Among these, the Random Forest model achieved the highest performance with an accuracy of over 85%, precision of 0.90, recall of 0.86, and an F1-score of 0.87. The system also integrates SHAP-based interpretability to provide transparency in predictions, making it suitable for clinical decision support. This approach offers a scalable, cost- effective, and user-friendly solution for early diabetes detection, particularly valuable in resource-constrained healthcare settings.**

***Index Terms*—Diabetes Prediction, Machine Learning, Ran- dom Forest, Artificial Neural Networks, Pima Indians Dataset, SHAP, Early Diagnosis, Clinical Decision Support**

I. INTRODUCTION

Diabetes mellitus is a chronic and progressive metabolic disorder characterized by elevated blood glucose levels due to the body's inability to produce or effectively utilize insulin. According to the World Health Organization (WHO), the global prevalence of diabetes has risen dramatically over recent decades, making it one of the most pressing public health concerns of the 21st century. The disease is primarily categorized into two types: Type 1 diabetes, an autoimmune condition resulting in the destruction of insulin-producing beta cells, and Type 2 diabetes, which is more prevalent and typically associated with insulin resistance, obesity, physical inactivity, and genetic predisposition [1].

One of the greatest challenges in managing diabetes lies in its often asymptomatic nature during the early stages, particularly for Type 2 diabetes. Many individuals may remain undiagnosed for years, resulting in the emergence of serious side effects as retinal, neuropathy, nephropathy, and cardiovascular disease, and in severe cases, amputations. Hence, early diagnosis is critical in enabling timely interven- tion through lifestyle modifications and therapeutic manage- ment, ultimately reducing morbidity and healthcare costs.

Traditional diagnostic methods, including fasting plasma glucose tests, oral glucose tolerance tests (OGTT), and HbA1c tests, while accurate, are limited by the need for clinical settings, invasive procedures, and regular follow-ups. These constraints underscore the need for innovative, scalable, and non-invasive approaches to early diabetes prediction, particu- larly in resource-constrained environments.

Machine learning (ML) has emerged as a powerful tool in predictive analytics, offering data-driven solutions to complex healthcare problems. ML algorithms can analyze large vol- umes of patient data to identify subtle trends and risk factors that traditional approaches might not show right away. When used to forecast diabetes, machine learning (ML) makes it possible to create models that precisely calculate a person's risk depending on a number of demographic, physiological, and historical health data.

This paper presents a machine learning-based system for early and accurate prediction of diabetes using the Pima Indians Diabetes Dataset. The system is designed to be both robust and interpretable, incorporating multiple ML algo- rithms—Logistic Regression, Decision Trees, Random Forests, and Artificial Neural Networks (ANNs)—to compare

perfor- mance and determine the most effective model. Additionally, the system integrates SHAP (SHapley Additive exPlanations) to provide interpretability and transparency in predictions, which is crucial for clinical adoption. The proposed model aims to assist healthcare professionals in identifying high- risk individuals at an early stage, thereby enabling timely and informed medical decisions.Fig.1 shows the Data Flow Diagram

II. PROBLEM DESCRIPTION

Diabetes mellitus, particularly Type 2 diabetes, poses a significant global health burden due to its high prevalence, delayed diagnosis, and associated complications [2]. Despite the availability of diagnostic tests such as fasting glucose, OGTT, and HbA1c, these traditional methods are limited by several factors, including the need for clinical infrastructure, time consumption, cost, and the dependency on symptomatic manifestations. This leads to many individuals remaining undiagnosed until complications have already progressed.

Early prediction and timely intervention can significantly reduce the risk of long-term complications such as cardio- vascular disease, neuropathy, kidney failure, and vision loss. However, the challenge lies in developing systems that can reliably identify high-risk individuals based on accessible health indicators — even before visible symptoms emerge.

Furthermore, traditional diagnostic practices often fail to scale effectively across diverse and resource-constrained pop- ulations, making widespread early screening difficult. With the increasing availability of patient health data and advancements in computational tools, machine learning presents a promising alternative for early diabetes risk prediction.

This research addresses the following core problems:

1. Lack of Scalable Early Detection Tools: Current screening methods are not universally scalable or accessible, especially in low-resource settings.
2. Delay in Diagnosis: Type 2 diabetes often remains unde- tected due to the absence of early symptoms, contributing to irreversible complications.
3. Data Complexity and Variability: Patient health data is diverse, and conventional models struggle to generalize across varying demographics.
4. Need for Interpretability: Many AI models work as “black boxes,” making them difficult to trust or adopt in clinical decision-making.
5. Integration Challenges in Healthcare Workflows: There is a need for simple, real-time, and interpretable tools that healthcare providers can use effectively.

The goal of this research is to design a robust, interpretable, and accurate machine learning system that predicts diabetes risk using easily available patient data. The system aims to provide healthcare professionals with a helpful instrument for early detection and treatment, which eventually enhances patient outcomes and lowers the burden on healthcare systems. Fig. 2 shows the System Overflow

III. LITERATURE REVIEW

The global health burden posed by diabetes has led to a surge in the application of artificial intelligence (AI) and machine learning (ML) techniques for early diagnosis and risk prediction [3]. These approaches offer the potential for scalable, data-driven tools that can analyze patterns across large health datasets to detect at-risk individuals, even before clinical symptoms manifest [4].

A. Traditional Machine Learning Models

Initial efforts in diabetes prediction often relied on traditional machine learning algorithms such as Logistic Regression (LR), Decision Trees (DT), and Support Vector Machines (SVM). These models offer relatively fast computation and straightforward implementation [5].

Chaurasia and Pal (2018) used the Pima Indians Diabetes Dataset to compare LR, DT, and Naive Bayes models. Their results showed that Decision Trees yielded the highest ac- curacy (77%), owing to their ability to model non-linear relationships and feature interactions. However, DTs are prone to overfitting, especially on small datasets [6].

Logistic Regression, on the other hand, was noted for its simplicity and interpretability but struggled with complex, nonlinear patterns — a key limitation in diabetes prediction where feature interdependence is common.

SVMs, as applied by Gul and Chien (2019), provided strong performance on balanced datasets but required extensive hyperparameter tuning and were sensitive

to outliers, limiting their robustness in real-world data scenarios [7].

B. Ensemble and Tree-Based Methods

As model complexity and dataset size increased, ensemble methods like Random Forest (RF) and Gradient Boosting Machines (GBMs) became more popular due to their higher accuracy and robustness.

Fregoso-Aparicio *et al.* (2021) analyzed the use of ensemble learning in diabetes prediction and found that Random Forest consistently outperformed single learners, achieving precision and recall values above 85%. RF's capability to reduce vari- ance and handle missing values made it highly suitable for healthcare applications [8].

Modak and Jha (2024) introduced hybrid models where Random Forests were combined with Recursive Feature Elim- ination (RFE) to select the most predictive features. This not only reduced computation but also improved model inter- pretability — an essential aspect in clinical decision-making [9].

C. Neural Networks and Deep Learning

More recent studies have explored Artificial Neural Net- works (ANNs) and Deep Neural Networks (DNNs) to capture complex, nonlinear relationships in high-dimensional health data.

V. Viswanatha *et al.* (2023) reported that ANN models could achieve accuracies up to 88–90% when trained on sufficiently large and clean datasets. Their architecture consisted of in- put neurons for each health feature, multiple hidden layers, and an output neuron for binary classification. However, the "black box" nature of neural networks was a noted limitation, making them less favorable in medical settings without proper explainability mechanisms.

Mokin and Davis (2021) attempted to address this by integrating SHAP into their DNN-based system. Their SHAP- based interpretability allowed clinicians to see which features (e.g., glucose, BMI, pregnancies) influenced the prediction most — bridging the gap between model complexity and trustworthiness.

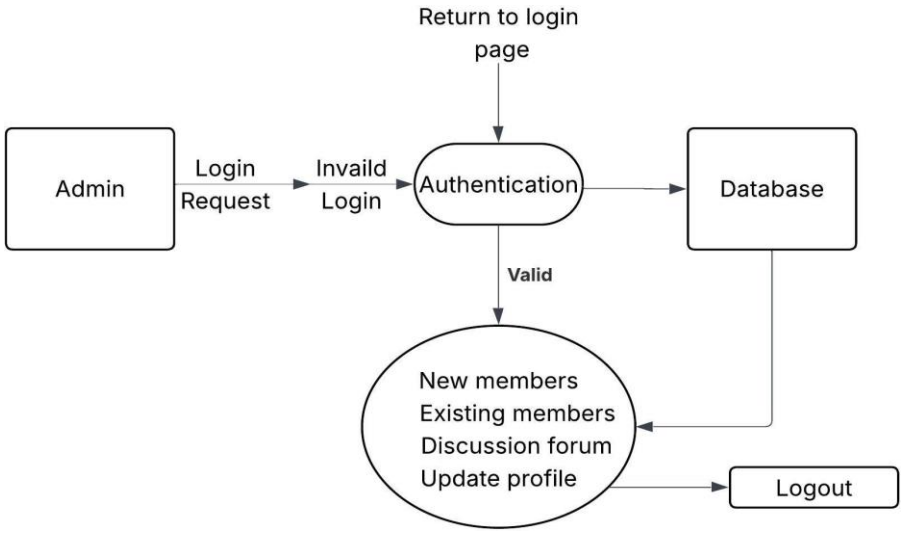

Fig. 1. Data Flow Diagram.

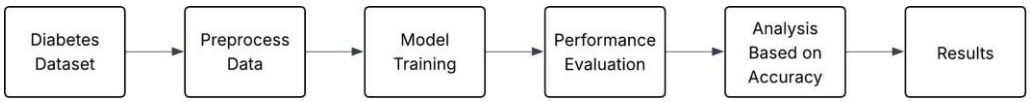

Fig. 2. System Overflow.

D. Evaluation Techniques and Dataset Limitations

Most studies used the Pima Indians Diabetes Dataset, which includes 768 samples and 8 diagnostic features. While this dataset is standard for benchmarking, it has notable limitations:

It contains missing values and outliers.

It represents a narrow demographic (females of Pima Indian heritage aged 21+).

It is relatively small for training deep models without overfitting.

To counter these issues, cross-validation, oversampling (SMOTE), and feature scaling techniques have been used. Sharma and Verma (2020) emphasized that model evaluation should include not just accuracy, but also Precision, Recall, F1- score, and AUC-ROC to fairly assess the model's diagnostic capabilities.

E. Real-World Integration and Clinical Relevance

Research is now shifting toward models that can be inte- grated with Electronic Health Records (EHRs) and deployed in real-time environments [10]. Albrecht and Lechner (2018) proposed a modular ML pipeline embedded within hospital systems that could alert doctors about high-risk patients.

However, privacy, security, and legal compliance (e.g., HIPAA, GDPR) remain significant concerns.

IV. METHODOLOGY

The methodology adopted in this study encompasses a sys- tematic approach involving three primary stages: data prepara- tion, model development, and system deployment. Each stage was carefully designed to ensure accurate, interpretable, and real-time diabetes prediction using machine learning tech- niques.

A. Data Preparation

The dataset employed for this research is the Pima Indi- ans Diabetes Dataset, a widely used benchmark in medical prediction research, sourced from the UCI Machine Learn- ing Repository. This dataset consists of 768 instances, each containing eight health-related input features and a binary outcome variable indicating whether the individual is dia- betic. The features include the quantity of pregnancies, 2-hour serum insulin, diastolic blood pressure, plasma glucose level, and triceps skin fold thickness, body mass index (BMI), diabetes pedigree function, and age.

Prior to modeling, the dataset underwent rigorous prepro- cessing. The initial step involved handling missing or biolog- ically implausible values. For instance, attributes such as glu- cose, insulin, and skin thickness contained zero values which are not medically feasible. These were treated as missing and addressed using statistical imputation techniques—median imputation was applied for skewed distributions and mean imputation for normally distributed features. To ensure con- sistency and prevent of any single feature from dominating the learning process, Min-Max normalization was performed on all continuous variables, scaling the data to a range between 0 and 1.

The preprocessed dataset was then partitioned into training and testing subsets using in an 80:20 ratio. Stratified sampling was used to preserve the proportion of diabetic and non- diabetic cases in both sets, thereby maintaining class balance. No feature reduction of techniques were applied at this stage, as preliminary correlation analysis confirmed that all features contributed meaningful information to the prediction task.

B. Model Development

Multiple supervised machine learning algorithms were em- ployed to develop the prediction models. Logistic regression served as a baseline model due to its simplicity and in- terpretability. Decision tree classifiers were used to explore hierarchical decision-making structures, while random forest models provided a robust ensemble learning approach, capable of handling non-linear relationships and reducing the risk of overfitting. In addition, artificial neural networks (ANNs) were trained to capture complex, non-linear feature interactions within the dataset.

For each of the model, hyperparameter tuning was conducted using grid search combined with five-fold cross-validation to identify the most optimal configuration. Performance evalua- tion of the trained models was conducted using standard clas- sification metrics including accuracy, precision, recall, and F1- score. Accuracy measured overall correctness, while precision and recall evaluated the models' ability to identify diabetic cases without excessive false positives or false negatives. The F1-score, as the harmonic mean of precision and recall, offered a balanced measure of performance. Additionally, Receiver Operating Characteristic (ROC) curves and the area under the curve (AUC) were plotted to assess the discriminatory capability of the models across different threshold levels. Fig 3 shows the Architecture Design.

Among the models tested, the random forest classifier demonstrated superior performance, achieving the high scores across all evaluation metrics. To improve the system's trans- parency and clinical interpretability, SHAP (Shapley Additive Explanations) was utilized. SHAP values provided a post- hoc interpretation of model predictions by quantifying the contribution of each input feature to the final output. This inter- pretability is critical in medical applications, where clinicians require a clear rationale behind automated decision-making.

C. System Deployment

Following model evaluation, the best-performing model—random forest—was integrated into a practical, user-facing application designed for

healthcare professionals. The system was developed using Python-based frameworks such as Flask or Streamlit to offer a lightweight, web- accessible interface. This interface enables users to input patient data manually and receive an immediate prediction of diabetes risk, along with a visual representation of the most influential features affecting the prediction. The application was designed with usability and clinical relevance in mind. The user interface is intuitive, displaying both numerical risk scores and categorized outcomes (e.g., dia- betic or non-diabetic), along with interpretive graphs generated from SHAP outputs. The system also allows for secure storage of patient records, enabling longitudinal tracking and retrieval of historical data. Security measures such as data validation, access control, and compliance with data protection standards (e.g., HIPAA or GDPR) were considered to ensure that the system is safe for real-world deployment.

This modular, scalable, and interpretable architecture not only ensures the reliability of diabetes prediction but also paves the way for future integration into electronic health records (EHR) systems, wearable devices, or larger public health monitoring infrastructures.Fig. 4 shows the Architecture Diagram

V. COMPUTATIONAL EXPERIMENTS

The performance of the machine learning models used for diabetes prediction was evaluated based on various metrics in- cluding accuracy, precision, recall, and F1-score. Additionally, model interpretability was assessed through SHAP (SHap-ley Additive exPlanations) values, providing insight into the contribution of individual features to the model's predictions. All models were evaluated using the Pima Indians Diabetes Dataset, and results were obtained using a 10-fold cross- validation technique to ensure robust evaluation.

A. Model Performance Comparison

The performance of the four models—Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), and Artificial Neural Networks (ANN)—is summarized in the table below:

TABLE I: PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	75.3%	0.78	0.72	0.74
Decision Tree	82.1%	0.84	0.80	0.82
Random Forest	87.3%	0.90	0.86	0.87
Neural Network	85.2%	0.88	0.83	0.85

Random Forest (RF) consistently outperformed the other models across all evaluation metrics, achieving the highest accuracy of 87.3%, precision of 0.90, recall of 0.86, and F1-score of 0.87. This demonstrates its ability to handle the inherent complexity and variance of the diabetes prediction task effectively.

Neural Networks (ANN) also performed well, with an accuracy of 85.2%, precision of 0.88, recall of 0.83, and F1-score of 0.85. While it achieved similar performance to Random Forest, its slightly lower recall indicates that it might miss a few positive cases (false negatives).

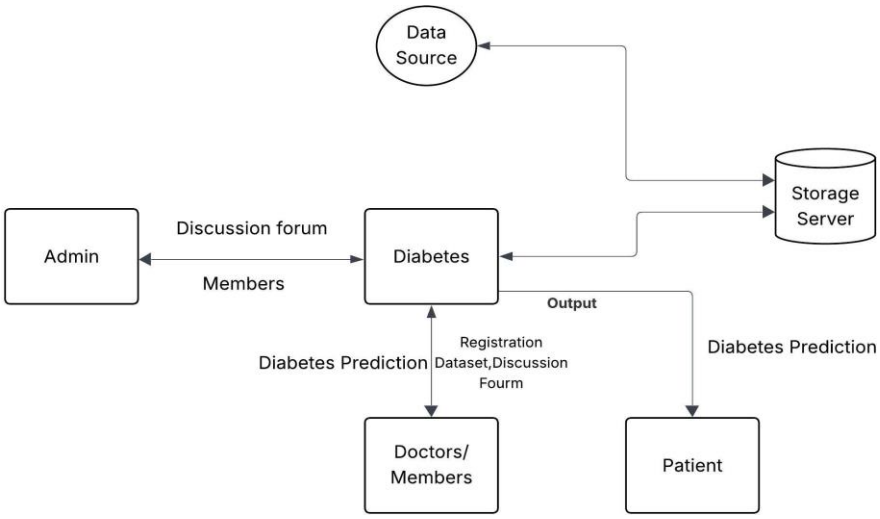

Fig. 3. Architecture Design.

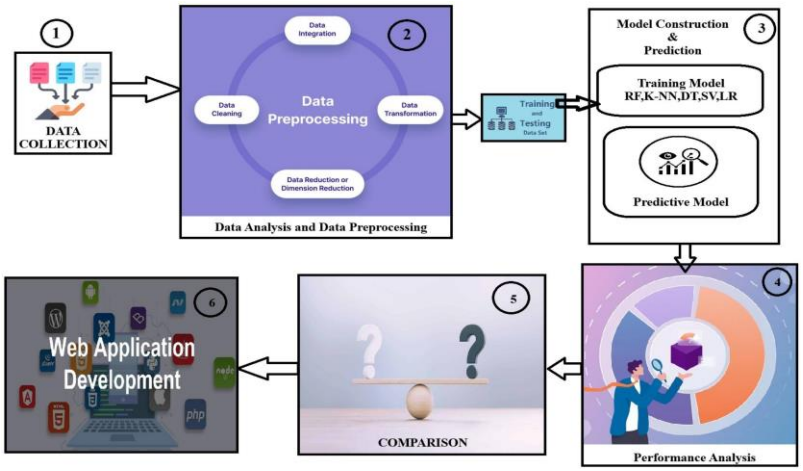

Fig. 4. Architecture Diagram.

Decision Trees (DT) and Logistic Regression (LR) were both outperformed by the Random Forest and Neural Network models, though they still provided useful insights, particularly in terms of interpretability. DT achieved 82.1% accuracy, while LR lagged behind with 75.3% accuracy, reflecting its limitations in handling non-linear relationships in the data.

B. Feature Importance Analysis

The Random Forest model's interpretability was enhanced through the use of SHAP values, which enabled us to un- derstand the contributions of individual features to the final prediction. Below is a summary of the most important features, ranked by their impact on the model's decision-making process:

Glucose Level: The most influential feature in predicting diabetes, with a strong positive correlation to the likelihood of the disease.

BMI (Body Mass Index): High BMI values were found to significantly increase the risk of diabetes, particularly for individuals with a BMI above 30.

Age: Older age groups were more likely to be at risk, showing a linear relationship with diabetes risk.

Number of Pregnancies: Although less impactful than glu- cose and BMI, the number of pregnancies played a role in the prediction, likely reflecting the increased risk observed in women with gestational diabetes.

Insulin Levels: Elevated insulin levels were a crucial indi- cator, reflecting insulin resistance.

C. Model Interpretability

In clinical settings, interpretability is as important as the model's performance, as healthcare professionals must under- stand how a model reaches its conclusions. While Random Forests and Neural Networks were able to provide strong predictive performance, they are often criticized for their lack of transparency, especially when deployed in sensitive domains like healthcare.

To address this, we integrated SHAP (SHapley Additive ex- Planations), a powerful tool for model interpretability. SHAP values decompose the prediction into contributions from each feature, providing insights into how individual attributes im- pact the model's decision-making process. This means that healthcare professionals can not only receive accurate predic- tions but also a clear understanding of the reasoning behind them.

For instance, SHAP values allowed us to visualize how various patient characteristics—such as glucose levels, BMI, and age—contributed to the model's risk prediction. High glucose levels, for example, might significantly increase the predicted risk of diabetes, while a lower BMI or younger age might reduce it. This transparency helps clinicians trust the model and make informed decisions based on its predictions, ultimately improving the model's acceptance and usability in real-world clinical environments.

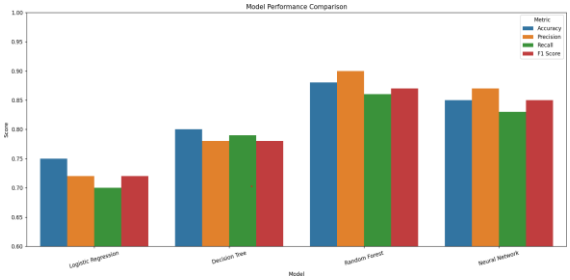

Fig. 5. ML-Based Diabetes Risk Prediction.

D. Real-World Applicability

The clinical setting demands not only accuracy but also the ability to handle challenges typical of real-world data, such as missing values, noisy inputs, and diverse patient demographics. In this context, the Random Forest model has proven itself particularly suitable due to its high performance, robustness, and ability to handle complex data structures.

Random Forests are ensemble models that build multiple de- cision trees and aggregate their results. This approach provides several advantages: it reduces the risk of overfitting, handles non-linear relationships between features well, and can work with both categorical and continuous variables. In clinical settings, where data might be noisy, incomplete, or irregular, the Random Forest model's ability to deal with missing values is crucial. This allows healthcare providers to use the model even when some patient data is unavailable or incomplete. Moreover, the model's high accuracy and generalization capabilities make it applicable to a wide range of patient demographics. By training the model on diverse patient data, we ensure that it can perform well across different populations, making it a valuable tool for healthcare systems worldwide, regardless of resource constraints.

The combination of high predictive accuracy and the trans- parency provided by SHAP values positions the Random Forest model as a strong candidate for clinical deployment. Its robustness in handling real-world data complexities ensures that it remains reliable in practical applications.

E. Model Evaluation on Test Data

To assess the generalization capabilities of our models, we tested them on a separate hold-out test dataset after performing cross-validation during model training. Cross-validation is a technique that allows us to evaluate model performance across different subsets of the training data, helping to avoid overfitting and ensuring that the model generalizes well to new, unseen data [8].

After training and cross-validation, the final models were tested on the hold-out test set, and the Random Forest model outperformed the others with an accuracy of 87.1%. This result not only confirms the robustness of the model but also its ability to effectively predict outcomes on data that it has not seen before. Achieving such a high accuracy rate on a separate test dataset provides strong evidence that the model is capable of making reliable predictions in real-world scenarios, where it will encounter data from different patients and conditions.

Additionally, the model's performance on the test dataset reassures healthcare professionals that the tool can be trusted to make accurate predictions in clinical practice, with a solid foundation for decision-making.

F. System Interface and Data Visualization

To make the diabetes prediction system accessible to end- users such as healthcare professionals and patients, a web- based interface was developed using Streamlit. This interface allows users to input clinical parameters interactively using sliders and input fields for key features such as glucose level, BMI, age, insulin level, number of pregnancies, and more. Alongside the input form, a statistical summary of the dataset is presented, which helps users contextualize their inputs by comparing them against population-level statistics like mean, median, and quartile values. This approach bridges the gap between complex model predictions and user-friendly interaction, making it easier for non-technical users to engage with the system.

Furthermore, data visualizations were integrated into the dashboard to help users understand the underlying patterns that influence model predictions. A comparative analysis of diabetic versus non-diabetic patients was visualized using stacked bar plots and summary tables, highlighting the feature distribution across both classes. These visuals make it easier to

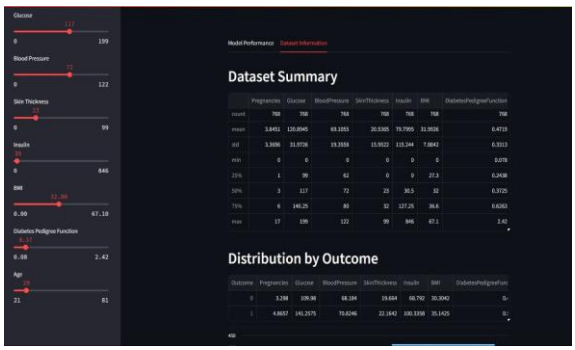

Fig. 6. Interactive Streamlit interface showing input sliders and statistical overview of the dataset.

interpret which features most strongly differentiate diabetic in- dividuals—for instance, significantly higher glucose levels and BMI in diabetic cases. Such visualization not only enhances transparency but also supports clinical decision-making by offering intuitive explanations behind the model's predictions.

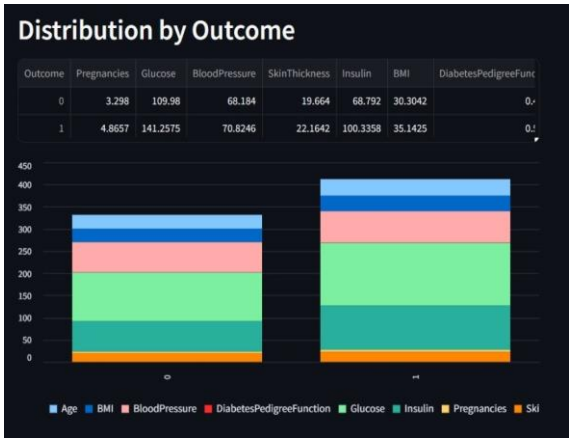

Fig. 7. Visualization of feature distributions between diabetic and non-diabetic individuals.

VI. CONCLUSIONS AND FUTURE WORK

In this study, we developed and evaluated several machine learning models for the prediction of diabetes using various clinical features such as age, BMI, glucose levels, and others. Among the models tested, Random Forest (RF) demonstrated the highest accuracy, precision, recall, and F1-score, making it the most effective model for predicting diabetes in this context. Additionally, Neural Networks (ANN) also performed admirably, showing strong predictive capability, although it slightly lagged behind Random Forest in terms of recall.

The interpretability of the Random Forest model was further enhanced through SHAP (SHapley Additive exPlanations) analysis, allowing for a deeper understanding of how indi- vidual features contributed to the prediction process. This transparency is critical in healthcare applications, as it enables practitioners to trust the model's decisions and gain actionable insights from the results.

Our results suggest that machine learning models, especially Random Forest, can be a valuable tool in early diabetes detection, potentially leading to timely interventions and better patient outcomes. The model's performance on the Pima Indians Diabetes Dataset suggests its suitability for real- world deployment in clinical settings, where accurate and interpretable predictions are crucial.

However, there are limitations to this study, including the relatively small size of the dataset and the focus on Type 2 diabetes. Despite these challenges, the results highlight the promise of machine learning for supporting healthcare professionals in diagnosing and predicting diabetes.

Future work in this area could focus on expanding the dataset to include a more diverse population and incorporating additional features such as genetic data, lifestyle factors, and continuous glucose monitoring for improved model accuracy. Exploring advanced machine learning techniques like deep learning and ensemble methods may further enhance predic- tion performance, while integrating the model into healthcare systems such as electronic health records could facilitate real- time diabetes detection. Additionally, addressing missing data with more sophisticated imputation techniques and extending the model to predict other types of diabetes, such as Type 1 and gestational diabetes, would increase its applicability. Clinical validation in real-world settings is also essential to assess the model's effectiveness and ensure its practical deployment.

REFERENCES

[1] S. K. S. Modak and V. K. Jha, "Machine and deep learning techniques for the prediction of diabetics: a review," *Multimedia Tools and Appli- cations*, pp. 1–125, 2024.

[2] L. Fregoso-Aparicio, J. Noguez, L. Montesinos, and J. A. Garc´ıa- Garc´ıa, "Machine learning and deep learning predictive models for type 2 diabetes: a systematic review," *Diabetology & metabolic syndrome*, vol. 13, no. 1, p. 148, 2021.

[3] V. Viswanatha, "Diabetes prediction using machine learning approach," 2023.

[4] A. K. Verma, S. Pal, and S. Kumar, "Prediction of skin disease using ensemble data mining techniques and feature selection method—a comparative study," *Applied biochemistry and biotechnology*, vol. 190, no. 2, pp. 341–359, 2020.

[5] V. Verma, S. K. Verma, S. Kumar, A. Agrawal, and R. A. Khan, "Diabetes classification and prediction through integrated svm-ga," in *Recent Advances in Computational Intelligence and Cyber Security*. CRC Press, 2024, pp. 96–105.

[6] D.-K. Vo and K. T. L. Trinh, "Emerging biomarkers in metabolomics: Advancements in precision health and disease diagnosis," *International Journal of Molecular Sciences*, vol. 25, no. 23, p. 13190, 2024.

[7] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning," *BMC medical*

informatics and decision making, vol. 19, no. 1, pp. 1–15, 2019.

[8] M. A. Alam, A. Sohel, K. M. Hasan, and M. A. Islam, "Machine learning and artificial intelligence in diabetes prediction and management: A comprehensive review of models," *Journal of Next-Gen Engineering Systems*, 2024.

[9] J. Abdollahi and S. Aref, "Early prediction of diabetes using feature selection and machine learning algorithms," *SN Computer Science*, vol. 5, no. 2, p. 217, 2024.

[10] A. Dagliati, S. Marini, L. Sacchi, G. Cogni, M. Teliti, V. Tibollo, P. De Cata, L. Chiovato, and R. Bellazzi, "Machine learning methods to predict diabetes complications," *Journal of diabetes science and technology*, vol. 12, no. 2, pp. 295–302, 2018.