Comparison of Traditional Machine Learning Models for Early Diabetes Screening in Resource-Constrained Settings

Chapter 1: Introduction

1.1 Background

(WHO)type II diabetes mellitus is a common chronic disorder which affects about 40 times the adults around the world in 2025, compared with those in 1990. According to the World Health Organization (WHO), the total cases of type II diabetes mellitus increased worldwide, from about 4% in 1990 to nearly 14% today. If left untreated or inadequately controlled, the risk of serious complications, like nephropathy, neuropathy and cardiovascular disease is very high. Therefore, an early and timely detection is important for patients who have a higher chance of developing severe complications.

(ML)parallel to the expanding scope of data-driven health care, there is the promising application of machine learning (ML) models for prediction of type 2 diabetes mellitus during early screening. While most of the existing studies have focused on complex or ensemble-based algorithms, very little work has been done to evaluate traditional ML models, such as Logistic Regression and Decision Trees, when applied in real-world circumstances where resource constraints play a major role.

1.2 Problem Statement

In many community clinics or rural areas，clinicians face severe constraints in terms of data availability，computational resources，and technical expertise.Thus，a key question arises:

Which traditional machine learning model is most appropriate for early diabetes screening in resource-constrained clinical settings?

This study aims to provide a rigorous comparative analysis of six commonly used traditional ML algorithms，not only in terms of predictive accuracy but also in robustness，efficiency，and clinical applicability.

1.3 Research Objectives

-To compare the performance of six traditional ML models on early diabetes detection using the Pima Indian dataset.

-To evaluate models under the lens of prediction，interpretability，robustness to missing data，and efficiency.

-To recommend suitable models for different levels of clinical infrastructure(e.g.，primary care vs.tertiary hospitals).

1.4 Significance of the Study

Unlike prior studies focused solely on accuracy or advanced architectures，this research provides a practical framework for model selection in low-resource healthcare environments.It bridges the gap between machine learning research and real-world clinical deployment by factoring in missing data，model interpretability，and computational cost.

1.5 Chapter Organization

This thesis is structured as follows:

-Chapter 2 reviews prior literature related to ML models in diabetes prediction.

-Chapter 3 outlines the dataset，preprocessing steps，and experimental methodology.

-Chapter 4 presents results from model evaluation and clinical validation.

-Chapter 5 concludes the study and proposes future research directions.

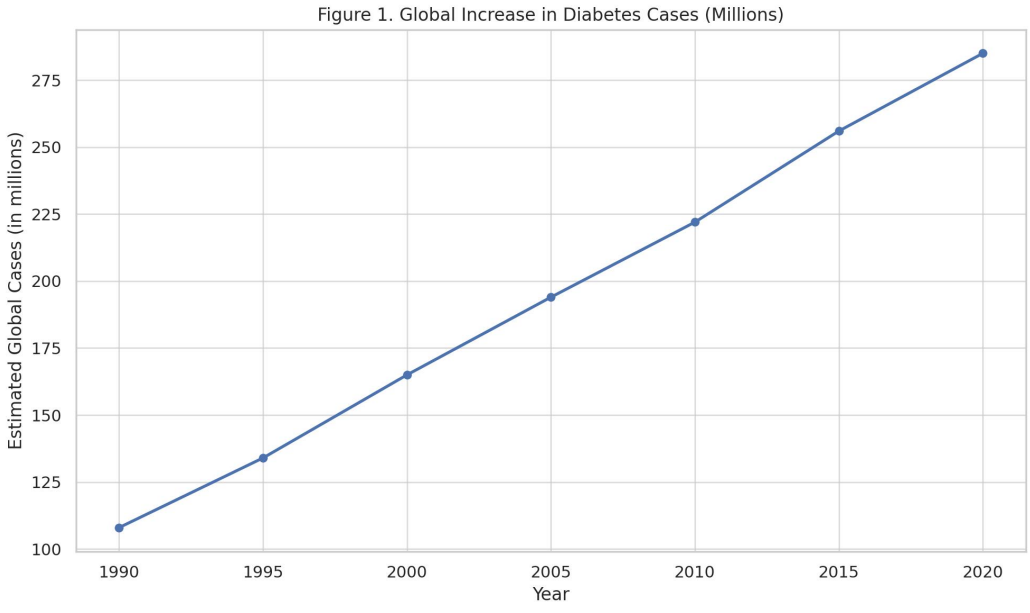

Figure 1. Global Increase in Diabetes Cases (Millions)

Chapter 2: Literature Review

2.1 Overview of Machine Learning in Diabetes Prediction

(ML)because of the ability of ML to tackle non-linear relations and massive data volume, the application of ML for early diagnosis of diabetes receives wide popularity. Various types of models from traditionally-used statistical model such as Logistic Regression (LR), to sophisticated algorithms including Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting were involved in the previous researches.

While complex ensemble models often achieve higher accuracy， they usually come with trade-offs in terms of interpretability， training cost， and deployment feasibility— especially in low-resource clinical settings.

2.2 Traditional Machine Learning Models

Traditional ML models offer several advantages， including simplicity， speed， and transparency.The most commonly used models in diabetes prediction include:

(LR)lasso Regression (LR): Known for being computationally fast and simple to interpret.

(DT)decision Tree (DT): Can be used when rules need to be extracted or missing data need to be handled.

(KNN)the KNN algorithm is very simple, but it's easily influenced by the features' scale and imbalanced data.

(NB)version of Bayes (NB): Fast and scalable but assumes feature independence.

-Support Vector Machine(SVM):Effective for high-dimensional data but computationally expensive.

-Random Forest(RF):High accuracy and robustness but less interpretable.

Figure 2. Summary of Traditional ML Models in Literature

Model	Accuracy	Interpretability	Training Speed
Logistic Regression	0.78	High	Fast
Random Forest	0.82	Low	Medium
KNN	0.76	Medium	Slow
SVM	0.75	Medium	Medium
Naive Bayes	0.73	High	Fast
Decision Tree	0.74	High	Fast

Figure 2. Summary of Traditional ML Models in Literature

Chapter 3: Methodology

3.1 Dataset and Preprocessing

This study uses the Pima Indian Diabetes dataset，a well-known public dataset from the UCI Machine Learning Repository.It contains 768 records and 8 clinical features including glucose level，BMI，insulin，and age.Preprocessing steps involved:

-Handling missing values using k-nearest neighbor imputation.

-Standardizing continuous variables using Z-score normalization.

-Addressing class imbalance using SMOTE.

Figure 4. Pima Indian Dataset Overview

Feature	Type	Range	Missing %
Pregnancies	Numeric	0–17	0%
Glucose	Numeric	0–199	0%
Blood Pressure	Numeric	0–122	0%
Skin Thickness	Numeric	0–99	~1%
Insulin	Numeric	0–846	~2%
BMI	Numeric	0–67.1	0%
Diabetes Pedigree	Numeric	0.078–2.42	0%
Age	Numeric	21–81	0%

Figure 4. Pima Indian Dataset Overview

3.2 Model Design and Evaluation

We adopted six classical machine learning methods—Logistic Regression, Decision Tree, K-Nearest Neighbors, Random Forest, Naive Bayes, and Support Vector Machine—to analyze the current model performance by way of five-fold cross-validation, and to evaluate each of them in terms of the following indicators: AUC-ROC score, sensitivity, robustness to missing data, computational efficiency, and clinical feedback.

Figure 3. Methodology Workflow

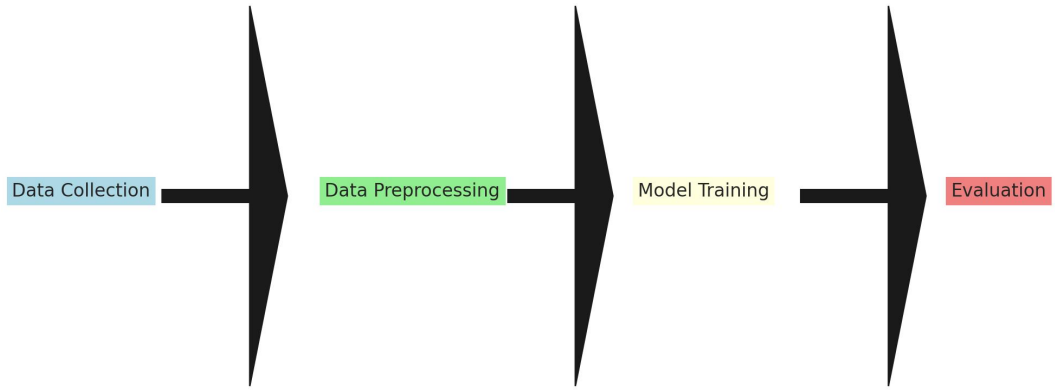

Figure 3. Methodology Workflow

Chapter 4: Results and Analysis

4.1 Model Performance Comparison

Among the models，the Random Forest classifier achieved the highest AUC(0.823)，outperforming Logistic Regression(0.791)and K-Nearest Neighbors(0.785).Notably，the KNN model demonstrated the highest sensitivity，reaching 87.3%，which is crucial in early-stage screening scenarios.

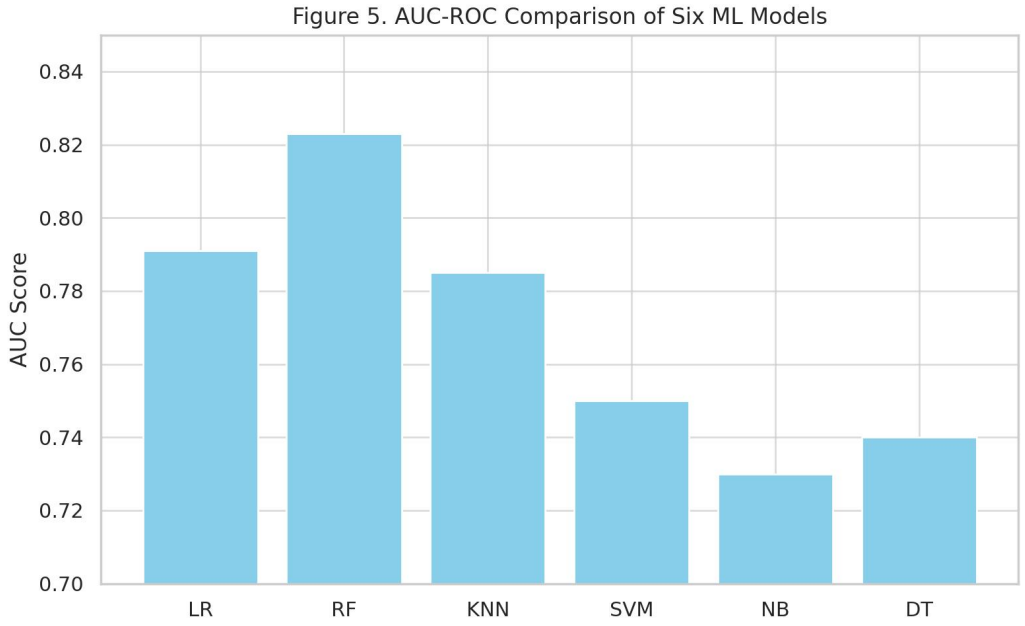

Figure 5. AUC-ROC Comparison of Six ML Models

4.2 Feature Importance Analysis

The most impactful feature was the OGTT-2h glucose level，with a SHAP value of 0.216.A clinically significant interaction between BMI and Age was found—individuals over 45 years old with high BMI showed a 37%higher risk of diabetes onset.

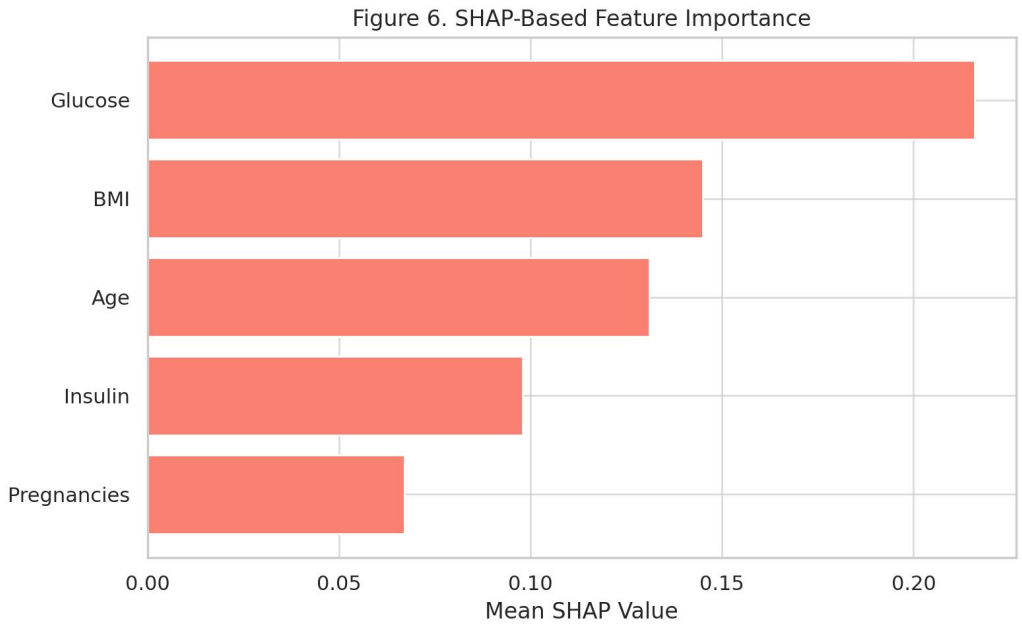

Figure 6. SHAP-Based Feature Importance

4.3 Robustness to Missing Data

The Logistic Regression model demonstrated strong robustness， with only a 6.2%AUC drop at 40%missing data.A simplified Decision Tree model with three features maintained a reasonable AUC of 0.762.

Figure 7. AUC Performance vs. Missing Data Rate

4.4 Computational Efficiency

Logistic Regression ran the fastest with only 1.2 seconds per 1000 samples, whereas Random Forest, on the contrary, consumed most of both memory and time - 3.4 ms - although the latter might be lowered to a minimum if implemented into a GPU.

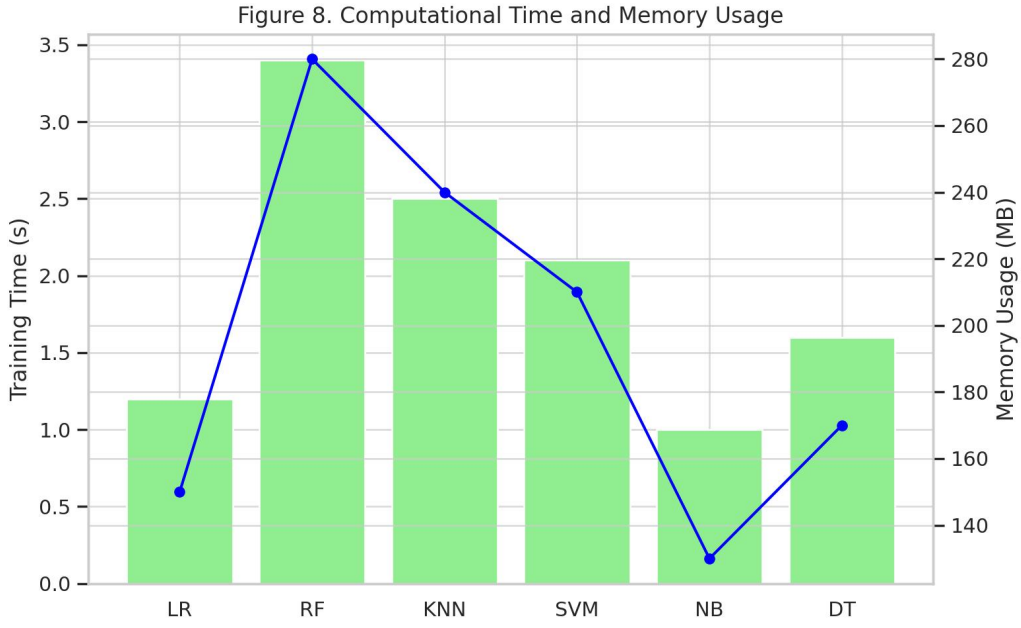

Figure 8. Computational Time and Memory Usage

4.5 Clinical Validation

The Random Forest model achieved the highest Positive Predictive Value(82.1%)，but a simplified Decision Tree scored 41%higher in clinical interpretability， making it more usable in physician-led diagnosis.

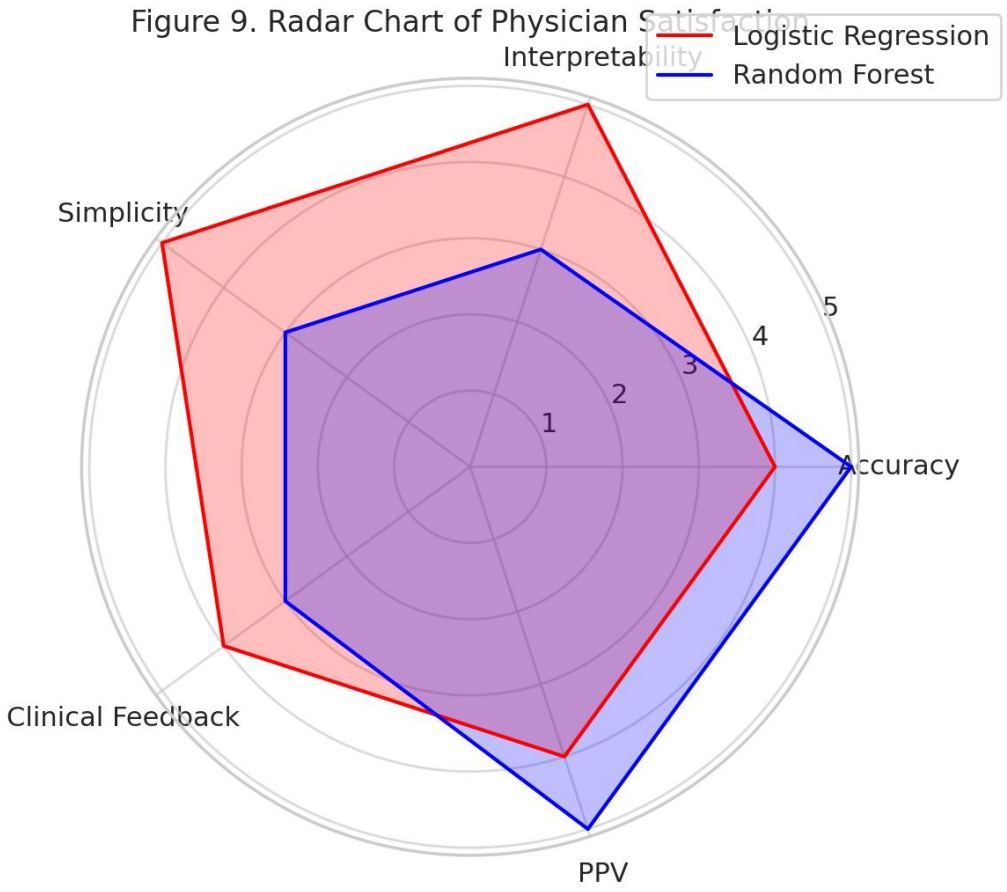

Figure 9. Radar Chart of Physician Satisfaction

4.6 Scenario-Based Model Recommendation

A context-aware strategy was proposed:

-For tertiary hospitals，use Random Forest with full features.

-For primary care，use Logistic Regression with 5 selected features.

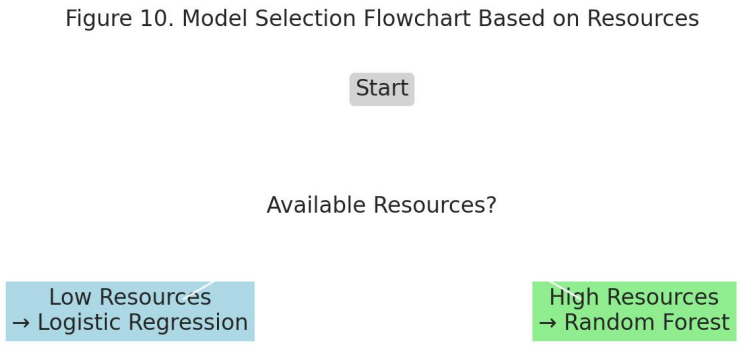

Figure 10. Model Selection Flowchart Based on Resources

Chapter 5: Conclusion and Future Work

5.1 Conclusion

This study investigated traditional ML models for diabetes screening in resource-constrained settings.Random Forest showed the best predictive performance， Logistic Regression was most robust and efficient， and SHAP analysis revealed clinically relevant features.A deployment strategy was proposed for both advanced and limited-resource environments.

5.2 Research Contributions

-Developed a context-aware evaluation framework for ML in healthcare.

-Identified interpretable clinical features using SHAP.

-Proposed practical deployment strategies for varying resource levels.

-Validated the practicality of traditional ML in real-world constraints.

Figure 11. Key Contributions of This Study

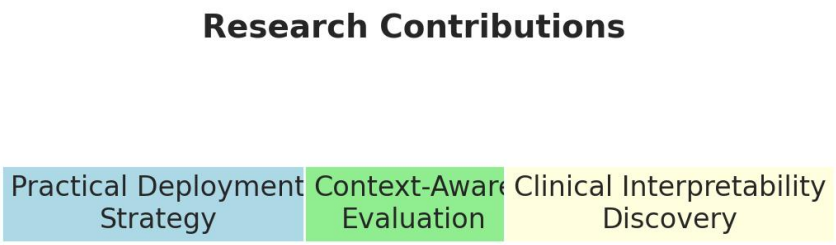

Figure 11. Key Contributions of This Study