ML Models for Diabetes Screening

Reporter: Zhang Yibo

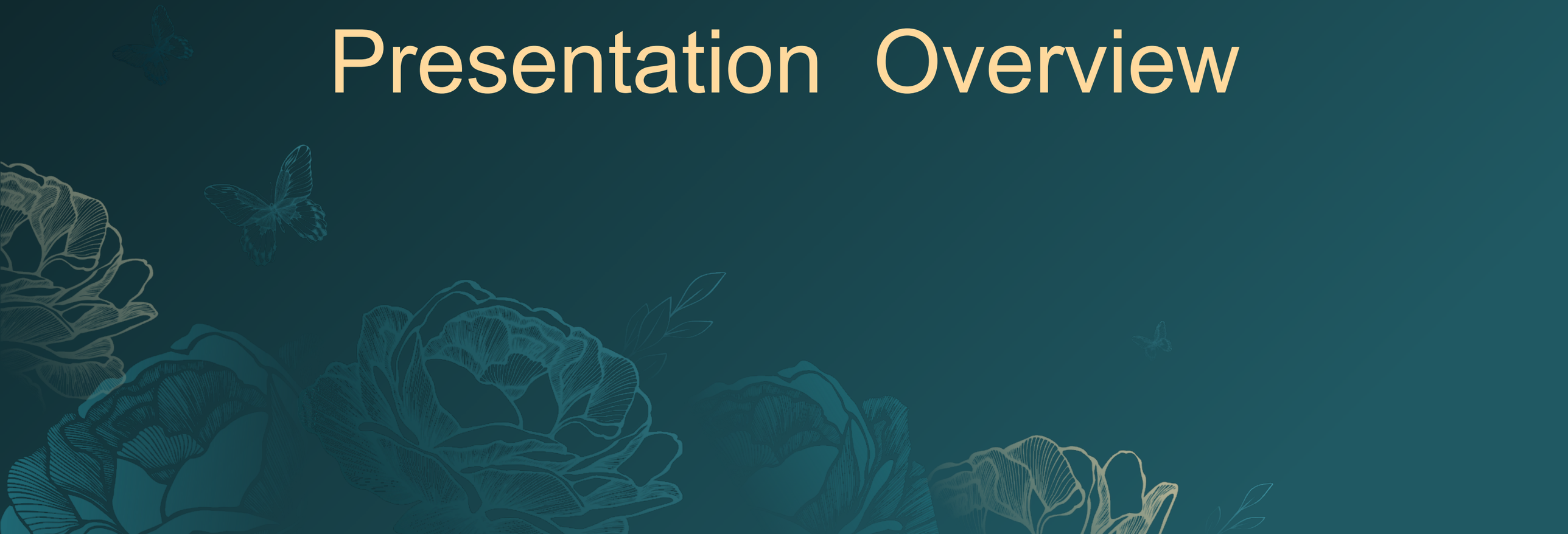
MCS241044

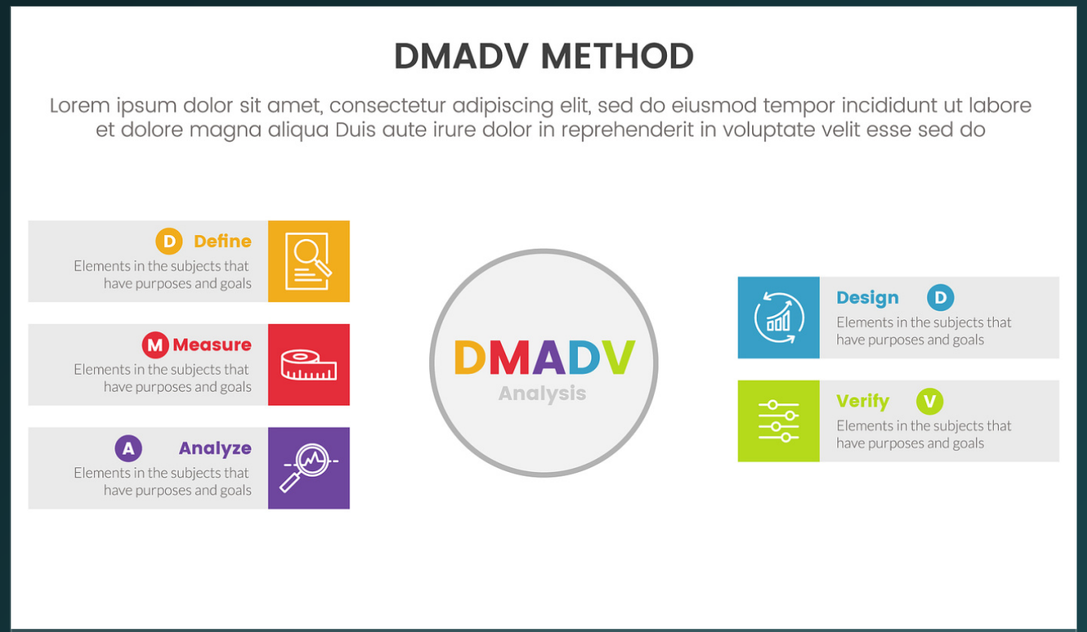
Instructors：Shahizan

Catalog

01

Presentation Overview





Introduction to Diabetes Screening

Overview of the importance and potential of ML models in early diabetes detection

Research Objectives

Evaluating ML models, testing in low-resource settings, and recommending solutions

01

Methodology and Findings

Review of literature, dataset, methodology workflow, and comparison of ML models' performance

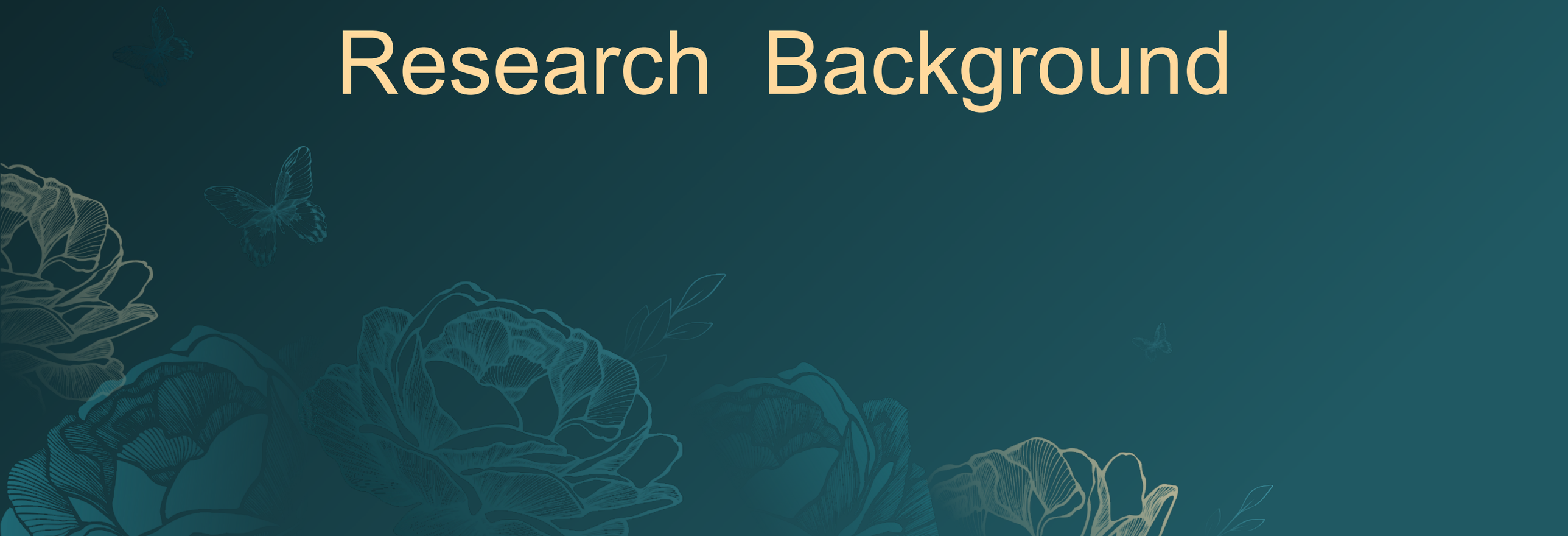
02

Conclusion and Contributions

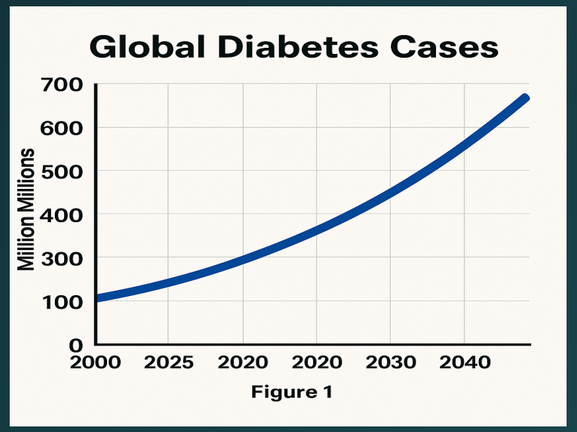
Summarizing the framework's effectiveness and its contributions to the field

02

Research Background



Global Diabetes Trend



Increasing Diabetes Prevalence

Figure 1 shows the upward trend of global diabetes cases.

Public Health Concern

The trend indicates a growing burden on healthcare systems.

Research Problem

01

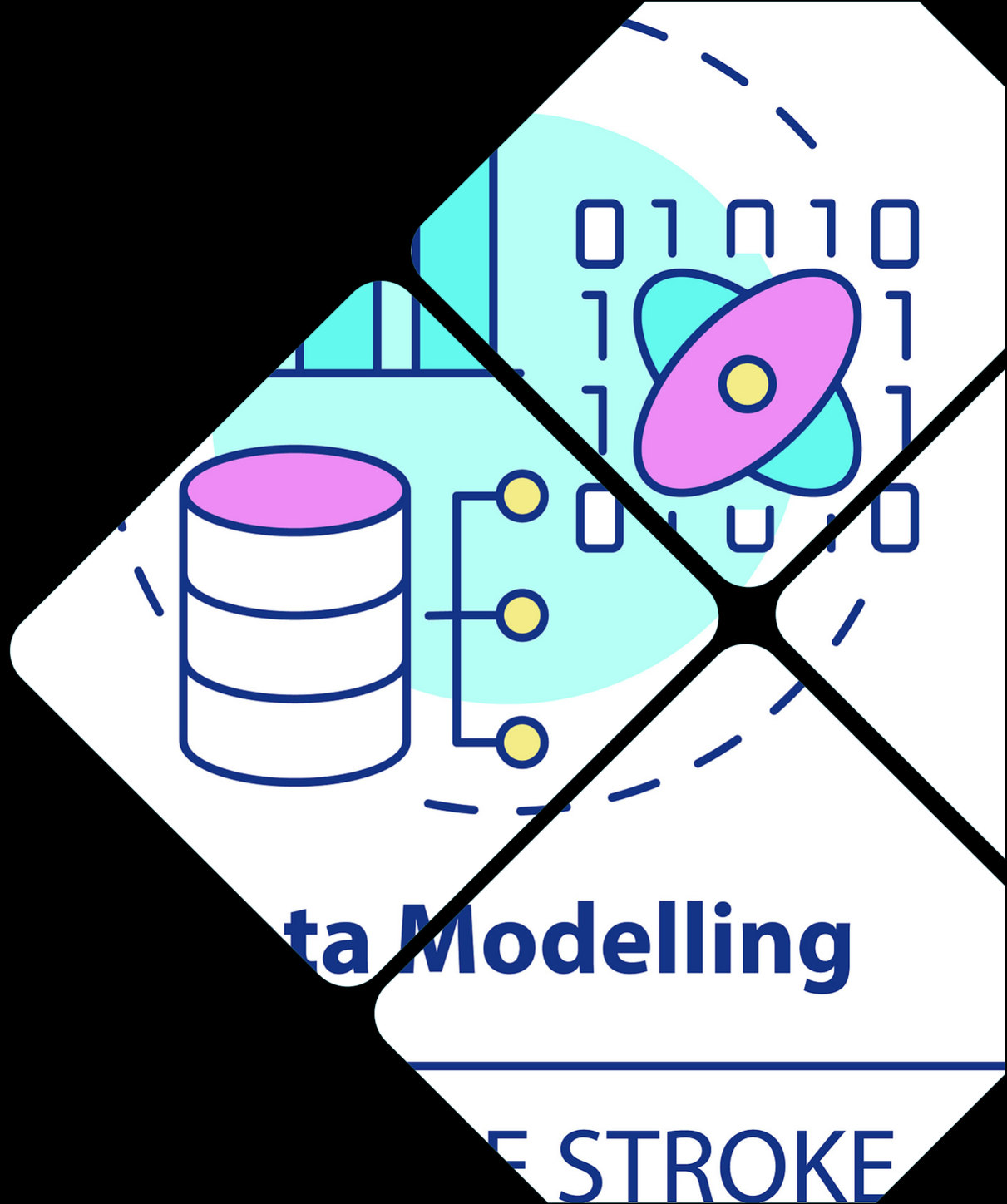
Early Screening Necessity

Identifying the best ML model for early detection of diabetes.

02

Resource-Constrained Settings

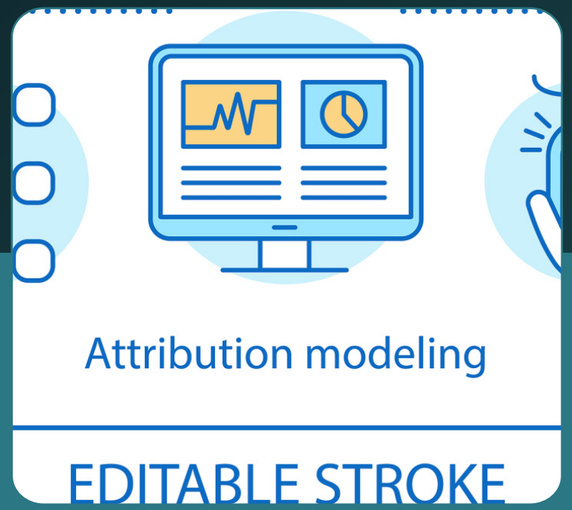
Focusing on ML models suitable for low-resource clinical environments.



03

Research Objectives

Evaluation of ML Models



Evaluate 6 ML models

The research aims to assess the performance of six traditional machine learning models for diabetes screening.

Consistency in model training

All models are trained using the same Pima dataset to ensure a fair comparison.

Testing Environment

Low-resource clinical settings

The models will be tested in resource-constrained clinical environments to evaluate their practicality.

Performance under missing data

The robustness of the models will be tested against varying levels of missing data.

Solution Recommendation

Context-based solutions

The study recommends specific ML models tailored to the constraints of different clinical settings.

Logistic Regression for low-resource

Logistic Regression is recommended for use in low-resource clinics due to its computational efficiency and interpretability.

Solution Recommendation

Random Forest for resource-rich

Random Forest is suggested if resources permit, due to its higher accuracy and predictive power.

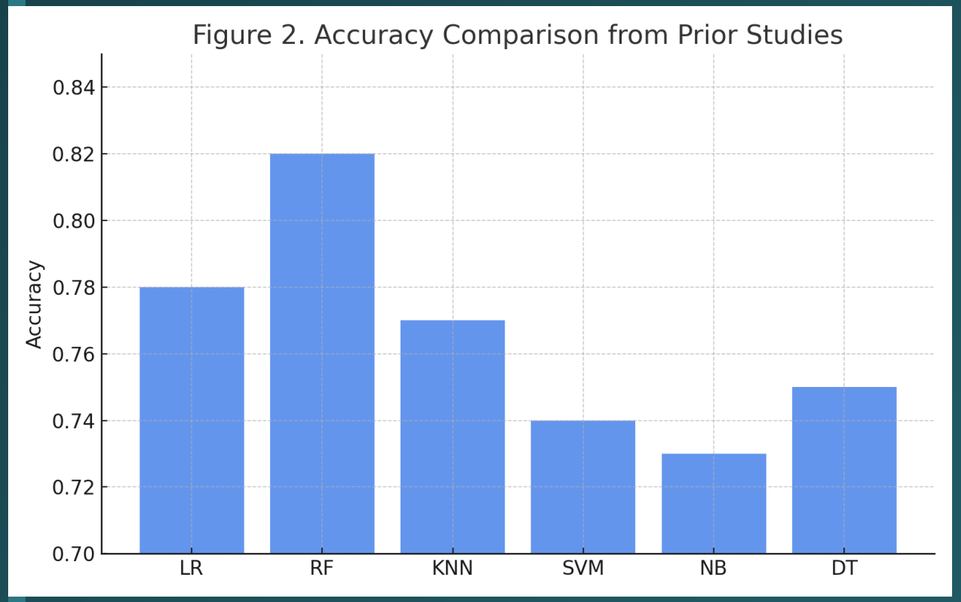
04

Methodology and Results

Literature Review

02

Model Performance Trends



01

Prior Studies Accuracy

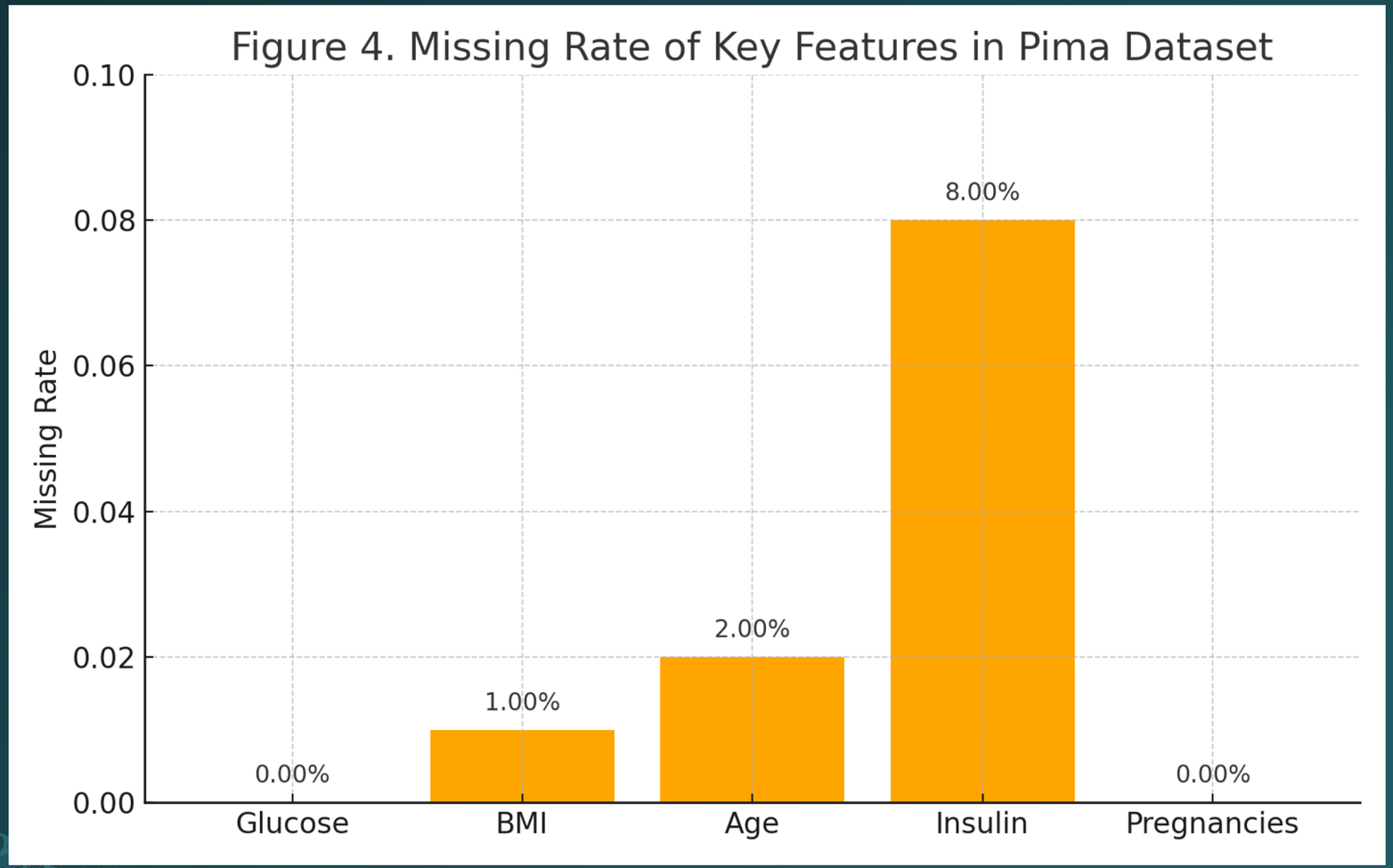
Random Forest consistently shows higher accuracy in prior studies.

Dataset Overview

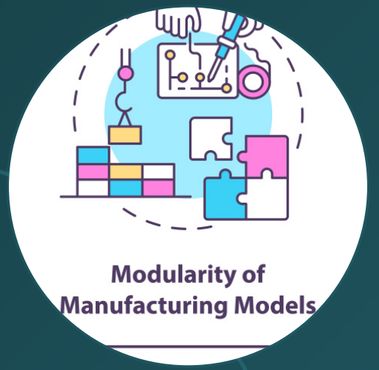
Feature Missing Rates

Only insulin shows notable missing values (~8%).

Data Set Characteristics



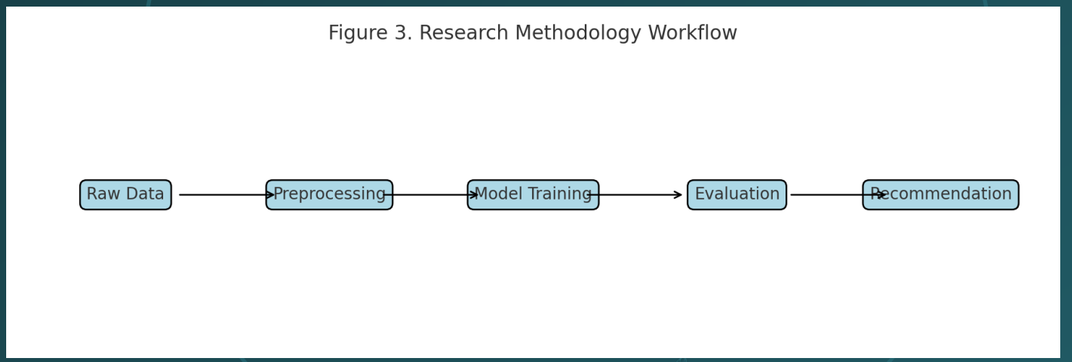
Methodology Workflow



Stepwise Pipeline Approach

Stepwise pipeline: preprocessing → model training → evaluation.

Workflow Visualization



Models Compared

01

ML Models Evaluated

Logistic Regression, Decision Tree, KNN, Naive Bayes, SVM, Random Forest.

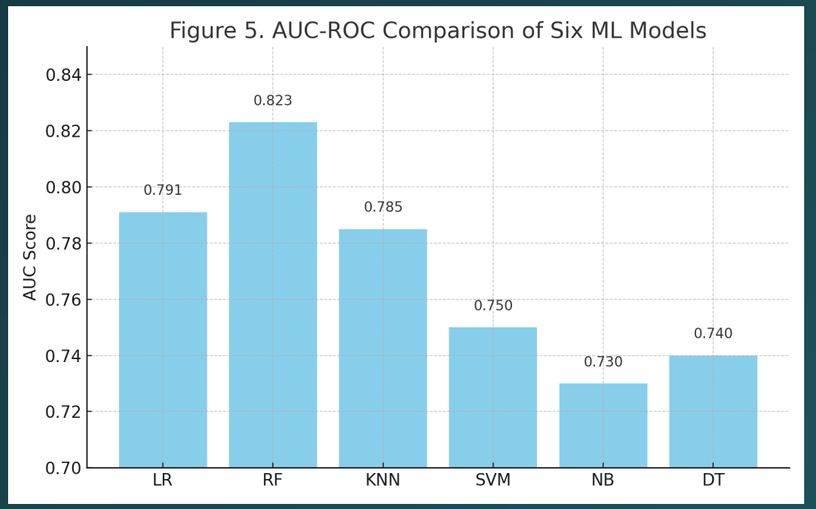
02

Training Consistency

All models are trained using the same Pima dataset for consistency.

Performance Comparison

Random Forest achieves the highest AUC (0.823).



Highest AUC Achieved

Performance Visualization

Feature Importance

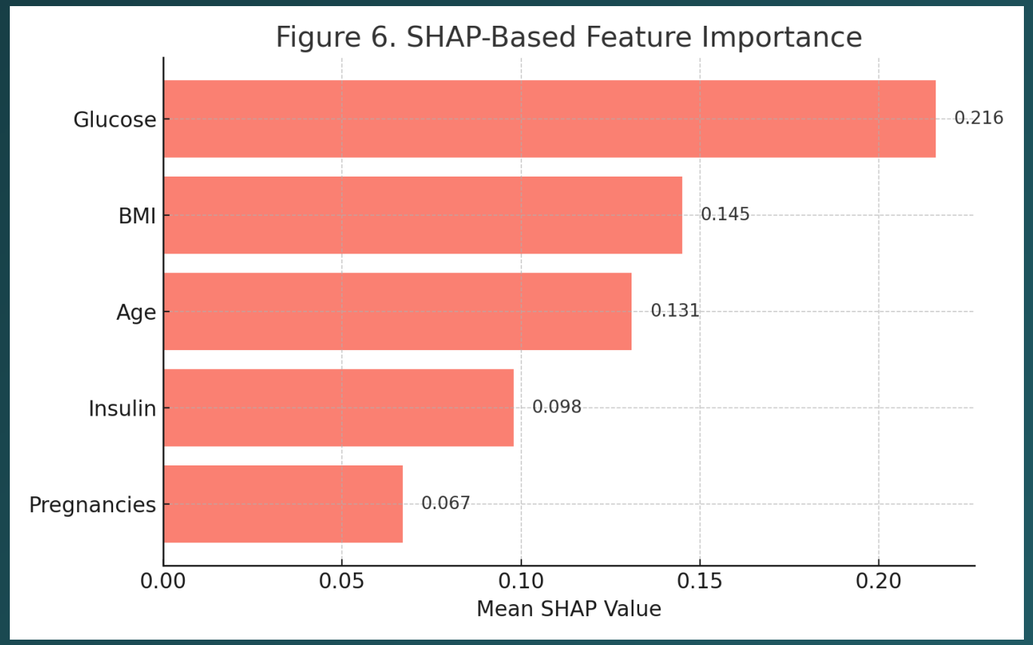
01

SHAP Value Analysis

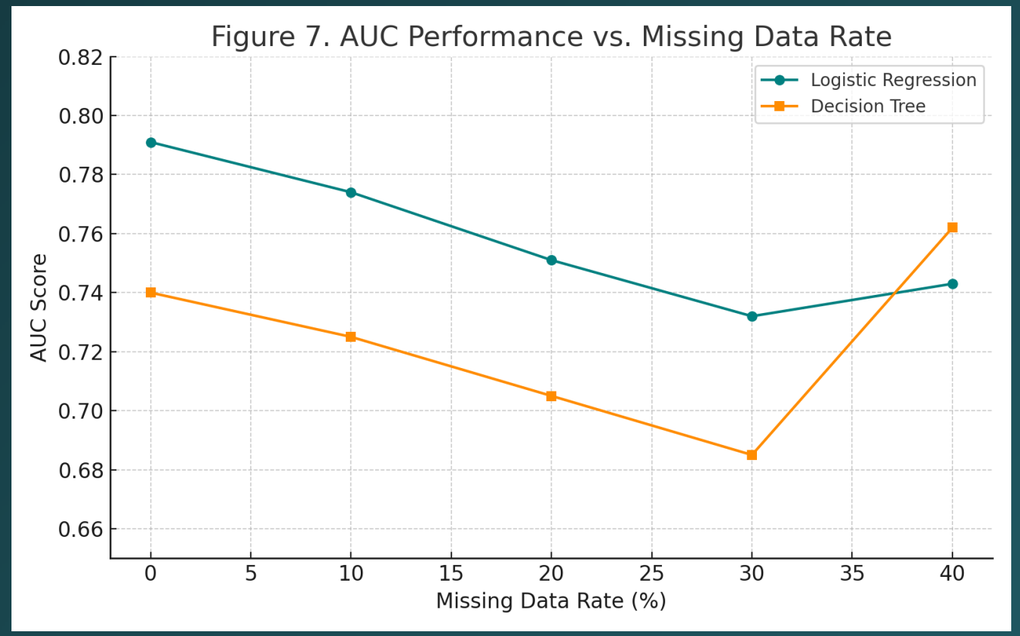
Glucose has the highest SHAP value, indicating strong predictive power.

02

Feature Impact Ranking



Missing Data Robustness



LR vs DT Stability

LR is more stable across missing data levels than DT.

Robustness Evaluation

Computational Efficiency

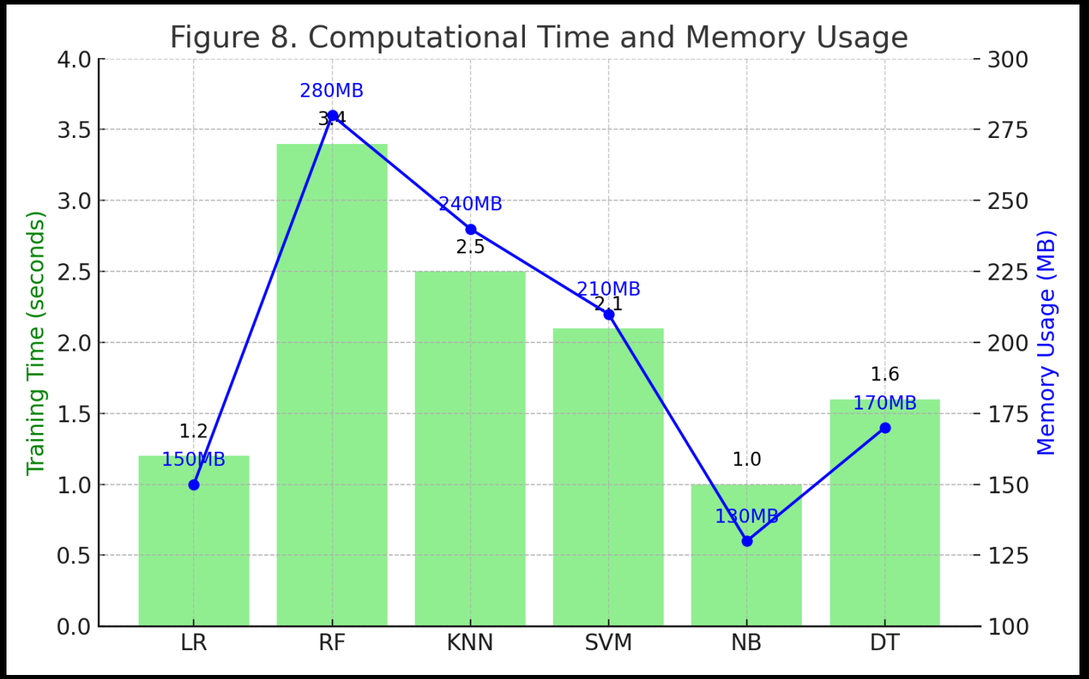
01

Naive Bayes Speed

Naive Bayes is the fastest; RF uses the most memory.

02

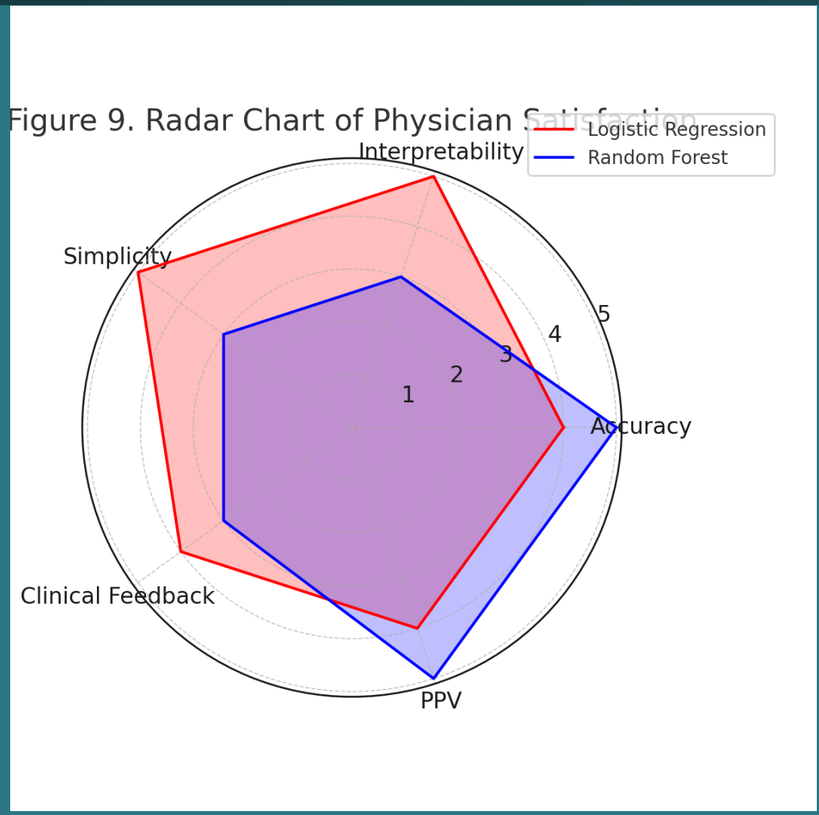
Resource Usage Comparison



Clinical Validation

Interpretability and Simplicity

LR scores highest in interpretability and simplicity.



Clinical Validation Results

Model Recommendation

Low-Resource Clinic Model

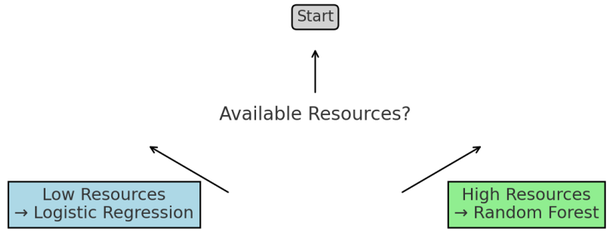
Use Logistic Regression in low-resource clinics.

Resourceful Setting Model

Use Random Forest if resources permit.

Recommendation Flowchart

Figure 10. Model Selection Flowchart Based on Resources



Video link

https://youtu.be/qGbj8_p9lCo

Thanks

Reporter: Zhang Yibo

MCS241044