The Challenge of Class Imbalance in Diabetes Prediction: A Performance-Driven Analysis of Feature and Model Selection

By

Tamal Golder

Student ID: 20221003011

Piyale Sarkar

Student ID: 20221009011

Md Alauddin Khan

Student ID: 20221012011

And

Md Mahidul Islam

Student ID: 20221015011

SUBMITTED IN PARTIAL FULLFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

BACHELOR OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

AT

NORTH WESTERN UNIVERSITY

KHULNA, BANGLADESH

13th September, 2025

NORTH WESTERN UNIVERSITY, KHULNA

DEPARMENT OF COMPUTER SCIENCE AND ENGINEERING

The undersigned hereby certify that they have read and recommended for acceptance a thesis entitled "**The Challenge of Class Imbalance in Diabetes Prediction: A Performance-Driven Analysis of Feature and Model Selection**" by Tamal Golder, Piyale Sarkar, Md Alauddin Khan, Md Mahidul Islam,in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering..

1. **Thesis Supervisor**

Nusrat Jahan Urme
Lecturer
Department of Computer Science and Engineering
North Western University , Khulna

2. **Second Examiner**

Nurzahan Akter Joly
Senior Lecturer
Department of Computer Science and Engineering
North Western University , Khulna

3. **Head of The Department**

Md. Mahedi Hasan
Assistant Professor
Department of Computer Science and Engineering
North Western University , Khulna

NORTH WESTERN UNIVERSITY, KHULNA

Date: September 13, 2025

Authors : **Tamal Golder, Piyale Sarkar, Md Alauddin Khan and Md Mahidul Islam**

Title : **The Challenge of Class Imbalance in Diabetes Prediction: A Performance-Driven Analysis of Feature and Model Selection**

Department : **Computer Science and Engineering**

Degree : **Bachelor of Science in Computer Science and Engineering**

Permission is herewith granted to North Western University to circulate and to have copied for non-commercial purpose, at its discretion, the above title upon the request of individuals or institutions.

Tamal Golder

Piyale Sarkar

Md Alauddin Khan

Md Mahidul Islam

Abstract

Diabetes has become a major health concern, especially in developing countries where the number of patients is increasing rapidly. If it is not detected early, it can lead to various complications. Therefore, building an effective prediction system is very important. In this study, we applied different machine learning algorithms to predict diabetes. The dataset used was Pima Indian Diabetes Dataset, which contains 768 samples and 9 attributes. During data preprocessing, we used the Median Method to remove missing and zero values from the dataset and used SMOTE to balance the dataset. Used Filter Method and Wrapper Method, to select the most relevant features. Five most relevant features were found from the dataset using filter method and they are: Pregnancies, Glucose, BMI, DiabetesPedigreeFunction and Age. Similarly, five most relevant features were found using the wrapper method, These are Glucose, BloodPressure, BMI, DiabetesPedigreeFunction, Age.

For splitting the data, we used Hold-Out Method, K-Fold Cross-Validation, Stratified K-Fold Cross Validation, Leave-One-Out Cross-Validation so that the generalization capability of the models could be properly tested. Several machine learning algorithms were applied, including Logistic Regression, Random Forest, XGBoost Classifier. To improve performance, we also applied hyperparameter tuning using Grid Search for each model.

The models were evaluated based on Accuracy, Precision, Recall, F1-score, and ROC-AUC. The experimental results showed that Performed best using Filter Method and Leave-One-Out Cross Validation method with XGboost Model and achieved the highest accuracy 87%, In this combination, we have balanced the class through SMOTE, which previously dropped the result down to 84%.Overall, this research indicates that a machine learning-based prediction system can be highly effective for early diabetes detection and can serve as a useful decision-support tool for healthcare professionals.

Keywords: Diabetes Prediction, Machine Learning, Data Preprocessing, Feature Engineering, Data Partitioning, Hyperparameter Tuning, Classification, Performance Evaluation

Acknowledgments

First, we express my heartiest thanks and gratefulness to almighty God for His divine blessing makes me possible to complete the final year project/internship successfully.

We express our sincere gratitude to our dedicated supervisor, **‘‘Nusrat Jahan Urme’’** (Lecturer, Department of Computer Science and Engineering, North Western University, Khulna). Our supervisor's deep knowledge and interest in implementing this project made our work easy. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this Thesis.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

Dedication

We dedicate this humble work to the Almighty, the Most Gracious and the Most Merciful,who has been our ultimate source of strength, patience, and guidance throughout this journey. His infinite blessings have enabled us to overcome challenges, stay determined, and achieve this milestone. We extend our heartfelt gratitude to our beloved parents and family,whose endless love, sacrifices, and constant encouragement have been our greatest inspiration. Their prayers and support have carried us through every step of this endeavor.

We also express our sincere appreciation to our respected supervisor, **‘‘Nusrat Jahan Urme’’**for her valuable guidance, encouragement, and unwavering support. Her mentorship and dedication have been instrumental in the successful completion of this thesis.

Table of Contents

List of Tables

List of Figures

Glossary of Terms

Terms		Full Form
ML	**:**	Machine Learning
BMI	**:**	Body Mass Index
DPF	**:**	Diabetes Pedigree Function
PIDD	**:**	Pima Indian Diabetes Dataset
CV	**:**	Cross Validation
LOOCV	**:**	Leave-One-Out Cross Validation
LR	**:**	Logistic Regression
RF	**:**	Random Forest
CNN	**:**	Convolutional Neural Network
RNN	**:**	Recurrent Neural Network
TP	**:**	True Positive
TN	**:**	True Negative
FN	**:**	False Negative
FP	**:**	False Positive
AUC	**:**	Accuracy
ROC	**:**	Receiver-operating characteristic curve
WHO	**:**	World Health Organization
SMOTE	**:**	Synthetic Minority Over-sampling Technique