

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
DEPARTMENT 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

THE AUTHORS RESERVE OTHER PUBLICATION RIGHTS, AND NEITHER THE THESIS NOR EXTENSIVE EXTRACTS FROM IT MAY BE PRINTED OR OTHERWISE REPRODUCED WITHOUT THE AUTHORS WRITTEN PERMISSION. THE AUTHORS ATTEST THAT PERMISSION HAS BEEN OBTAINED FOR THE USE OF ANY COPYRIGHTED MATERIAL APPEARING IN THIS PROJECT (OTHER THAN BRIEF EXCERPTS REQUIRING ONLY PROPER ACKNOWLEDGEMENT IN SCHOLARLY WRITING) AND THAT ALL SUCH USE IS CLEARLY ACKNOWLEDGED.

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

Title page	i
Abstract	iv
Acknowledgment	v
Dedication	vi
Table of Contents	vii
List of Tables	xi
List of Figures	x
Glossary of Terms	xi
1 Introduction	1
1.1 Background	1
1.2 Problem Statement	2
1.3 Motivation	2
1.4 Research Objecties	3
1.5 Thesis Organization	3
2 Related Works	4
2.1 Introduction	4
2.2 Comparison of Machine Learning Models for Predicting Type 2 Diabetes Risk Using the Pima Indians Diabetes Dataset.	4
2.3 Diabetes Prediction Using Machine Learning.	4
2.4 Diabetes Prediction using Machine Learning Techniques.	4
2.5 Diabetes Disease Prediction Using Machine Learning Algorithms.	5
2.6 Diabetes prediction using supervised machine learning.	5
2.7 Prediction of Diabetes using Classification Algorithms.	5
2.8 Early Prediction of Diabetes Mellitus Using Machine Learning.	5
2.9 Techniques of Machine Learning for the Purpose of Predicting Diabetes Risk in PIMA Indians	5
2.10 Diabetes Prediction Using Machine Learning Algorithms and Ontology.	6
2.11 Analisis Perbandingan Ensemble Machine Learning Dengan Teknik Somote Untuk Prediksi Diabetes.	6
2.12 Diabetes Prediction Using Machine Learning Techniques.	6
2.13 Early Diabetes Detection using Novel Hybird Machine Learning Algorithm.	6
2.14 SHAP-Based Explainable Framework for Disease Prediction and Comorbidity Risk Assessment in Healthcare.	7
2.15 Comparative Analysis of Random Forest and Support Vector Machine for Classifying Pima Indians Diabetes Dataset.	7
2.16 Literature Review Summary	7

3	Methodology	11
3.1	Introduction	11
3.2	System Architecture	11
3.3	Data Collection	12
3.4	Dataset Description	12
3.5	Descriptive Statistics of the Dataset	12
3.6	Data Visualization	13
3.6.1	Histogram	14
3.6.2	Correlation Matrix	17
3.6.3	Boxplots	18
3.7	Data Preprocessing	21
3.7.1	Median Method	21
3.7.2	SMOTE	22
3.8	Feature Extraction	22
3.8.1	Filter Method	22
3.8.2	Wrapper Method	22
3.9	Data Partitioning	23
3.9.1	Hold-out Method	23
3.9.2	Stratified K-Fold Validation Method	24
3.9.3	K-Fold Cross Validation Method	24
3.9.4	LOOCV (Leave-One-Out Cross Validation) Method	25
3.10	Model Selection	25
3.10.1	Logistic Regression	25
3.10.2	Random Forest	26
3.10.3	XGBoost	27
3.11	Hyperparameter Tuning	27
3.11.1	Grid Search	27
3.12	The hypermat parameters used in various machine learning algorithms are given in the table below.	28
3.13	Evaluation Metrics	29
4	Results	32
4.1	Results Analysis	32
4.2	Logistic Regression	32
4.3	XGBoost	36
4.4	Random Forest	40
4.5	Handling Class Imbalances	44
4.6	Comparison with Previous Work	45
5	Conclusion and Future works	46
5.1	Conclusion	46
5.2	Future Works	46
	References	47

List of Tables

Table No	Table Name	Page No
2.1	Summary of existing works on diabetes prediction using machine learning.	10
3.1	Dataset Feature Description.	12
3.2	Summary Stats for Diabetes Dataset	13
3.3	Missing Value Count.	21
3.4	Zero Value Count.	21
3.5	Hyperparameters of Different ML Model	29
4.1	Logistic Regression Performance of Different Feature Selection and Data Partitioning Combinations	32
4.2	Classification Reports(Logistic Regression uses Filter Methods with Hold-Out methods)	34
4.3	Classification Reports(Logistic Regression uses Wrapper Method with Stratified K-fold Cross Validation)	34
4.4	XGBoost Performance of Different Feature Selection and Data Partitioning Combinations	36
4.5	Classification Report(XGBoost Model using Filter Method with Leave-One- Out Cross Validation)	38
4.6	Classification Report (XGBoost Model using Wrapper Method with Stratified k-fold cross validation method)	38
4.7	Random Forest Performance of Different Feature Selection and Data Partitioning Combinations	40
4.8	Classification Report (Random Forest Model using Filter Method with k-fold cross validation method)	42
4.9	Classification Report (Random Forest Model using Wrapper Method with Stratified K-fold Cross Validation)	42
4.10	Comparative Analysis with Other Related Work the Using Same Dataset	45

List of Figures

Figure No	Figure Name	Page No
1.1	Diabetes Impact Statistics.	1
3.1	Operational Process of Diabetes Prediction.	11
3.2	Outcome Cases.	13
3.3	Histogram Relation Between Single Attribute.	15
3.4	Correlation Matrix.	17
3.5	Boxplots.	19
3.6	Process of a Wrapper Method.	23
3.7	Hold-Out Method.	23
3.8	Stratified K-Fold Cross Validation Method.	24
3.9	K-Fold Cross Validation Method	24
3.10	LOOCV Method.	25
3.11	Logistic Regression.	26
3.12	Random Forest	26
3.13	XGBoost.	27
3.14	Grid-Search Architecture.	28
3.15	Confusion Matrix.	30
4.1	Confusion Matrix (Logistic Regression uses Filter Methods with Hold-Out methods)	33
4.2	Confusion Matrix (Logistic Regression uses Wrapper Method with Stratified K-fold Cross Validation)	33
4.3	ROC Curves (Logistic Regression uses Filter Methods with Hold-Out methods)	35
4.4	ROC Curves (Logistic Regression uses Wrapper Method with Stratified K-fold Cross Validation)	35
4.5	Confusion Matrix (XGBoost Model using Filter Method with Leave-One-Out Cross Validation)	37
4.6	Confusion Matrix (XGBoost Model using Wrapper Method with Stratified k-fold cross validation method)	37
4.7	ROC Curves (XGBoost Model using Filter Method with Leave-One- Out Cross Validation)	39
4.8	ROC Curves (XGBoost Model using Wrapper Method with Stratified k-fold cross validation method)	39
4.9	Confusion Matrix (Random Forest Model using Filter Method with k-fold cross validation method)	41
4.10	Confusion Matrix (Random Forest Model using Wrapper Method with Stratified K-fold Cross Validation)	41
4.11	ROC Curves (Random Forest Model using Filter Method with Filter Method with k-fold cross validation)	43
4.12	ROC Curves (Wrapper Method Model using Filter Method with Wrapper Method with Stratified K-fold Cross Validation)	43
4.13	Confusion Matrix (XGBoost-SMOTE)	44
4.14	ROC Curves (XGBoost-SMOTE)	44

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