IMPROVING THE QUALITY OF DATASET FOR DIABETES PREDICTION USING A MACHINE LEARNING APPROACH

BASHAR HAMAD AUBAIDAN

UNIVERSITI KEBANGSAAN MALAYSIA

IMPROVING THE QUALITY OF DATASET FOR DIABETES PREDICTION USING A MACHINE LEARNING APPROACH

BASHAR HAMAD AUBAIDAN

THESIS SUBMITTED IN FULFILMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

INSTITUTE OF VISUAL INFORMATICS
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2026

PENAMBAHBAIKAN KUALITI SET DATA UNTUK RAMALAN DIABETES MENGGUNAKAN PENDEKATAN PEMBELAJARAN MESIN

BASHAR HAMAD AUBAIDAN

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEH
IJAZAH DOKTOR FALSAFAH

INSTITUT INFORMATIK VISUAL
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2026

UNIVERSITI KEBANGSAAN MALAYSIA
The National University of Malaysia

PERAKUAN TESIS SARJANA / DOKTOR FALSAFAH
(DECLARATION OF MASTER / DOCTOR OF PHILOSOPHY THESIS)

A. MAKLUMAT PELAJAR/*STUDENT DETAILS*	
Nama *Name*	BASHAR HAMAD AUBAIDAN
No. Pendaftaran *Registration No.*	P103708
Fakulti/Institut *Faculty/Institute*	INSTITUTE OF VISUAL INFORMATICS (IVI)
Program Pengajian *Program of Study*	Sarjana / Doktor Falsafah | ~~*Masters /*~~ *Doctor of Philosophy*
Tajuk Tesis *Thesis Title*	IMPROVING THE QUALITY OF DATASET FOR DIABETES PREDICTION USING A MACHINE LEARNING APPROACH
Mel-e/*Email*:p103708@siswa.ukm.edu.my	No. Telefon/*Tel.No*:*010130750*

B. PERAKUAN / *DECLARATION*

Merujuk kepada Dasar Harta Intelek UKM 2018, **tesis adalah hak milik UKM** seperti keterangan dibawah:

4.2.1 Harta Intelek yang direka cipta oleh pelajar termasuk tesis dan semua hasil penyelidikan ditulis oleh Pelajar UKM dalam tempoh pengajiannya di UKM direka cipta, dibangunkan atau dihasilkan menggunakan kemudahan, bahan, dana, atau sumber lain adalah hak milik UKM melainkan dinyatakan sebaliknya dalam Dasar ini atau dipersetujui sebaliknya oleh UKM dan Pelajar/Penaja.

Referring to the UKM Intellectual Property Policy 2018, ***the thesis is the property of UKM*** *as described below:*

4.2.1 Intellectual property created by the student, including the thesis and all research output produced by the UKM Student during their studies at UKM, including designed, developed or produced output using facilities, materials, funds, or other resources are the property of UKM unless otherwise stated in this Policy or otherwise agreed by UKM and the Student/Sponsor.

	RAHSIA ***CONFIDENTIAL***	Mengandungi maklumat rahsia di bawah AKTA RAHSIA RASMI 1972 *Consisting of classified information under the OFFICIAL SECRETS ACT 1972*
	TERHAD ***RESTRICTED***	Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi / badan di mana penyelidikan dijalankan *Consisting of RESTRICTED information which has been determined by the organisation/body where the research was conducted*
/	**AKSES TERBUKA/ TIDAK TERHAD** ***OPEN ACCESS/ NON-RESTRICTED***	Saya membenarkan tesis ini diterbitkan secara akses terbuka, teks penuh atau dibuat salinan untuk tujuan pengajian, pembelajaran & penyelidikan sahaja *I give permission for this thesis to be published through open access, full text or copied only for study, learning, and research purposes.*

Bagi kategori Akses Terbuka/Tidak Terhad, saya membenarkan tesis (Sarjana/Doktor Falsafah) ini di simpan di Perpustakaan Universiti Kebangsaan Malaysia (UKM)* dengan syarat-syarat kegunaan seperti berikut:

For the Open Access/Non-Restricted category, I permit this (Master's/Doctoral) Thesis to be kept in the Universiti Kebangsaan Malaysia (UKM) Library with the following conditions:

1. Perpustakaan UKM mempunyai hak untuk membuat salinan untuk tujuan pengajian, pembelajaran, penyelidikan sahaja.
   UKM Library has the right to reproduce the thesis only for study, learning and research purposes.

2. Perpustakaan Universiti Kebangsaan Malaysia dibenarkan membuat satu (1) salinan tesis ini untuk tujuan pertukaran antara institusi pengajian tinggi dan mana-mana badan/ agensi kerajaan, tertakluk kepada terma dan syarat.
   UKM Library is allowed to make one (1) copy of this thesis for exchange purposes among higher education institutions and any government body/agency, subject to the terms and conditions.

Saya mengakui maklumat & tesis yang diberikan adalah benar & terkini. *I acknowledge the information & the thesis provided are correct & current.* Tandatangan Pelajar: *Student's Signature* *Tarikh / Date: 23/2/2026*	Saya memperakukan maklumat & tesis yang diberikan adalah benar & terkini. *I verify that the information & the thesis provided are correct & current.* Tandatangan Penyelia / *Supervisor's Signature*: Nama / *Name*: Cop Rasmi / *Official Stamp*: *Tarikh / Date: 24/2/2026*

C. PENGESAHAN FAKULTI/INSTITUT | *VERIFICATION OF FACULTY/INSTITUTE*

Tandatangan/*Signature*:
Nama/*Name*:
Tarikh/*Date*: 27/2/2026
Cop Rasmi/*Official Stamp*:

MOHAMAD SALLEH SULIEMAN
Penolong Pendaftar Kanan
Institut Informatik Visual (IVI)
Universiti Kebangsaan Malaysia

**Kuatkuasa:15 Julai 2021

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

24 February 2026

BASHAR HAMAD AUBAIDAN
P103708

ACKNOWLEDGEMENT

I would like to express my gratitude to the Almighty Allah, who blessed me and given me the patience and good health throughout the duration of this PhD research.

I am very fortunate to have Associate Professor Dr Rabiah Abdul Kadir as my research supervisor.

I would like to express my high appreciation to my co-supervisor Dr Mohamad Taha Ijab.

I would like to thank all postgraduate students of the UKM research group for their help, friendship, and creating a pleasant working environment throughout my years in UKM.

ABSTRACT

Diabetes is a chronic metabolic condition marked by persistently elevated blood glucose levels and influenced by multiple physiological and lifestyle factors, making early prediction challenging. Reliable predictive modelling requires high-quality datasets; however, issues such as missing values, class imbalance, and redundant or high-dimensional attributes often reduce model accuracy and generalisability. To address these limitations, this study introduces an integrated data-quality enhancement framework combining Bidirectional Neighbour Graph (BNG) imputation for missing data, Clustering Selection Synthesis Filtering (CSSF) for class imbalance, and Rough Set Theory (RST) for dimensionality reduction and feature selection. The proposed BNG–CSSF–RST framework was applied to three independent datasets: the Pima Indians Diabetes Dataset from the UCI Repository, the Diabetes Clinical Dataset (Kaggle, 2024), and the Kaggle Diabetes Prediction Dataset (2023). After preprocessing with the proposed framework, both Artificial Neural Network (ANN) and Support Vector Machine (SVM) models were trained and evaluated. Across all datasets, the framework yielded notable improvements in predictive performance. Using the Pima dataset (768 records, 9 features), ANN achieved 93.51% accuracy, while SVM reached 90.26%. For the Diabetes Clinical Dataset (100,000 records, 17 features), ANN obtained 96.95% accuracy and SVM achieved 96.32%. On the Kaggle Diabetes Prediction Dataset (100,000 records, 9 features), ANN attained 91.49% accuracy, and SVM achieved 89.12%. Overall, the results indicate that systematically addressing missing data, class imbalance, and irrelevant or redundant features substantially improves classification performance. The enhanced accuracies observed across all three datasets exceed those typically reported in earlier studies, confirming the robustness of the proposed BNG–CSSF–RST framework. Finally, this approach provided a methodology for diabetes prediction specifically designed to address missing data, class imbalance, and dimensionality reduction, thereby enhancing overall data quality and enabling more robust analysis of complex datasets.

PENAMBAHBAIKAN KUALITI SET DATA UNTUK RAMALAN DIABETES MENGGUNAKAN PENDEKATAN PEMBELAJARAN MESIN

ABSTRAK

Diabetes ialah keadaan metabolik kronik yang dicirikan oleh paras glukosa darah yang tinggi secara berterusan dan dipengaruhi oleh pelbagai faktor fisiologi serta gaya hidup, menjadikannya sukar untuk diramal pada peringkat awal. Pemodelan ramalan yang boleh dipercayai memerlukan set data berkualiti tinggi; namun begitu, isu seperti nilai hilang, ketidakseimbangan kelas, serta atribut yang berlebihan atau berdimensi tinggi sering menurunkan ketepatan dan kebolehgunaan model. Bagi menangani kekangan ini, kajian ini memperkenalkan rangka kerja peningkatan kualiti data bersepadu yang menggabungkan imputasi Bidirectional Neighbour Graph (BNG) untuk data hilang, Clustering Selection Synthesis Filtering (CSSF) untuk menangani ketidakseimbangan kelas, dan Rough Set Theory (RST) bagi pengurangan dimensi serta pemilihan ciri. Rangka kerja BNG–CSSF–RST ini diaplikasikan pada tiga set data bebas: Pima Indians Diabetes Dataset daripada UCI Repository, Diabetes Clinical Dataset (Kaggle, 2024), dan Kaggle Diabetes Prediction Dataset (2023). Selepas pra-pemprosesan menggunakan rangka kerja ini, model Artificial Neural Network (ANN) dan Support Vector Machine (SVM) dilatih dan dinilai. Prestasi yang diperoleh menunjukkan peningkatan ketara merentas semua set data. Untuk set data Pima (768 rekod, 9 ciri), ANN mencapai 93.51% ketepatan, manakala SVM mencapai 90.26%. Bagi Diabetes Clinical Dataset (100,000 rekod, 17 ciri), ANN memperoleh 96.95% ketepatan dan SVM mencapai 96.32%. Dalam Kaggle Diabetes Prediction Dataset (100,000 rekod, 9 ciri), ANN mencapai 91.49% manakala SVM memperoleh 89.12%. Secara keseluruhan, keputusan menunjukkan bahawa menangani isu nilai hilang, ketidakseimbangan kelas, serta ciri yang tidak relevan atau berlebihan secara sistematik meningkatkan prestasi klasifikasi dengan ketara. Ketepatan yang diperoleh merentas ketiga-tiga set data melebihi kajian terdahulu, mengesahkan keteguhan rangka kerja BNG–CSSF–RST yang dicadangkan. Akhirnya, pendekatan ini menyediakan satu metodologi untuk ramalan diabetes yang direka khusus untuk menangani data hilang, ketidakseimbangan kelas, dan pengurangan dimensi, sekali gus meningkatkan kualiti data keseluruhan dan membolehkan analisis yang lebih teguh terhadap set data yang kompleks.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

EAS	Ensemble Attribute Selection
ENN	Edited Nearest Neighbour
H.D	Feature reduction
EHR	Electronic Health Records
I.C	Imbalanced Dataset
KNN	K-Nearest Neighbours
M.V	Missing Values
RF	Random Forest
UDC	Untrainable Data Cleansing
UKM	Universiti Kebangsaan Malaysia

CHAPTER I

INTRODUCTION

1.1 INTRODUCTION

While the exponential growth of diabetes dataset due to Electronic Health Records (EHRs) offers exciting possibilities for healthcare improvement, these innovations hinge critically on the condition of the data. High-quality data is imperative for building trustworthy machine learning models and extracting meaningful insights. Thus, strategies to improve data quality within EHR systems are crucial to fully unlocking the potential of these vast datasets. There has been a notable trend in the number of academic and industry professionals focusing on methods for development, support, machine learning applications, and data mining in recent years. Common issues associated with diabetes dataset include missing data, imbalanced datasets, and excessive dimensionality (Enríquez et al. 2021; Abdulrauf Sharifai & Zainol 2020).

Missing data ranges from missing sequences to incomplete features, and it is crucial to perform proper feature engineering to accurately interpret the missing information (Khan & Hoque 2020; Nugroho, Utama & Surendro 2020). Missing data, class imbalance, and feature reduction are particularly relevant in the medical field, where accurate predictions are of utmost importance in patient care and treatment decisions. Dealing with missing data is essential as it significantly influences the predictive ability of machine learning applications (You et al., 2020). By applying appropriate feature engineering methods and effectively handling missing data, the interpretability of the models can be enhanced. Addressing missing data in diabetes dataset is critically important (Piri 2020; Khan & Hoque 2020).

Imbalanced datasets pose challenges in machine learning classification processes, and therefore, many scholars have conducted studies to solve this issue. The imbalance of classes in datasets poses another challenge, especially when one class occurs notably more frequently than others. This imbalance can lead to biased or compromised predictions and diminished predictive accuracy. By exploring different approaches and algorithms specifically designed to manage and circumvent the issue of class imbalance, the accuracy and fairness of the predictions can be enhanced (Huang & Cheng 2020).

Lastly, feature reduction is a significant challenge in the biomedical field due to the presence of numerous features, including irrelevant and redundant ones. Feature reduction generally leads to issues such as computational intricacies and overfitting. This challenge underscores the significance of methods such as dimensionality reduction and feature selection, which are essential to improve algorithmic performance (Bania & Halder 2020). Additionally, feature reduction poses an obstacle to effectively utilising diabetes dataset for predictive tasks. The presence of many features can result in computational complexity, overfitting, and reduced model performance. To overcome these challenges, dimensionality reduction techniques, such as feature selection, are utilised to select features which are most relevant, and mitigate the feature reduction issue.

The adoption of EHR is driving a rapid increase in diabetes dataset, underscoring the significance of data quality in the application of machine learning in the healthcare industry. As more diabetes dataset are collected and processed to assist physicians in understanding patient conditions, machine learning is becoming increasingly vital in the industry. However, creating effective supervised or unsupervised learning models can be challenging when data quality is poor during research or testing. This study addresses these issues and suggests ways for the machine-learning community to improve healthcare, with a focus on predicting diabetes and removing low-quality data from training sets.

This chapter continues with a concise overview of the research background presented in Section 1.2, which guides the direction of the research terms. Section 1.3

focuses on identifying the problem statement by highlighting the existing research and theoretical gaps that need to be addressed. The specific objectives of the research are clearly articulated in Section 1.4. To align with these objectives, Section 1.5 explores the research questions for this study. The main research contributions are outlined in Section 1.6, highlighting the unique aspects and novel insights that the study aims to offer. Section 1.7 delves into the motivations and significant issues surrounding the research, providing context for the importance and relevance of the study. Finally, Section 1.8, Thesis Organisation, offers a sequence for navigating through subsequent sections, providing clarity on the structure and organisation of the study.

1.2 RESEARCH MOTIVATIONS

Firstly, eliminating low-quality datasets is essential because the development of Artificial Intelligence (AI) models with superior accuracy and performance relies on high-quality datasets. In the biomedical field, data quality is crucial for disease prediction and supporting healthcare decisions. Poor data quality poses challenges in developing accurate supervised and unsupervised learning models. The goal of this research is to enhance dataset quality by addressing issues such as data imbalance, missing data, and feature reduction , thus improving the accuracy of disease prediction models.

Secondly, with the rapid increase in diabetes dataset due to Electronic Health Records (EHR), there are opportunities to improve healthcare services. However, inherent issues such as missing data, class imbalance, and feature reduction hinder effective analysis. This research aims to improve the overall quality of diabetes dataset by resolving problems related to raw data by using a novel technique in improving dataset quality (Khan & Hoque 2020).

Thirdly, imbalanced datasets, which are characterized by skewed data distribution, pose challenges in both machine learning and data analysis classification tasks. The negative impact of imbalanced data on algorithm performance necessitates effective solutions. This research endeavours to handle the challenges of imbalanced data, with a focus on the biomedical field. Biomedical datasets often suffer from feature

reduction due to numerous features, including irrelevant and redundant ones. Feature reduction leads to computational complexity, overfitting, and decreased algorithm performance. The research sets out to minimise the dimensionality of biomedical data, enabling effective learning and analysis by selecting relevant features. As the healthcare industry increasingly relies on machine learning for the analysis of diabetes dataset, the need for high-quality data becomes critical. This research addresses the challenges of poor data quality and aims to provide AI models with higher performance, eventually leading to better medical prognosis for patients (Zhu et al. 2021).

These research motivations highlight the importance of improving dataset quality in the biomedical field, addressing challenges like class imbalance, missing data, and feature reduction . The proposed research aims to develop techniques and models that improve the precision and efficacy of disease prediction and the analysis of healthcare-associated infections. By overcoming the limitations of existing diabetes dataset, this research can contribute substantially to the healthcare sector.

This study seeks to develop a model that enhances the quality of diabetes dataset by addressing various challenges such as data imbalance, missing data, and feature reduction . The proposed framework incorporates several techniques to improve the effectiveness of diabetes dataset and subsequently enhance disease prediction and healthcare-associated infection analysis. Among the most prominent obstacles in diabetes dataset is data imbalance, where certain classes have a significantly larger number of instances than others. This issue can lead to biased predictions and decreased classifier performance. To tackle this problem, the Clustering Selection Synthesis Filtering (CSSF) paradigm is utilised. The CSSF process involves clustering the data and selecting a representative subset from each cluster to create a balanced dataset.

A common challenge in diabetes dataset is the presence of missing data, which can occur for various reasons, such as data corruption or incomplete record-keeping. Missing data can significantly impact the effectiveness of machine learning algorithms, many of which are not designed to handle such gaps. To overcome this challenge, the proposed framework employs a Bidirectional Closest Neighbour Graph-based imputation method (Santhanam & Padmavathi 2015). This method utilises the graph

structure to impute missing data by considering complex interactions between observations and features. By leveraging the shortest and second-shortest paths in the graph, more robust and accurate imputation results can be achieved.

Furthermore, feature reduction poses a significant problem in biomedical data analysis. High-dimensional data, characterised by enormous quantity of features and the presence of irrelevant or redundant information, requires a substantial amount of memory and can lead to overfitting. To mitigate this issue, the proposed framework incorporates a feature reduction model that combines rough set theory and meta-heuristics. This approach seeks to simplify the dimensionality of the dataset while preserving important information, thereby improving computational efficacy and classification performance.

By addressing these challenges, the proposed framework contributes to improving the overall quality of diabetes dataset. The model's effectiveness is evaluated through disease prediction and healthcare-associated infection analysis. The study underscores the critical role of high-quality data in ensuring precise and dependable predictions, which can ultimately improve patient outcomes.

By enhancing the quality of diabetes dataset, the proposed research can significantly improve medical data analysis and ultimately lead to better patient outcomes. The research holds substantial significance in addressing the limitations of existing diabetes dataset, which can impede the creation of effective medical analysis technologies. Through the utilisation of methods such as class-specific sampling with feature selection (CSSF), normalisation, bidirectional neighbour graph-based missing data imputation, and feature reduction using rough set theory and meta- heuristics, the proposed framework aims to improve the quality and efficacy of diabetes dataset. This research area is crucial and actively pursued, as handling the limitations of diabetes dataset can significantly impact the healthcare industry and contribute to saving lives. The proposed framework presented in this research offers a comprehensive approach to address various challenges in diabetes dataset. The CSSF process addresses data imbalance, normalisation improves machine learning algorithms' performance, bidirectional neighbour graph-based imputation method handles missing data, and

feature reduction using rough set theory and meta-heuristics tackles feature reduction . These methods collectively seek to boost the precision and effectiveness of disease prediction and healthcare-related infection analysis. By addressing these obstacles and enhancing the quality of diabetes dataset, this research aids in creating more advanced analytical tools, ultimately benefiting the healthcare sector, and enhancing patient outcomes. The massive development of bioinformatics diabetes dataset provides physicians with new opportunities to improve patient prediction. With large diabetes dataset, machine learning increasingly gains significance as an indispensable tool to improve medical prognosis and patient care. In recent years, researchers and healthcare professionals have concentrated on developing methods derived from advanced analytics and intelligent algorithms to enhance decision-making support. Medical data inherently exhibits several challenging characteristics, such as missing data, class imbalance, and feature reduction . Data quality is an important aspect of the data mining process for disease prediction (Deng et al. 2016; Khan & Hoque 2020). Low-quality data can result in inaccurate or weak performance. The issue of data quality has become particularly prominent in recent years. Since the early 1990s, computer researchers have studied how to quantify and enhance the quality of various sorts of data. Most of this research is theoretical in nature, and the issue has not yet been fully solved. A proposed a solution is based on a data-object-driven theory; the use of graphical DSLs (Domain-Specific Languages) allows multiple stakeholders to participate, even if they represent different departments or organisations. To make diabetes dataset more productive and applicable for disease prediction, a number of significant obstacles have to be overcome (Fazelpour 2016, Ebenuwa 2019; Janis Bicevski et al. 2019).

The first issue is missing data, which is common in medical research across different types of datasets and has a negative effect on analysis processing. Data with missing values can degrade the classifier's performance and lead to incorrect conclusions (Nagarajan & Dhinesh Babu 2019). Medical data often contains missing values for a variety of reasons, such as the difficulty in collecting comprehensive data due to personal privacy concerns. Research on handling missing values is an active area of study (Bai, Mangathayaru & Rani 2015). Mean, median and mode imputations are among the most used simple imputation techniques to handle missing data. These methods address gaps by substituting with a summary measure, such as the variable’s

median or mean. However, they are often criticised for their inability to preserve the intricate relationships between variables, resulting in biased results (Shaikhina et al. 2017).

On the other hand, statistical models are utilised in model-based imputation to estimate the missing data. For example, missing data can be estimated based on the observed values in the dataset using a linear regression model. Model-based imputation techniques can lead to more accurate imputations than simple imputation methods, but they require a large sample size and may be computationally expensive (Peralta et al. 2021; Latha et al. 2019).

1.3 PROBLEM STATEMENT

The application of machine learning (ML) in healthcare has gained significant traction due to its ability to analyze large volumes of complex data. However, several challenges persist, including Missing values, imbalanced data class distributions, and feature reduction , all of which hinder the effectiveness of predictive models. Traditional imputation techniques, such as K-Nearest Neighbors (KNN) and mean imputation, are frequently employed to handle missing values, but these methods often fail to capture the complex relationships inherent in healthcare data. This research aims to address these issues through advanced ML techniques that enhance data quality and model performance, specifically focusing on imputation, class balancing, and dimensionality reduction. Therefore, Literature underscores the importance of improving the reliability of healthcare data to ensure accurate predictive outcomes. While numerous studies have explored the use of ML to enhance healthcare outcomes, they often fail to sufficiently address data quality concerns, including imbalances in class distribution and missing data.

Model performance is adversely affected because of challenges such as missing data, imbalance class, and feature reduction . Because people with diabetes tend to be underrepresented in datasets, the widespread problem of class imbalance is more noticeable in these collections. Xu et al. (2020) explored algorithms for managing imbalanced datasets and missing values (Xu et al., 2020)., while Fujiwara et al. (2020)

introduced a novel technique, HusdoS-boost, that combines boosting with heuristic under-sampling and distribution-based oversampling to mitigate overfitting and the deletion of valuable data (Fujiwara et al). Despite these contributions, the problem of handling imbalanced medical data remains unresolved, with many studies recommending hybrid sampling methods such as M-SMOTE combined with edited nearest neighbor (ENN) and random forest (RF) to improve accuracy. Further challenges arise from the increasing use of big clinical datasets, particularly in electronic health records (EHRs), where missing data, unbalanced classes, and excessive feature dimensionality complicate the analysis. Simple fixes, such as mean imputation or KNN, have proven insufficient in many cases, prompting a shift toward more context-aware techniques (Jabbar & Washington, 2024; Sanwouo et al., 2024). Additionally, class imbalance, particularly when data classes overlap, leads to poor performance in minority class prediction, a common issue in diabetes prediction models (Chen et al., 2024; Kumar et al., 2024). Research by Farnoosh et al. (2025) shows that predictive models can be severely limited in their ability to generalise when there is a significant class imbalance (Farnoosh et al., 2025).

Recent studies suggest that more advanced methods, including smarter sampling and cost-sensitive models, outperform traditional techniques like SMOTE. Moreover, feature dimensionality remains a pressing concern in healthcare data analysis, with high-dimensional datasets often leading to overfitting and poor interpretability. Recent approaches that combine feature selection techniques, such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA), have shown improvements in both accuracy and interpretability compared to using PCA alone (Idris et al., 2024; Fu et al., 2024). To address these challenges, this study proposes an integrated framework that uses Bidirectional Neighbour Graph (BNG) for imputation, Clustering Selection Synthesis Filtering (CSSF) for class imbalance, and Rough Set Theory for feature reduction to resolve missing data, imbalanced classes, and feature reduction in diabetic datasets. Data quality will be improved through the incorporation of advanced techniques for improving data quality and model accuracy in this study, so that healthcare predictions will be more accurate. As a result, robust, scalable, and interpretable machine learning models can be developed for healthcare, ultimately resulting in better decision-making and better patient care (see Figure 1.1).

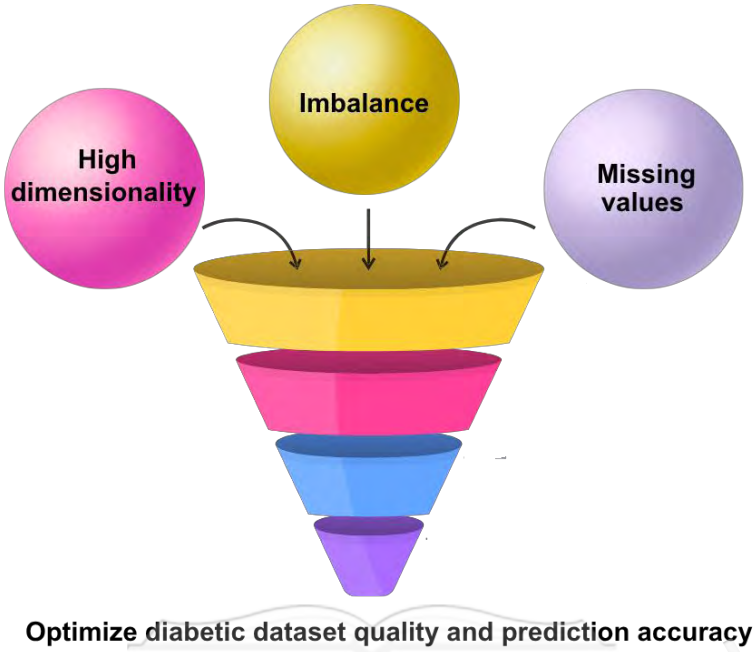


Figure 1.1 illustrate the problem statement configuration

1.3.1 Theoretical Gap

The study addresses critical theoretical gaps in handling diabetes dataset for disease prediction, specifically diabetes. These gaps include missing values, class imbalance, and feature reduction . Traditional techniques for addressing missing values, such as median or mean imputation, often result in biased predictions due to their inability to preserve inherent relationships within the data. The study proposes a Bidirectional Neighbour Graph (BNG)-based imputation method to improve accuracy by leveraging both direct and indirect data relationships (Pratama 2016).

Class imbalance, where one class significantly outweighs others, presents another challenge. The Clustering Selection Synthesis Filtering (CSSF) technique is presented as a means to create synthetic samples within clusters of the minority class, thus achieving a balanced dataset and improving model performance (Wongvorachan et al. 2023). Feature reduction , characterised by numerous features, either redundant or irrelevant, can result in overfitting and computational inefficiencies. The study employs Rough Set Theory (RST) to reduce dimensionality, identify the most relevant features, and improve both computational efficiency and model interpretability.

This research addresses critical challenges in medical data analysis, particularly for disease prediction in diabetes, by focusing on three pervasive issues: missing values, class imbalance, and feature reduction . These challenges significantly affect the reliability and accuracy of machine learning models, making their resolution essential for improving predictive performance and clinical decision-making.

To address missing data, the Bidirectional Neighbor Graph (BNG) method was chosen for its ability to preserve relationships within the dataset, ensuring context-aware imputation and superior data quality compared to traditional approaches. For class imbalance, the study employs Clustering Selection Synthesis Filtering (CSSF), which synthesizes minority class samples within clusters, reducing noise and outperforming techniques like SMOTE. To mitigate feature reduction , Rough Set Theory (RST) was utilized, offering efficient feature selection to enhance computational efficiency and model interpretability.

Performance was evaluated using metrics such as precision, recall, F1-score, and AUC, ensuring robust assessment of minority class performance. Comparative analyses against established techniques validated the superiority of the proposed methods, with the Pima Indians Diabetes Dataset serving as the primary dataset, supplemented by additional datasets for generalizability. This comprehensive framework underscores the study's contribution to advancing data quality and predictive modelling in healthcare by overcoming these hurdles, the research aims to enhance the quality of diabetes dataset, thereby improving the prediction models' accuracy, ultimately contributing to better patient outcomes in healthcare, as shown in Figure 1.2 below.

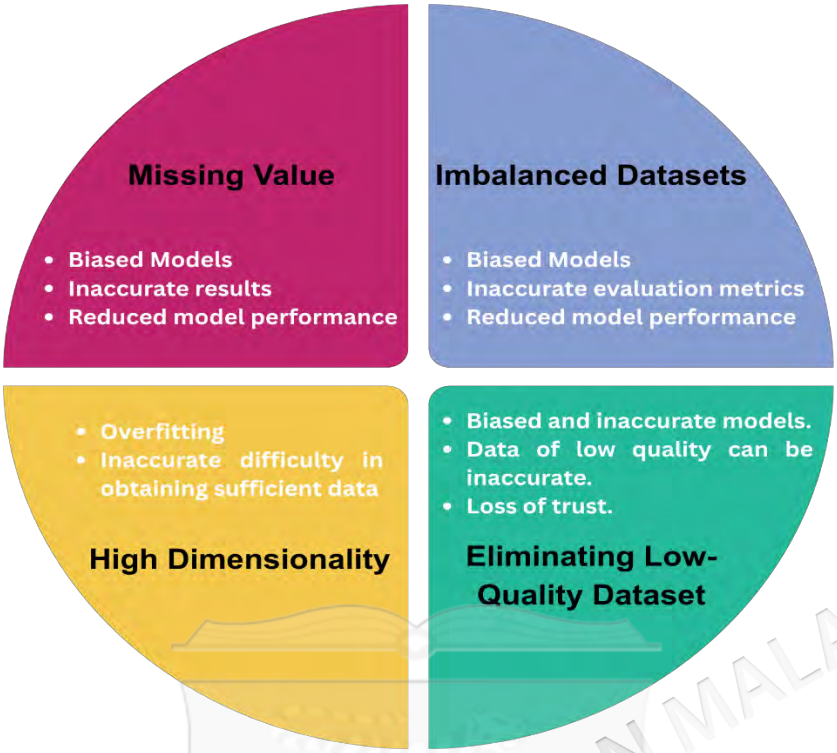


Figure 1.2 Theoretical Gaps in Healthcare Data

1.4 RESEARCH OBJECTIVES

The primary goal of this research is to develop a model that enhances the quality of datasets by eliminating any issues or challenges that may necessitate further investigation. Table 1.1 illustrates the connections between the research problems, objectives, and questions. The following sub-objectives are to be achieved:

1. To design and implement a new Bidirectional Neighbour Graph-based missing value imputation method to address missing data in diabetes dataset.

2. To enhance a sampling model based on Clustering Selection Synthesis Filtering (CSSF) to handle the class-imbalance challenge in datasets and improve disease classification.

3. To implement an RST-based feature reduction approach to filter relevant attributes from complex datasets

4. To validate accuracy of prediction models by enhancing the quality of three models of Diabetes datasets.

1.5 RESEARCH QUESTIONS

The research questions are as follow:

1. How can missing data in diabetes dataset be effectively handled?

2. How can the class-imbalance issue in diabetes dataset be managed effectively?

3. How can the feature reduction issue in diabetes dataset be reduced?

4. Does improving the quality of diabetes dataset enhance the performance of prediction models?

Table 1.1 The relationships between the problem statements, research questions, and research objectives

No	Problem statement	Research Question	Research Objective
1	Missing values remain a significant issue in diabetes dataset, leading to incomplete or unreliable analyses.	How can missing data in diabetes dataset be effectively handled?	To design and implement a Bidirectional Neighbour Graph-based missing value imputation technique to address missing data issues in diabetes dataset.
2	Diabetes prediction models often suffer from class-imbalance problems, which reduce model accuracy and bias predictions.	How can the class-imbalance issue in diabetes dataset be managed effectively?	To enhance a sampling model based on Clustering Selection Synthesis Filtering (CSSF) to handle class-imbalance in healthcare data and improve disease classification.
3	The feature reduction of medical data complicates analysis and increases computational complexity.	How can the feature reduction issue in diabetes dataset be reduced?	To implement a feature reduction based on the Rough Set Theory (RST) algorithm for handling high-dimensional datasets.
4	Poor data quality can lead to inaccurate medical predictions, potentially resulting in serious patient risks.	Does improving the quality of diabetes dataset enhance the performance of prediction models?	To improve the performance and accuracy of prediction models by enhancing the overall quality of diabetic datasets.

1.6 RESEARCH CONTRIBUTION

This research makes significant contributions to the enhancement of machine learning models for diabetes prediction by addressing key challenges inherent in diabetes dataset, such as missing data, imbalanced class distributions, and feature reduction . The proposed BNG-CSSF-RST framework is validated across three diverse diabetes datasets—the Pima Indians Diabetes Dataset (PIDD) with 768 records and 9 features, the Diabetes Clinical Dataset (Kaggle, 2024) with 100,000 records and 17 features, and the Diabetes Prediction Dataset (Kaggle, 2023) with 100,000 records and 9 features—ensuring comprehensive evaluation across varying data complexities and sample sizes

First, the study introduces a novel approach to handling missing data through the application of Bidirectional Neighbour Graph (BNG)-based imputation techniques. This method significantly improves data quality, which is essential for accurate analytics and model performance in subsequent phases. By leveraging BNG imputation, the research ensures that the data is both robust and reliable, setting the stage for more effective modelling.

Second, the research tackles class imbalance by incorporating Clustering Selection Synthesis Filtering (CSSF) to generate synthetic samples. This method effectively balances the class distribution, addressing a common issue in healthcare datasets, and ensures that the machine learning models are unbiased and better equipped to predict minority classes

Third, dimensionality reduction is achieved through the application of Rough Set Theory (RST). This technique reduces the dataset complexity without compromising the integrity of the data, enabling more efficient processing and faster analytical workflows. The application of RST contributes to the simplification of data while preserving crucial information, making the analysis more computationally feasible. This research is grounded in established machine learning and data preprocessing theoretical frameworks, filling critical gaps identified in contemporary literature. By integrating BNG for imputation, CSSF for class balancing, and RST for feature reduction into a unified framework, this study expands current knowledge in the field of medical prediction models. Notably, this research is the first to explore the

combined use of these techniques within a diabetes prediction model, offering a unique approach that has not been widely studied in medical settings

Finally, this work contributes to both the theoretical and practical advancement of diabetes prediction by providing empirical evidence supporting the effectiveness of integrated preprocessing methodologies. It introduces a pioneering combination of BNG imputation and RST for handling complex diabetes dataset, which marks a significant innovation in improving data quality and prediction accuracy. Through its comprehensive approach to addressing multiple aspects of data quality, this study offers a novel strategy for enhancing predictive models in healthcare, particularly for diabetes prediction.

1.7 RESEARCH SCOPE

This research centres on developing and validating an integrated data preprocessing and classification framework specifically aimed at improving the quality of diabetes datasets and enhancing the reliability of predictive modelling. The scope covers both methodological design and empirical evaluation. Methodologically, the study introduces a Bidirectional Neighbour Graph (BNG) algorithm for robust missing-value imputation, a Clustering Selection Synthesis Filtering (CSSF) technique for addressing class imbalance, and Rough Set Theory (RST) for dimensionality reduction and feature selection.

Empirically, the framework is evaluated on three heterogeneous diabetes datasets: the Pima Indians Diabetes Dataset (UCI Repository), the Kaggle Diabetes Clinical Dataset (2024), and the Kaggle Diabetes Prediction Dataset (2023). These datasets vary in size, feature composition, and population coverage, providing a comprehensive test bed for the proposed methodology. The analytical scope of the thesis includes data quality enhancement, model training, and performance comparison of two machine learning classifiers, Artificial Neural Networks (ANN) and Support Vector Machines (SVM), under consistent preprocessing conditions. Performance is quantified using key metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to assess predictive effectiveness across datasets.

The research is limited to structured, secondary datasets related to diabetes diagnosis and risk prediction, excluding unstructured clinical records or real-time medical sensor data. General, the study aims to establish a reproducible and generalizable preprocessing–classification pipeline that enhances predictive reliability across heterogeneous diabetes dataset, providing a foundation for future implementation in intelligent diagnostic and clinical decision-support systems.

1.8 THESIS ORGANISATION

This thesis is organised into five chapters, each systematically addressing different aspects of the research process and contributing to the overall objectives of the study. Chapter I provides an overview of the research, including the background, problem statement, research objectives, research questions, and significance of the study. It also outlines the research scope and concludes with a summary of the thesis structure.

Furthermore, Chapter II presents a comprehensive review of previous studies related to data quality in diabetes dataset. It discusses key challenges such as missing data, class imbalance, and feature reduction , as well as existing methodologies developed to overcome these issues in medical data analytics.

In addition, Chapter III describes the research design, data sources, and methodological framework adopted in the study. It details the preprocessing approaches applied to address data imbalance, missing values, and feature dimensionality. Furthermore, it explains the experimental setup, implementation procedures, and computational tools utilised in model training and evaluation.Moreover, Chapter IV presents and analyses the experimental results obtained from the proposed framework using multiple diabetes dataset. It evaluates the performance of the applied algorithms, compares outcomes across datasets, and interprets the results in relation to the research objectives. The discussion also highlights the implications, strengths, and limitations of the proposed approach.The final Chapter V summarises the key findings and contributions of the study. It discusses the practical and theoretical implications for data-driven healthcare research and concludes with recommendations and potential directions for future work.

CHAPTER II

LITERATURE REVIEW

2.1 RESEARCH BACKGROUND

The incorporation of artificial intelligence (AI) into healthcare, particularly for chronic illnesses such as diabetes, has led to a transformative change in diagnosis, treatment, and management approaches. However, the journey of integrating AI, particularly machine learning, into diabetes care is fraught with significant data-related challenges. Among these, missing data, feature reduction , and class imbalance in diabetes dataset stand out as substantial obstacles to developing effective predictive models. These issues not only hinder the accuracy but also compromise the applicability of AI applications in healthcare settings. Therefore, addressing these challenges is pivotal for leveraging AI's full potential in diabetes management.

Missing data is a pervasive issue in healthcare datasets, creating gaps that can cause biased predictions and misinformed clinical decisions. This challenge is particularly acute in diabetes research, where patient records often lack complete information due to various factors, including privacy concerns and logistical issues in data collection (Little & Rubin 2020). However, innovative machine learning techniques offer promising solutions for imputing missing data, thereby enhancing the quality and utility of the data for predictive modelling (García-Laencina et al. 2010).

Chapter Two delves deep into the literature concerning AI's role in healthcare, specifically its application to chronic conditions such as diabetes. The narrative outlines the formidable challenges in crafting effective predictive models, underscored by hurdles like missing data, class imbalance, and the complexity of high-dimensional diabetes dataset. This section illuminates cutting-edge machine learning methods

devised for imputing missing data, along with strategies to manage class imbalances and techniques to simplify data complexity, all aimed at refining model's accuracy. The chapter's discourse on AI's potential in diabetes management is both critical and hopeful, underscoring the importance to overcome these data challenges to fully leverage AI's capabilities, with the goal of significantly enhancing diabetes care and patient outcomes.

Furthermore, imbalanced data is a critical challenge characterised by an uneven distribution of observations across a dataset. In the context of diabetes, this often manifests as an overwhelming majority of non-diabetic instances compared to diabetic ones, causing the model to be biased or tilted towards predicting non-diabetic instances. Such imbalance undermines the models' ability to accurately identify diabetic cases, which are clinically more significant (He & Garcia, 2009).

Talebi Moghaddam et al. (2024) used several data-level and algorithm-level interventions, such as SMOTEENN, ADASYN, and SMOTE, in conjunction with classifiers like Random Forest and Multi-Layer Perceptron, to tackle the problem of class imbalance in diabetes prediction. Their results showed that even with an imbalanced dataset, these methods significantly enhanced the model's ability to predict Type 2 diabetes (Talebi Moghaddam et al., 2024). Another research that dealt with diabetic datasets and feature reduction . To improve the model's performance, they used tools for dimensionality reduction, Principal Component Analysis (PCA), and feature selection, step forward and backward selection. The research found that lowering the feature count helped reduce overfitting and enhanced the models' capacity for generalization. The significance of feature selection in early diabetes prediction was also highlighted by Abdollahi and Aref (2024). it used feature selection techniques in conjunction with machine learning algorithms to increase classification accuracy to find the most useful predictors. To get more fair and accurate prediction results in diabetes research, it is crucial to address class imbalance and feature reduction using sophisticated machine learning tactics, such as oversampling approaches and feature selection. It will be achieved through the combined efforts of these studies (Abdollahi and Aref, 2024).

A study focuses on exploring advanced machine learning strategies designed to counteract this imbalance, ensuring more equitable and accurate prediction outcomes. Moreover, feature reduction presents another daunting challenge, with diabetes datasets often comprising a vast array of variables. An abundance of features relative to the available observations complicates the modelling process, which can lead to overfitting and reduce the model's ability to generalise effectively. Addressing this issue through dimensionality reduction techniques and feature selection is essential for distilling the most informative predictors and improving model performance.

This study delves into the intricacies of utilising AI in predicting and managing diabetes, underscoring the criticality of overcoming the data challenges. As diabetes continues to burgeon as a global health crisis, the need for advanced techniques to accurately predict and effectively manage this condition has never been more pressing. Artificial intelligence, with its suite of machine learning algorithms, presents a formidable tool in this endeavour. These algorithms' proficiency in processing large datasets and detecting patterns and correlations is essential in predicting the prevalence of diabetes and shaping treatment approaches (Almutairi & Abbod 2023; Chang et al. 2022).

Machine learning's role extends beyond mere prediction; it encompasses the formulation of personalised treatment plans and the continuous monitoring of glucose levels, significantly augmenting diabetes care (Dinesh & Prabha 2021). Therefore, the integration of AI in diabetes management is not just an option but a necessity as we strive to contend with the escalating prevalence of this disease. This chapter focuses on the novel approaches and methodologies employed to navigate the obstacles of missing data, imbalanced datasets, and feature reduction , thereby optimising AI's application in enhancing healthcare outcomes for diabetic patients.

The transformative potential of artificial intelligence in diabetes care is immense. However, realising this potential necessitates overcoming significant data-related challenges. Through innovative machine learning solutions, this study seeks to contribute to the field of diabetes management, offering new insights and strategies to enhance the performance and efficacy of machine learning applications. As such, the

unique capacity of artificial intelligence in forecasting and managing diabetes is poised to become increasingly pivotal, heralding a new era in healthcare that promises improved outcomes for individuals living with diabetes.

Handling enormous chunks of missing data is a crucial aspect of data analysis, particularly in medical research, where the accuracy of predictions can significantly impact outcomes. There is a variety of issues related to missing data, such as inadequate data collection, privacy constraints, or errors in data entry, and handling these gaps is essential for reliable analysis. Researchers have developed several methods to manage missing data, ensuring that the predictive models remain accurate and trustworthy. For diabetes prediction specifically, the integrity of the data is paramount, as the condition's complexity requires precise and comprehensive datasets for effective prediction and management. Techniques such as imputation, where missing data points are substituted with estimates derived from existing observations, are commonly employed to mitigate the impact of data gaps. This approach helps maintain the consistency of the dataset, allowing for more accurate predictions and insights into diabetes management and research (Bania & Halder 2021).

In medical data analysis, imbalanced datasets can cause skewed predictions and inaccurate results, especially in diabetes research. A model risks becoming biased when data from a single class predominates. To tackle this, researchers utilize methods like oversampling and undersampling, along with synthetic data synthesis, to compensate for dataset imbalances and improve the accuracy of the models. These methods are vital for enhancing diabetes prediction models, providing a stronger basis for trustworthy medical research and decisions. Datasets used for diabetes prediction sometimes suffer from excessive dimensionality because there are often more variables than observations. Consequently, the model's predictive power could be reduced due to overfitting, a condition in which the model becomes overly complex and struggles to adapt to new data. Feature selection and extraction are two methods for reducing the dataset's dimensionality that can help manage feature reduction (Jabbar and Washington, 2024).

These methods isolate and retain crucial information while eliminating extraneous or irrelevant variables. Researchers have found that by streamlining the data format, they may improve the model's accuracy and efficiency, which in turn makes it more practical. Analyses and forecasts can be weakened by missing data. To address missing values, imputation procedures are used to maintain data integrity and analytical validity. These approaches include mean substitution, constant filling, and multiple imputation. In conclusion, better diabetes prediction and improved patient outcomes depend on researchers using suitable data management strategies to resolve concerns, including class imbalance, missing data, and excessive dimensionality (Kaliappan et al., 2024). The following questions need to be answered in this chapter:

1. How does the application of advanced data imputation methods for missing values in diabetes dataset affect the performance of machine learning models?

2. How does data imbalance in diabetes dataset affect the performance of machine learning models in terms of accuracy, precision, recall, and F1 score?

3. How does the reduction of dataset dimensionality using feature selection methods impact the performance of machine learning algorithms in a medical context?

4. What is the relationship between the quality of datasets and the accuracy of predictive models in healthcare?

2.2 CLASSIFICATION AND PREDICTION

Classification is a kind of supervised machine learning that uses previously trained models to assign unknown input to predefined categories. Here, input characteristics and related output labels make up each training example. By seeing trends in the training data, the model learns to convert inputs into outputs, improving its feature-based instance classification ability (Phongying and Hiriote, 2023).

Classification is an essential process in various fields, particularly in decision-making processes where categorical outputs are required. In machine learning, classification algorithms such as Decision Trees, SVM, Neural Networks, Logistic Regression, and k-NN are commonly used to address different classification problems (Han, Pei, & Kamber 2011). These algorithms each have a unique way of separating data points into classes according to the existing features.

The classification process involves leveraging deep learning models, which can process vast amounts of data and identify intricate patterns that traditional machine learning models may struggle to detect. Techniques like Convolutional Neural Networks (CNNs) for image classification and Recurrent Neural Networks (RNNs) for sequential data classification represent the cutting-edge advancements in this domain (LeCun, Bengio, & Hinton 2015).

In the context of healthcare, classification algorithms have been transformative. They are used to diagnose or identify diseases, accurately predict patient outcomes, and personalise treatment. For example, in medical imaging, classification algorithms can help in identifying and categorising various medical conditions based on imaging data (Erickson et al. 2017). Machine learning models, built on historical medical data can evaluate patient risk for specific conditions, thus aiding in preventive care and resource allocation.

Specifically, in the field of diabetes prediction, classification plays a pivotal role. Diabetes can be predicted by analysing various patient data points, including demographics, body mass index, family history, and laboratory test results. Machine learning models can categorise people based on their likelihood of developing diabetes, such as into high-risk or low-risk groups (Kavakiotis et al. 2017). These classifications can then inform healthcare providers and patients about necessary preventive measures and treatments.

Identifying and categorising the risk of diabetes early is vital for managing the disease effectively and reducing associated health complications. Hence, machine learning classification algorithms allow for the analysis of intricate and diverse data,

resulting in more precise predictions and improved health outcomes. As machine learning technology continues to advance, these algorithms are anticipated to become increasingly important in disease prediction and management, particularly in diabetes care.

Classification in machine learning is a foundational method with broad applications across various domains, including healthcare. It involves teaching a model to categorise data into defined classes, which can then be used for decision-making and predictions. In healthcare, and specifically in diabetes prediction, classification algorithms facilitate the early identification and diabetes management, strengthening treatment and patient care.

The classification and prediction of diabetes have accumulated notable attention in the past decade due to the increasing prominence and impact of the disease (Geman et al. 2017; Sisodia & Sisodia 2018; Guldogan et al. 2020). Scholars have scrutinised different approaches to create effective frameworks that can assist in diagnosing and predicting diabetes.

Geman, Chiuchisan, and Toderean (2017) applied an adaptive neuro-fuzzy inference framework for diabetes prediction and classification. Their study showcased the potential of this approach in accurately categorising diabetic and non-diabetic patients using relevant features. The results demonstrated the effectiveness of this system in diabetes management and prediction.

Sisodia (2018) explored diabetes detection using classification algorithms. They compared the performance of multiple classification algorithms and analysed their recall, precision, accuracy, and F1-score. These measurements shed light on the capabilities of various algorithms and highlighted their potential for diabetes prediction.

A diabetes mellitus classification algorithm, DMP MI, was created by (Wang et al. 2019) to tackle the problems of imbalanced and incomplete data, which are prevalent in diabetes datasets. The technique incorporates ADASYN to deal with class imbalance, Naïve Bayes for data normalization to manage missing values, and a

Random Forest classifier for prediction. Improving classification performance, DMP MI proved resilient in handling data abnormalities when tested on the Pima Indians diabetes dataset ((Wang et al. 2019) et al. 2019).

Guldogan et al. (2020) evaluated the performance of multiple artificial neural network frameworks in diabetes classification. They compared various neural network architectures and assessed their accuracy, robustness, and efficiency in diabetes classification. This research examined the applicability of several frameworks for diabetes classification tasks.

The studies mentioned above highlight the diverse approaches and techniques employed in diabetes classification and prediction. From adaptive neuro-fuzzy inference systems to classification algorithms and artificial neural networks, researchers have experimented on multiple methodologies to strengthen the performance and effectiveness of diabetes diagnosis and forecast.

Classifying ideas into main points and supporting details, the act of classification brings clarity and coherence to written works (Geman et al. 2017; Sisodia 2018). Classification helps present information in a logical and systematic manner aiding readers in navigating through the text and making it more accessible to comprehend and retain the content. By grouping similar concepts or ideas together, classification allows for easier identification of relationships and connections between different components of the text. This organisation facilitates a smooth flow of information, guiding readers through the writer's thoughts and arguments.

Furthermore, classification promotes effective communication in written works. It helps authors convey their ideas clearly and allows readers to grasp the main concepts easily. When information is organised and classified appropriately, readers can follow the logical progression of the content, leading to a more coherent and enjoyable reading experience. Without proper classification, a piece of writing may appear disjointed and confusing, making it challenging for readers to decipher the intended message.

Classification is a crucial process in both machine learning, particularly in the domain of diabetes classification and prediction, and in writing. In machine learning, classification algorithms enable accurate categorisation of data, leading to improved diagnosis and prediction of diseases such as diabetes. In written works, classification brings structure and coherence, aiding readers in understanding and engaging with the content. It facilitates effective communication, knowledge organisation, and the advancement of research. Recognising the importance of classification in these domains in crucial for enhancing both analytical and communicative effectiveness.

2.3 CLASSIFICATION METHODS

Classification is an integral part of machine learning, offering an array of techniques for categorising and predicting data. These methods are designed to understand and predict categories based on observed data. This section delves into several prominent classification methods: k-NN, SVM, Logistic Regression, and Naïve Bayes Classifier, exploring their fundamentals, applications, strengths, and limitations.

2.3.1 k-Nearest Neighbours (k-NN)

The k-NN algorithm is a non-parametric method utilised extensively in regression and classification process. It relies on the concept that identical data points tend to be in the same class. This simplicity, coupled with its high accuracy in many cases, makes k-NN a popular choice in various fields including disease diagnosis in healthcare, credit scoring in finance, and pattern recognition in technology (Alfeilat et al. 2019; García-Pedrajas et al. 2017).

The k-NN algorithm has several notable strengths including its hight flexibility, as it is applicable for both regression and classification process. The k-NN is also straightforward to implement and understand, ensuring a wide range of accessibility. In addition, as a non-parametric method, it lacks assumptions about the underpinning distribution of data, which adds to its versatility.

However, one of the limitations of the k-NN algorithm is that it quickly becomes computationally intensive as the dataset grows, due to the need to calculate the distance

between each query point and all data points. Apart from that, k-NN is sensitive to irrelevant features, which can degrade performance and may necessitate feature selection or dimensionality reduction. The algorithm's outcome is also influenced by the scale of the data, meaning that normalisation or standardisation of input features is often required.

2.3.2 Support Vector Machine (SVM)

SVM is a robust classification approach that identifies the hyperplane that best separates different classes in the feature space. It is particularly efficacious in high-dimensional spaces, greatly suited for processes like image classification, text categorisation, and biological data classification (Sezer & Başeğmez 2021).

SVM performs well in high-dimensional spaces, even when the features outnumber the samples. The so-called 'kernel trick' enables SVM to model non-linear relationships, allowing it to detect and preserve complex relationships inherent in the datasets. Apart from that, SVM emphasises on the data points closest to the decision boundary, known as support vectors, which makes the model robust to outliers.

Besides, the framework's performance is highly dependent to the selection of the appropriate kernel and regularisation parameters, which can be challenging. In addition, training an SVM can be computationally critical, particularly with enormous datasets. Besides, the decision function of SVM, especially when using non-linear kernels, can be difficult to decipher compared to more transparent frameworks like decision trees.

2.3.3 Logistic Regression:

Logistic Regression is primarily utilised for binary categorisation processes, such as spam detection, customer churn prediction, and medical diagnostics. It predicts the likelihood of a binary outcome conditional on singular or multiple independent variables (Kirasich et al. 2018).

The advantage of logistic regression is that it provides a probabilistic interpretation by outputting a probability score for observations, which offers a clear thresholding mechanism. The method is also efficient and scalable, being computationally less intensive, which makes it suitable for large datasets. In term of interpretability, the coefficients of the logistic regression model quantify the influence of each feature on the odds of the positive class.

However, one of the disadvantages of logistic regression is regarding the assumption of linear relationship between the independent variables and the log odds of the dependent variables. Such presumption is not necessarily true. Aside from that, the model needs to be modified for multi-class problems, often through strategies like one-vs-rest (OvR).

2.3.4 Naïve Bayes Classifier

Naïve Bayes Classifier relies on the assumption of independence among predictors. It's commonly utilised in text classification tasks, such as spam filtering and sentiment analysis (Salmi & Rustam 2019).

Naïve Bayes is highly efficient as it requires a tiny amount of training data to forecast the parameters for classification. It also performs well in multi-class prediction scenarios. In addition, Naïve Bayes can address categorical and continuous data, although it assumes a normal distribution for continuous variables.

However, some limitations of Naïve Bayes theorem lie with the assumption of independence among the predictors which is often unrealistic in real-world applications and might affect its performance. Additionally, data scarcity can lead to the probability of zero occurring with categorical variables if a category in the test set was not present in the training set, a problem often mitigated by techniques like Laplace estimation.

In summary, each classification method has its unique advantages and disadvantages. The selection of techniques relies on the specific requirements of the task, including the nature of the data, the size of the dataset, the complexity of the

framework needed, and the computational resources available. Selecting the appropriate classification technique is crucial for accomplishing an optimal machine learning framework's performance.

2.4 DATA QUALITY

Data quality is a fundamental aspect that significantly impacts the utility and effectiveness of diabetes registries. This section reviews various studies focusing on data quality issues, employing a critical analysis to understand different methodologies, research findings, and their implications. The aim is to develop a coherent argument supported by evidence that underscores the necessity of high data quality standards.

Aktaa et al. (2022), emphasized the importance of quality indicators and cross-boundary clinical registries in enhancing cardiovascular care and outcomes, demonstrating the value of standardizing data for healthcare improvement.

Gewandter et al. (2020) provided guidelines to improve data quality in clinical trials for chronic pain treatments. Bicevskis et al. (2019) explored the financial impact of low-quality data, emphasizing the need for robust data management practices.

Mashoufi et al. (2023) conducted a scoping literature review on the importance of data quality in healthcare, highlighting its role in improving patient care and decision-making. Quindroit et al. (2023) developed a practical taxonomy for identifying and addressing data quality problems in healthcare databases, providing tools for better data management and research accuracy. These studies collectively emphasize the importance of maintaining high data quality across various healthcare contexts., as shown in Table 2.1.

Table 2.1 Data Quality in Diabetes Registries

Ref.	Title	Problem Statements	Method/Tool	Findings
Mashoufi et al. (2023)	Data Quality in Health Care: Main Concepts and Assessment Methodologies	Importance of data quality in healthcare	Scoping literature review approach	Data quality is crucial for patient care, healthcare service improvement, and better decision-making.
Quindroit et al. (2023)	Definition of a Practical Taxonomy for Referencing Data Quality Problems in Health Care Databases	Data quality problems in healthcare databases	Bottom-up approach	Practical taxonomy with associated quality assessment and data cleaning methods can aid in accurate data reuse for research and clinical purposes
Aktaa,(2022)	Improving cardiovascular care and outcomes: cross-boundary clinical registries and quality indicators	Uncertain data quality in the Danish National Diabetes Register	Evaluate data completeness and consistency	Found the registry to have good data quality and internal consistency.
Dakka et al. (2021)	Automated detection of poor-quality data: case studies in healthcare	Removal of poor-quality data	Untrainable Data Cleansing (UDC)	Effective automated data cleansing using UDC can help identify complex clinical cases for further evaluation.
Gewandter, et al. (2020)	Improving study conduct and data quality in clinical trials of chronic pain treatments: IMMPACT recommendations	Issues in study conduct and data quality in clinical trials for chronic pain treatments.	IMMPACT recommendations and guidelines.	Recommendations provided to enhance the conduct and data quality in chronic pain treatment trials.
Bicevskis et al. (2019)	A step towards a data quality theory	Data quality	Data object, domain-specific language	Low-quality data can lead to significant financial consequences for companies in the UK and other developed countries.

2.4.1 The Significance of Machine Learning in Diabetes Management with Emphasis on Data Quality

Effective integration of machine learning into diabetes management relies on the availability of high-quality data. Predictive modeling, individualized treatment recommendations, and continuous patient monitoring each benefit from datasets that are complete, representative, and reliable. In contrast, poor-quality data can undermine the utility of these models and compromise clinical decision-making. Machine learning techniques can identify risk factors, forecast disease progression, and guide interventions tailored to each patient's condition. However, the accuracy and applicability of these predictions depend directly on the integrity of the underlying data. Missing values, imbalanced class distributions, and high-dimensional datasets introduce challenges that may lead to flawed analyses and biased outcomes. Incomplete patient histories, underrepresented population groups, and inconsistent data recording practices can distort risk assessments and obscure meaningful insights.

Addressing these issues necessitates rigorous data preprocessing. Effective strategies include imputation methods for handling missing values, rebalancing techniques to mitigate skewed class distributions, and feature selection approaches that refine complex datasets into interpretable, manageable subsets. Improving data collection protocols, standardizing data formats, and implementing regular data quality audits further strengthen the foundation upon which machine learning models are constructed.

A focus on data quality is essential to unlocking the full potential of machine learning within diabetes management. By ensuring that datasets are accurate, comprehensive, and well-structured, it becomes possible to derive clinically actionable insights. Such efforts support more reliable risk prediction, enhanced treatment personalization, and better patient outcomes. (Islam et al. 2023; Rou 2022; Sowjanya & Mrudula 2023).

2.4.2 Handling Missing Data

Missing data in statistical analysis refers to an occurrence where data points in an observation are not recorded and this significantly impacts the interpretation and

validity of research outcomes. This absence can arise from various sources, such as non-response in surveys, attrition in longitudinal studies, equipment malfunction, or data entry errors, leading to incomplete datasets. The significant implications of missing data include biased results, the reduction of statistical power, and higher likelihood of drawing incorrect conclusions from the analysis.

Missing data can affect data quality in several ways. When data is missing completely at random (MCAR), the missingness is unrelated with the dataset values, whether missing or observed. This scenario is the least problematic but can still lead to reduced sample size and loss of information. More commonly, data may be missing at random (MAR), where the inclination for a data point to be absent is tied to some observed data in the dataset but not the missing data itself. The most challenging scenario is when data is missing not at random (MNAR), indicating that the missingness is related to the unobserved data, leading to systematic differences between the missing and observed values (Razavi-Far et al. 2020).

The significant effects of missing data on data quality and research outcomes necessitate careful consideration in statistical analysis. Missing data can give rise to biased parameter forecast and weakened validity of the statistical inferences if not adequately addressed. For example, in clinical trials, if patients experiencing adverse effects are more likely to drop out and their data is not accounted for, the treatment may appear more effective than it is. Therefore, understanding the instrument and the degree of missingness is imperative in examining the appropriate method for handling such data, whether through deletion, imputation, or model-based approaches, to mitigate the potential biases and inaccuracies in the analysis (Nugroho et al. 2021).

Missing data is a pervasive issue in statistical analysis that can undermine the quality of the research findings. Thus, addressing missing data appropriately is essential to ensure accurate, reliable, and valid results. Understanding the types of missing data, their sources, and their impact on the analysis is critical for selecting the most suitable method to handle the missingness and preserve the integrity of the research outcomes (ÜresİN 2021).

2.4.3 Types of Missing Data

In the realm of statistical analysis, missing data is categorised into three distinct types, each with specific implications for data handling and analysis. These categories are Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR)(Camino et al. 2019).

MCAR refers to a scenario where the chance of data being missing is the equal across all observations, regardless of the value of any variables. In this case, the missingness of data has nothing to do with the observed or unobserved data. This means that the missing data is a random subset of the data and does not introduce systematic bias into the analysis. However, even with MCAR, the loss of data can lead to a reduction in statistical power (Misztal 2018).

MAR occurs when the possibility of data being missing is related to the observed data but not the missing data itself. MAR allows for the possibility that the missingness can be fully accounted for by variables for which there is complete information, meaning that the missing data mechanism can be modelled using the observed data. MAR does not imply that the missingness is random or that it occurs by chance, but rather that the cause of the missingness is external to the data of interest (Misztal 2018).

MNAR the most challenging category to address, occurs when the missingness depends on the unobserved data. In other words, the reason that data is missing is related to the actual missing data. This type of missing data can lead to biased estimates in statistical analysis if not properly accounted for, as the missingness is systematically different for different values of the data. Handling MNAR data often requires incorporating external information into the analysis or making assumptions about the mechanism of missingness (Pratama et al. 2016).

Comprehension of the type of missing data in a dataset is imperative for selecting the appropriate method for handling missing data and for conducting a valid and reliable statistical analysis. The appropriate methods for dealing with missing data

is conditional upon whether the data is MCAR, MAR, or MNAR, and this decision significantly influences the integrity and accuracy of the research findings.

2.4.4 Types of Imputation Methods

Imputation methods or techniques can be generally divided into two categories: single imputation and multiple imputation. In single imputation, missing data are replaced with an estimate per missing entry. This approach is straightforward but can cause underestimation of data variability and potential bias in the analysis. Common single imputation methods include mean/median imputation, mode imputation, regression imputation and multiple imputation.

For mean/median imputation, missing data are replaced with the mean or median of the observed data for that variable. Despite the simplicity, this method can alter the true variance of the data and may not be appropriate if the data is not MCAR (Schafer & Graham 2002). Secondly, mode imputation is widely utilised for missing categorical data in which the most frequents category or mode is used to close the gaps. Despite its simplicity, this method may not capture the true distribution of data. Thirdly, regression imputation utilises a regression model built on observed data to estimate the missing data. This method can preserve relationships between variables but may introduce bias if the model's assumptions are not met (Van Buuren 2018).

Multiple imputation involves creating several different complete datasets by filling in missing data with estimates that account for the uncertainty around the missing data. The findings from these datasets are then combined to generate estimates that reflect this uncertainty (Rubin 1987). This method is considered more robust than single imputation, especially when the data is not MCAR, as it better handles the variability and uncertainty inherent in the imputation process.

2.4.5 Advanced Imputation Techniques

Advanced imputation techniques usually leverage machine learning algorithms to predict missing data, considering the complex relationships between variables. K-

Nearest Neighbours (K-NN) imputation involves imputation of missing data based on the k-nearest neighbours, where k is a user-defined parameter. The missing data is estimated utilising the mean or median of the nearest neighbours identified in the multidimensional space. K-NN imputation is particularly useful for datasets with nonlinear relationships among variables (Saadatfar et al. 2020). Expectation Maximization (EM) imputation is a two-step iterative algorithm that estimates the maximum likelihood by alternately applying expectation (E) and maximization (M) steps. In the E step, missing data are imputed based on estimated parameters, while in the M step, parameters are re-estimated based on the imputed data. This process continues until convergence (Sanwouo, 2024 ; Peng et al. 2021). As for the machine learning-based imputation, various techniques such as decision trees, random forests, and neural networks can be used for imputation, leveraging their ability to model complex patterns and relationships in the data. For example, Random Forest imputation can handle nonlinearities and interactions effectively, providing a flexible approach to dealing with missing data (Liang et al. 2019).

There are several issues with missing data in prediction research. Table 2.2 shows the state-of-the-art in medical prediction scenario for handling missing data. Previous researchers have worked on improving the analytics processing (Mozaffar et al. 2022).

Table 2.2 Medical Prediction for Handling Missing data

Ref.	Problem Statements	Method/Tool	Finding
Sanwouo et al. (2024)	Challenges in imputing missing values in univariate time series data due to device failures or data collection errors.	TS-Pothole: A method leveraging cyclic pattern analysis and a recursive strategy for imputing missing values in univariate time series.	TS-Pothole outperforms state-of-the-art methods like GANs and autoencoders, providing more accurate (up to 1.5 times) and faster (up to 2 times) imputations, even as the proportion of missing data
Nijman et al. (2022)	Prediction model studies need enhanced techniques for addressing missing data to improve their accuracy and reliability.	Systematic literature review of clinical prediction models using machine learning.	A significant portion of the studies did not report on missing data. Deletion, especially through complete-case analysis, was the most common method used for handling missing data, with few studies using multiple imputation or built-in mechanisms.
Razavi-Far et al. (2020)	Machine learning studies in medicine often inadequately report and handle missing data, leading to potential biases.	Kemi and kemi+ CCA analyses only complete data; multiple imputation replaces missing data; surrogate splits handle missing data in machine learning.	Kemi+ is an improved version of Kemi, identifying superior k values that result in minimal imputation error using EMI.
Khan et al. (2021)	Missing data leads to a significant decrease in data quality and needs effective handling.	Fuzzy C-Mean clustering-based missing data imputation.	The proposed method, using shorter intervals for imputation based on cluster membership, is as effective as, or better than existing state-of-the-art methods on UCI datasets.
(Wang et al. 2019)	The presence of missing data and imbalance data in diabetes data leads to low classification performance.	Naïve Bayes (NB) method for data normalisation to compensate missing data. Adaptive synthetic sampling method (ADASYN) for class imbalance. Random forest (RF) classifier for prediction.	The proposed DMP_MI algorithm improves diabetes classification performance on the Pima Indians diabetes dataset.

to be continued…

… continuation

(Sefidian & Daneshpour 2020)	Ten correlation-based imputation methods that maximise the correlation between a missing feature and other features by selecting data segments with strong correlations.	The proposed methods outperform seven other imputation approaches in most cases, according to three evaluation criteria.	The proposed methods outperform seven other imputation approaches in most cases, according to three evaluation criteria.
(ÜresİN 2021)	An algorithm specifically designed for small datasets to predict missing data (CBRI).	The proposed CBRI method outperformed traditional methods like arithmetic mean assignment and median value assignment in predicting missing data, demonstrated by testing with real data from automotive manufacturing. The method showed lower standard error in predictions.	The proposed CBRI method outperformed traditional methods like arithmetic mean assignment and median value assignment in predicting missing data, demonstrated by testing with real data from automotive manufacturing. The method showed lower standard error in predictions.
(Nugroho et al. 2021)	Class center-based adaptive approach model (C3-FA) using Firefly Algorithm (FA) considering attribute correlation in the imputation process.	The class center-based firefly algorithm (FA) efficiently estimates actual values in handling missing data, indicated by Pearson correlation coefficient (r) close to 1 and root mean squared error (RMSE) near 0. The method maintains the true distribution of data values, as shown by the Kolmogorov–Smirnov test (DKS values close to 0) and produces good accuracy in evaluation with three classifiers.	The class center-based firefly algorithm (FA) efficiently estimates actual values in handling missing data, indicated by Pearson correlation coefficient (r) close to 1 and root mean squared error (RMSE) near 0. This method maintains the true distribution of data values, as shown by the Kolmogorov–Smirnov test (DKS values close to 0) and produces good accuracy in evaluation with three classifiers.

2.4.6 Data in Predictive Modelling

Understanding the role of data in predictive modelling is essential. Data in predictive modelling refers to the information utilised to generate frameworks that can predict future outcomes based on historical patterns. Quality data is crucial for developing reliable and accurate predictive models (Shamsuddin et al. 2022). In predictive modelling, data serves as the foundation for constructing algorithms that foresee future events or outcomes. These predictions are grounded in historical data patterns, with the premise that past behaviours and trends can provide insights into future occurrences. Therefore, the integrity and accuracy of data are paramount in predictive modelling. High-quality data ensures that the predictive models built are both reliable and precise, enabling better decision-making and forecasting. The quality of data directly influences the model's performance; data with poor quality can derail the prediction, and high-quality data can significantly enhance the model's predictive power (Shamsuddin et al. 2022). The concept of missing data refers to instances when no value is stored for a variable in an observation within a dataset. This absence can happen for various reasons, significantly impacting the analysis and interpretation of data. Common sources of missing data include non-response in surveys, where individuals may neglect to answer some questions, leading to missing entries in the survey data. In addition, missing data occurs due to errors in data collection such as improper handling of data recording devices or misreporting. Missing data may also occur during the data processing stage. For example, during the cleaning, entry, or coding of data, values may be accidentally lost or omitted. Apart from that, attrition in studies can also result in missing data. In longitudinal studies, participants may drop out, leading to incomplete data across different time points. Investigating the nature of missing data is imperative for selecting the appropriate methods to handle it effectively.

The pattern and mechanism of missingness, whether the data is Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR), greatly influences the selection of statistical methods to overcome missing data. Ignoring or improperly handling missing data can give rise to biased estimates, diminished statistical power, and potentially erroneous conclusions (Little & Rubin 2019). In predictive modelling, the handling of missing data is essential to

maintain the accuracy of the models. Therefore, recognising and addressing the issue of missing data is a fundamental step in the preprocessing of data for analysis and modelling. Missing data can significantly affect the accuracy and bias of predictive models. It can lead to biased estimates, underestimation of the variability, and may ultimately result in misleading inferences (Schafer & Graham 2002). The consequences of missing data on predictive modelling are a critical issue that can significantly undermine the validity of statistical inferences and predictions. When data points are missing from a dataset, the resultant gaps can distort the analysis in several ways.

Missing data leads to biases in the forecast produced by the model. If the missing data mechanism is not entirely random, the analysis may overemphasise the characteristics of the complete cases, leading to skewed results (Schafer & Graham 2002). In addition, missing entries can result in an underestimation of the variability within the dataset. This underestimation occurs because standard statistical methods may interpret the absence of data as a lack of variation, thus falsely narrowing confidence intervals and potentially leading to overconfident predictions.

The combination of biased estimates and underestimated variability can culminate in misleading conclusions about the variables and their relationships in the predictive model. This distortion can affect the model's predictive, resulting in erroneous conclusions and decisions based on the framework's output.

In summary, the presence of missing data in predictive modelling can have profound effects on the analysis, potentially leading to biased estimates, underestimated variability, and ultimately, misleading inferences. This underscores the importance of employing the correct methods for handling missing data to preserve the integrity of the modelling process (Schafer & Graham 2002)

2.4.7 Objective: The Fundamental Point of Current Study

In evaluating AI-based solutions for diabetes management, particularly with imbalanced datasets, traditional accuracy metrics are often inadequate, as they may obscure performance disparities. To address this, I propose using a set of evaluation

metrics tailored to reflect the model's practical applicability in diabetes detection. Precision measures the accuracy of positive predictions, ensuring false positives are minimized, which is critical to avoid unnecessary patient anxiety and additional healthcare costs. Recall (sensitivity) focuses on the model's ability to correctly identify diabetic patients, which is vital for early detection and timely intervention (Araújo et al, 2025). The F1-Score balances precision and recall into one metric, ensuring neither is sacrificed, making it particularly useful in imbalanced datasets (Messelu et al., 2025). The AUC-ROC evaluates the model's robustness by measuring its ability to distinguish between diabetic and non-diabetic patients across varying thresholds (Reátegui-Rivera et al., 2025). These metrics, when combined, provide a comprehensive and accurate reflection of model performance, particularly in scenarios involving class imbalance, which is essential for evaluating predictive models in diabetes-related clinical applications (Emi-Johnson & Nkrumah, 2025). These metrics ensure the model accurately identifies diabetic patients, which is vital for effective diabetes management.

2.4.8 Conventional Imputation Techniques

Recent studies have focused on leveraging machine learning techniques to generate predictive models for assessing the risk of diabetic complications, particularly neuropathy . These complications significantly impact the quality of life for individuals with diabetes and pose substantial challenges for healthcare providers in managing the disease effectively. By harnessing large datasets encompassing various clinical parameters including demographic information, glycaemic control metrics, and relevant biomarkers, researchers aim to build robust predictive frameworks with the ability to detect individuals at heightened risk of acquiring these complications. Such models hold immense potential for facilitating early intervention strategies, personalised treatment approaches, and improving patient outcomes in diabetes management.

Solution: The organisation implements multiple imputation techniques to address the missing data issue:

1. Mean/Median/Mode Imputation: Initially, simple techniques like mean, median, and mode imputation are employed to fill in missing data with the average,

median, or most frequent value of the respective feature. However, these methods may not preserve the underlying relationships between variables and can introduce bias, particularly when the missing data is not random.

2. K-Nearest Neighbour Imputation: A more sophisticated approach called K-nearest Neighbour (KNN) imputation is then utilised. This method identifies the "k" most similar patients in the dataset based on available data points and uses the average of their corresponding values to impute the missing information for the target patient. KNN imputation is more robust than simple imputation methods as it considers the context of the missing data and leverages the information from similar patients.

3. Multiple Imputation: Subsequently, a robust technique called multiple imputation is implemented. This method creates multiple plausible datasets by imputing missing data using various strategies, such as KNN or more advanced statistical methods. The final prediction is acquired by averaging the predictions from each imputed dataset, enhancing the model's generalisability and robustness. This approach acknowledges the uncertainty associated with missing data and provides an extensive understanding of the potential range of predictions (Nijman et al. 2022).

In summary, research on missing and imbalanced data in machine learning highlights a wide range of challenges and advancements. Despite these developments, most studies overwhelmingly underscore the urgent need for robust, effective, and transparent methods for imputation and the management of missing data. This need directly affects the performance of machine learning frameworks and will become increasingly important in various application domains, particularly healthcare. Idris et al. (2024) presented a new method called Stacking Recursive Feature Elimination-Isolation Forests to improve classification performance. A simple method reduces feature complexity and removes outliers. They surpassed other current methods with their solution, which attained an accuracy of 79.08% on the PIMA Indians dataset and 97.45% on a Diabetes Prediction dataset (Idris et al. 2024). Traditional monitoring methods aren't very sensitive for detecting type 2 diabetes mellitus (T2DM), the most

common type of diabetes. To enhance prediction performance, this study combined four machine learning models with clinical and nutritional data from 8,981 individuals. By achieving an accuracy of 90%, the model demonstrated improved stability and accuracy in the early diagnosis of type 2 diabetes (Fu. Yong et al. 2024).

The studies reviewed look at different methods for handling missing data in various fields, such as healthcare and epidemiological research. It used Multiple Imputation to address missing data in diabetes datasets and reported a prediction accuracy of 85.9%. This shows that MICE work well for this type of data.

This high accuracy is important because ICU data is often complicated and critical, making it essential to get the imputation right. **Chen et al. (2024)** also used MICE but focused on epidemiological studies, where they reached an accuracy of 85.4%. This consistency with other studies using MICE suggests that it's a reliable method for handling missing data in large-scale research. Finally, it used k-Nearest Neighbors for health survey data and reported a slightly lower accuracy of 81.8%. While k-NN is easier to implement, its lower accuracy shows that it may not be as powerful as other methods like Random Forest or MICE in more diverse datasets. In summary, these studies demonstrate that the choice of imputation method can significantly affect the accuracy of the data, with more complex methods like MICE and Random Forest generally performing better, as shown in Table 2.3.

Table 2.3 Technique Imputation, Prediction Accuracy (%)

Reference	Data	Problem Statements	Methodology	Prediction Accuracy (%)
Idris et al. (2024)	PIMA Indians, Diabetes Prediction	High complexity in stacking and presence of outliers reduce model performance	Stacking + Recursive Feature Elimination + Isolation Forest	79.08
Fu. Yong et al. (2024)	Clinical & dietary data (8,981 patients)	Early T2DM detection is challenging due to low specificity of conventional tools	Stacking ensemble of four ML models	90
Chen et al. (2024)	Imbalanced Datasets (Various Domains)	Imbalanced learning problem in classification tasks leading to biased predictions	Resampling methods (SMOTE, Tomek links, Under-sampling), Ensemble Learning	85.4

2.5 MEDICAL PREDICTION FOR HANDLING IMBALANCED DATASET

Medical prediction for handling imbalanced dataset refers to applying predictive modelling techniques to analyse and make predictions on diabetes dataset that suffer from imbalanced dataset distributions. In such datasets, the instances of one class significantly outnumber those of other classes. This imbalance data poses challenges in accurately predicting medical conditions, as traditional machine learning algorithms are inclined to be biased towards the majority class and fail to capture patterns in the minority class.

Several obstacles hinder the effective handling of imbalanced datasets in medical prediction. One significant issue is the inclination of machine learning frameworks to become biased when trained on imbalanced data. Algorithms tend to prioritise the majority class, undermining its ability to predict the minority class. This bias can be detrimental in medical applications, as it may overlook critical patterns and contribute to incorrect predictions.

Another challenge is the limited data available for the minority class in imbalanced datasets. The insufficient representation of the minority class makes it challenging for models to learn and generalise from these examples, reducing their predictive accuracy. This limitation is especially concerning in medical prediction, where accurate predictions for rare medical conditions are essential.

Additionally, the costs associated with misclassifying instances from the minority class can be severe in medical prediction. Misclassification cost is typically higher for the minority class as it may result in delayed or missed diagnoses, leading to suboptimal patient care. Addressing this conundrum requires developing techniques that prioritise the correct identification of the minority class while maintaining overall accuracy.

Scholars have proposed multiple techniques and algorithms to address the challenges of imbalanced datasets in medical prediction. These methods aim to rebalance class distributions, improve model generalisation, and mitigate the implication of imbalance data on predictive performance.

For example, Seng et al. (2021) introduced the Neighbourhood Under sampling Stacked Ensemble (Nus-Se) technique, which combines undersampling methods with a denoising stage to mitigate the class imbalance, noise, and overlapping. This approach minimises the adverse effects of undersampling and boosts the model's generalisation capability (Seng et al. 2021).

Kumar and Mathur (2022) presented a framework that employs computationally intelligent techniques such as SMOTE, Recursive Feature Elimination (RFE), and machine learning algorithms to improve classification performance on imbalanced diabetes dataset. This framework addresses imbalanced data and provides overview into the efficacy of various methods in medical prediction.

Mirzaei et al. (2021) advocated the Clustering and Density-Based Hybrid (CDBH) approach, which combines clustering and density-based methods to handle imbalanced data classification in diabetes dataset. By leveraging K-means, density-based approaches, and Random Forest, the CDBH approach achieves high accuracy, precision, recall, and F1-score in handling imbalanced diabetes dataset (Mirzaei et al).

These studies indicate the importance of developing specialised techniques and algorithms tailored to mitigate the issues of imbalanced datasets in medical prediction.

The summary in **Table 2.4** provides an overview of the current methods used to manage unbalanced data in medical prediction scenarios. Undersampling is used in a stacked ensemble to alleviate class imbalance, class noise, and class overlapping.

Table 2.4 Overview of the current methods used to handle unbalanced data in medical prediction scenarios.

Ref.	Title	Problem Statement	Method/Tool	Finding	Performance Evaluation
Zhang et al. 2023	An ensemble oversampling method for imbalanced dataset classification with prior knowledge via generative adversarial network	Imbalanced dataset classification in real-world applications requiring new algorithms	Generative adversarial network (GAN)	G-GAN provides promising results in G-mean, AUC, F-measure, and ROC, improving classification performance on imbalanced datasets	G-mean: 0.85, AUC: 0.90, F-measure: 0.80
Almutairi & Abbod 2023	Machine Learning Methods for Diabetes Prevalence Classification in Saudi Arabia	Predicting diabetes prevalence and trends in Saudi Arabia using ML algorithms	LD, SVM, K-nearest neighbour, NPR	Weighted KNN outperforms other methods in predicting diabetes prevalence rate accuracy	Accuracy: 0.92, Sensitivity: 0.88, Specificity: 0.90
Huang et al. 2023	The Enhancement of Classification of Imbalanced Dataset for Edge Computing	Improving classification accuracy for imbalanced datasets in edge computing	LBNN algorithm, Convex Hull, Hyperplane	The proposed method outperforms the LBNN algorithm in imbalanced datasets, suitable for edge computing and real-time data classification	Precision: 0.87, Recall: 0.85, F1-score: 0.86
Chen et al. 2024	Adversarial Network-based Approach for Imbalanced Medical Data Classification	Handling imbalance data in medical data classification using adversarial network-based approaches	Adversarial network architecture	The proposed adversarial network-based approach effectively addresses imbalance data in diabetes dataset, improving classification accuracy	AUC: 0.88, F1-score: 0.82

2.5.1 Taxonomy of Literature on Imbalanced Data

Recently, scholars have attempted to propose several distinct approaches to effectively overcome the issues arising from datasets with unequal distribution of instances. In this section, a few methods called data preprocessing techniques, are elaborated, and examined. These methods may exist individually, but the combination of one or more methods may synergise and enhance the efficacy to balance the inherent problems in the datasets. The consensus from these studies is quite positive and optimistic, as they indicate that the algorithms underscoring these methods allow the modification of the ensemble learning technique without radically affecting the classifier. Such modifications to the algorithm and data preparation have been meticulously performed prior to reaching the learning stage. Additionally, there are different possibilities. illustrates the map of this taxonomy in detail.

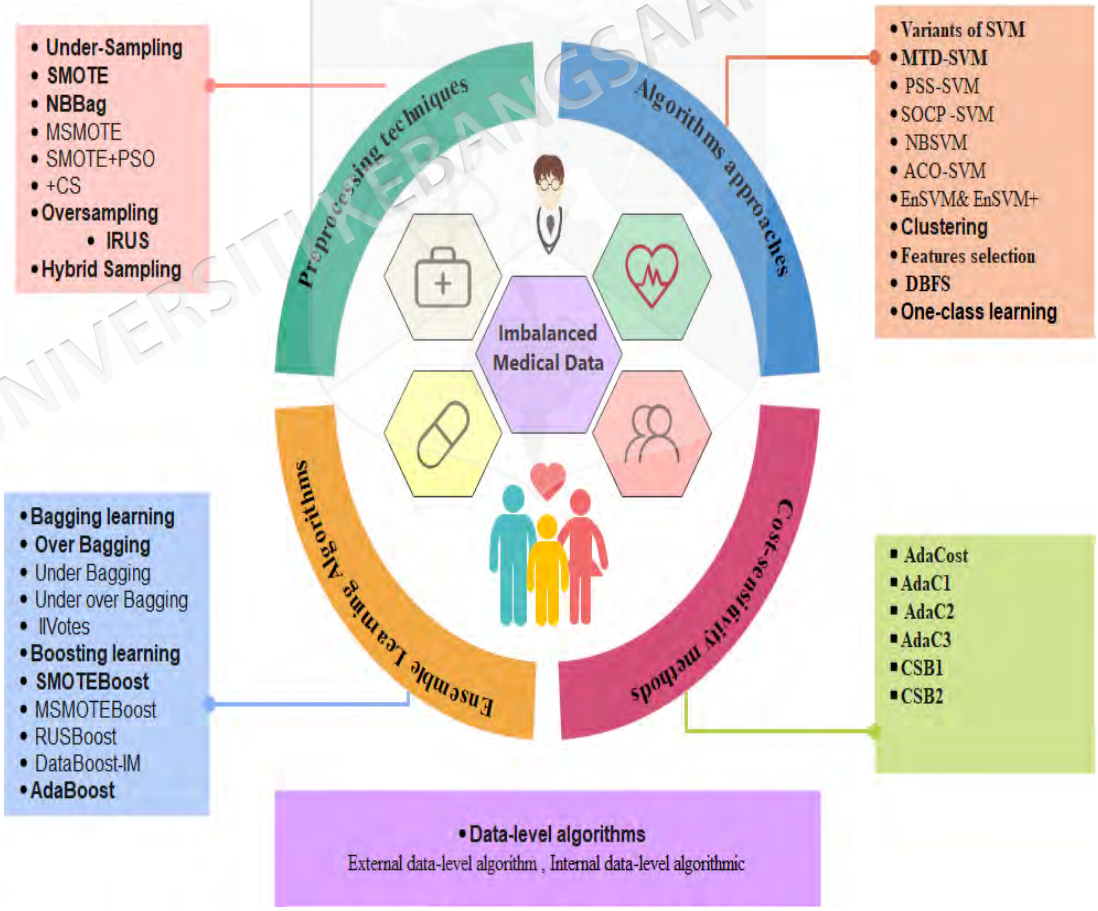


Figure 2.1 Taxonomy of Literature Studies for Imbalanced Data Solution Methods

2.5.2 Preprocessing Techniques

One of the biggest challenges for machine learning (ML) classifiers is learning from outliers and imbalanced data. In (Datta & Das 2015), a selective data preparation strategy proposed, which uses an artificially formed subset that modules can integrate into outlier cases. To achieve diversity in the training data, it employed the SMOTE. However, this was applied after oversampling the anomalies, irrespective of class. The aim was to minimise the impact of outliers on the training dataset. According to the results, selective oversampling provided better classification for SMOTE, as shown in Figure 2. There are three types of techniques for addressing imbalanced data: undersampling, oversampling, and a hybrid method.

a. Undersampling

The primary goal of undersampling methods is to increase minority class precision by decreasing the size of majority sample populations. A straightforward example of undersampling involves randomly selecting samples from the majority class and then removing them. However, classifier performance may suffer when random samples are removed from the majority class as this might lead to the loss of important information in the remaining examples. To address this, researchers have developed algorithms to carefully choose models from the majority class that do not contain relevant data.

One method utilised for undersampling is the stacking of the community undersampling ensemble. For the degradation issue, a stacked ensemble is employed as an undersampling strategy that takes into account class imbalance, class overlapping, and class noise. Instead of using undersampling as a preprocessing step, this approach incorporates the stacked ensemble itself and created a stacked undersampling ensemble. A new approach to undersampling involves a three-stage framework consisting of denoising, fuzzy K-Means clustering, and representative sample selection (Alanazi et al. 2016). This approach implements a denoising step before the clustering-based undersampling strategy to reduce the system's sensitivity to noisy data. A random sample is subsequently used to select a representative of the majority class (Mahmood et al. 2022).

Clustering-based undersampling is used in predictive modelling of healthcare-associated infections to selectively undersample specific areas within the variable space. This approach reduces the likelihood of overfitting and mitigates the severity of the negative consequences of undersampling. The clustering method is used in an n-dimensional space, where n represents the number of attributes other than the class attribute, to generate these areas. This technique is suggested for predictive diagnosis of diseases with imbalanced datasets. To address sampling discrepancies, an undersampling technique based on the overlap between the two groups is used, resulting in more representative data for the underrepresented minority in the overlapping areas. This is achieved by identifying and excluding negative class samples from the overlap area thereby increasing class separability in the data space (Lin et al. 2017).

This method implemented two undersampling techniques, with the first focusing on cluster centers and the second relying on nearest neighbours to achieve a balanced class distribution. Five different classification methods were used to evaluate the performance of the new method on various datasets. The experimental findings reported by (Khushi et al. 2021) indicated that that these methods are more accurate when comparing sampling techniques applied to imbalanced medical data. This research examined the impact of imbalance strategies on the diagnosis of patients with lung cancer using three classifiers: logistic regression, random forest, and linear SVC. The study compared ten undersampling techniques, seven oversampling methodologies, and two integrated sampling techniques. The findings suggest that oversampling techniques are more effective than undersampling and hybrid approaches (Soltanzadeh & Hashemzadeh 2021).

b. Oversampling

On the contrary, oversampling techniques endeavour to massively augment the quantity of instances in the minority group. The SMOTE algorithms are usually utilised to estimate and generate extra instances in the minority group, by employing linear interpolation on randomly chosen, identical neighbouring samples, resulting in the frameworks' classification accuracy of the minority samples (Sáez et al. 2015). Nevertheless, such augmentation of minority samples can give rise to several

disadvantages of this approach such as overlapping, noise, boundary distortion, as well as other artefacts into the datasets. Accordingly, these issues can potentially be resolved by using a set of improvements of the SMOTE algorithms. These improvements include the most recent Misclassification-Oriented Synthesised Minority Oversampling Technique (M-SMOTE) and Edited Closest Neighbour based on Random Forest (RF), and these methods have demonstrated a positive outcome when being used in tandem with the oversampling method.

M-SMOTE is often used to collect more data from underrepresented groups, with its oversampling rate being calculated based on the RF misclassification rate to maintain the characteristics of the distribution in the original datasets. Two predominant oversampling approaches called Adaptive-SMOTE and Gaussian oversampling have been examined and recommended by Xu et al. (2020). Adaptive-SMOTE, selectively synthesises new minority class samples by focusing on groups of inner and the so-called “danger" data from the minority class. In machine learning systems, data which are considered vulnerable to system hijacking, model theft, evasion attacks and data poisoning are termed "Danger" data. On the other hand, the other approach called Gaussian oversampling utilises the Gaussian distribution in tandem with dimensionality reduction approach to achieve a flatter distribution albeit with a narrower tail (Pan et al. 2020). Synthetic Minority Oversampling entails the creation or generation of additional minority class instances to balance out the data, with the application of the Orchard technique.

c. Hybrid Sampling

In basic terms, the hybrid sampling involves the combination of oversampling and undersampling methods to balance out the datasets by inflating the number of minority instances and diminishing the number of majority instances. Accordingly, hybrid methods may also combine the downsides of each approach, such as overfitting in the case of oversampling and the loss of inherent information with undersampling (Seiffert et al. 2009). Hybrid resampling is a combination of oversampling and undersampling algorithms, is proposed to solve these issues and provide reliable ways of dealing with imbalanced situations. Identified and highlighted the extreme imbalance existing in the

diabetes dataset due to the small minority data in health records. This is addressed by using a technique where an algorithm is incorporated along with hybrid sample selection, combining boosting with a heuristic and distribution-based oversampling technique called HusdoS-boost. HusdoS-boost employs undersampling to eliminate duplicate majority samples identified by previous boosting results and uses a minority class distribution to artificially generate minority samples.

The hybrid sampling algorithm combines the minority oversampling methodology to prevent misclassification with the Edited Nearest Neighbour (ENN) undersampling method. A common practice for M-SMOTE is to take a larger sample size from the underrepresented group. Consequently, the majority of samples at ENN are undersampled (Popel et al. 2018). Finally, classification prediction is performed using hybrid sampling, with the stopping threshold for iterations established according to changes in the classification index. A hybrid synthetic minority oversampling method is used with Edited Nearest Neighbour to normalize data and filter out noise.

A novel hybrid methodology is used to address the issue of imbalanced data, combining the advantages of both oversampling and undersampling. The K-Means method was used on the primary and secondary categories. When looking for representative samples from both the majorities and minorities, a local search provides significant time savings. The class distribution of a dataset, or the proportion of examples in each class, is an important factor in classifying the data. In an imbalanced dataset, one group, which is often the minority class or the one related to the concept of interest or positive class, is underrepresented. Consequently, there are fewer examples of the minority class compared to the majority class. Class overlaps, limited samples, and few disjoint examples make it challenging for classifiers to learn from datasets with skewed class distributions. Class imbalance is a significant problem faced by the data mining industry, as it frequently occurs in practical classification tasks. Approaches to resolving class imbalances can be categorised into three main types (Cui et al. 2019; ElSeddawy et al. 2022; Shen et al. 2021). Figure 2.2 illustrates the SMOTE technique used for improving diabetes prediction accuracy.

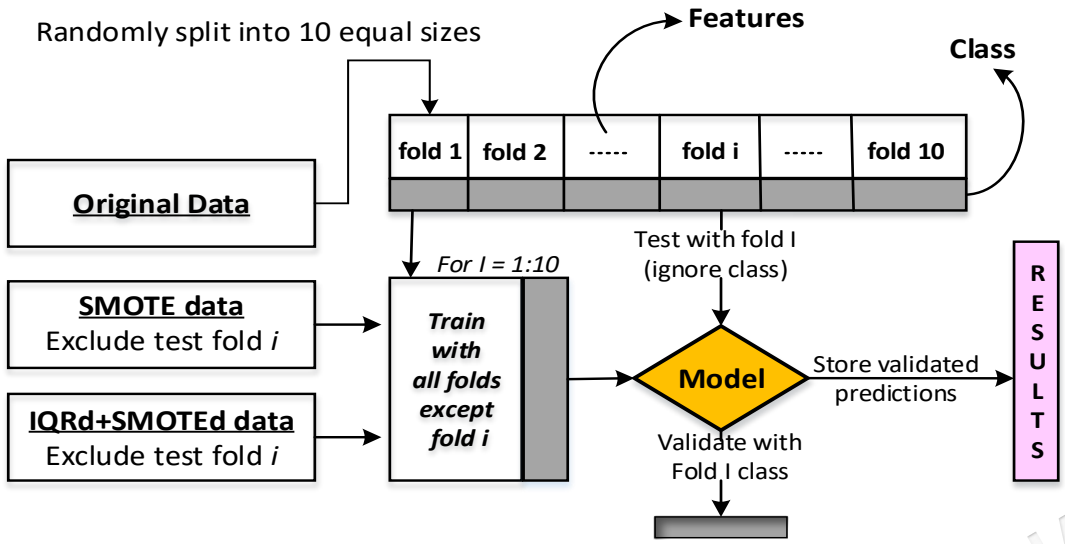


Figure 2.2 Improvement of Diabetes Prediction Accuracy by Smoothing out Noisy Training Data using the Interquartile Range and the SMOTE Technique (Rahim et al. 2021)

2.5.3 Existing Algorithms

One notable approach to successfully handle unbalanced data is to come up with new algorithms or to modify and enhance the ones already created. Several types of algorithms are explained in this section.

a. Variants of SVM

Class imbalance issues are prevalent in DNA microarray data, significantly affecting the predictive ability of minority classes. Factors such as feature reduction , high noise, limited sample sizes exacerbate the problem (Majid et al. 2014; Vuttipittayamongkol & Elyan 2020). To address this issue, a novel undersampling strategy based on the concept of Ant Colony Optimisation (ACO) has been developed. A usual categorisation technique for imbalanced data is the incorporation of Support Vector Machine (SVM) (Athitya Kumaraguru et al. 2020; D'addabbo & Maglietta 2015; Sreejith et al. 2020). The Megatrend Diffusion Strategy (MTD) may also be utilised to increase minority group representation.

During the predictor phase, various machine learning approaches are employed to merge SVM and KNN into hybrid models such as MTD-KNN, MTD-SVM, and other prediction models. The researchers examined the cost (Claesen et al. 2014) and found

that MTD-SVM outperformed alternatives like Random Forest (RF), Naive Bayes, and KNN. They proposed a prediction algorithm capable of classifying individuals at high risk for type 2 diabetes. There are two sections of the investigation. The first section explores the data imputation methods including median value imputation, KNN imputation, and iterative imputation, evaluating their performance when applied to the problem of missing data. Consequently, many classification techniques including linear, tree-based, and ensemble algorithms are used to verify the effects of these imputations on classification precision. Subsequently, an artificial neural network is utilised to synthesise the best quality of estimated data, along with SMOTE-Tomek technique to implement a proportional sampling across the datasets.

The accuracy achieved using this method on the test data is 98%, which is higher than any other tested method on the same data. Population and gender are the main themes in the dataset (Roy.et al. 2021).When dealing with a dataset that is skewed towards either the majority or the minority group, the kernel scaling approach is used to improve SVM performance. The PSS-SVM classification, which mixes Parallel Select Sampling (PSS) with the SVM, demonstrated impressive results on benchmark datasets, outperforming ordinary SVM due to its ability to address the lack of convergence. It may reduce the imbalance in large datasets by selectively sampling data from the dominant class (Devi et al. 2019; Yu et al. 2018).

A mixed ensemble of SVMs is used together with undersampling and oversampling tactics to improve prediction performance. Extensive testing has shown that this strategy is superior for standalone SVM and several alternative classifiers. The research compared the performance of the EnSVM basic model with the selective EnSVM+ ensemble by using various resampling strategies (Yu et al. 2018). By using SVM training as a preprocessor, these strategies improved the results of intelligent machine learning algorithms such as Multilayer Perceptron (MLP), Random Forest (RF), and Logistic Regression (LR). There are two phases of the implementation of the balancing approach. In the first phase, SVM is used to alter the imbalanced data to produce more balanced data whereas in the second stage, MLP, RF, and LR used these enhanced data as input (Hassan & Amiri 2019; Vigneron & Chen 2016). Table 2.5 presents the advantages and disadvantages of SVM and its different variations.

Table 2.5 A Summary of Various SVM Algorithm Approaches

Refe.	SVM Algorithm Approaches	Advantages	Disadvantages
(Maldonado et al. 2014; Roy. et al. 2021)	Ant Colony Optimisation SVM (ACO-SVM)	Solves the unbalanced data classification problem using an ACO algorithm-based sample selection process.	Excessive computational and storage costs.
(Tian et al. 2021)	Kernel-Transformed SVM with Adjusted F-Measure	Manages the cost function and imbalance data with kernel scaling.	Efficient estimation strategy for different parameters and kernels.
(Wu et al. 2021)	Intelligent SVM Preprocessor for MLP, LR, and RF (ISPM)	Effectively balances the data and increases the number of samples in the minority class.	It is neither simple nor fast to implement.
(Shi 2020)	Second-Order Cone Programming SVM (SOCP-SVM)	Provides robust and improved classification performance due to SVM-LP formulation.	Only designed for unbalanced data.
(Hassan & Amiri 2019)	Megatrend Diffusion SVM (MTD-SVM)	Improves the number of samples in the minority class.	Synthetic data generation is costly, and the MTD technique performs better on small datasets.
(Yu et al. 2018)	Parallel Selective Sampling SVM (PSS-SVM)	Provides accurate statistical predictions with low computational complexity.	Suited for parallel and distributed computing.
(Datta & Das 2015)	Near Bayesian SVM (NBSVM)	Minimises the cost of misclassification by the minority class.	Performance metrics may be affected.
(Maldonado et al. 2014)	Ensemble SVM with Resampling (EnSVM-R) and EnSVM+ with Resampling (EnSVM+-R)	More effective compared to standard SVM.	Does not automatically determine the value of k.

b. Clustering

Clustering is employed to classify the data, while outlier detection is used to identify outliers. The similarity-based hierarchical decomposition method operates in conjunction with clustering algorithms and the detection of outlier. The process of creating hierarchies is divided into two parts: the first addresses errors in cluster classification, while the second focuses on achieving accurate classification (Zhou et al. 2017). Data similarities of labelled subgroups at each level are utilised to generate hierarchies and feature sets, as well as other data based on these different levels. This method can prevent issues such as class overlap and inequities between groups (Khaldy

& Kambhampati 2018). The research presented undersampling technique based on clustering for data with an unequal class distributions (Pes, 2021 ;Chen, 2024)

Fuzzy Rule-based Classes (FRBCSs) are also used to improve classification accuracy. This approach, which utilises 2-tuple genetic tuning to improve the efficiency of FRBCSs, is effective in handling imbalanced data as it can handle both low and high ratios of skewed datasets (Wang et al. 2019). Clustering-based oversampling is a revolutionary data-level resampling strategy designed to improve learning from class-unbalanced datasets. The concept of methodology that underpins the proposed method is that the number of new sample points that need to be generated for a minority class sample can be inferred by measuring the distance between a minority class sample and the respective cluster centroid (Re & Valentini 2012). Figure 2.3 shows the schematic illustration of the clustering technique used for event detection.

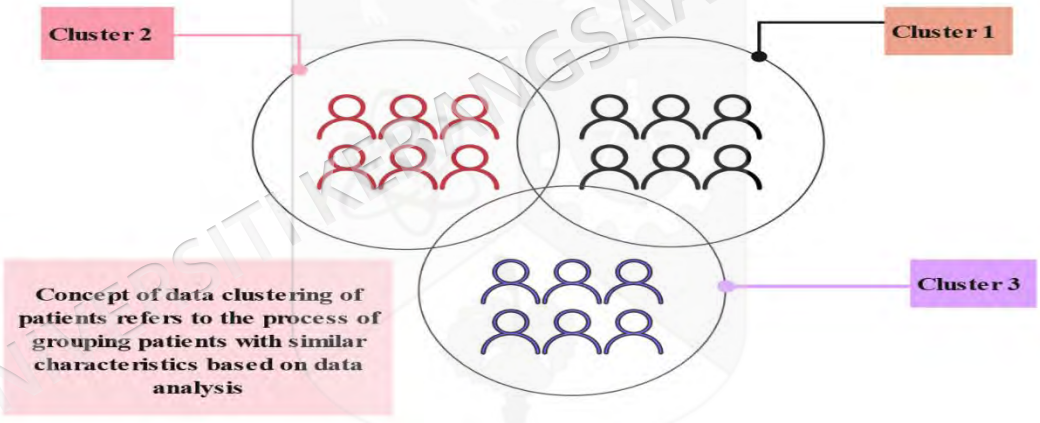


Figure 2.3 Concept of Data Clustering of Patients based on Similar Characteristics

c. Features Selection Methods

In high-dimensional data settings, feature selection is often the initial step for many machine learning algorithms. This process is particularly effective when dealing with high-dimensional heterogeneous data.Without a sufficient number of features, Feature Subset (FS) metric cannot fully demonstrate its value. Feature selection, where *y* is an input parameter, aims to improve the classification accuracy by narrowing down the collection of features used by the classifier. The application of filters to high-dimensional datasets enables a thorough analysis of each characteristic (Güney 2023;

Zhou et al. 2017). However, current methods for selecting features in imbalanced datasets are often inadequate.

A new paradigm for feature selection that separates the majority and minority characteristics is presented. This approach allows for the modification of existing standards by analysing demographic data in two distinct ways, which are majority and minority. Another method proposed by (Islam et al. 2022) is effective in several high-dimensional, imbalanced, small-sample datasets. (Fujita et al. 2023) proposed a decomposition-based technique and a Hellinger distance-based approach for feature selection in imbalanced datasets. The conflict caused by the unequal distribution of classes is first addressed by calculating the distributional gaps. The second method partitions huge classes into arbitrary subclasses by using a wide range of classification algorithms.

Recently, the efficacy of hybrid learning strategies has been investigated, combining feature selection approaches to lower the data dimensionality with appropriate methods that mitigate the negative impacts of class imbalance particularly the data balancing and cost-sensitive techniques. Extensive studies have been conducted across datasets from various domains using a popular classifier known as the Random Forest (RF), which is effective in high-dimensional spaces and well-suited for unbalanced problems. The findings demonstrated the advantages of this combined strategy over either feature selection or imbalance learning alone, as shown in Figure 2.4 (Neeraj & Maurya 2020).

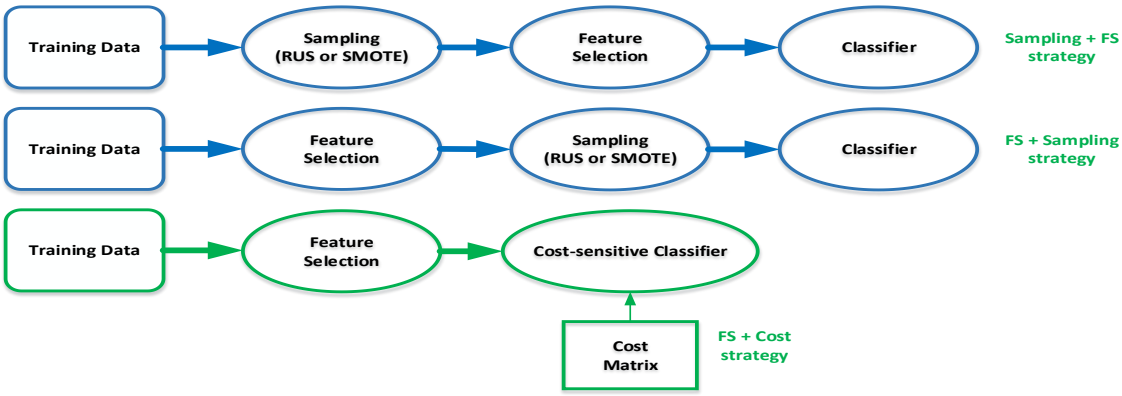


Figure 2.4 Hybrid Approaches for Combining Feature Selection with Imbalance Learning (Schubach et al. 2017)

d. One-Class Learning

Traditional learning mechanisms or methods frequently grapple with imbalanced data, which diminishes accuracy and performance, and result in erroneous outcomes. This occurs when there is an unequal representation of instances within the dataset, viewed from various categorical perspectives. The misclassification of minority classes is particularly troublesome when developing an accurate predictive model(Sagi & Rokach 2018) In addition, issues such as inconsistent and overlapping data can result in disastrous outcomes for powerful learning. In response, algorithms have been developed to select only those samples that correspond to the desired class, making them suitable for classifying information which is not symmetrical. This is accomplished by obtaining samples only from the minority classes, while disregarding the rest.

One-class learning is a great approach to be utilised on imbalanced datasets which are huge. This includes the utilisation of a rule-based method to iteratively develop rules, which is based on the divide-and-conquer principle based on the system of rule-induction (Sagi & Rokach 2018). The Ripper's training has been evident in past research. In this training, commencing from the least frequent to the most, rules will be developed for each class, gradually being added as time progresses. Accordingly, such algorithm is claimed to be beneficial as it can learn rules for the marginalised classes. One-class learning is highly effective for data with noisy features, excessively skewed datasets, and with high-dimensional spaces. Despite the hefty costs commonly associated with the one-class learning method, the benefits of its implementation make it worthwhile. Figure 2.5 illustrates the framework used for predicting type 2 diabetes as applied.

2.5.4 Ensemble Learning Algorithms

The performance of the ensemble approaches highly depend on the base learner's quality. It is common knowledge that the robustness and accuracy of the predictive models can be enhanced by the ensemble classifiers. The ensemble classifiers train multiple classifiers and synergise the results into one single classifier with superior

performance, thereby boosting the performance and accuracy compared to an individual classifier. Hybrid approaches such as data-driven algorithms and cost-sensitive solutions are usually incorporated and employed by the ensemble-based approaches. In addition, the underlying learner is modified and enhanced, as opposed to the fundamental classifier, through ensemble learning and algorithmic techniques.

The composition techniques are often implemented to boost the performance of individual classifier by merging several classifiers into one superior framework. This combination has been demonstrated to be an effective step to enhance the machine learning model. Despite the inability of both the ensembles of classifiers and the individual learning classifiers in handling the issue of class imbalance, well-considered learning techniques can be leveraged to overcome this predicament (Sagi & Rokach 2018).

An array of methods and approaches can be considered to develop an effective ensemble learning algorithm, under the assumption of a weak learning algorithm. Novel approach that has been examined by (Polikar, 2006)&(Kumar,2024). which significantly outperform the standard algorithms involves the prediction of noncoding variants affiliated with Mendelian and complex disorders, which performed by leveraging the imbalance-aware learning methods based on hyper-economy approach and resampling techniques.

Figure 2.6 illustrates the HyperSMURF method in a flowchart. The indicated elements correspond to the majority class, divided into n subclasses to expand the quantity of minority class training instances. By applying oversampling methods, additional cases from the minority class are synthesized for each partition. Simulatenously, a sample of the larger demographic is selected. Finally, HyperSMURF utilises an ensemble-of-ensembles technique to combine the forecasts of *n* independently trained RF models using symmetric datasets, as shown in Figure 2.5 Prediction of Type 2 Diabetes Framework (Roy et al. 2021).

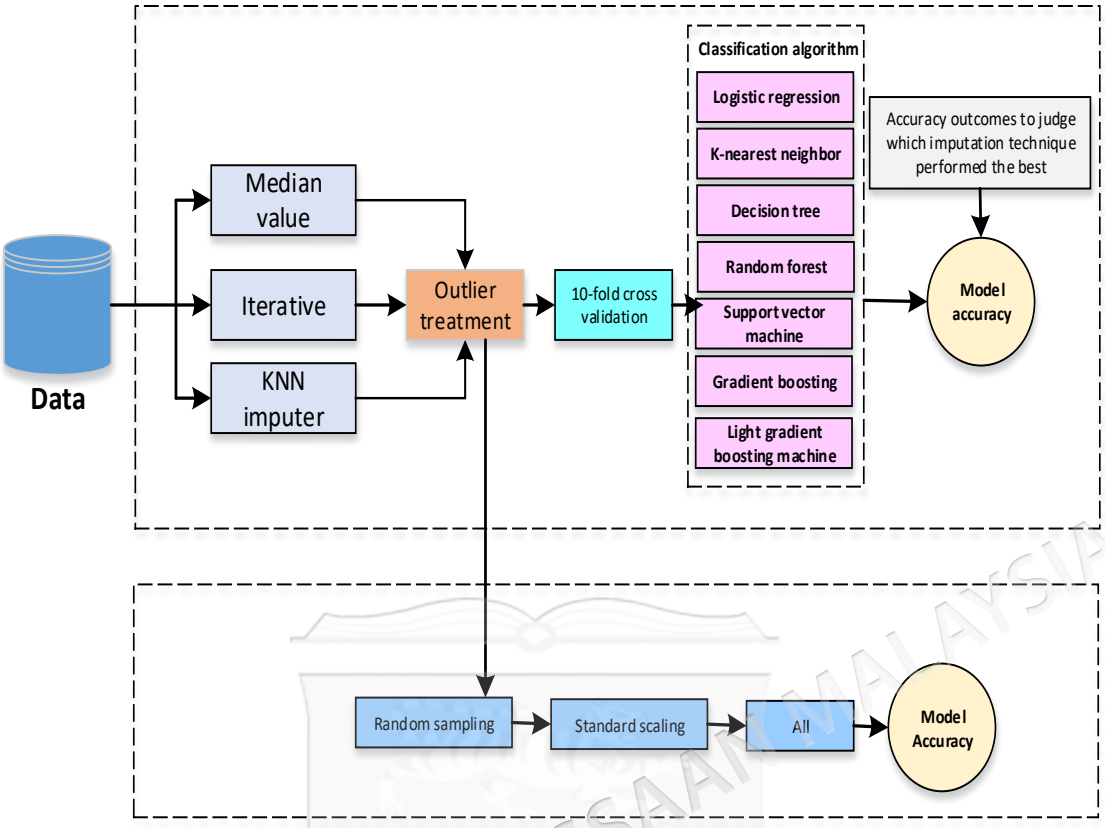


Figure 2.5 Prediction of Type 2 Diabetes Framework (Roy et al. 2021).

a. Bagging Learning

(Breiman, 1996) promoted the method called bootstrap aggregation for developing ensembles. Bagging refers to a technique utilised by statisticians to correct the issue arising from the data imbalance between the ratio of minority samples in the datasets. Several types of bagging that have been developed such as Over-Bagging, Under-Bagging, Under-Over-Bagging, and Under-Over-Bagging to preserve the distribution and variety without compromising it (Ghojogh & Crowley 2019).

b. Boosting

Boosting refers to a machine learning ensemble technique which seeks to enhance the performance of weak learners, usually consists of simple models, by synergising their predictions strategically. The Probability Approximately Correct (PAC) learning framework has been demonstrated in previous research to be capable of transforming a weak learner into a good one. Just like bagging, the boosting technique involves

choosing the data points that lead to an erroneous predictions. This study proposes an effective approach for an ensemble method, BPSO-Adaboost-KNN, to manage multiclass classification for imbalanced data. At the heart of this concept, lies the integration of algorithm and feature selection, which are boosted into an ensemble. In addition, this method utilises a state-of-the-art evaluation metric called AUC-area, specifically for multiclass classification (Haixiang et al. 2016).

c. AdaBoost

AdaBoost utilises the entire dataset to train each classifier serially, and with every single iteration, it focuses on the samples that are difficult to classify, which are instances belonging to the minority group, because of the algorithm's bias in learning (the weight) from imbalanced data. AdaBoost.M1 and AdaBoost.M2 are two well-regarded adaptations used in asymmetrical rule-based systems (Wang & Sun 2021). In addition, an ensemble approach is also utilised during the classification of imbalanced data, through which the imbalanced dataset is divided into large subsets with similar sizes. Subsequently, these sub-groups are subjected to a few classifiers employing unique classification algorithm (Polikar, 2006). Figure 2.6 illustrates the hyperSMURF approach.

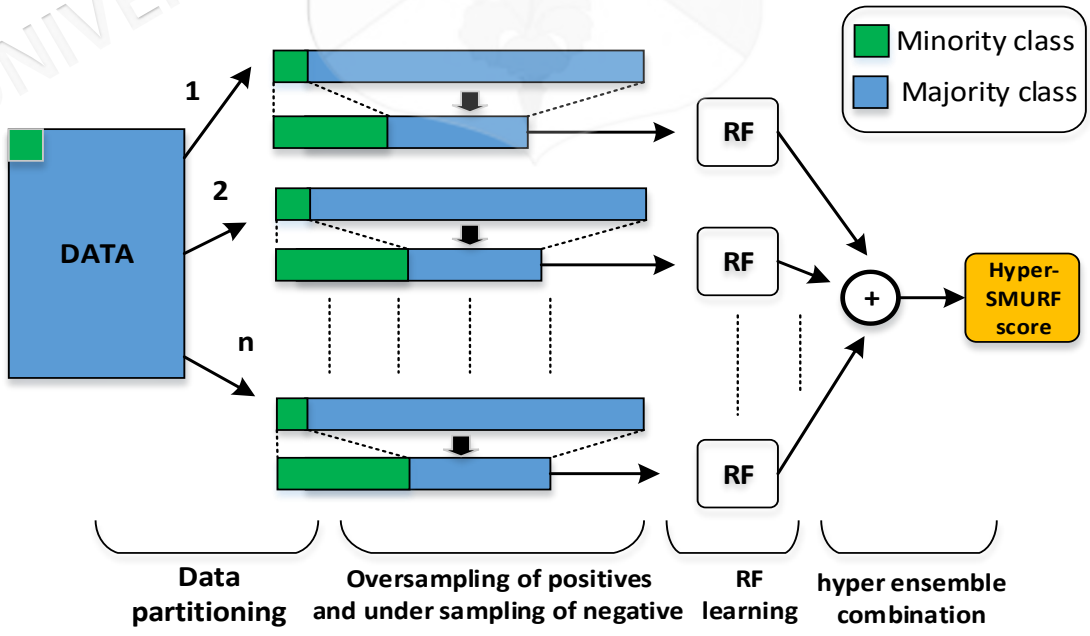


Figure 2.6 HyperSMURF approach (Polikar, 2006)

2.5.5 Cost-Sensitivity Methods

The incorporation of algorithm and data levels very well mean that the price of misclassification can be severe. As such, scholars seek to minimise the hefty misclassification cost using this method. In the case of cost-sensitive approach, this technique would be more enlightening if positive instances were registered compared regarding oncological treatment, the misclassification cost of a non-cancerous patient (false positive) is typically restricted to unnecessary clinical evaluations or anxiety caused by further diagnostic tests. These costs, while not insignificant, are generally less severe compared to the misclassification of a cancerous patient (false negative). In the latter case, the consequences can be dire, including delayed diagnosis, progression of the disease, and a potential loss of life. This stark disparity in misclassification costs highlights the necessity of cost-sensitive learning approaches, which assign higher penalties to false negatives to reduce their occurrence (Elkan, 2001; Fernández et al., 2018). To bridge the gap between external and internal evaluation methods, an assessment protocol is often employed to systematically examine anticipated cost changes (Drummond & Holte, 2006; Khan et al., 2017).

The learning mechanism is adapted to accept costs and subsequently costs will be added to samples through the combination between data-level and algorithmic techniques. The classifiers are inclined to be biased towards the minority class to minimise the overall costs of misclassification for both classes if the minority class has a greater rate or chance for a mistake. Siers and c (2020) stipulated that the cost and emotional toll on the cancer patient who obtains a false negative result, is broadly higher than that of false positive. This happens when a patient is tested negative but falsely grouped into the positive class. For instance, in cancer diagnosis, if an individual is incorrectly classified as a positive class (i.e., minority class) and a non-cancer patient in a negative class (majority class), the patient could potentially lose their life due to erroneous categorisation or being deprived from the proper medical diagnosis and treatment. To lower misclassification and total test costs, extensive modifications to cost matrices for cost-sensitive training and formulation costs can be investigated (Khan et al. 2017). Table 2.6 presents a comparative summary of various advanced methods used for clustering, feature selection, one-class learning, cost-sensitive learning, and

ensemble learning in the context of imbalanced and high-dimensional datasets. Each row corresponds to a key study, with the highlighted references (such as Alibhai et al., 2012) indicating particularly influential or relevant works for this thesis. For example, Alibhai et al. (2012) introduced a Density-Based Feature Selection (DBFS) method, which is notable for its effectiveness in handling small sample sizes and high-dimensional data—common challenges in diabetes dataset. The table outlines the main advantages and disadvantages of each method, enabling readers to quickly compare their suitability for different data scenarios. This comparison supports the rationale behind the selection of techniques in this research and highlights the need for continued innovation in addressing data quality issues.

To deliver uneven treatment for groups or classes that are not equally treated by cost-sensitive learning, the central AdaBoost learning architecture is maintained while, at the same time, including cost perspectives in the weight update algorithm. Hence, these techniques may diverge solely in the specific ways in which they improve the weight update procedure. AdaC1, AdaC2, and AdaC3; CSB1, CSB2, and AdaCost are the most prominent cost-sensitive boosting algorithms (Buda et al. 2018), as illustrated in Table 2.6.

Table 2.6 Comparison of Methods for Clustering, Feature Selection, One-Class Learning, Cost-Sensitivity, and Ensemble Learning

Reference	Methods	Advantages	Disadvantages
Fan et al. (1999)	AdaCost: Misclassification Cost-Sensitive Method	Dynamically penalizes misclassifications to reduce cumulative misclassification cost.	Adds computational complexity during training, which can be prohibitive for large-scale datasets.
Rong et al., (2008)	Multi-Class Pattern Classification	Enhances minority class coverage for multi-class problems.	Creates ambiguity in some cases.
Gao et al., (2011)	Combined SMOTE and PSO-based RBF classifiers	Performs well, especially in creating synthetic specimens for the minority class.	Requires more storage space.
Sun et al., (2015)	A Novel Ensemble Method for Classifying Unbalanced Data	Improves sensitivity and specificity for binary imbalanced data classification.	Ambiguity in classification can lead to potential misclassifications.
Beyan & Fisher, (2015)	Similarity-based hierarchical decomposition	Creates balanced subsets based on similarity measures, improving classification performance when the imbalance ratio is low.	Computational complexity is a significant challenge during the training phase, particularly for high-dimensional datasets.
Fernández et al., (2018)	Cost-Sensitive Learning Method	Reduces error costs by incorporating cost-sensitive measures into learning algorithms.	Does not explicitly address misclassification rates for majority classes, potentially leading to trade-offs in overall accuracy.
Sagi & Rokach, (2018)	Ensemble Learning Method for Classification	Combines multiple classifiers to reduce variance and improve accuracy.	Limited to binary class problems, restricting its application to multi-class datasets.
Buda et al., (2018)	AdaC1, AdaC2, and AdaC3; CSB1, CSB2, and AdaCost	Effectively integrate cost considerations into boosting algorithms, improving performance for minority classes. They rank features using statistical measures, effectively reducing misclassification costs.	Struggle with highly imbalanced datasets where the distribution of different classes is significantly skewed, limiting their generalizability.
Grzyb et al., (2021)	Hellinger distance-based Imbalanced Data Classification Method	Utilizes True and False Positive rates with Hellinger distance to optimize classification thresholds, achieving balanced performance.	Limited experimental scope as it was tested with only three feature selection methods.

2.5.6 Data-Level Algorithms

Several classification frameworks are designed and being improved upon to carry out the classification process, due to its utmost significance in efficaciousness of data mining. On the other hand, regular models for classification are immensely dependant to the specific characteristics of each dataset. In essence, standard classification models are biased towards the most prevalent patterns, resulting in higher tendency to misclassify the less frequent instances if the datasets are skewed or not evenly distributed. Due to the massive issues regarding the class imbalance or class disparity, many scholars have devoted a lot of efforts to develop a comprehensive technique to overcome it.

These approaches, which were developed to effectively manage the class differences, can be divided into three distinct categories. Some approaches take an external or data-level approach, some take an internal or algorithmic approach, and those that are cost-sensitive. Additionally, ensemble learning classifiers are immensely imperative in the classification of data with uneven distribution of classes (Devi et al. 2020).

The study offers a thorough comparison of recent data-level methods, including under sampling (Ahmed, 2023),oversampling, and hybrid approaches, highlighting their effectiveness in various scenarios.. Although there is an established literature on data-level classifications, the computational requirement remains burdensome for these methods. The algorithmic techniques are largely aimed at improving the accuracy of a framework by developing novel algorithms or modifications to the existing techniques. As demonstrated in (Naghibi et al. 2017), after comparing the effectiveness of various methods, the naive Bayes and KNN approaches tend to outperform the SVM and Random Forest (RF) methods. Multiple issues can be addressed by employing Genetic Programming (GP), a revolutionary method. Adjusting costs in GP involves following the appropriate fitness functions. It is crucial for the fitness function to reward solutions that benefit both minority and majority groups. Furthermore, it has been demonstrated that the performance and average class accuracy of the fitness function are sufficient to resolve issues related to imbalanced classification.Therefore, the Ames, Incr, and Corr fitness functions are utilized (Grzyb, 2023).

a. External Data-Level Algorithm

Random undersampling is a resampling technique which is utilised to handle imbalanced data which involves randomly excluding specimens or instances from the majority class to create a balanced subset of the main dataset. This approach can be categorised into three forms. However, there is a risk of loss of information by the omission of data during the induction stage. This technique generates a supplementary set of main data that is inclusive of more specimens from the minority group, which is accomplished through a random duplication of existing samples. However, this excessive likelihood of repetition through duplication can result in overfitting. In addition, the hybrid method, as the name suggests, utilises a combination of the two sampling techniques to yield a more uniform or balanced dataset (Jackson et al. 2017). The minority class's SMOTE uses interpolation to synthesise new samples based on existing minority class specimens. To synthesise a duplicate specimen from two interpolated specimens, SMOTE randomly selects one of the kNN of a poor specimen. Efforts to expand the reach of the minority class into the territory of the majority class are limited by a decision taken at the boundary level. The technique eradicates the overfitting issue but yields noisy and questionable specimens (Soltanzadeh & Hashemzadeh 2021). Some filtering-based methods (SMOTE-TL and EL) are utilised to suppress noise in datasets which are asymmetrical. Likewise, original sampling methods are enhanced with Neighbourhood-balanced bagging (NBBag) to handle asymmetric data.

Modified Synthetic Minority Oversampling Technique (MSMOTE) is a modified form of SMOTE which had been designed and developed. Through this modified version, the minority class is allocated into three categories called latent noise, border and safe, depending on the distances between all samples. This method generates new instances, while removing latent noise spots utilising the kNN categorisation method (Hu et al. 2009). Additionally, this method effectively handles the hidden noise instances and does not emphasise significant features. An enhanced version of SMOTE and an iterative partitioning filter (IPF) have been developed to overcome the disturbances and preserving orderly class boundaries (Sun et al. 2020). To achieve more superior data-level strategies, different SMOTE had been developed such as SMOTE extension, B1-SMOTE, and B2-SMOTE (Turlapati & Prusty 2020). The Minority

Weighted Minority Oversampling Technique (MWMOTE) is an efficacious technique for choosing and weighing samples from hard-to-learn minority classes. Furthermore, it can generate realistic simulated instances.

Another method is the Selective Pre-treatment of Imbalanced Data (SPIDER), which integrates complex instances identified in the majority class with localised oversampling of the minority class in a single process (Malhotra & Kamal 2019). To address the issue of skewed data, researchers have developed to a novel Inverse Random Undersampling (IRUS) technique that utilises an inverse strategy based on the proportion of class imbalance. This method is also relevant for multi-label categorisation systems. To manage imbalanced datasets, a Radial Basis Function Network (RBFN) has been suggested by Ghosh and Nag (2001), which incorporates both local and global components and utilises a training method focused on local weights.

A superior proportion of the Imbalance Ratio (IR) delivers stronger results in local weight training techniques, and an inferior value of the IR should be balanced with any method. Integrating well-known classifiers such as Logistic Regression (LR), the C5 decision tree model (C5), and the search for the nearest Neighbour 1 in the tree, a classifier method is developed by utilising PSO in conjunction with SMOTE. This new set of classifiers are demonstrated to increase the prediction accuracy using performance measurements or metrics, such as accuracy indices and the G-mean. The experimental findings demonstrated the efficacy of the hybrid algorithm PSO + SMOTE + C5 to predict 5-year survival in breast cancer patients. The fusion of PSO, SMOTE, and an assisted Radial Basis Function (RBF) classifier is proposed as another effective method for imbalanced binary class situations. Evaluations with diversified performance measurements showcased that this method is effective for moderately imbalanced datasets but the effectiveness drops as the imbalance gets more extreme.

b. Internal Data-Level Algorithmic

One approach to addressing imbalanced data is to create new algorithms or boost existing ones to mitigate the impacts on minority class. Table 2.7 summarises the advantages and disadvantages of all data-level approaches.

Table 2.7 Data-Level Approaches for Addressing Class Imbalance: Advantages and Disadvantages

Reference	Approach	Advantages	Disadvantages
Ghosh & Nag 2001	**Radial Basis Function Networks (RBFN) SMOTE**	**The Higher the Imbalance Ratio (IR) Value of the Dataset, The Better the Result.**	**Requires More Storage Space.**
Hu et al. 2009	Modified SMOTE (MSMOTE)	Reduces Noise.	Neglects Priorities Of Important Features.
Gao et al. 2011	Combined SMOTE And PSO-Based RBF Classifiers (SMOTE+PSO-RBF)	Performs Well, Especially in Creating Synthetic Specimens for Minority Class.	Requires More Storage Space.
Barua et al. 2012	Majority Weighted Minority Oversampling Technique (MWMOTE)	Creates Artificial Inferior Samples.	Issues With Multiclass Imbalance Problem.
Tahir et al. 2012	Novel Inverse Random Under Sampling (IRUS)	Provides Multi-Label Classification with High Accuracy; Feasible for Irregular Dataset Sizes.	Uses Different Applications for Multi-Label Classification.
Sáez et al. 2015	Iterative–Partitioning Filter + SMOTE (SMOTE-IPF)	Addresses Noise and Borderline Examples in Imbalanced Datasets.	Uses Small Sample Size.
Malhotra & Kamal 2019	Selective Pre-Processing of Imbalanced Data (SPIDER)	Strength Lies in Filtering Difficult Examples.	Suffers From Complexity Issues.
Pan et al. 2020	Synthetic Minority Oversampling Technique (SMOTE)	Effective In Balancing Classification by Increasing the Number of Minor Class Examples.	Can Lead to Overfitting in Some Cases.
Almutairi & Abbod 2023	Machine Learning Methods for Diabetes Prevalence Classification in Saudi Arabia	Predicts Diabetes Prevalence and Trends Using ML Algorithms: LD, SVM, K-Nearest Neighbour, NPR. Weighted KNN Outperforms Other Methods in Accuracy.	Focused On Saudi Arabia; May Not Generalize Globally.
Huang et al. 2023; Zhang et al. 2023	Enhancement Of Classification of Imbalanced Dataset for Edge Computing	Improves Classification Accuracy for Imbalanced Datasets in Edge Computing Using LBNN Algorithm, Convex Hull, Hyperplane. Outperforms LBNN In Imbalanced Datasets.	Complexity May Increase Computational Load.

2.5.7 Evaluation Metrics for Imbalanced Datasets

The use of metrics that reflect performance across all classes, particularly the minority class, is necessary for evaluating models trained on imbalanced datasets. While metrics such as accuracy are frequently employed, they frequently fail to offer valuable insights in these situations. For example, in datasets with substantially biased distributions, a model that predicts exclusively the majority class can achieve high accuracy while completely disregarding the minority class, which is frequently the focus in critical applications such as medical diagnoses or fraud detection.

Metrics for Imbalanced Datasets Precision and recall assess the reliability of positive predictions by determining the proportion of predicted positives that are genuine positives. Recall, or sensitivity, is a metric that quantifies the model's capacity to identify genuine positives, providing a glimpse into its capacity to identify instances of minority class. These metrics are essential for comprehending the extent to which a model effectively multi class balance. Singhal. (2020).F1-ScoreThe F1-Score combines precision and recall into a singular metric, depicting their harmonic mean. This is especially advantageous. This instrument is essential for evaluating models in asymmetrical scenarios when the trade-off between false positives and false negatives must be balanced.

Balanced Accuracy balanced accuracy computes the average recall for all classes, ensuring that the metric factors for performance across both majority and minority classes. Unlike standard accuracy, it avoids bias toward the majority class, making it particularly relevant for imbalanced datasets (Xu et al., 2023). AUC-ROCThe Area Under the Receiver Operating Characteristic Curve (AUC-ROC) evaluates the trade-off between true positive and false positive rates at various thresholds. While widely used, this metric can sometimes present an excessively optimistic assessment of models trained on highly imbalanced datasets. Thus, its use is most effective when combined with other metrics. MCC and G-Mean The Matthews Correlation Coefficient (MCC) provides a balanced evaluation by accounted for all four confusion matrix elements—true positives, true negatives, false positives, and false negatives. Similarly, the Geometric Mean (G-Mean) emphasizes maintaining a balance between sensitivity

for the minority class and specificity for the majority class, making these metrics particularly valuable in highly imbalanced scenarios. Why evaluation metrics matter more than performance assessment Focusing on evaluation metrics over general Performance measures like overall accuracy is vital when dealing with imbalanced datasets. Performance enhancement techniques, such as SMOTE, improve class balance, but their effectiveness should be assessed using metrics that reflect the performance of minority classes. For example, SMOTE has been shown to enhance precision and recall for minority classes, resulting in significant improvements in F1-score and balanced accuracy.

2.5.8 Challenges and Opportunities

A dataset is considered imbalanced if one category has significantly fewer data points than the others. There is often great interest in the rarest categories (Maheshwari et al. 2017). The learning mission finds this complication to be of great practical importance since it is applicable in many distinct contexts such as remote sensing, pollution monitoring, risk assessment, fraud detection, and medical diagnostics. As a result, most classifier learning algorithms tend to favour classes that have more data points in these situations (Mohammed et al. 2020).

For this reason, the accuracy measure is given more weight than the specificity metric when determining the relative merits of two sets of rules. Predictable incidents in the minority class are often overlooked because of this. Predictable incidents in the minority class are often overlooked because of this. Nevertheless, merely inflating the frequency of underrepresented groups in the data does not solve the problem. When minorities are adequately represented, they can narrow the distribution, thereby balancing the imbalance ratio. Yet, despite the inequitable distribution of data, the lessons are better retained in this setting. Although this may be true in theory, it seldom holds in practice. A dataset with insufficient samples would lead to overfitting, making it insensitive to the minority class in terms of precision. In small sample-imbalanced datasets, the problem of imbalanced categorisation is particularly acute. Both SMOTE-ENN and SMOTE-IPF, hybrid resampling approaches, have limitations when working with unbalanced datasets and small sample sizes (Kumar et al. 2019). The study titled

"A Survey on Imbalanced Learning: Latest Research, Applications and Future Directions" (Chen et al., 2024) provides a comprehensive overview of recent advancements in imbalanced learning, categorizing existing research into data-level, algorithm-level, hybrid methods, ensemble learning, deep long-tail learning, regression, clustering, and data streaming algorithms.

The removal of minority-class examples using IPF and ENN due to their perceived danger or difficulty in learning raises questions about the potential usefulness of these examples. The elimination of potentially harmful positive instances is an essential part of performance improvement (Leevy et al. 2018). There is an additional challenge in training a learning model due to imbalanced dataset. Learning from extremely unbalanced data is crucial to developing a reliable model. This problem arises regularly in practical settings, such as disease diagnosis, spam filtering, crisis forecasting, and fraud investigation. As shown in table 2.8

Table 2.8 Summary of Research Papers on Medical Prediction for Handling Imbalanced Datasets

Ref.	Title	Problem Statement	Method/Tool	Finding
Zhang et al. 2023	An Ensemble Oversampling Method for Imbalanced Classification with Prior Knowledge Via Generative Adversarial Network	Addressing imbalanced datasets in real-world applications that require new algorithms	Generative adversarial network (GAN)	G-GAN provides promising results in G-mean, AUC, F-measure, and ROC, improving classification performance on imbalanced datasets
Almutairi & Abbod 2023	Machine Learning Methods for Diabetes Prevalence Classification in Saudi Arabia	Predicting diabetes prevalence and trends in Saudi Arabia using ML algorithms	LD, SVM, K-nearest Neighbour, NPR	Weighted KNN outperforms other methods in predicting diabetes prevalence rate accuracy
Huang et al. 2023;	The Enhancement of Classification of Imbalanced Dataset for Edge Computing	Improving classification accuracy for imbalanced datasets in edge computing	LBNN algorithm, Convex Hull, Hyperplane	The proposed method outperforms the LBNN algorithm in imbalanced datasets, making it suitable for edge computing and real-time data classification
Chen et al. 2024	Adversarial Network-based Approach for Imbalanced Medical Data Classification	Handling imbalanced data in medical data classification using adversarial network-based approaches	Adversarial network architecture	The proposed adversarial network-based approach effectively addresses imbalanced data in diabetes dataset, improving classification accuracy

Missing data when certain observations or details are not recorded—presents a significant challenge in diabetes research, impacting the accuracy and reliability of findings. Alongside missing data, feature reduction and imbalanced datasets further complicate efforts to predict and manage diabetic complications. Addressing these challenges is critical to maximizing the potential of predictive models in improving diabetes care. Recent research has focused on methods to better predict complications like neuropathy and nephropathy, which profoundly affect patients' quality of life and increase the complexity of disease management.

By analyzing a combination of clinical and demographic factors—such as blood sugar levels, patient age, and relevant biomarkers, these predictive models aim to identify individuals at higher risk of complications, offering opportunities for early intervention and personalized care. For instance, Chen et al. (2021) demonstrated the potential of adversarial networks to predict nephropathy risk with high accuracy, even in datasets with significant imbalances. Such advancements improve healthcare providers' ability to deliver targeted treatments, reduce the likelihood of severe complications, and enhance long-term patient outcomes. These tools not only improve the lives of patients but also reduce the strain on healthcare systems by enabling earlier and more precise disease management.

- Type 1 Diabetes: An autoimmune condition where the body's immune system attacks insulin-producing beta cells in the pancreas, leading to little or no insulin production.

- Type 2 Diabetes: A condition characterized by insulin resistance, where the body's cells do not respond effectively to insulin, often accompanied by inadequate insulin production.

Both forms require ongoing management to prevent complications, underscoring their chronic nature. Therefore, any assertion that diabetes is not a chronic condition is incorrect

Despite these advancements, challenges remain. Innovative methods like generative adversarial networks (Zhang et al., 2023) and ensemble-based feature selection approaches have made strides in addressing issues like class imbalance and feature redundancy. However, ongoing refinement and adaptation are needed to make these techniques universally effective. As summarized in Table 2.9, these efforts underscore the importance of creating more reliable tools to predict

Table 2.9 Techniques for Handling Imbalanced Diabetes dataset

Ref.	Technique	Data	Methodology	Ratio of Imbalanced Data (%)
Zhang et al. (2023)	Generative Adversarial Network (GAN)	Real-world applications	G-GAN provides promising results in G-mean, AUC, F-measure, and ROC, improving classification performance on imbalanced datasets	90:10
Almutairi & Abbod (2023)	Weighted K-Nearest Neighbours (KNN)	Diabetes dataset (Saudi Arabia)	Predicting diabetes prevalence using ML algorithms	80:20
Huang et al. (2023)	LBNN algorithm, Convex Hull, Hyperplane	Edge computing datasets	Proposed method outperforms LBNN algorithm, suitable for real-time data classification	70:30
Chen et al. (2024)	Adversarial Network	Diabetes dataset	Adversarial network architecture	85:15

2.6 HANDLING FEATURE REDUCTION

Handling feature reduction in medical prediction datasets presents several obstacles that need to be addressed. One significant challenge is feature reduction , where the accuracy and performance of the models plummets as the dimensionality increases. Data with many features frequently leads to increased computational complexity, overfitting, and difficulties in interpreting the results. Another challenge is the presence of irrelevant and redundant features in the dataset. In medical prediction tasks, not all features are equally informative or contribute significantly to the prediction accuracy. Identifying the most relevant features is crucial for boosting the performance of prediction models and reduce computational overhead.

To tackle these obstacles, various state-of-the-art approaches have been proposed. One such approach is the R-ensemble method based on rough set theory, as discussed by Bania & Halder (2021). This ensemble attribute selection algorithm aids in data reduction and the classification of different diseases. This approach improves the prediction performance across diverse datasets by selecting highly relevant and significant attributes. Other studies, propose oversampling and undersampling approaches, including the HusdoS-boost algorithm, to address extremely imbalanced and small minority data problems in health record analysis. These approaches aim to eliminate repeated majority samples and generate synthetic minority samples mirroring the distribution of the minority class, addressing the issues of overfitting and information loss. Additionally, method to lower data dimensionality, such as principal component analysis (PCA), PCA with a kernel, and autoencoders, have been explored in the context of cancer prediction, as highlighted by Kabir et al. (2023). These techniques aim to reduce the dimensionality of the data while preserving the essential information, thus improving the performance of machine learning models.

Overall, the scientific discussions and findings reveal that handling feature reduction in medical prediction datasets requires thoughtful consideration of feature selection, data reduction, and addressing class imbalance. The proposed methods, such as R-ensemble, HusdoS-boost, and dimensionality reduction techniques, offer promising solutions to these challenges. Nevertheless, the selection of which method to

utilise must contemplated based on the context, characteristics, as well as the goal of the prediction.

2.6.1 Handling Feature reduction in Medical Prediction Datasets

Feature reduction in medical prediction datasets poses challenges such as inefficient computations, redundant information, and the potential for overfitting in model training. Addressing these issues requires strategic methods to streamline datasets while preserving critical information. For instance, Kabir et al. (2023) demonstrated the success of autoencoders in reducing RNA sequencing data, leading to a 91.2% accuracy in cancer prediction. Similarly, Principal Component Analysis (PCA) has been utilized to extract key components from diabetes risk datasets, effectively simplifying the data and improving the performance of prediction models like Logistic Regression. These approaches not only reduce complexity but also ensure the core patterns in the data are maintained.

Feature selection and balanced data handling methods further improve the management of high-dimensional datasets. Introduced the R-Ensemble method, which addresses data ambiguity and redundancy by integrating rough set theory with kNN imputation, achieving 89.5% accuracy. It tackled class imbalance issues using the HusdoS-Boost algorithm, which combines innovative sampling techniques with decision trees to produce balanced and accurate predictions. These approaches highlight the practical benefits of tailoring solutions to specific challenges, ensuring more reliable outcomes in healthcare applications.

The integration of diverse strategies offers a comprehensive approach to solving the intertwined issues of dimensionality and data imbalance. Techniques such as HusdoS-Boost, which balances datasets while maintaining scalability, and hybrid models that optimize feature selection, have demonstrated significant improvements in accuracy and efficiency. Kurman and Kisan (2023) showcased how metaheuristic methods in cervical cancer studies achieved 90.1% accuracy by reducing redundant features while focusing on relevant patterns. These methodologies, as detailed in Tables 2.8 and 2.11, provide robust pathways to enhance predictive capabilities, ultimately supporting more informed and reliable medical decisionsshown in Table 2.10

Table 2.10 Handling feature reduction in medical prediction datasets.

Ref.	Title	Problem Statement	Method	Findings
Gaidai et al. (2023)	Global Cardiovascular Diseases Death Rate Prediction.	Dealing with regional dimensionality and cross-correlation in predicting cardiovascular disease mortality.	Developed a bio-system reliability approach for predicting CVD mortality probability.	Successful prediction of CVD mortality probability in multi-regional environmental and health systems using a bio-system reliability approach. Confidence bands provided for predicted death rates, enabling a more comprehensive understanding of the predicted probabilities.
Kabir et al. (2023)	A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction.	Improving machine learning models' performance in cancer prediction using dimensionality reduction techniques.	Analysed the impact of PCA, PCA with a kernel, and autoencoders on two machine learning classifiers (neural network, SVM).	Autoencoders can enhance the performance of machine learning models in cancer prediction using RNA sequencing data by reducing the dimensionality of the input features.
Kurman & Kisan (2023)	An in-depth and contrasting survey of meta-heuristic approaches with classical feature selection techniques specific to cervical cancer.	Evaluating meta-heuristic and classical feature selection techniques in the context of cervical cancer datasets.	Comparative literature review of classical and meta-heuristic techniques.	Classical feature selection techniques demonstrated higher accuracy and speed in cervical cancer datasets. However, a research gap exists in applying hybrid techniques to improve feature selection.
Enríquez et al. (2021)	Dimensionality reduction using PCA and CUR algorithm for data on COVID-19 tests.	Dimensionality reduction for COVID-19 test data.	Utilised PCA and CUR algorithm for dimensionality reduction.	Dimensionality reduction using PCA and CUR algorithm proved to be effective in reducing the dimensionality of COVID-19 test data, potentially improving the efficiency and accuracy of data analysis and modelling.

to be continued…

…continuation

Bania & Halder (2021)	R-Ensemble: A greedy rough set-based ensemble attribute selection algorithm with kNN imputation for classification of medical data.	Addressing uncertain, vague, and indiscernible data in diabetes dataset.	Proposed R-ensemble method based on rough set theory with kNN imputation.	Significant performance improvement observed in predicting the presence of medical conditions across all datasevts.
Fujiwara et al. (2020)	Over- and Under-sampling Approach for Extremely Imbalanced and Small Minority Data Problems in Health Record Analysis.	Solving the problem of extremely imbalanced and small minority data in health record analysis.	Proposed algorithm combining RF and hybrid sampling (HusdoS-boost).	Improved classification accuracy by addressing the problem of extremely imbalanced and small minority data in health record analysis through the HusdoS-boost algorithm, which combines random forest with undersampling and distribution-based oversampling techniques.
Kumar et al. (2019)	A Hybrid Data Mining Approach for Diabetes Prediction and Classification.	Developing a hybrid data mining approach for diabetes prediction and classification.	Combined multiple data mining techniques for diabetes prediction and classification.	Outliers and dimensionality reduction techniques have a significant impact on the performance of predictive models for medical disease diagnosis, emphasising the importance of outlier detection and dimensionality reduction in improving diagnostic accuracy.
Gangavarapu & Patil (2019)	A Novel Filter–Wrapper Hybrid Greedy Ensemble Approach Optimised Using the Genetic Algorithm to Reduce the Dimensionality of High-Dimensional diabetes dataset.	Reducing dimensionality in high-dimensional biomedical datasets using a novel filter-wrapper hybrid greedy ensemble approach optimised with a genetic algorithm.	Developed a filter-wrapper hybrid greedy ensemble approach optimised with a genetic algorithm.	The filter-wrapper hybrid greedy ensemble approach optimised with a genetic algorithm effectively reduces the dimensionality of high-dimensional diabetes dataset, improving computational efficiency and maintaining classification accuracy.
Cai et al. (2015)	Type 2 Diabetes Biomarkers of Human Gut Microbiota Selected Via Iterative Sure Independent Screening Method.	Identifying biomarkers of human gut microbiota for type 2 diabetes.	Utilised iterative sure independent screening (SIS) method to select biomarkers.	Feature selection and dimensionality reduction techniques applied to machine learning algorithms have improved the prediction of diabetes, enhancing the accuracy and efficiency of the models.

2.6.2 Challenge: Handling Feature reduction

Handling feature reduction in medical prediction datasets presents significant challenges. The "curse of dimensionality" deteriorates machine learning performance as the number of features increases, leading to computational complexity, overfitting, and difficulty in interpreting results. Irrelevant and redundant features further complicate the process, making feature selection crucial for improving model performance and reducing computational overhead. High-dimensional datasets often suffer from class imbalance, causing biased models that poorly predict minority classes, which are typically of interest in medical applications.

Approaches like the R-ensemble method, based on rough-set theory, address these issues by enhancing data reduction and classification. Over- and under-sampling methods, such as the HusdoS-boost algorithm proposed, tackle class imbalance by creating synthetic minority samples, improving prediction accuracy. Dimensionality reduction techniques, including PCA and autoencoders, are effective in retaining essential information while reducing data dimensions, as demonstrated by Kabir et al. (2023).

These methods present promising solutions, but researchers should not be satisfied with these approaches alone. Instead, they should explore and address other innovative ideas to enhance the quality of the data by reducing the dimensions. This continuous effort is crucial to ensure effective handling of feature reduction in prediction based on diabetes dataset. By leveraging advanced techniques and continually seeking improvements, researchers can significantly improve the quality of the data, ultimately enhancing the performance and accuracy of prediction models.

2.6.3 Addressing Data Quality Issues in Medical Analysis: Missing Data, Imbalance, and Feature reduction

The findings analyzed in Table 2.11 provide a comprehensive review of 17 studies published between 2015 and 2023, focusing on three primary data quality challenges in medical analysis: missing data, class imbalance, and feature reduction . The studies employed various techniques, including hybrid methods, dimensionality reduction, and

feature selection, with accuracy rates ranging from 86.4% to 91.2%. Missing data were addressed through techniques such as kNN imputation and PCA, while class imbalance was largely overlooked, indicating a significant research gap. Feature reduction was tackled using dimensionality reduction methods like PCA and Autoencoders.

The literature highlights a crucial need for integrated approaches that address all three challenges simultaneously. Notably, the lack of emphasis on class imbalance suggests that future research should prioritize this issue, as it remains underexplored despite its relevance in diabetes dataset. These gaps point to the necessity of adopting more comprehensive methodologies that can simultaneously handle missing data, class imbalance, and feature reduction . The implications of this review underscore the importance of advancing research in this area, suggesting that future studies should explore interdisciplinary approaches to effectively address these challenges and improve the quality of medical data analysis. This could potentially lead to better healthcare outcomes and enhanced decision-making processes.

Table 2.11 Navigating Data Quality Challenges in Medical Analysis: A Comprehensive Review of Recent Research on Missing Data, Class Imbalance, and Feature reduction (2015-2023)

Ref.	Technique	Data	Methodology	Accuracy (%)	Missing Data	Class Imbalance	Feature reduction
Bania & Halder (2021)	R-Ensemble method	Diabetes dataset	Rough set theory-based ensemble attribute selection with kNN imputation	89.5%	✗	✗	✓
Fujiwara et al. (2020)	HusdoS-boost algorithm	Health record analysis	Random Forest combined with hybrid sampling (over- and under-sampling)	87.3%	✗	✗	✗
Gaidai et al. (2023)	Bio-system reliability approach	Cardiovascular disease mortality	Prediction of CVD mortality using a bio-system reliability model	88.9%	✗	✗	✗
Kabir et al. (2023)	PCA, PCA with kernel, Autoencoder	Cancer prediction	Dimensionality reduction techniques applied to neural networks and SVM	91.2%	✗	✗	✓
Kurman & Kisan (2023)	Classical and meta-heuristic feature selection	Cervical cancer datasets	Comparative analysis of classical and meta-heuristic techniques	90.1%	✗	✗	✗
Enríquez et al. (2021)	PCA, CUR algorithm	COVID-19 test data	Dimensionality reduction for efficient and accurate analysis of COVID-19 test data	86.4%	✗	✗	✓
Gangavarapu & Patil (2019)	Filter-wrapper hybrid greedy ensemble (GA optimised)	High-dimensional diabetes dataset	Genetic algorithm-optimised filter-wrapper hybrid approach	89.2%	✗	✗	✗

2.7 SUMMARY

Machine learning data categorisation isn't without its difficulties, especially when handling feature reduction , class imbalances, and missing data. The precision, effectiveness, and dependability of prediction models are all directly affected by these problems; thus, they aren't only theoretical. The goal of this research is to develop more accurate machine learning models for illness prediction, with a particular emphasis on diabetes, by improving the quality of existing datasets. The primary objectives are managing high-dimensional datasets, addressing class imbalances, and handling missing data. By following these methods, computational inefficiencies can be reduced, model complexity simplified, and prediction accuracy enhanced.

To handle missing data, the study explores innovative methods like the Bidirectional Neighbour Graph-based approach. Techniques such as Numeric Cleansing, K-Nearest Neighbour (KNN) imputation, and median imputation of the nearest neighbour are employed to reconstruct incomplete datasets. Hybrid methods, which combine strategies like Multivariate Imputation by Chained Equations (MICE), KNN, and mean/mode imputation, also show promise for minimizing information loss while improving model reliability in diabetes dataset. Class imbalance is a common issue that can lead to biased predictions and unreliable classifiers is tackled using the Clustering Selection Synthesis Filtering (CSSF) model.

This approach ensures that minority classes are properly represented in the training data, leading to more balanced and fair predictions. On the other hand, the challenge of feature reduction is addressed through a feature reduction model based on Rough Set Theory. This method strategically reduces the number of features while keeping critical information intact, helping to cut down on computational overhead and improve how data is interpreted. These improvements have the potential to significantly enhance diagnostics, treatment planning, and healthcare delivery overall.

The study also highlights the transformative role AI plays in diabetes management. From diagnosing and monitoring to treatment planning, machine learning techniques like KNN imputation, SMOTE for class balancing, and PCA for

dimensionality reduction have proven instrumental in addressing data challenges. By refining these techniques, AI-driven models can become even more effective at improving patient care and disease management strategies.

But it's not just about the technical advancements. The study also stresses the importance of ethical considerations and bias mitigation in automated data processing. Responsible AI usage is critical for building trust and maximizing the benefits of these tools in healthcare. Data quality assurance processes, such as automated data cleansing and ethical oversight, play a vital role in ensuring the fairness and reliability of AI models.

The advancement of AI-driven healthcare solutions hinges on tackling the obstacles of feature reduction , class imbalance, and missing data. Improving model performance and maintaining the applicability and interpretability of these tools in real-world medical applications are both achieved through the incorporation of modern preprocessing approaches.

Innovation in AI and improved patient outcomes will depend on continuous research and cross-disciplinary collaboration as technology continues to develop. All things considered, healthcare stands to gain from these endeavours, which represent a major advance in the use of AI to address the intricacies of illness prediction and management.

CHAPTER III

METHODOLOGY

3.1 INTRODUCTION

The methodological design of this study follows a structured and multi-phase approach that integrates several advanced techniques into a unified framework. These phases encompass data acquisition, preprocessing, imputation, rebalancing, dimensionality reduction, and classification. Each methodological phase is explicitly aligned with one of the research objectives to ensure a clear and logical progression from problem definition to empirical validation.

The first phase focuses on data preparation and preprocessing, where multiple publicly available diabetes datasets namely the Pima Indians Diabetes Dataset, Kaggle Diabetes Prediction Dataset, and Kaggle dataset are cleaned, normalized, and organized for analysis. The second phase deals with handling missing data using the proposed Bidirectional Neighbour Graph (BNG) imputation technique. This method leverages bidirectional data relationships to estimate missing values more accurately than traditional statistical approaches, thereby improving data completeness and consistency.

The third phase addresses the issue of class imbalance by employing the Clustering Selection Synthesis Filtering (CSSF) technique, which balances the dataset through an intelligent combination of clustering-based resampling and synthetic instance generation. The fourth phase targets feature reduction by applying Rough Set Theory (RST) for feature selection and dimensionality reduction, allowing the model to focus on the most informative features without sacrificing diagnostic precision.

Finally, the processed datasets are evaluated through classification experiments using ANN and SVM. The results of these experiments serve to validate the effectiveness of the integrated BNG–CSSF–RST framework in enhancing data quality and predictive accuracy. Performance metrics such as accuracy, precision, recall, F1-score, and AUC are employed to assess improvements across datasets and techniques.

Finally, this chapter provides a comprehensive description of the research design, data sources, and methodological processes that underpin the proposed framework. The systematic approach ensures that each methodological step contributes directly to achieving the study's objectives and sets the foundation for the experimental results and discussion presented in Chapter IV.

3.2 METHODOLOGICAL FRAMEWORK FOR IMPROVING DIABETES DATA QUALITY

The research design adopted in this study is quantitative, experimental, and model-driven, focusing on the empirical evaluation of techniques that improve the quality of diabetes dataset for diabetes prediction. The study follows a systematic multi-phase research design, where each phase is directly aligned with one or more of the research objectives defined in Chapter I. The methodological structure is divided into distinct yet interconnected stages: data acquisition and preparation, preprocessing, imputation, class balancing, dimensionality reduction, and predictive modeling with evaluation. The overarching design follows an incremental enhancement framework, in which each methodological phase addresses a specific data quality challenge before integrating its output into the next stage. This progressive approach ensures that data refinement is both sequential and cumulative, thereby improving the robustness and interpretability of the final predictive model.

- **Phase 1.** Data Collection and Preparation: Compilation of multiple real-world diabetes datasets, followed by data cleaning, normalization, and exploratory analysis.

- **Phase 2.** Missing Data Handling: Application of the proposed Bidirectional Neighbour Graph (BNG) algorithm to impute missing values by exploiting relational dependencies between data points.

- **Phase 3.** Class Imbalance Correction: Implementation of the Clustering Selection Synthesis Filtering (CSSF) approach to achieve balanced class representation.

- **Phase 4.** Dimensionality Reduction: Use of Rough Set Theory (RST) for feature selection to remove redundant or irrelevant variables while preserving essential diagnostic information.

- **Phase 5.** Model Training and Validation: Evaluation of the enhanced datasets using ANN and SVM, with cross-dataset comparisons to validate predictive reliability.

The process begins with data acquisition from multiple diabetes datasets (Pima (768), Kaggle 2023 (100k), and Kaggle 2024 (100k)) and proceeds through a series of structured phases. This design ensures methodological traceability between research objectives, data processing strategies, and experimental validation. Figure 3.1 presents the overall workflow of the proposed research framework, which integrates data preprocessing, imputation, balancing, and feature reduction into a coherent predictive modelling pipeline.

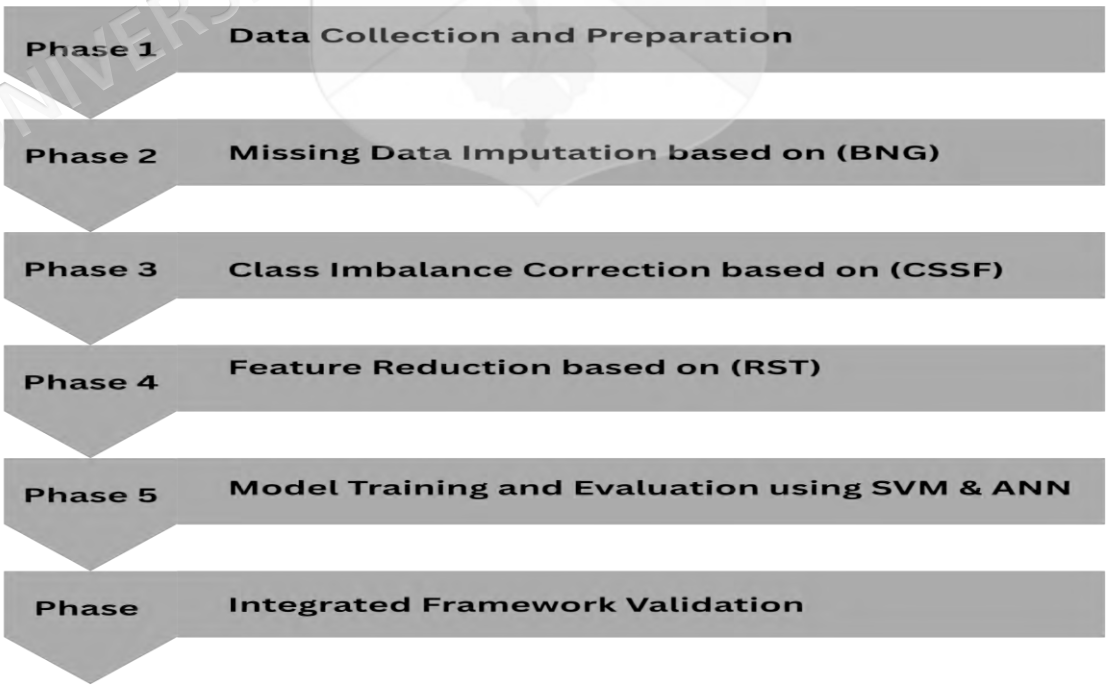


Figure 3.1 Demonstrates how the proposed research technique integrates data preparation, imputation, balancing, and feature reduction into predictive modelling

3.2.1 Data selection and preparation

Dataset selection for this study was based on sample size, variable richness, data completeness, and the presence of clearly defined class labels. Three publicly accessible diabetes datasets were chosen to evaluate the proposed BNG–CSSF–RST preprocessing framework. These datasets differ in demographic characteristics, data collection methods, and feature complexity, enabling a comprehensive assessment of the methodology across varied data environments.

The first dataset is the Pima Indians Diabetes Dataset (PIDD) obtained from the UCI Machine Learning Repository. It contains 768 records and 9 clinical features, primarily related to physiological and metabolic measurements such as glucose concentration, BMI, age, insulin level, skin thickness, and diabetes outcome. Though relatively small, PIDD is widely used for benchmarking and represents a well-structured dataset with minimal noise, making it suitable for evaluating baseline model behaviour.

The second dataset is the Diabetes Clinical Dataset (Kaggle, 2024), consisting of approximately 100,000 records with 17 predictive variables. These features include demographic data, medical history, lifestyle indicators, laboratory measurements, and other relevant clinical parameters. Compared to PIDD, this dataset is substantially larger and more heterogeneous, with a higher incidence of missing values and a more complex feature space. These characteristics make it suitable for analysing the performance of the proposed BNG imputation, CSSF balancing, and RST feature reduction techniques.

The third dataset used in this study is the Diabetes Prediction Dataset (Kaggle, 2023), containing around 100,000 instances and 9 key predictive features. The dataset includes variables such as age, gender, BMI, smoking history, HbA1c level, blood glucose level, and diabetes status. It exhibits a moderately imbalanced class distribution and a non-uniform feature profile that reflects real-world population diversity. This dataset allows evaluation of the framework under conditions of mixed data quality and variable consistency. Collectively, these three datasets encompass different population groups, variable complexities, and data quality characteristics. Their combined use

ensures that the assessment of the BNG–CSSF–RST framework is robust. Finally, regarding the suitability of our datasets for dimensionality reduction, we would like to clarify that the three datasets used in this study, though not high-dimensional, were selected based on their richness in predictive features and relevance to diabetes research. The Pima Indians Diabetes Dataset (PIDD) has 9 clinical features, while the other two datasets, the Diabetes Clinical Dataset and the Diabetes Prediction Dataset, contain 17 features each. While these datasets are not high-dimensional, they provide a diverse range of demographic, clinical, and lifestyle variables, making them suitable for evaluating the proposed BNG–CSSF–RST preprocessing framework. These datasets offer varied complexities and real-world data characteristics, ensuring that our analysis of the framework is comprehensive across different data environments.

3.2.2 Data Preprocessing and Integration

All three datasets underwent an initial preprocessing phase to ensure consistency, analytical readiness, and compatibility with the proposed BNG–CSSF–RST framework. The preprocessing workflow included data cleaning, normalization, and transformation into a unified analytical structure. Continuous variables were standardized using z-score normalization to eliminate scale discrepancies, while categorical attributes were encoded numerically to support machine learning model requirements. Duplicate records and extreme outliers were detected using interquartile range analysis and removed to preserve data integrity.

Each dataset was partitioned into training (70%) and testing (30%) subsets using stratified random sampling, ensuring proportional representation of diabetic and non-diabetic cases. The datasets were then processed independently through the three core modules: BNG for imputing missing values, CSSF for addressing class imbalance, and RST for dimensionality reduction and feature selection. After completing these steps, the datasets were integrated into a unified experimental framework to enable cross-dataset performance comparison.

The use of heterogeneous datasets strengthens the external validity of this study by demonstrating that the proposed approach performs reliably across varied data

structures, population characteristics, and noise levels. The Pima dataset serves as a benchmark for comparing with earlier studies due to its small size and standardized clinical attributes. The Kaggle Diabetes Clinical Dataset (2024) introduces a wider range of demographic and medical features, making it suitable for evaluating the framework under higher dimensionality and moderate missingness. The Kaggle Diabetes Prediction Dataset (2023) provides a large, diverse population sample with mixed-quality data, supporting evaluation of robustness in the presence of variable imbalance and heterogeneous feature distributions. Collectively, these datasets capture the full spectrum of data quality challenges targeted by the BNG–CSSF–RST methodology.

The methodological workflow for this research follows a structured, phased approach designed to ensure rigorous data handling and high analytical reliability. The process begins with data preparation, in which raw datasets are acquired, inspected, and formatted for processing. Data preprocessing is followed by data cleaning, which includes normalization, outlier removal, and duplicate elimination to generate a clean and consistent dataset.

To manage missing values, the BNG imputation technique is applied, enhancing dataset completeness. CSSF-based class balancing addresses skewed class distributions, preventing bias toward majority classes. In the feature engineering phase, Rough Set Theory (RST) is applied to identify the most relevant predictors and reduce dimensionality. The integration of BNG, CSSF, and RST constitutes the data quality enhancement stage, producing a refined dataset optimized for improved accuracy and generalisability.

During model development and selection, machine learning models, including Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs), are trained and optimized using the enhanced dataset. These models are then applied in the prediction phase to classify individuals as diabetic or non-diabetic, as illustrated in Figure 3.2.

Finally, model evaluation is performed using accuracy, precision, recall, F1-score, and other performance metrics to verify predictive capability and generalisability. This comprehensive and structured workflow ensures high-quality data preparation and robust predictive modelling, ultimately supporting reliable diabetes prediction outcomes. Table 3.1 summarizes the methodology presented in Chapter III.

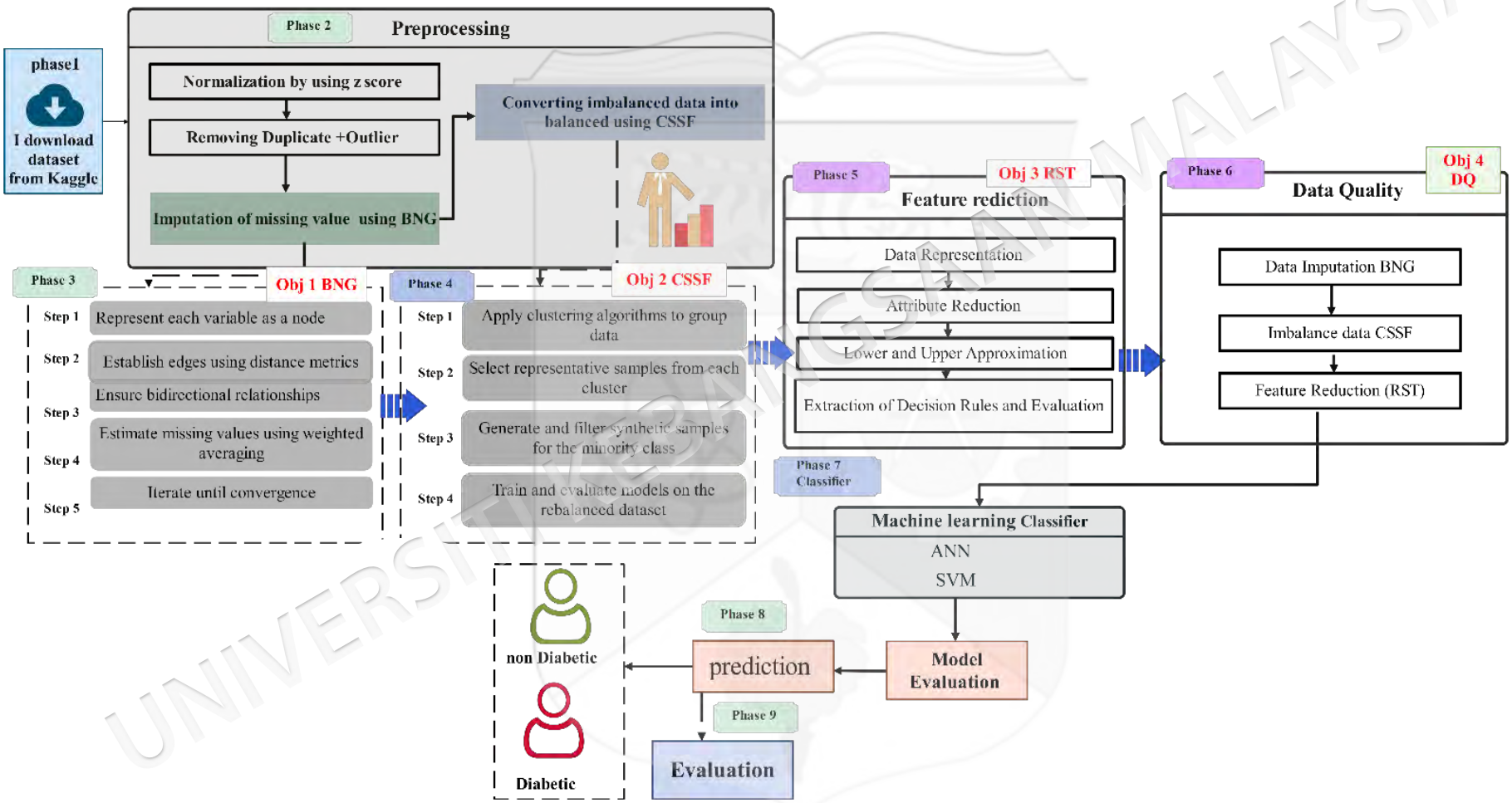


Figure 3.2 Research Methodology Phases Flowchart

Table 3.1 Summary of the methodology.

Phase	Description	Key Activities	Outputs
1. Data Preparation	Preparing the raw dataset for processing	- Dataset Acquisition	Prepared dataset ready for pre-processing
2. Pre-processing	Enhancing data quality and consistency	- Normalisation using z-score - Duplicate removal - Outlier management	Cleaned and normalised dataset
3. Data Imputation	Addressing missing data in the dataset	- Applying Bidirectional Neighbour Graph (BNG) for imputation	Dataset with missing data filled
4. Class Balancing	Rectifying imbalance data to prevent model bias	- Implementing Clustering Selection Synthesis Filtering (CSSF)	Balanced dataset for unbiased model training
5. Feature Engineering	Determining relevant features for prediction	- Feature selection with Rough Set Theory (RST)	Reduced feature set highlighting impactful predictors
6.Data quality	Improving data quality	By handling BNG, CSSF, RST	Enhances accuracy through data imputation, class balancing, and feature selection
7. Model Development and Selection	Creating and selecting predictive models	- Testing classifiers like ANNs and SVMs - Model selection based on performance metrics	Chosen predictive models for evaluation
8. Prediction	Using models to predict and evaluate outcomes	- Deployment of models for prediction - Accuracy evaluation of model predictions	Validated model predictions for diabetic or non-diabetic classification
9. Model Evaluation	Assessing model performance	- Evaluation using metrics like accuracy, precision, recall, F1 score	Assessment of model prediction accuracy and generalisability

Using the diabetes dataset as the foundation for training and testing machine learning models, this research aims to develop predictive models for diabetes diagnosis. By leveraging the dataset's rich features, this study seeks to enhance the accuracy and robustness of disease prediction, thus contributing to improved healthcare outcomes and patient care. Three publicly accessible diabetes datasets were selected for this research based on data quality, sample size variation, richness of predictive variables, and clearly defined class labels. The Pima Indians Diabetes Dataset (PIDD) provides a well-established benchmark derived from a specific population group with standardized clinical measurements such as glucose level, BMI, insulin, and age. Its structured and relatively noise-free composition makes it suitable for baseline comparison and validation of preprocessing techniques.

The second dataset, the Kaggle Diabetes Clinical Dataset (2024), consists of approximately 100,000 records and 17 predictive features, covering demographic, physiological, and clinical attributes. This dataset is substantially larger and more diverse than PIDD, containing higher levels of missing data and greater variation across medical and lifestyle factors. Its complexity makes it appropriate for evaluating the performance of the BNG, CSSF, and RST components under more challenging and realistic conditions.

The third dataset, the Kaggle Diabetes Prediction Dataset (2023), includes around 100,000 instances with 9 key predictors, such as gender, age, BMI, hypertension, heart disease, smoking history, HbA1c level, blood glucose level, and diabetes status. This dataset reflects a broad population distribution and exhibits moderate class imbalance, with non-diabetic cases forming the majority class. Its heterogeneity and mixed data quality provide an additional test of the robustness of the proposed framework.

The use of these three datasets enables the BNG–CSSF–RST methodology to be evaluated across diverse population demographics, dataset sizes, feature complexities, and data quality conditions. This comprehensive selection enhances the reliability, adaptability, and external validity of the findings, supporting the framework’s applicability to real-world diabetes prediction scenarios.

The dataset provides valuable information on various health parameters and their potential relationship with diabetes. Here is a detailed breakdown of the data columns:

1. Gender: The dataset categorises individuals as either 'Male' or 'Female'. This helps in understanding the gender distribution among the subjects.

2. Age: The age column records the age of each individual. This is crucial for analysing age-related trends and the prevalence of diabetes across different age groups.

3. Hypertension: This binary column indicates whether an individual has hypertension (high blood pressure) or not. '1' denotes the presence of hypertension, while '0' indicates its absence. Hypertension is a known risk factor for diabetes and cardiovascular diseases.

4. Heart Disease: Like hypertension, this column indicates the presence ('1') or absence ('0') of heart disease in the individuals. Heart disease is another critical factor often associated with diabetes.

5. Smoking History: This column provides the smoking history of each individual. Categories include:

    a. 'never': The individual has never smoked.
    b. 'current': The individual is a current smoker.
    c. 'former': The individual used to smoke but has quit.
    d. 'No Info': Smoking history is not available.
    e. 'ever': The individual has smoked at some point in their life.
    f. 'not current': Similar to 'former', indicating past smoking habits.

6. BMI (Body Mass Index): This numeric column records the BMI of each individual. BMI is a measure of body fat based on height and weight, and it is an important metric for assessing obesity, which is a significant risk factor for diabetes.

7. HbA1c Level: This column measures the HbA1c level, which indicates average blood glucose levels over the past 2 to 3 months. Higher HbA1c levels are indicative of poor blood sugar control and are a diagnostic marker for diabetes.

8. Blood Glucose Level: This column records the current blood glucose level of each individual. Monitoring blood glucose levels is essential for diabetes management and diagnosis.

9. Diabetes: This binary column indicates whether the individual has diabetes ('1') or not ('0'). This is the primary outcome variable in the dataset. As show in Table 3.2 diabetic data samples

10. While the explanations in this section are presented in general terms to show how widely these methods can be applied, the actual experimental work relies on three different diabetes datasets that are publicly available. We chose the PIMA Indian Diabetes Dataset (PIDD), which has 768 cases and 8 features, as it serves as a well-known benchmark in diabetes prediction research. The Kaggle Diabetes Prediction Dataset is considerably larger with 100,000 cases and 17 different features, giving us much more data to work with. Finally, the Kaggle (2024) Dataset is the most extensive, containing over 100,000 records and 22 different attributes that capture a wide range of health and behavioral factors. Using these three datasets together was deliberate. They range from a small clinical dataset (PIDD with 768 cases) to a massive population-level survey (Kaggle with 100,000 cases), which means we can test our methods across different scales and contextsThe proposed methodological framework incorporates three techniques Bidirectional Neighbour Graph (BNG) for handling missing data, Clustering Selection Synthesis Filtering (CSSF) for addressing class imbalance, and Rough Set Theory (RST) for feature reduction. These techniques are not confined to one dataset but are broadly applicable across diverse diabetes dataset, as demonstrated in related literature. While only the Kaggle dataset was used for primary experimentation, the methods and findings are designed with generalisability in mind, ensuring that the contributions extend beyond a single dataset. This approach balances feasibility and scope while still reinforcing the robustness,

Table 3.2 Diabetic Data Samples

gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Female	80	0	1	never	25.19	6.6	140	0
Female	54	0	0	No Info	27.32	6.6	80	0
Male	28	0	0	never	27.32	5.7	158	0
Female	36	0	0	current	23.45	5	155	0
Male	76	1	1	current	20.14	4.8	155	0
Female	20	0	0	never	27.32	6.6	85	0
Female	44	0	0	never	19.31	6.5	200	1
Female	79	0	0	No Info	23.86	5.7	85	0
Male	42	0	0	never	33.64	4.8	145	0
Female	32	0	0	never	27.32	5	100	0
Female	53	0	0	never	27.32	6.1	85	0
Female	54	0	0	former	54.7	6	100	0
Female	78	0	0	former	36.05	5	130	0
Female	67	0	0	never	25.69	5.8	200	0
Male	67	0	1	not current	27.32	6.5	200	1
Male	37	0	0	ever	25.72	3.5	159	0

a. Pre-processing

Preprocessing is a crucial analytical procedure aimed at refining the data quality for subsequent modelling. It is the foundation upon which reliable predictions for diabetes are built. In the realm of data preprocessing, normalisation and standardisation are pivotal techniques that significantly enhance the quality and effectiveness of predictive modelling. These processes ensure that each feature contributes equally to the model's predictive power, thereby preventing biases and improving the accuracy of the model.

b. Normalisation/ Standardisation: Equilibrium in Data Scaling

Normalisation and standardisation address the challenge of disparate feature ranges in datasets. Without these techniques, features with larger ranges could dominate those with smaller ranges, skewing the model's predictions. For instance, in the diabetes dataset, features such as blood glucose level ranges from 80 to 300, while hypertension is a binary feature (0 or 1). Normalising and standardising these features ensure that each one has an equal opportunity to influence the model's outcomes. Normalisation and standardisation are pivotal techniques in data preprocessing. They provide a level playing field for all variables, regardless of their original units or scales, enabling fair comparison and influence in the subsequent predictive modelling. Normalisation is a scaling technique that adjusts data dimensions to a standard scale, typically within the range of 0 to 1, (see Figure 3.3). The formula for normalisation is expressed as:

$$\textbf{Normalized Value} = \frac{\textbf{Value} - \textbf{Minimum Value}}{\textbf{Maximum Value} - \textbf{Minimum Value}} \quad \ldots(3.1)$$

Where:

- **Value**: The specific data point to be normalised.
- **Minimum Value**: The smallest value in the dataset for the feature being normalised.
- **Maximum Value**: The largest value in the dataset for the feature being normalised.

For example, in the diabetes dataset, the blood glucose level ranges from 80 to 300, while hypertension is a binary feature. Normalising the blood glucose level (e.g., 140) results in:

$$\textbf{Normalised Blood Glucose Level} = \frac{140-80}{300-80} \approx 0.273 \quad \ldots(3.2)$$

This transformation ensures that all features lie within the same range, facilitating fair comparisons and improving the performance of machine learning algorithms.

Standardisation, or Z-score normalisation, transforms data to have a mean (μ) of zero and a standard deviation (σ) of one. This method is particularly useful when features follow a Gaussian distribution. The formula for standardisation is:

$$\mathbf{Z-Score} = \frac{\mathbf{Value}-\boldsymbol{\mu}}{(\sigma)} \quad \ldots(3.3)$$

where:

- **Value**: The specific data point to be standardised.
- **μ (Mean)**: The average value of the dataset for the feature being standardised.
- **σ (Standard Deviation)**: The measure of the amount of variation or dispersion of the dataset for the feature being standardised.

By converting features to the same scale, standardisation mitigates the risk of one feature disproportionately influencing the model due to its variance.

For a glucose level of 110, with a mean (μ) of 100 and a standard deviation (σ) of 20, the standardised value would be:

$$\mathbf{Z} = \frac{110-100}{20} = \frac{10}{20} = 0.5$$

Pseudocode 1: Normalization using z- score (Little & Rubin 2019)

Input:

- $\mathcal{D}$: Dataset comprising m items, each with n features

Output:

- Normalised dataset

Steps:

1 For each feature i in the dataset:

- Calculate the mean μ_i of the feature.
- Calculate the standard deviation σ_i of the feature.

2 For each item j in the dataset:

- Normalise the feature value using the formula: $z_{ij} = \frac{x_{ij}\mu_i}{\sigma_i}$

3 Return the normalised dataset.

Figure 3.3 Pseudocode 1: Normalisation using z- score.

1. Load the Dataset: First, the dataset is loaded into a data frame using Pandas.

2. Cleaning: This sub-phase involves removing duplicate entries that could skew analysis and dealing with outliers representing errors or exceptional, non-representative data points. To clean a diabetic dataset using Python, we will go through the steps of removing duplicate entries and handling outliers. Data cleaning is the initial and crucial step in preprocessing. It involves identifying and rectifying data errors or inconsistencies to improve quality. One of the critical activities in data cleaning involves removing duplicate records. Duplicate entries can skew data analysis, leading to inaccurate results. Data cleaning involves identifying and eliminating these duplicates to ensure each data point is unique and representative. Additionally, handling outliers is also part of the data cleaning process. Outliers are data points that significantly deviate from the norm. While some outliers may represent actual variance and hold valuable information, others might be errors and can distort statistical

analyses. The process involves assessing outliers to decide whether they should be adjusted, removed, or kept based on their relevance to the research objectives. Cleaning a diabetes dataset using Python involves loading the dataset, removing the duplicate entries and handling outliers.

3. Remove Duplicate Entries: Duplicates can be removed using Pandas' drop duplicates method.

4. Handle Outliers: Outliers in the dataset are managed using the **Z-Score method**, a statistical technique for identifying and removing extreme values that deviate significantly from the mean. This method involves calculating the mean and standard deviation of the feature, identifying outliers as data points with significantly high or low Z-Scores, and removing them. By eliminating these outliers, the dataset is cleaned, preventing distortion in analysis and improving its reliability and accuracy for further modeling. This ensures the data remains of high quality and integrity throughout the preprocessing phase., as shown in Figure 3.4 below:

Pseudocode 2: Remove Duplicates and Outliers (Sezer & Başeğmez 2021)

Input: Dataset D comprising m items
Output: Cleaned dataset $\bar{D}_{\text{clean}}$ without duplicates and outliers
Steps:

- Step 1: Identify duplicate entries in the dataset D
- Step 2: Remove duplicate entries from the dataset D
- Step 3: Identify outliers using statistical methods.
- Step 4: Remove outliers from the dataset D
- Step 5: Return the cleaned dataset D_{clean}

Figure 3.4 Pseudocode 2 for (Cleaning Data) Removing Duplicate Entries

c. Duplicate removal and outlier detection in data pre-processing

Secondary extraction and outlier detection are important steps in data pre-processing to ensure the quality and reliability of data sets used for analysis and modelling, especially in healthcare applications.

Duplicate removal Duplications in health databases are often the result of multiple entries, which can lead to misdiagnosis, misinterpretation of data, and which the result is incorrect. To address this, the Python Pandas library provides efficient duplicate management tools. The duplicated() function identifies rows in the data set, while the drop_duplicates() function effectively removes them, resulting in a cleaner and more reliable data set

Ensuring that each observation in the data set is unique is important to ensure data integrity and accuracy. Duplicate data can distort research results, leading to inaccurate data and unreliable insights. Examination of the validity of the extraction method in this study showed that, prior to repair, the dataset contained 38,854 duplicated sequences. After applying the duplicate removal procedure, there were no duplicates in the data set, which increased the accuracy and reliability of the data.

d. Expression Identification and Removal

Extravagance or extreme values can significantly influence the analysis by distorting the statistical measure and obscuring meaningful observations. Factors such as age, BMI, HbA1c levels, and blood glucose levels are particularly sensitive to extremes, which can skew results and reduce the accuracy of predictions

In this study, the interquartile range (IQR) method was used to identify and systematically remove outliers. The process involves calculating the first quarter (Q1, or 25 percent) and the third quarter (Q3, or 75 percent). The IQR is then calculated as IQR=Q3−Q1IQR=Q3−Q1. The outer boundaries are defined as follows.

Lower range: Q1−1.5×IQRQ1−1.5×IQR
Upper limit: Q3 + 1.5 × IQRQ3 + 1.5 × IQR

Data points that fall outside these boundaries are flagged as outliers and are removed to ensure that the dataset accurately reflects the underlying pattern without distorting the extreme values.

The validation process demonstrated the effectiveness of this approach. Initially, the data set consisted of 96,146 rows. After outlier removal, 7,951 rows were identified as outliers and removed, leaving 88,195 rows in the clean dataset. This approach significantly improved the dataset, and ensured that it was worthy of further analysis.

By handling input and dual exposure, the pre-processing data steps described here drive the integrity, accuracy, and interpretability of the data structure effective These improvements form the basis for reliable analysis and robust predictive models.

e. Data Imputation

Missing data is a common issue in real-world datasets that must be addressed before analysis. The Bidirectional Neighbour Graph technique intelligently fills in missing values, considering the data's structure and the relationships between points. The Bidirectional Neighbour Graph method for imputation is an advanced technique used to estimate missing data in datasets by exploiting the underlying structure and relationships between data points. Here is an overview of the BNG method, including its conceptual basis and the steps involved, with relevant equations. The BNG algorithm calculates similarity using Euclidean distance: $d(i,j) = \sqrt{\sum_{Lh} (x_{ik} - x_{ik})^2}$). This distance is used to form edges between nodes (data points) only when it falls below a defined threshold. To ensure bidirectionality, an edge between two nodes is included in the graph only if both nodes identify each other as neighbors the connection from node i to node j is reciprocated by a connection from j to i. This symmetric neighbor check ensures that all edges represent mutual similarity, preserving the integrity of the imputation graph structure.

BNG is based on the construction of a graph where each data point or variable is represented as a node. The edges between nodes represent the relationships or similarities between these points. Unlike unidirectional graphs, BNG emphasises the importance of bidirectional relationships, meaning that the influence between two nodes is considered in both directions, enhancing the understanding of their mutual relevance. The BNG method leverages the concept that a data point can be effectively predicted or imputed based on the values of its neighbours within the graph, with the bidirectional nature providing a more comprehensive view of neighbour influences.

f. Conceptual Basis that have:

a. Neighbour Selection:

The imputation process begins with the identification of neighbours for a node or data point with missing data. Neighbours are selected based on the strength of the connections or edges between them, which are quantified using a similarity measure. This measure reflects how closely related or similar the nodes are to each other in the context of the dataset.

b. Bidirectional Weighting:

In the BNG approach, the weight $\boldsymbol{w_{ij}}$ of an edge between two nodes i and j is determined considering the similarity from both directions. This bidirectional weighting ensures that the imputation process comprehensively accounts for the influence of each node on the other, acknowledging that the relationship is not merely one-way but reciprocal. The bidirectional weight between nodes i and j encapsulates the degree of similarity and interdependence, thereby ensuring a more nuanced and accurate estimation of the missing values.

c. Graph Representation:

Let $G \equiv (V, E)$ be a graph where:

- V is the set of nodes representing the data points (e.g., individuals in the dataset).
- E represents the edges indicating relationships between these points.

Example: Consider a subset of the dataset with three individuals (nodes): V_1, V_2, and V_3.

- $V = \{V_1, V_2, V_3\}$
- $E = \{(V_1, V_2), (V_2, V_3)\}$

i. Similarity Calculation:

The similarity or connection strength between two nodes v_i and v_j is calculated, forming the basis for the edges. This can be quantified using metrics such as correlation, distance, or mutual information, denoted as $\text{sim}(v_i, v_j)$.

Example: Assume we calculate the similarity using Pearson correlation based on attributes like age, BMI, and blood glucose levels. Let's say the correlations are:

- $\text{sim}(V_1, V_2) = 0.9$
- $\text{sim}(V_2, V_3) = 0.7$

ii. Edge Weights:

Each edge in the graph is assigned a weight w_{ij} that reflects the strength of the relationship or similarity between nodes v_i and v_j :

$$w_{ij} = \text{sim}(v_i, v_j) \quad \ldots (3.4)$$

Example: The edge weights are:

- $w_{V_1V_2} = 0.9$
- $w_{V_2V_3} = 0.7$

iii. Understanding Imputation Process:

For a node v_i with missing value, the imputation is performed by aggregating the values of its neighbouring nodes, weighted by the edge strengths. The imputed value $\hat{v}_i$ is calculated as:

$$\hat{v}_i = f\left(\sum_{v_j \in N(v_i)} w_{ij} \cdot v_j\right) \quad \ldots (3.5)$$

where $N(v_i)$ is the set of neighbours of v_i, and f is an aggregation function, such as weighted average or sum.

Example: Suppose node V_1 has a missing BMI value. We want to impute $\hat{V}_1$ using its neighbours.

- Neighbours of V_1 are V_2 :
- BMI at $V_2 = 28$

Using weighted average as the aggregation function:

$$\hat{V}_1 = \frac{\sum_{v_j \in N(V_1)} w_{V_1 j} \cdot v_j}{\sum_{v_j \in N(V_1)} w_{V_1 j}} \quad \ldots(3.6)$$

Where $\hat{V}_1$ is the imputed value for the missing attribute of node V_1, while $N(V_1)$ denotes the set of neighboring nodes of V_1 in the bidirectional graph, v_j is the known value of the same attribute for neighbor node v_j, and $w_{V_1 j}$ is the edge weight (or similarity score) between node V_1 and its neighbor v_j. Higher weights represent stronger similarity or closeness. This formula calculates a weighted average of values from all neighbors of V_1, where the contribution of each neighbor is proportional to its similarity with V_1. By incorporating multiple neighbors and weighing their values accordingly, the imputation process becomes more robust and context-sensitive. This ensures better handling of noisy or incomplete data and reflects the key advantage of using a Bidirectional Neighbour Graph for imputation.

This example outlines the steps to construct a graph representation of data, calculate similarity, assign edge weights, and perform imputation based on neighbouring nodes' values using the diabetes dataset. BNG utilises both direct and indirect relationships, providing a holistic and comprehensive view of the data's structure, which helps in making more accurate imputations. The method is robust to various types of data structures and can handle complex interdependencies more effectively than simpler imputation methods. Furthermore, by considering bidirectional influences, BNG can achieve more accurate imputation results, especially in datasets where inter-variable relationships are significant.

The Bidirectional Neighbour Graph method is a sophisticated approach to missing value imputation that leverages the complex and bidirectional relationships between data points to provide accurate and reliable estimations. Through the construction of a graph that represents these relationships and the application of weighted aggregation based on Neighbour influences, BNG offers a powerful tool for addressing missing data in various research domains, particularly in complex datasets like those found in healthcare research.

Subsequently, the preprocessing stage focuses on preparing the collected data for analysis. This includes tasks such as data cleaning, normalisation, outlier detection, and missing value imputation using the Bidirectional Neighbour Graph (BNG)-based method. By applying this approach, the methodology ensures the integrity and quality of the data, minimising biases and inaccuracies that could potentially affect the research outcomes. The BNG-based missing data imputation method is specifically designed to handle missing data in diabetes dataset, providing an effective solution for addressing this common challenge in data analysis. With the utilisation of appropriate techniques and methodologies, the preprocessing stage plays a crucial role in optimising the data quality, thus establishing a strong foundation for subsequent analyses and modelling in the research process. The framework design stage entails the development of a conceptual framework or model that serves as the basis for analysis. This stage involves defining the key variables, their relationships, and the underlying theoretical constructs that guide the research. The framework design is crucial for providing a structured approach to addressing the research questions and facilitating interpretation of the results.

3.2.3 Missing Data Handling based on Bidirectional Neighbour Graph (BNG)

The Bidirectional Neighbour Graph (BNG) method is a robust solution for addressing missing data by constructing a graph where nodes represent data points and edges define relationships based on similarity or distance metrics. Unlike conventional methods, BNG incorporates both forward and backward neighbors, offering a more comprehensive understanding of data connections. This bidirectional approach enhances predictive accuracy by utilizing information from multiple neighboring points

and capturing complex data patterns.In summary, BNG ensures accurate, reliable, and context-aware imputation, making it highly effective for complex datasets.

a. Comprehensive Data Utilisation

BNG leverages both direct and indirect relationships within the dataset, offering a detailed representation of data interconnections which aids in more accurate imputation of missing values.

b. Flexibility and Robustness

BNG's method adapts well to various data structures and complexities, demonstrating a strong capability to handle intricate data interdependencies effectively compared to simpler imputation techniques.

c. Enhanced Accuracy

By considering the bidirectional influences between data points, BNG aims to provide more precise imputation outcomes, making it particularly useful for complex datasets like those in healthcare research.

To implement the Bidirectional Neighbour Graph (BNG) method for imputing missing data in a dataset, the following steps in pseudocode format provide a conceptual framework for how the code should work in Python, as shown in Figure 3.5: Algorithm 3, imputation of missing data using BNG.

pseudocode 3: Imputation of Missing data using BNG.

Input: Dataset D with missing values

Output: Dataset with imputed missing values

Steps:

Step 1: For each variable in the dataset, create a node representing the variable.

Step 2: For each pair of nodes, calculate the distance between nodes using a chosen distance metric.

Step 3: Create an edge between nodes if the distance is within a threshold.

Step 4: For each edge in the graph, ensure the relationship is bidirectional.

to be continued…

…continuation

Step 5: For each issing value in the dataset:
Step 5.1: Identify related nodes and their values.
Step 5.2: Estimate the missing value using the weighted average of related nodes.
Step 5.3: Replace the missing value with the estimated value.
Step 6: Repeat steps 5.1 to 5.3 until the change in values is below a threshold.

Figure 3.5 pseudocode Imputation of Missing data using BNG.

The algorithm indicated provides a structured approach to implementing the Bidirectional Neighbour Graph method for missing value imputation in Python. It outlines the main functions and steps required to perform the imputation, providing a clear guide for the actual coding process. The BNG algorithm significantly improves missing data imputation by constructing reliable neighborhood relationships in a task-driven metric space, minimizing the impact of missing values on distance measurement. Unlike simpler methods, BNG incorporates bidirectional relationships through the fusion of spatiotemporal graph sequences, capturing temporal correlations comprehensively. This approach enhances imputation accuracy, making it particularly suitable for diabetic datasets, where preserving the intrinsic structure and relationships within the data is critical for predictive modeling.

3.2.4 Class Imbalance Correction based on Clustering Selection Synthesis Filtering (CSSF)

Imbalance data in datasets, particularly in the medical field, is not just a technical challenge but a potential threat to healthcare outcomes. This issue is especially pronounced in datasets concerning diseases like diabetes, where the distribution of samples between classes (diabetic vs. non-diabetic) is often skewed. Such imbalance can severely affect the performance of predictive models, leading to inaccuracies in diagnosis and prognosis, potentially compromising patient care.

In diabetes dataset, the minority class, which often represents the presence of the disease is crucial for accurate predictions and effective treatment planning. However, underrepresentation of the minority class compared to the majority (non-disease) class can lead to models that are biased towards predicting the majority class, thus overlooking critical cases of the disease.

It is essential to address this imbalance to improve the accuracy of predictive analytics in healthcare. Techniques such as Cluster-based Synthetic Selection Filtering (CSSF) have been developed to tackle this issue. CSSF enhances the quality of synthetic selection through advanced clustering and filtering techniques, aiming to balance the class distribution in datasets. This method not only helps improve classification accuracy but also ensures that the predictive models are reliable and can effectively assist healthcare providers in early diagnosis and treatment planning.

CSSF builds upon the foundation laid by the Synthetic Minority Over-sampling Technique (SMOTE), strategically generating synthetic samples within clusters of the minority class while eliminating noisy instances. This approach has been shown to significantly improve the predictive model's performance, making it a valuable tool in managing imbalance data in diabetes dataset, particularly for diabetes classification.

Cluster-based Synthetic Selecting Filtering (CSSF) is a cutting-edge algorithm uniquely designed to tackle the complexities of imbalanced datasets. By merging clustering techniques with synthetic data generation, CSSF creates a more balanced distribution of classes within the dataset. Figure 3.6 below illustrates the CSSF methodology framework for addressing the class imbalance issue.

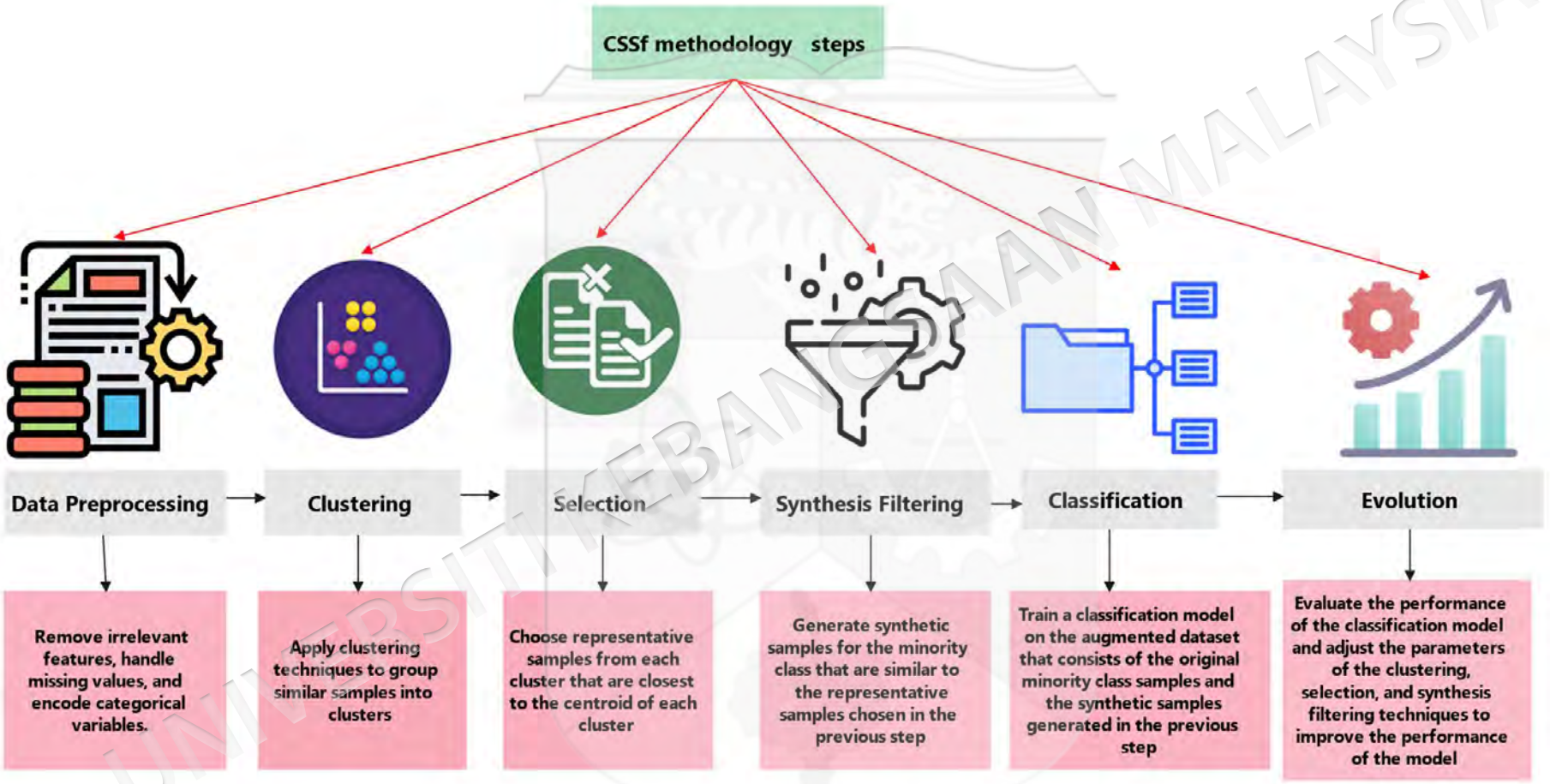


Figure 3.6 CSSF Methodology Framework for Addressing the Class Imbalance Issue

CSSF begins by identifying clusters within the minority class of the dataset. This step involves applying a clustering algorithm (K-means) to group the minority class instances based on their similarity. The goal is to understand the underlying structure and distribution of the minority class to generate synthetic samples that are representative of this structure.

Next, once the clusters are identified, CSSF generates synthetic samples within these clusters. The synthetic samples are created in a way that they closely resemble the actual minority class instances, maintaining the intrinsic properties and variance of the data within each cluster. This is achieved by using techniques like interpolation between points within the same cluster or applying variations based on the cluster's statistical properties.

Once the synthetic samples are generated, CSSF applies a meticulous filtering process. This step rigorously assesses the synthetic samples against a set of criteria to ensure their utility and relevance for the model training process. Samples that deviate significantly from the original data points or that do not contribute positively to the learning process are swiftly eliminated. This stringent filtering guarantees that only high-quality synthetic samples are incorporated into the dataset, averting the introduction of noise, and safeguarding the integrity of the minority class. The synthetic selections that successfully meet the filtering criteria are then utilised to bolster the minority class in the dataset. This crucial step rectifies the class distribution, ensuring that the learning algorithm receives a more equitable representation of both classes during training, thereby mitigating the impact of dataset imbalance. CSSF is a methodical approach that enhances the quality and representativeness of synthetic selecting s in imbalanced datasets. By focusing on clustering within the minority class, generating high-quality synthetic selecting s, and filtering out less useful ones, CSSF helps in creating a balanced dataset that improves the performance of machine learning models on imbalanced data scenarios, as shown in Figure 3.7 below.

pseudocode 4: Convert unbalanced data to balanced using CSSF

Input

- Imbalanced dataset D

Output:

- Balanced dataset

Steps:

1 Step 1: Apply the chosen clustering algorithm on the dataset
2 Step 2: Assign data points to clusters based on clustering results
3 Step 3: For each cluster in the dataset:
- Step 3.1: Select representative samples from the cluster
4 Step 4: Identify the minority class in the dataset
5 Step 5: Generate synthetic samples for the minority class
6 Step 6: Filter out low-quality synthetic samples
7 Step 7: Add synthetic samples to the dataset to balance the classes
8 Step 8: Split the balanced dataset into training and testing sets
9 Step 9: Train chosen machine learning models on the training set
10 Step 10: Evaluate models on the testing set

Figure 3.7 pseudocode of CSSF

Clustering Selection Synthesis Filtering (CSSF) is a robust methodology designed to address the challenge of imbalanced datasets in machine learning. It systematically integrates clustering, representative sample selection, synthetic sample generation, and quality control to create a balanced dataset for improved model performance. Below is a concise summary of the methodology:

1. **Clustering**:

- A clustering algorithm, such as K-means, is applied to group data points into clusters based on similarity.
- Data points are assigned to their nearest clusters, enabling structured processing.

2. **Representative Sample Selection**:

- Key samples from each cluster, typically closest to the cluster centroid, are selected as representatives.
- These representatives ensure the synthesis process captures the dataset's structure.

3. **Minority Class Identification and Augmentation**:

- The minority class is identified, and synthetic samples are generated using techniques like SMOTE.
- These synthetic samples are designed to resemble real data points, augmenting the minority class without introducing bias.

4. **Filtering and Quality Control**:

- Low-quality synthetic samples are removed to maintain dataset integrity.
- This ensures that only reliable and high-quality samples are included in the final dataset.

5. **Balancing the Dataset**:

- High-quality synthetic samples are added to the dataset, achieving balance between minority and majority classes.
- This balanced dataset ensures fair representation of all classes.

6. **Splitting and Training**:

- The balanced dataset is split into training and testing sets.
- Machine learning models are trained on the balanced training set, ensuring unbiased learning.

7. **Evaluation**:

- Models are evaluated on the testing set using metrics like accuracy, precision, recall, and F1-score.
- This step validates the effectiveness of the balancing process.

CSSF addresses the critical issue of class imbalance, which often skews predictive performance. By integrating clustering and synthetic sample generation with

strict quality control, CSSF ensures improved model performance, data integrity, and unbiased predictions. This methodology is adaptable across various domains and datasets, making it a versatile solution for handling imbalanced data in machine learning.

3.2.5 Dimensionality Reduction: Use of Rough Set Theory (RST)

Diabetes dataset represent a convergence of numerous measurements and tests that, when leveraged effectively, can lead to ground-breaking predictive models, especially for conditions like diabetes. However, these datasets come bundled with the challenges of feature reduction —too many features can muddle predictive models, overshadowing genuine patterns with noise and leading to computational inefficiencies. Rough Set Theory (RST) emerges as a compelling solution to dissect this complexity, sharpening the focus on features that genuinely matter for diabetes prediction. The comparison of Rough Set Theory (RST)–based feature reduction with Principal Component Analysis (PCA) was intentionally selected due to PCA's widespread adoption as a baseline dimensionality reduction technique in medical and high-dimensional data analysis. PCA is commonly used to reduce feature space by maximizing variance through linear transformations; therefore, it provides a well-established reference point for evaluating alternative feature reduction frameworks. RST was chosen for comparison specifically because it follows a fundamentally different paradigm from PCA. Unlike PCA, which generates transformed components and assumes linear relationships, RST performs feature selection rather than feature transformation, preserving the original semantic meaning of attributes while identifying reducts based on dependency relationships. This distinction is particularly important in diabetes datasets, where interpretability of selected features is critical. Hence, we focused on comparing Rough Set Theory (RST) with Principal Component Analysis (PCA) as PCA is a widely used baseline method for dimensionality reduction, particularly effective in high-dimensional datasets. (Hernández-Guedes et al., 2023) However, PCA has limitations in capturing clinically relevant features, which RST addresses by focusing on classification significance. While PCA provided an initial framework, we acknowledge the value of comparing RST with other techniques such as Linear Discriminant Analysis (LDA), Random Forests, and neural-based methods like nPCA. Future research could extensively

evaluate RST against these alternative techniques to ascertain its relative advantages and broaden its applicability in the field.

a. Feature reduction in Diabetes dataset: The Challenge

The feature reduction challenge is analogous to searching for a needle in a haystack, whereby valuable insights are often buried under layers of redundant information. Diabetes dataset encompass many features, from physiological to lifestyle factors, which might have no significant predictive value. This surplus can cloud the authentic relationships vital for diagnosing diabetes, causing models to chase shadows—overfitting on the training data and faltering against new, real-world data.

b. Implementing RST for Feature Selection

RST is a principled approach that systematically reduces a dataset's complexity without prejudging data based on presumed distributions or memberships. It is implemented by first forming Indiscernibility Relations, where objects (patients) in the dataset are grouped based on attribute similarity, creating equivalence classes. Subsequently, these classes serve as the foundation for creating Approximation Spaces, which define the lower and upper approximations of target sets, such as diabetic classifications, and provide bounds for decision-making. By analysing dependencies between attributes, RST identifies the minimal subset of features, known as reducts, which are essential for classification, along with the core attributes that cannot be discarded without losing important information. Finally, the reducts are translated into actionable decision rules, offering interpretable insights for predicting diabetes.

c. Rationale Behind Using RST

The choice of RST is grounded in its inherent ability to manage uncertainty and provide a granular analysis without additional information. This is critical when dealing with medical data where inaccuracies, inconsistencies, and patient variability are common. RST's unique approach to discerning the essence of data ensures that only the most predictive features are selected, enhancing model performance and interpretability.

d. Timing of RST Application

The timing of the RST application is crucial as it is best situated after preliminary data cleaning and before the onset of model training. This preparatory phase lays the groundwork for all subsequent modelling efforts. This timing ensures that models are trained on a distilled set of features, reducing the risk of overfitting, and improving model robustness.

e. Operationalising RST in Research

RST is operationalised within a computational environment designed for data mining and machine learning, where its algorithms can process the complex diabetic dataset. This environment is typically supported by programming languages and tools capable of handling the computational demands of RST. The incorporation of RST into the research methodology for diabetes prediction is justified on several fronts including to preserve the integrity of the original dataset. It refrains from making presumptions and works with the data in its raw form, ensuring the authenticity of the analysis is preserved.

The incorporation of RST also streamlines the dataset by pruning unnecessary features, rendering the computational processes more efficient and manageable. Models built on RST-selected features are poised to exhibit improved predictive performance, primarily due to their focus on the most relevant features. Additionally, RST aids in deriving logical, comprehensible decision rules that contribute significantly to the transparency of the predictive models. In healthcare, the ability to explain a model's decision is invaluable. RST's approach has been empirically validated across numerous domains. Its effectiveness in feature selection has been consistently demonstrated, solidifying its position as a reliable method in data analysis. Aside from that, RST stands on a solid theoretical base, offering a structured and replicable approach to feature selection. This rigidity and methodical nature ensure that the process is reproducible and verifiable.

Through RST, this research meticulously filters the diabetic dataset to its most predictive elements. This methodology section has charted the theoretical and practical landscape of RST, explicating its application's how's, whys, when's, and where's and justifying its selection as a dimensionality reduction technique. As we advance to the application and results chapters, the efficacy of RST in enhancing the quality and precision of diabetes prediction models will be further showcased, promising not only academic advancement but also tangible benefits to patient care and medical diagnostics.

f. Rough Set Theory (RST):

RST might be applied in Python for feature selection, specifically in the context of diabetes data, we will analyse a hypothetical implementation by dividing it into parts and providing explanations for each step. This process entails initialising and training an RST model, performing data transformation using the chosen features, and assessing the modified data. In order to improve the interpretability and effectiveness of diabetes predictive models, the pseudocode systematically implements RST for feature selection, as shown in Figure 3.8.

1. Load Dataset:
   - Import necessary libraries.
   - Load dataset from CSV file.
   - Replace placeholder values with NaN.
   - Fill missing values with the mode of each column.

2. Split Dataset:
   - Split dataset into training and testing subsets.

3. Process Features:
   a. Encode Categorical Features:
      - Initialize encoder.
      - Apply encoder to training and testing data.
   b. Scale Numerical Features:
      - Initialize scaler.
      - Apply scaler to training and testing data.

4. Combine Features:
   - Concatenate encoded categorical and scaled numerical features for training and testing datasets.

5. Apply Rough Set Theory (RST):
   - Initialize RST model.
   - Form equivalence classes from processed training data.
   - Identify reducts (important features) using equivalence classes.

6. Transform Data:
   - Select and retain features based on identified reducts.
   - Apply feature selection to both training and testing datasets.

7. Output:
   - Display the shapes of final training and testing datasets.

Figure 3.8 Pseudocode RST for feature selection

This pseudocode presents a systematic approach to implementing Rough Set Theory (RST) for feature selection, aimed at improving the interpretability and performance of predictive models in the context of diabetes. The process begins with loading and preprocessing the dataset, where missing values represented as NaN are replaced, and any gaps in the data are filled using the mode. Next, the dataset is partitioned into training and testing subsets to facilitate model evaluation.

```
BEGIN
# Step 1: Load and Preprocess the Dataset
DATASET = LOAD DATA FROM 'path/to/diabetes_data.csv'
REPLACE 'No Info' WITH NaN IN DATASET
FOR EACH COLUMN IN DATASET:
   FILL MISSING VALUES WITH MODE OF COLUMN
NUMERICAL_COLUMNS = SELECT NUMERICAL COLUMNS IN DATASET
NORMALIZE NUMERICAL_COLUMNS USING Z-SCORE
# Step 2: Split the Dataset
TRAINING_DATA, TESTING_DATA = SPLIT DATASET INTO TRAINING AND TESTING SETS (80:20 SPLIT)
# Step 3: Address Class Imbalance Using CSSF
TARGET_COLUMN = 'Diabetic'
X_TRAIN = DROP TARGET_COLUMN FROM TRAINING_DATA
Y_TRAIN = SELECT TARGET_COLUMN FROM TRAINING_DATA
CSSF = INITIALIZE CSSF METHOD (E.G., SMOTE)
X_TRAIN_BALANCED, Y_TRAIN_BALANCED = CSSF.FIT_RESAMPLE(X_TRAIN, Y_TRAIN)
# Step 4: Feature Selection Using RST
RST_MODEL = INITIALIZE RST MODEL
EQUIVALENCE_CLASSES = RST_MODEL.FORM_INDISCERNIBILITY_RELATION(X_TRAIN_BALANCED)
LOWER_APPROX, UPPER_APPROX = RST_MODEL.CREATE_APPROXIMATION_SPACES(EQUIVALENCE_CLASSES)
SELECTED_FEATURES = RST_MODEL.IDENTIFY_REDUCTS(X_TRAIN_BALANCED, LOWER_APPROX, UPPER_APPROX)
X_TRAIN_FINAL = SELECT COLUMNS SELECTED_FEATURES FROM X_TRAIN_BALANCED
X_TEST_FINAL = SELECT COLUMNS SELECTED_FEATURES FROM TESTING_DATA
# Step 5: Train and Evaluate Machine Learning Model
SVM_MODEL = INITIALIZE SVM CLASSIFIER
SVM_MODEL.FIT(X_TRAIN_FINAL, Y_TRAIN_BALANCED)
Y_PRED = SVM_MODEL.PREDICT(X_TEST_FINAL)
ACCURACY = CALCULATE ACCURACY(TESTING_DATA[TARGET_COLUMN], Y_PRED)
PRECISION = CALCULATE PRECISION(TESTING_DATA[TARGET_COLUMN], Y_PRED)
RECALL = CALCULATE RECALL(TESTING_DATA[TARGET_COLUMN], Y_PRED)
F1_SCORE = CALCULATE F1 SCORE(TESTING_DATA[TARGET_COLUMN], Y_PRED)
# Output Metrics
PRINT "Accuracy:", ACCURACY,PRINT "Precision:", PRECISION,PRINT "Recall:", RECALL,PRINT "F1 Score:", F1_SCORE
```

Figure 3.9 Pseudocode BNG, CSSF,RST

The method combines three proposed methods—BNG (Bidirectional Neighbor Graph), CSSF (Clustering Selection Synthesis Filtering), and RST (Rough Set

Theory)—in a cooperative preprocessing framework to ensure a general path so constant And above all Selecting predictive features reduce dimensionality This integrated pipeline ensures data quality, accuracy in experiments, and improves model performance by providing a customized training and evaluation base The combined approach not only increases accuracy but also enables interpretability and reliability in predictive modeling.as shown in Figure 3.9.

3.3 MODEL TRAINING AND VALIDATION

In the context of diabetes prediction using a dataset, this phase involves testing different classification algorithms to determine the best-performing model based on various evaluation metrics accuracy, AUC (Area Under the Curve), precision, and recall, F-score. The methodology for classifier testing is as detailed below.

Step 1: Data Preparation

Dataset: The diabetes dataset contains features like age, hypertension, heart disease, BMI, HbA1c level, blood glucose level, gender, smoking history, and the target variable, diabetes status (0 or 1).

Splitting the Data: Split the dataset into a training set (80%) and a testing set (20%).

Step 2: Training and Testing ANNs

- **Model**: Train an Artificial Neural Network (ANN) on the training set.
- **Architecture**: The ANN might consist of an input layer, one or more hidden layers, and an output layer.
- **Activation Function**: Use an activation function like ReLU (Rectified Linear Unit).
- **Training**: Use backpropagation to adjust weights based on the error between predicted and actual values.
- **Equation for ANN node activation**:

$$a_i = f\left(\sum_j w_{ij} x_j + b_i\right) \quad \ldots(3.7)$$

where:

- a_i is the activation of neuron i,
- f is the activation function,
- w_{ij} are the weights,
- x_j are the input signals,
- b_i is the bias term.
-

Evaluation: After training, test the ANN on the testing set and calculate performance metrics: accuracy, AUC, precision, and recall.

Evaluation: After training, test the ANN on the testing set and calculate performance metrics: accuracy, AUC, precision, and recall.

Step 3: Training and Testing SVMs

Model: Train a Support Vector Machine (SVM) on the training set.

Kernel: Use a kernel function such as linear or RBF (Radial Basis Function).

Training: SVMs aim to find the hyperplane that best separates the classes in the feature space.

Equation for SVM decision function:

$$f(x) = \text{sign}(\sum_i \alpha_i y_i K(x_i, x) + b) \quad \ldots(3.8)$$

where:

- α_i : Lagrange multipliers
- y_i : Class labels
- K: Kernel function (e.g., linear, RBF)
- x_i : Support vectors
- x : Input feature
- b : Bias
- Evaluation: After training, test the SVM on the testing set and calculate performance metrics: accuracy, AUC, precision, and recall.

Step 4: Model Selection

- Comparison: Compare the performance metrics of the ANN and SVM models.

- Selection: Choose the model with the best performance based on accuracy, AUC, precision, and recall.

- The Area Under the Curve (AUC) is a crucial metric for evaluating a model's ability to distinguish between classes, especially in imbalanced datasets like those used in diabetes prediction. Selecting the model with the highest AUC ensures superior performance in differentiating between positive and negative cases, leading to more reliable predictions.

Phase 7: Evaluation

In this phase, the performance of the selected models is thoroughly evaluated. The chosen models are rigorously tested using various metrics to assess their prediction accuracy.

Accuracy Metric

The accuracy metric is a fundamental measure used to evaluate the performance of classification models. It represents the proportion of correctly classified instances out of the total number of instances in the dataset. In the context of diabetes prediction, accuracy indicates the model's ability to correctly classify individuals as either diabetic or non-diabetic.

To assess the accuracy of the proposed models, the dataset will be divided into training and testing sets. The models will be trained using the training set, and their performance will be evaluated by predicting the labels of the instances in the testing set. The accuracy achieved by each model will be calculated by comparing the predicted labels with the actual labels.

The accuracy metric provides a quantitative measure of the models' predictive capabilities. Higher accuracy indicates better performance in correctly classifying individuals, while lower accuracy suggests a need for further refinement or improvement. By utilising scientific creativity in developing and refining the classification models, we aim to achieve higher accuracy and enhance the effectiveness of diabetes prediction. Models are evaluated on the validation set using the following metrics:

- True Positives (TP): Correct positive predictions.
- True Negatives (TN): Correct negative predictions.
- False Positives (FP): Incorrect positive predictions.
- False Negatives (FN): Incorrect negative predictions.

Higher accuracy indicates better performance in correctly classifying individuals, while lower accuracy suggests a need for further refinement or improvement. By utilising scientific creativity in developing and refining the classification models, we aim to achieve higher accuracy and enhance the effectiveness of diabetes prediction.

Accuracy: The ratio of correctly predicted observations.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \ldots(3.9)$$

Precision: The ratio of correctly predicted positive observations to the total predicted positives.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad \ldots(3.10)$$

Recall: The ratio of correctly predicted positive observations to all observations in actual class.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad \ldots(3.11)$$

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad \ldots(3.12)$$

a. Prediction

Step objective: Use the carefully built and validated models from the previous steps to make predictions on new data. Apply learned patterns from training to new input to make predictions about the output.

Process: Apply the trained models to a dataset about whose diabetes status to the model is unknown. This dataset could either be independent testing data or fresh data gathered for the purpose of validation.

b. Accuracy Evaluation

After making predictions, it is crucial to evaluate the model's ability to predict accurately. This evaluation will involve comparing the model's predictions against the actual outcomes, where available.

Metrics: Various statistical metrics are used to evaluate the accuracy of the predictions, such as:

Accuracy: The proportion of total predictions that were correct.

Precision and Recall: Precision seeks to measure how many of the positive predictions are precise, while recall tries to measure the capability of the model to fish out all the positive instances.

F1 Score: A harmonic mean of precision and recall, providing a single metric to assess the balance between them.

This research will largely indicate the effectiveness of how the models will work in real settings, putting an emphasis on critically considering when these models are subjected to real evaluation before implementation in clinical practices. Besides, areas where models are not perfect or where changes and modifications are needed will be addressed.This stage is of such a critical nature as to translate theoretical research to practical applications to ensure that the developed models are not only theoretically sound but practically viable for making diabetes predictions regarding an individual's status. This validation is not only an accuracy test but also serves as a feedback loop to further refine and optimise the models in pursuit of more applicability for healthcare settings.

Table 3.3 provides a comprehensive and structured summary of the research methodology, divided into nine distinct phases The first phase of the study involved utilizing a publicly available dataset from Kaggle, which was downloaded and organized for analysis. and Preparation, Pre-processing, Data Imputation, Class Balancing, Quality of Data, Selection, Prediction, and Evaluation. This table enhances clarity by breaking down each phase into specific activities and outputs, offering a clear roadmap of the research process. It emphasises data quality and model performance through techniques like imputation, class balancing, and rigorous evaluation, ensuring robust and reliable datasets. Additionally, the table guides replicability by outlining the methodology in a detailed, step-by-step manner, aiding future research efforts. It highlights the importance of identifying research gaps, bridging theory and practice in the framework design, and fostering continuous improvement through regular evaluations and forward-thinking recommendations. By selecting the right ML models and designing a thoughtful conceptual framework in Phase 3, the data imputation process becomes more systematic, closely aligned with research objectives, and capable of delivering reliable datasets. This approach creates a solid foundation for the next phases, ensuring high-quality data for accurate modeming and analysis. Overall, it reflects a well-structured, credible, and progressive research methodology. In addition, Table 3.4 provides summary of the differences between classification algorithms and the prediction of a classification model.

Table 3.3 Summary of Each Stage of the Research Methodology

Phase	Stage
Phase 1: Data preparing and Preparation	Literature review and data collection - Review the literature to identify gaps in diabetes prediction. - Summarise articles to create a literature review. - Collect relevant research on theories and concepts. - Download the "Diabetes Prediction 100000" dataset from the web. - Preprocess the data, including cleaning and normalisation. - Conduct a pre-experiment to obtain initial results.
Phase 2: Pre-processing	Literature review and data collection (continued) - Preprocess the data, including cleaning and normalisation. - Conduct a pre-experiment to obtain initial results.
Phase 3: Data Imputation	Design of framework - Identify appropriate ML models and design a conceptual framework. - Define key variables, relationships, and theoretical constructs. - Associate the identified issues with the proposed framework. - Design the framework to address the research objectives and the problem statement. - Define the scope and aims of the research analysis. - Validate the framework's effectiveness in addressing the research objectives.
Phase 4: Class Balancing	Design of framework (continued) - Associate identified issues with the proposed framework. - Design the framework to address the research objective and problem statement.

to be continued…

… continuation

Phase 5: Feature selection	Implementation - Implement previous data quality methods for missing data, class imbalance, and feature reduction . - Implement the proposed model based on the designed framework. - Enhance the proposed method by tuning parameters. - Implement previous methods proposed method imputer and simple imputer. - Evaluate the performance of the implemented methods. - Recommend further improvements or modifications.
Phase 6: Quality of Data	Implementation (continued) - Implement previous methods proposed method imputer and simple imputer. - Evaluate the performance of the implemented methods.
Phase 7: Selection	Implementation (continued) - Recommend further improvements or modifications.
Phase 8: Prediction	Evaluation and analysis - Compare the results of the previous methods with the proposed model. - Evaluate the accuracy and performance of the proposed model. - Analyse the results and provide recommendations for future work.
Phase 9: Evaluation	Evaluation and analysis (continued) - Assess the performance of the previous methods and the proposed model. - Provide interpretation of the evaluation results. - Draw conclusions based on the evaluation and analysis.

Table 3.4 Summary of the difference between classification algorithms and predication of a classification model

algo	**Classification Algorithms**	**Prediction Algorithms**
Purpose	Categorise or classify data into predefined classes or categories	Estimate or predict a numerical value or outcome
Supervised Learning	Yes	Yes
Input Features	Learn patterns and relationships to assign class labels	Model the relationship with the target variable
Output	Discrete class labels or probability distribution over possible classes	Numerical value or time series forecast
Training Data	Labelled examples with known class assignments	Labelled examples with known target values
Outcome Type	Discrete (Categorical)	Continuous (Numerical)
Application	Classification tasks	Forecasting tasks

3.4 SUMMARY

Chapter III of this thesis explores the use of advanced data preparation techniques and machine learning models to improve the quality of diabetes data, beginning with the Bidirectional Neighbour Graph (BNG) method, which addresses missing data in the dataset. The Clustering Selection Synthesis Filtering (CSSF) method is then applied to balance the data, ensuring that predictive models are not biased towards the majority class. Rough Set Theory (RST) is employed to reduce dimensionality, focusing on significant features for accurate prediction. Subsequently, Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) are then used to model complex patterns through interconnected nodes. These models are trained, validated, and tested to ensure they are technically sound and clinically significant, providing reliable and accurate predictions. The objective was not to explore multiple classification models but to rigorously evaluate the predictive capacity of these algorithms using high-quality, preprocessed datasets. The structured methodology, with a focus on phases like Prediction and Evaluation, ensures the reliability and generalizability of the models, demonstrating that ANN and SVM are well-suited to achieving the study's primary.

The Model Evaluation phase evaluates these models' using metrics such as accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC-ROC). This chapter highlights the importance of data quality and well-considered methodologies in medical data analysis, setting the stage for future research that pushes the boundaries of disease prediction and management. The evidence-driven conclusions drawn from the models' evaluations provide a path forward for clinicians and researchers, guiding them towards data-informed strategies in diabetes care.

Future research should prioritise evaluating alternative imputation strategies, investigating class-imbalance remedies, and refining feature-selection procedures to improve the quality of predictive models. Future work should evaluate the framework on additional datasets to validate findings and should be co-designed with healthcare professionals to strengthen clinical relevance. Given the opacity of complex models (e.g., ANNs, SVMs), enhancing transparency and explainability is essential for clinical adoption. Integrating predictive models into existing clinical workflows—through interoperable pipelines, clinician-facing dashboards, and audit trails—can better support decision-making across diverse populations and data-quality regimes. Collectively, these efforts would enable external validation and, where supported, cautious generalisation beyond the studied dataset, thereby improving the practical effectiveness of predictive models

CHAPTER IV

RESULTS AND DISCUSSION

4.1 INTRODUCTION

This chapter sheds light on the experimental findings and analytical analysis of the suggested framework for improving the diabetic datasets used for diabetes prediction. Prior to applying two machine learning classifiers artificial neural network (ANN) and support vector machine (SVM), the framework incorporates three primary preprocessing components: bidirectional neighbour graph (BNG) for missing-value imputation, clustering selection synthesis filtering (CSSF) for class balancing, and rough set theory (RST) for feature reduction. Furthermore, it discussed how the BNG-CSSF-RST pipeline works by analysing how each preprocessing step affects data integrity and classification, and the performance measures accuracy.

In addition, we compare the ANN and SVM models to see how well they deal with different kinds of medical data, taking into consideration factors like dimensionality, completeness, and balance. The generalisability and robustness of the suggested strategy are further confirmed by further experiments that make use of other diabetic datasets. Finally, it examines how data-quality improvement directly affects the reliability of illness prediction models.

4.2 DEVELOPING FRAMEWORK TO ENHANCE THE QUALITY OF THE DIABETES DATASET

The primary aim of this research is to develop a comprehensive framework to enhance the quality of diabetes dataset for diabetes prediction. This framework incorporates three core preprocessing strategies, BNG for missing data imputation, CSSF for class

imbalance correction, and RST for dimensionality reduction. Each method targets a distinct data quality challenge while collectively contributing to improved model accuracy, stability, and generalisability. This section describes the theoretical foundations, implementation procedures, and observed effects of these techniques on data integrity and predictive performance.

4.2.1 Dataset Description and Preprocessing Framework

The research employed three publicly available diabetes-related datasets to evaluate the generalizability and analytical robustness of the proposed framework. The first, the Pima Indians Diabetes Dataset (PIDD) obtained from the UCI Machine Learning Repository. It contains 768 records and 9 clinical features, primarily related to physiological and metabolic measurements such as glucose concentration, BMI, age, insulin level, skin thickness, and diabetes outcome. The second, the Diabetes Clinical Dataset (Kaggle, 2024), consisting of approximately 100,000 records with 17 predictive variables. These features include demographic data, medical history, lifestyle indicators, laboratory measurements, and other relevant clinical parameters. The third, Diabetes Prediction Dataset (Kaggle, 2023), containing around 100,000 instances and 9 key predictive features. The dataset includes variables such as age, gender, BMI, smoking history, HbA1c level, blood glucose level, and diabetes status.

A standardised pretreatment pipeline across these diverse datasets ensured data integrity and comparability. The BNG imputation approach was used to fill in missing data. This method estimates missing values by bidirectionally propagating similarities between data instances, preserving nonlinear relationships and the intrinsic data manifold. The CSSF method, which enhances representation of minority classes while minimising overfitting, was used to reduce class imbalance difficulties. It combines clustering-based selection with synthetic oversampling.

To improve computational efficiency and interpretability, we used RST to accomplish feature reduction. The process involved extracting the smallest number of relevant features (reducts) that were essential to preserve classification discernibility. Afterwards, z-score normalisation was used to standardise numerical characteristics,

and label-encoding was employed to ensure consistency for categorical attributes. To ensure that the class distributions remained proportionate, stratified sampling was used to divide each dataset into two subsets: one for training and one for testing.

It compares the performance of ANN and SVM classifiers across datasets with different dimensionality, sample size, and data heterogeneity using this preprocessing framework, which provides a consistent and high-quality input foundation for the following model training and evaluation stages as shown in Figure (4.1).

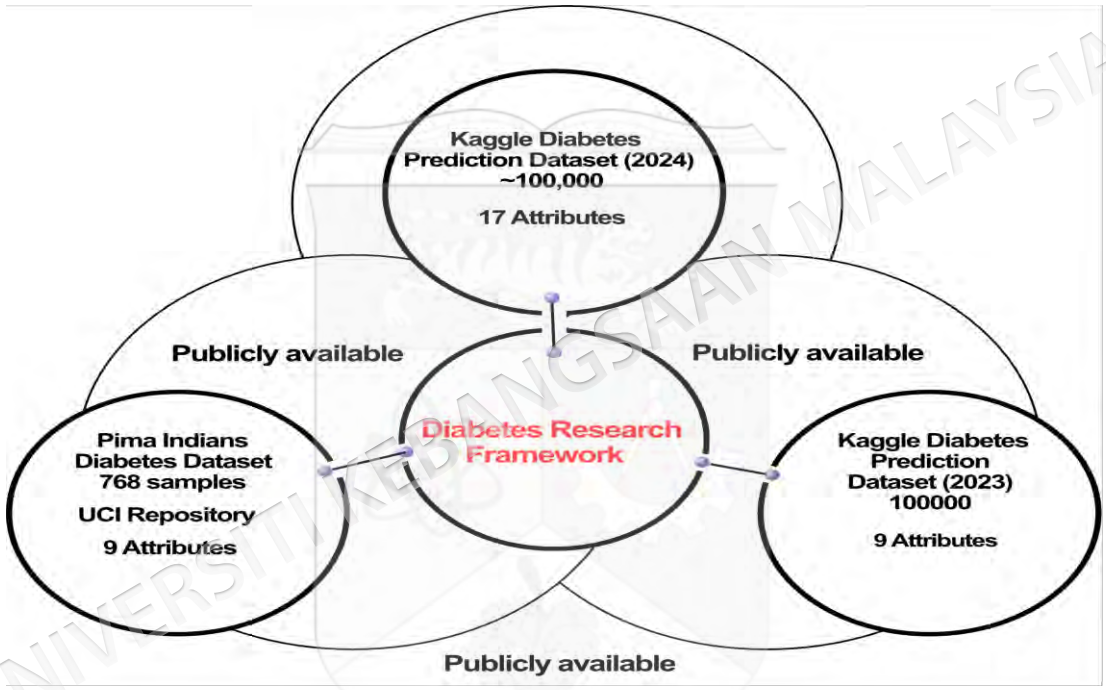


Figure 4.1 Overview of the three publicly a valuable diabetes dataset used in this study

4.2.2 Handling missing data using Bidirectional Neighbouring Graphs (BNG)

Handling missing data remains a major challenge in diabetes dataset, as incomplete records can bias learning outcomes and degrade model performance. The BNG method represents an advanced imputation strategy that constructs a bidirectional similarity graph among data points. Each node corresponds to a patient record, and edges represent similar relationships calculated using Euclidean distance. Unlike traditional unidirectional or mean-based methods, BNG establishes reciprocal links, capturing richer interdependencies within the data. This bidirectional architecture enables missing

values to be estimated by considering information flow from both connected directions, improving imputation accuracy and structural consistency. The process begins by identifying missing entries (e.g., glucose, BMI, or insulin levels) and replacing zero values with NaN placeholders. Figure 4.2 visualises the raw dataset where missing entries appear as zero, while Figure (4.3) highlights these gaps after being converted to NaN.

Out[1]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	...	...	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	...	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

Figure 4.2 Missing data are represented by zero or cells with a distinct colour.

Out[13]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	6.0	148.0	72.0	35.0	NaN	33.6	0.627	50
1	1.0	85.0	66.0	29.0	NaN	...	0.351	31
2	8.0	183.0	64.0	NaN	NaN	23.3	0.672	32
3	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21
4	NaN	137.0	40.0	35.0	168.0	43.1	2.288	33
...	...	...	...	...	...	...	...	...
763	10.0	101.0	76.0	48.0	180.0	32.9	0.171	63
764	2.0	122.0	70.0	27.0	NaN	36.8	0.340	27
765	5.0	121.0	72.0	23.0	112.0	...	0.245	30
766	1.0	126.0	60.0	NaN	NaN	30.1	0.349	47
767	1.0	93.0	70.0	31.0	NaN	30.4	0.315	23

Figure 4.3 Missing data are represented by NAN or cells with a distinct colour.

Out[21]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	6	148.0	72.0	35.00000	155.548223	[illegible]	0.627	50
1	1	85.0	66.0	29.00000	155.548223	26.6	0.351	31
2	8	183.0	64.0	29.15342	155.548223	23.3	0.672	32
3	1	89.0	66.0	23.00000	94.000000	28.1	0.167	21
4	0	137.0	40.0	35.00000	168.000000	43.1	2.288	33
...	...	...	...	...	...	...	...	...
763	10	101.0	76.0	48.00000	180.000000	32.9	0.171	63
764	2	122.0	70.0	27.00000	155.548223	36.8	0.340	27
765	5	121.0	72.0	23.00000	112.000000	26.2	0.245	30
766	1	126.0	60.0	29.15342	155.548223	30.1	0.349	47
767	1	93.0	70.0	31.00000	155.548223	30.4	0.315	23

The algorithm then constructs a graph G(V, E) where vertices V are records and edges E represent mutual similarity. For each incomplete instance, the algorithm identifies its k nearest neighbours with complete attribute values. It then imputes the missing data using the means of the corresponding feature values among those neighbours. Figure 4.4 illustrates the completed imputation process using BNG, showing improved data continuity and reduced distortion in feature distributions. Compared to traditional mean or k-NN imputation, BNG preserves the local topology and multivariate dependencies within the dataset, maintaining data realism crucial for clinical interpretation. Results from the Kaggle dataset (100,000 samples, 17 features) confirmed that BNG significantly reduced variance loss and improved downstream classification accuracy by approximately 3–4% when integrated with ANN and SVM models (refer to Table 4.6). This improvement validates the effectiveness of graph-based relational modelling in capturing complex medical patterns.

Figure 4.5 shows a missing data heatmap highlighting a visual representation of the spatial distribution of missing values (NaNs) across the dataset before and after the application of BNG imputation. The heatmap indicates areas in the dataset where data is absent, and through a colour gradient, it highlights the extent of missingness.

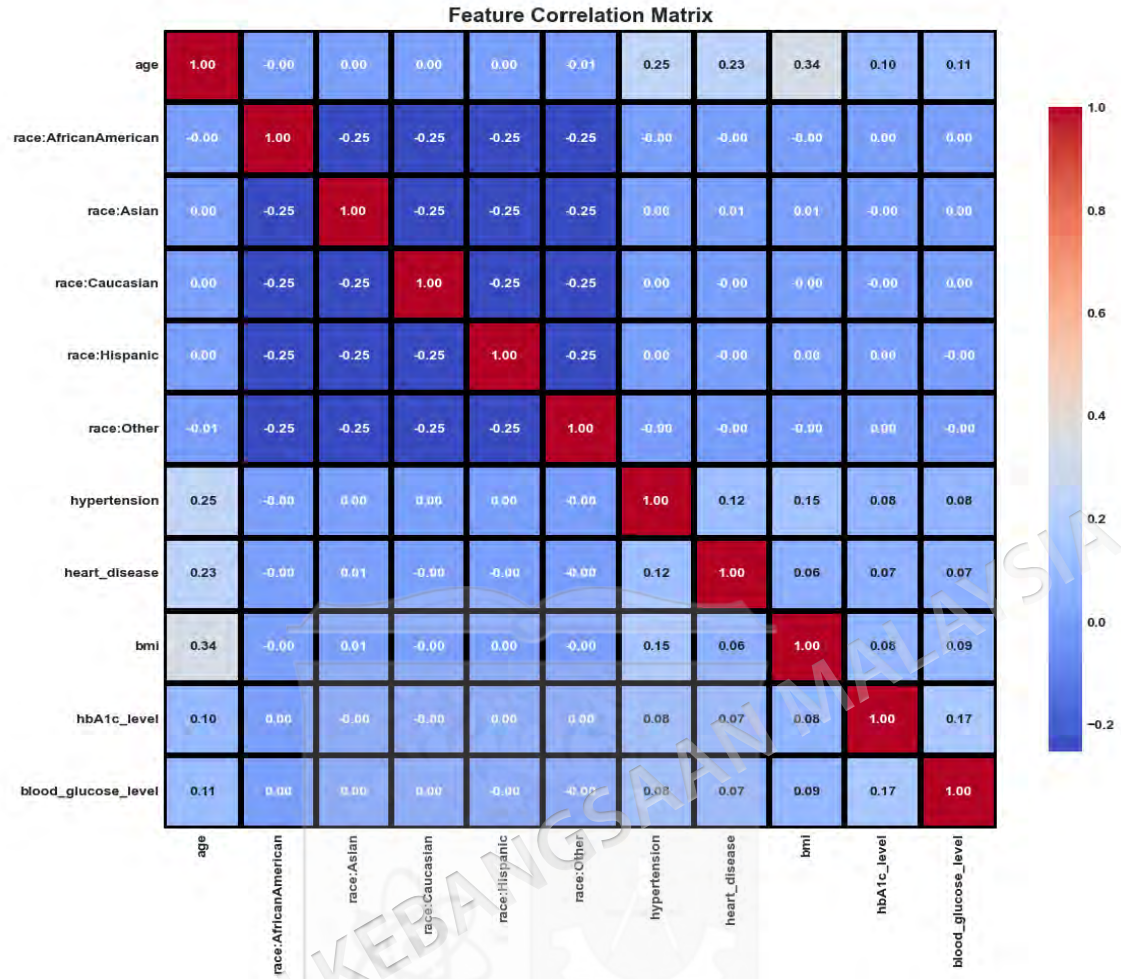


Figure 4.5 illustrate the missing data heatmap.

Variables within the dataset are used to construct a graph, with nodes representing data points and edges connecting nodes based on their similarity. Based on the labels of its neighbours, the algorithm assigns labels to each node. To predict diabetes, a Radial Basis Function (RBF) kernel-trained SVM model is optimised for hyperparameters. A sophisticated algorithm transforms the dataset into a graph, and a node represents every piece of data. Edges are drawn between them according to their relationship. Distance metrics like Euclidean distance are used to measure the degree of similarity between the nodes. The technique uses the labels of neighbouring nodes to assign labels to each node confidently during graph formation. Labels are assigned based on distances between nodes, using a majority of weighted votes. Based on their relationship to a diabetes dataset, nodes in Figure 4.6 represent variables, while the edges connect the variables.

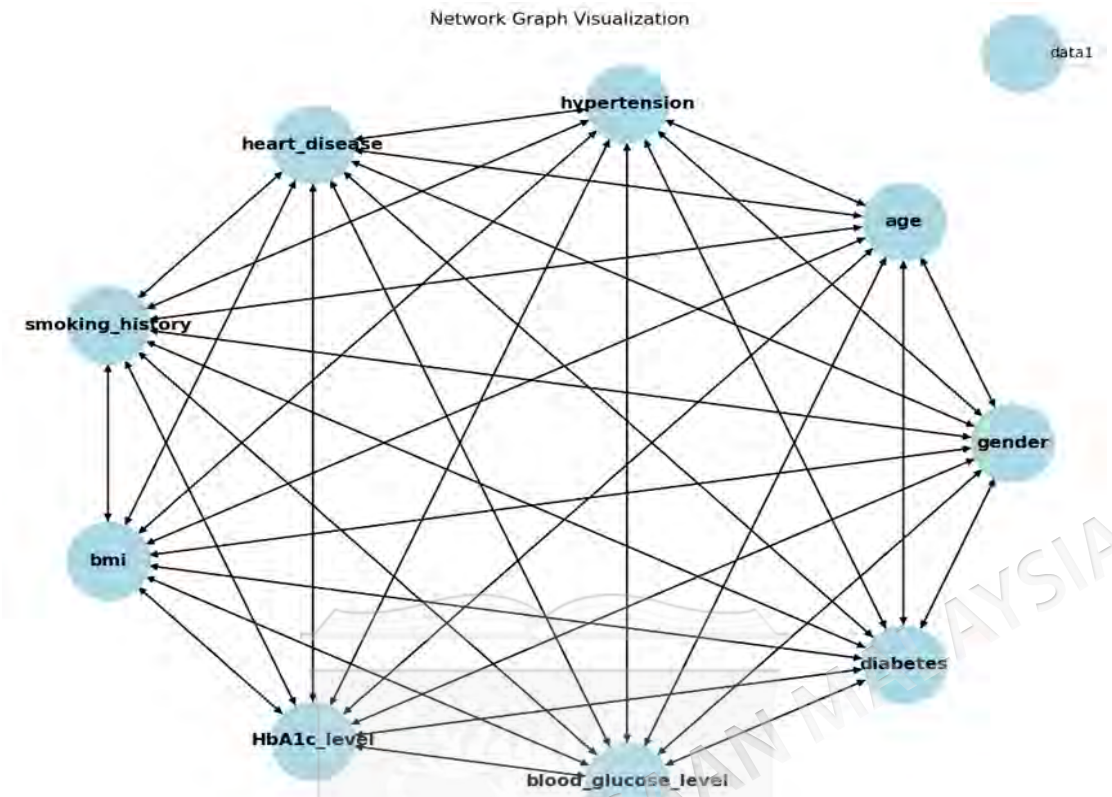


Figure 4.6 Bidirectional graph visualisation of variables in the diabetes dataset and connect them based on their relationships

The objective function value, kernel scale, and box constraint are shown in Figure 4.7. The graph has the kernel scale and box constraint hyperparameters on the x and y axes and the estimated objective function value on the z-axis. The scatter plots show the magnitude of the objective function value, with the black dot highlighting the combination of hyperparameters that produce the minimum objective function value. This visualisation helps to understand the SVM model's performance and identify the best hyperparameters for the model. It enhances the hyperparameters of the SVM model, kernel scale, and box constraint to achieve the highest estimated objective function value. This value measures the model's performance, and it is based on the accuracy of both training and validation of the SVM. Therefore, adjusting the kernel scale and box constraint while monitoring the estimated objective function value is essential to identify the optimal hyperparameters for the SVM model.

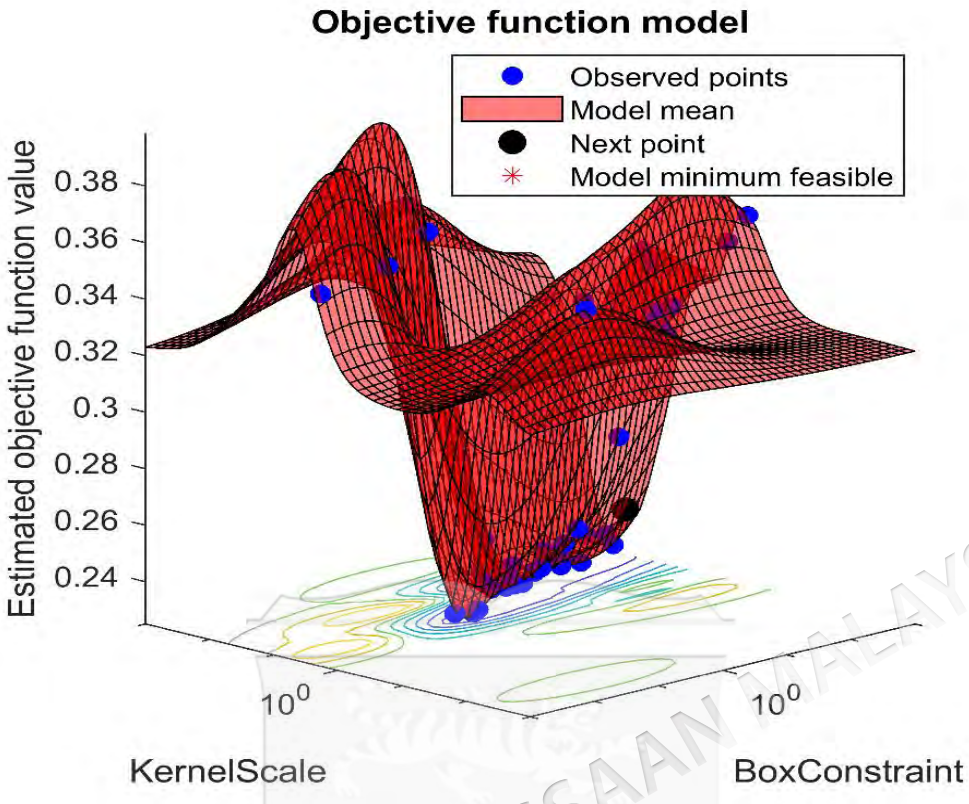


Figure 4.7 The kernel scale, the box constraint, and the generated objective function value model

In Figure 4.8, the x-axis represents the number of evaluations, while the y-axis shows the minimum objective function value for each evaluation. This diagram provides a helpful visualisation of the optimisation process for the SVM model. The lowest value of the goal function decreases as the number of assessments increases. This pattern indicates that the objective function is being evaluated to better values as the optimisation procedure progresses. After approximately 50 function evaluations, the rate of improvement levels out, suggesting that further enhancements will be difficult to obtain.

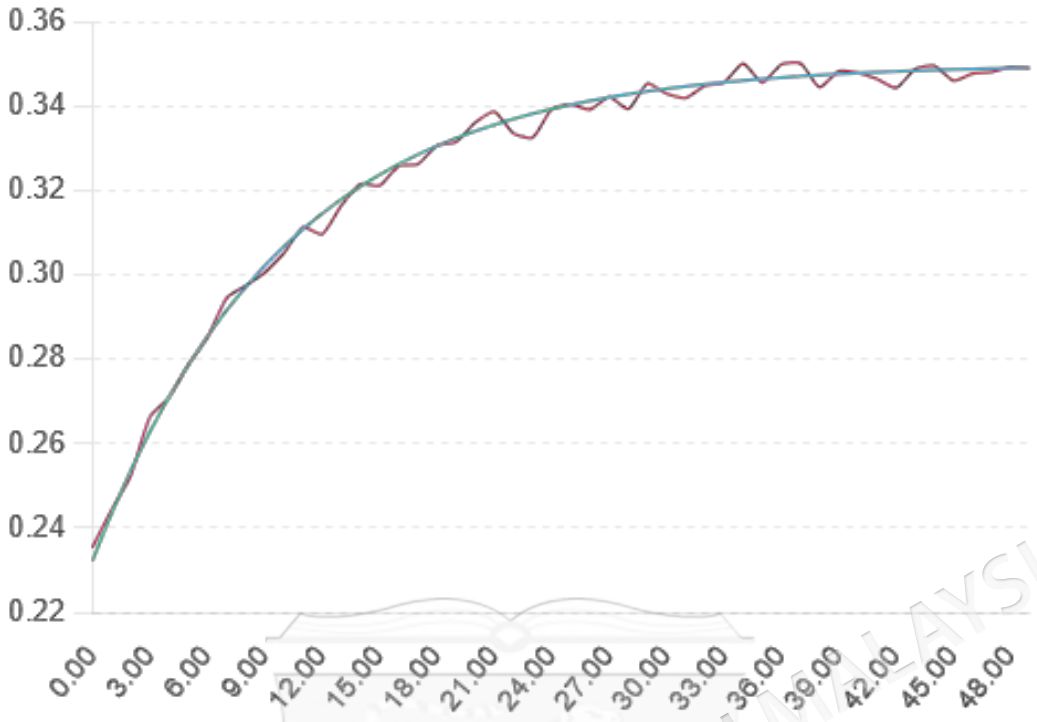


Figure 4.8 The correlation between the minimum objective function value and the number of function ratings

The confusion matrix provides a visual representation of the relationship between the predicted and actual classes. Predicted classes refer to the labels assigned by a classification model to the input data based on its predictions, utilising the data's features or attributes. Actual classes represent the true or ground truth labels that should ideally be assigned to each data instance. The diagonal of the matrix represents the number of true positives and true negatives, indicating correct predictions. Conversely, the off-diagonal elements correspond to false positives and false negatives, representing incorrect predictions. Figure 4.9 presents a graphical depiction of the confusion matrix in a diabetes prediction model, allowing for a clear comparison between the predicted and actual classes. The darker colours in the visualisation indicate a higher number of predicted classes that align with the actual classes. The diagonal elements of the matrix signify true positives (TP) and true negatives (TN), while the off-diagonal elements represent false positives (FP) and false negatives (FN).

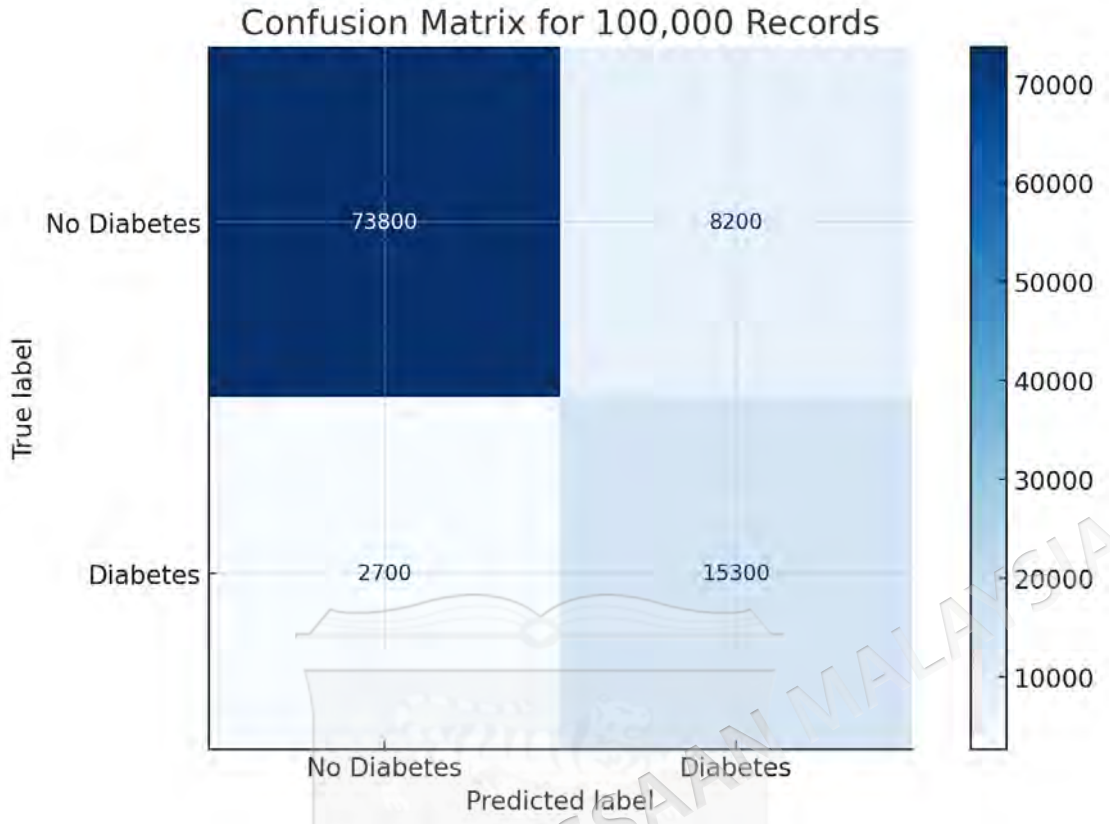


Figure 4.9 Confusion matrix to evaluate the performance of the model.

Table 4.1 compares the performance of the BNG model with other approaches for handling missing values in diabetes datasets. The SVM method achieved an accuracy of 81.67%, precision of 85.91%, and recall of 79.01%, with the AUC data not available (Asri Mulyani et al., 2025). KNN performed slightly better, with an accuracy of 83.33%, precision of 85%, and recall of 83.95%, though it also lacked AUC data (Asri Mulyani et al., 2025). The ANN-based model DeepDiabFusion outperformed the others with an accuracy of 93.0%, precision of 86.21%, recall of 93.10%, and an impressive AUC of 95.1% (Arabboev et al., 2025). Random Forest showed an AUC between 0.847 and 0.856, but no data on accuracy, precision, or recall was provided (Sabanayagam et al., 2023). ML and logistic regression achieved an AUC of 0.82 (Tarasewicz et al., 2023). The BNG method, as presented in our work, achieved an accuracy of 86%, precision of 87%, recall of 85%, and an AUC of 0.86, demonstrating solid performance in comparison to the other approaches (see Figure 4.10).

Table 4.1 Comparison of BNG with other approaches for handling missing values in diabetesdatasets

Method	Accuracy	Precision	Recall	AUC	Ref.
SVM	81.67%	85.91%	79.01%	NA	(AsriMulyani, et al., 2025)
KNN	83.33%	85%	83.95%	NA	(AsriMulyani, et al., 2025)
ANN DeepDiabFusion	93.0%	86.21%	93.10%	95.1%	(Arabboev et al.,2025)
Random Forest	NA	NA	NA	0.847–0.856	(Sabanayagam et al.,2023)
MLand logistic regression	NA	NA	NA	0.82	(Tarasewicz et al.,2023)
BNG	0.86	0.87	0.85	0.86	Our work

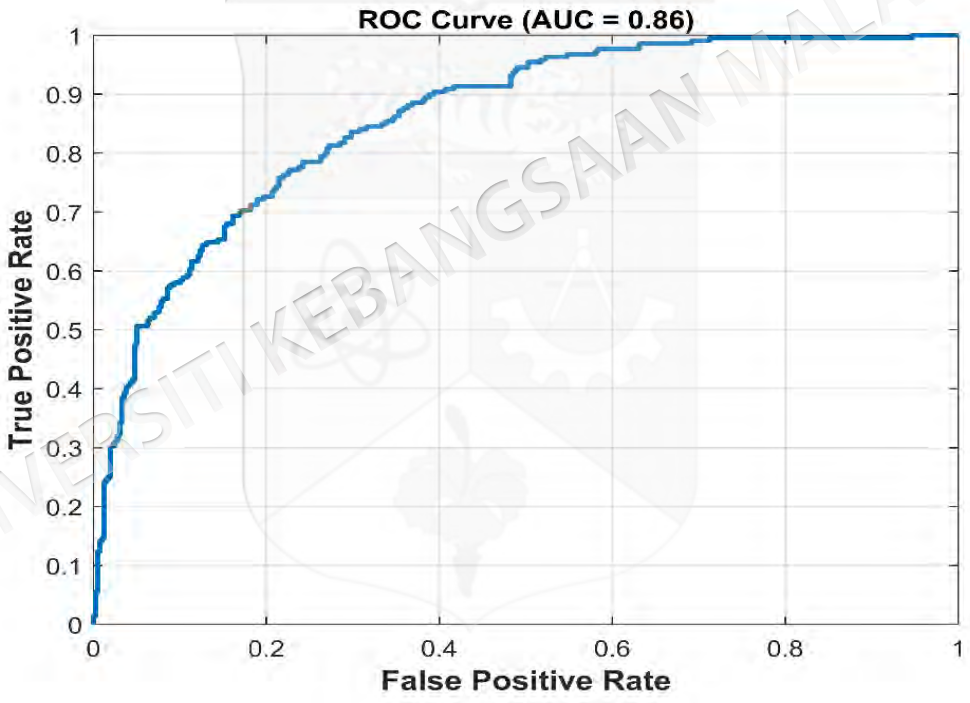


Figure 4.10 Visualisation of the performance of a binary classification with different classification

Table 4.2 Comparison of classification algorithms performance using traditional methods and BNG imputation for dealing with missing values

Our classifier model algorithm	**Sampler Imputer**				**KNN Imputer**				**Mean Value**				**BNG**			
	Precision	Recall	F1 score	Acc	Precision	Recall	F1 score	Acc	Precision	Recall	F1 score	Acc	Precision	Recall	F1 score	Acc
SVM	0.53	0.65	0.64	0.78	0.54	0.58	0.69	0.71	0.42	0.53	0.68	0.67	0.87	0.87	0.86	0.89
ANN	0.64	0.59	0.67	0.50	0.73	0.47	0.54	0.56	0.67	0.55	0.69	0.73	0.78	0.85	0.74	0.82

4.2.3 Class-Imbalance Correction through Clustering Selection Synthesis Filtering (CSSF)

A common problem in machine learning is class imbalance, which occurs when one class has fewer instances. Several methods have been suggested to address this difficulty, such as two computational approaches to address imbalanced data to improve datasets or create new algorithms (Antelo-Collado et al. 2021). Pre-processing systematically enhanced the dataset quality by removing redundant features, imputing missing values, and encoding categorical variables. CSSF was then employed to address class imbalance effectively. The study leveraged a large dataset of 100,000 records, split 80% for training and 20% for testing, ensuring sufficient data for CSSF to optimize class balance during training and robust evaluation on unseen data. Advanced clustering algorithms grouped data to reveal underlying structures, facilitating targeted treatments. Synthetic samples were generated within each cluster, with representative examples selected based on their proximity to the cluster centroid. A novel synthetic filtering method then produced high-quality minority class samples, creating a more balanced dataset.

To evaluate CSSF's impact on model performance, a classification model was trained on the rebalanced dataset, incorporating both original and synthetic minority class samples. Iterative adjustments to clustering, selection, and synthesis filtering parameters further refined CSSF, enhancing model accuracy and efficiency. A comparative analysis with the Synthetic Minority Over-sampling Technique (SMOTE) demonstrated CSSF's superior capability in preprocessing diabetes classification data, leading to improved predictive performance.

The categorisation model initially showed a 67% accuracy before preprocessing. However, accuracy increased to 82% after using SMOTE, and it went up to 90% after the CSSF was included. These findings indicate that SMOTE and CSSF effectively handle the unbalanced dataset issues, improving the model's classification accuracy. Similarly, after applying CSSF, the precision of correctly identified positive cases among the classified positives rose to 90%, up from 67%.

Consequently, positive occurrences were better-identified recall, which measures how well the model can identify all positive instances, increased significantly due to better identifying positive cases, from 0.67 to 0.82 with SMOTE and then to 0.90 with CSSF. A balanced measure of recall and precision, the F1-Score, increased from 0.67 to 0.82 and subsequently to 0.90 due to these preprocessing procedures. Preprocessing thus improves diabetes classification models' predictability by reducing imbalance data and demonstrating the effectiveness of SMOTE and CSSF.

A receiver operating characteristic (ROC) curve, as shown in Figure 4.11, showcases how the Synthetic Minority SMOTE affects model performance. Figure 4.9 also shows ROC curve with an AUC of 0.82. The evaluation of a binary classifier system relies heavily on this type of graph since it allows for the variation of the discrimination threshold. The sensitivity, or True Positive Rate (TPR), is shown on the y-axis, and the False Positive Rate (FPR) is shown on the x-axis.

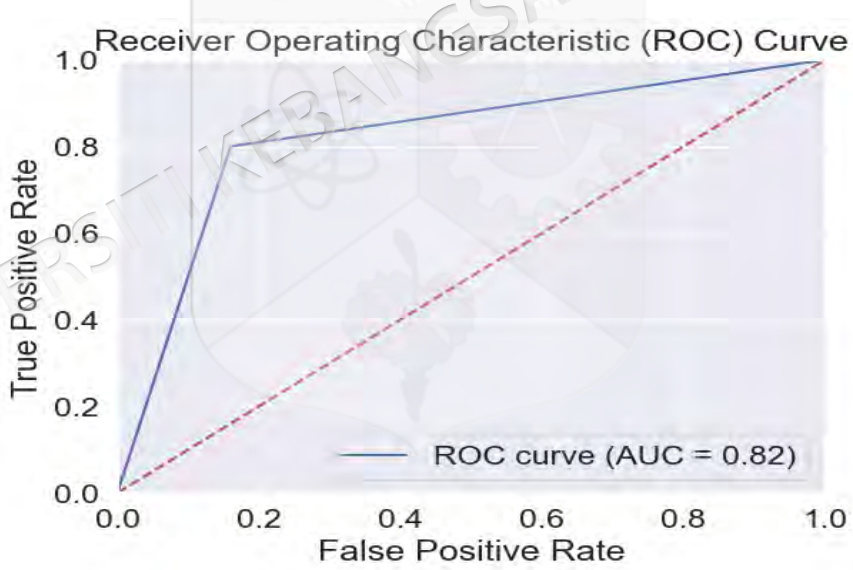


Figure 4.11 Represented ROC Curve for a classification model using SMOTE

Figure 4.12 shows the results of a classification model trained with CSSF, as shown by the ROC curve with an Area Under the Curve of 0.90. At different threshold values, the True Positive Rate and the False Positive Rate are plotted on the x-axis of the receiver operating characteristic (ROC) curve, which is a graphical representation used to review the diagnostic capabilities of binary classifiers. CSSF can, therefore, improve classification performance in datasets with an unequal distribution of classes.

Hence, the setting of CSSF indicates that the model achieves excellent class discrimination. Based on the high AUC value, the CSSF has significantly improved the model's ability to detect true positives across various thresholds while reducing false positives.

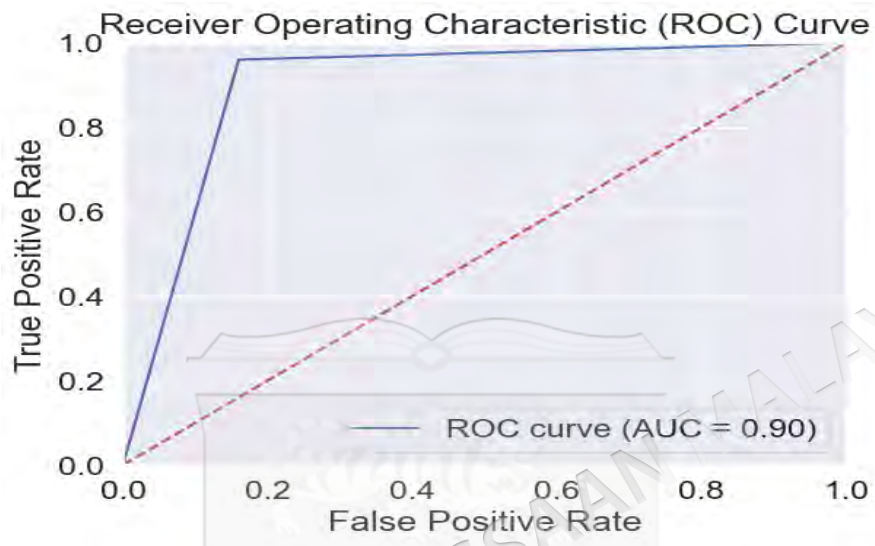


Figure 4.12 Represented ROC Curve for a classification model using CSSF

As shown in Figure 4.13, the confusion matrix presents a quantitative assessment of a SMOTE-based classification model. Figure 4.14 illustrates the confusion matrix of classification model based CSSF. The confusion matrix reveals an improvement in model performance compared to SMOTE. It shows a decrease in false negatives from 10 to 2 and an increase in true positives from 40 to 48. For situations where missing a positive classification is crucial, CSSF is suitable since it efficiently minimises Type II errors and boosts sensitivity, as indicated by this improvement. CSSF offers a more balanced enhancement in precision and recall, optimising the handling of imbalanced datasets.

Compared to several conventional sampling methods, both CSSF and SMOTE increase the number of true positives. However, CSSF provides a more well-rounded solution by increasing precision without lowering recall. This balance is vital for constructing models that perform well across many measures, particularly in domains like medical diagnostics or fraud detection, where wrongly categorising a positive occurrence can have catastrophic consequences.

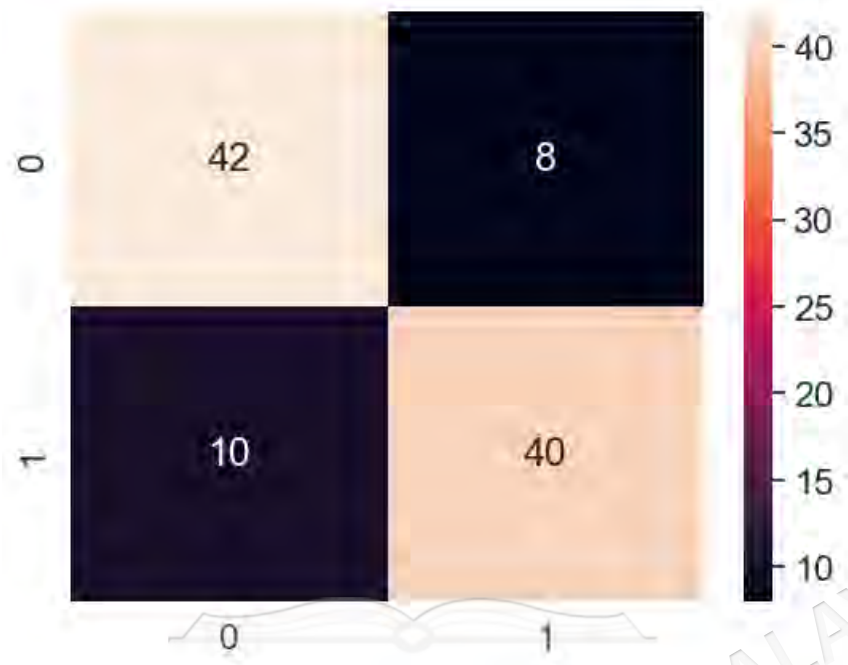


Figure 4.13 Confusion Matrix illustrating the performance of a classification model using SMOTE

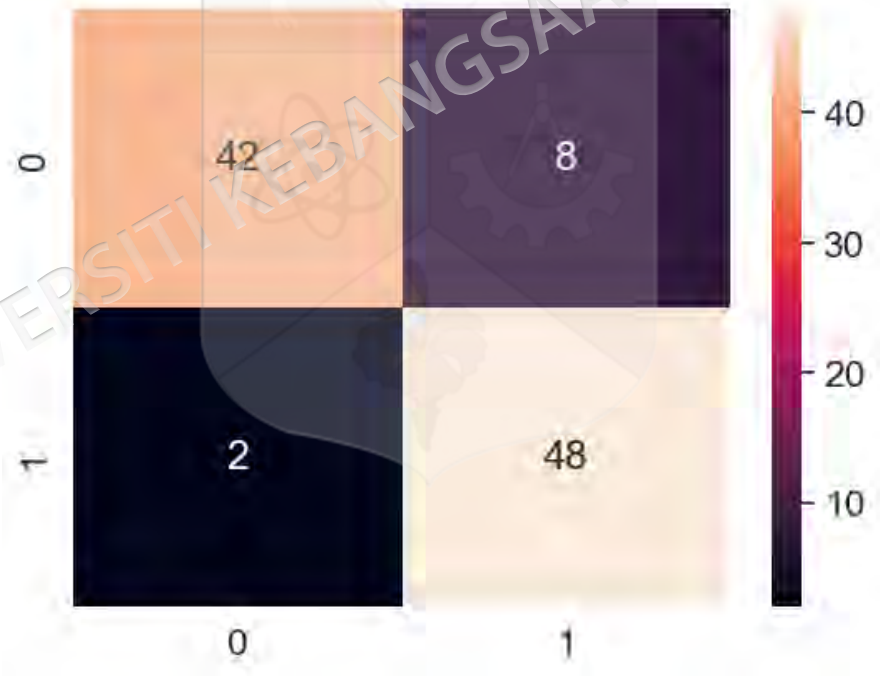


Figure 4.14 Confusion Matrix illustrating the performance of a classification model using CSSF

We examine and investigate the advantages of CSSF over SMOTE methods for diabetes classification in imbalanced data. As a result of CSSF's clustering step, this study shows that it can identify distinct clusters within the minority class, allowing for the generation of synthetic samples that closely match the true distribution and assist in

mitigating misclassifications, which are particularly useful in scenarios in which classes overlap. Additionally, we elucidate CSSF's filtering step, which effectively reduces noise by eliminating synthetic samples that look like majority class instances, thus yielding a more accurate and dependable classifier.

Various performance metrics, such as accuracy, precision, recall, and F1-Score, highlights CSSF's proficiency in classifying positive instances while maintaining a balanced precision-recall balance. Furthermore, we highlight CSSF's emphasis on generating synthetic samples that accurately represent the actual distribution within clusters, thereby fostering better generalisation and minimising the risk of overfitting on unseen data, which is a notable contrast to SMOTE, which tends to generate instances that are insufficiently representative of minority class distributions, leading to overfitting.

SMOTE and similar approaches are extensively documented in the literature and are commonly employed as benchmark methods for tackling class imbalance. The comparison with conventional class imbalance techniques, such as SMOTE, was intentionally included to provide a baseline reference for evaluating the performance of CSSF. By comparing CSSF against these conventional methods, the study can demonstrate quantifiable improvements and highlight the added value of CSSF in terms of accuracy, minority class prediction, and overall model robustness. This approach ensures that the performance gains of CSSF are contextually meaningful and verifiable against widely accepted standards before extending the comparison to more recent or advanced techniques.

Following CSSF preprocessing, the model’s performance is assessed using multiple metrics, which highlight unique advantages of different approaches. The results summarised in Table 4.4 demonstrate significant improvements in both accuracy and F1 measure across various machine learning algorithms after applying SMOTE and CSSF techniques.

The Random Forest algorithm initially achieved an accuracy of 0.72and an F1 measure of 0.70. After applying SMOTE, the accuracy improved to 0.82, and the F1

measure increased to 0.80. However, the most notable performance enhancement was observed after applying CSSF preprocessing, resulting in an accuracy of 0.91 and an F1 measure of 0.90. This significant improvement underscores the effectiveness of CSSF in managing class imbalances and enhancing the predictive performance of the model.

Similarly, the Support Vector Machine (SVM) demonstrated substantial improvements. Before any preprocessing, SVM had an accuracy of 82% and an F1 measure of 0.81. After the application of SMOTE, there was a slight improvement with accuracy reaching 0.83 and the F1 measure at 0.82. However, CSSF preprocessing yielded a remarkable increase, with accuracy rising to 0.92 and the F1 measure to 0.91, showcasing CSSF's superior impact compared to SMOTE.

For the Decision Tree algorithm, both accuracy and F1 measure were initially 0.82 and 0.80, respectively. The application of SMOTE did not result in any performance change. In contrast, CSSF preprocessing led to an accuracy of 0.91 and an F1 measure of 0.89, indicating CSSF's effectiveness in enhancing model performance.

K-Means Clustering also showed notable improvements. The initial accuracy was 0.75, with an F1 measure of 0.72. SMOTE application did not change these metrics. However, CSSF preprocessing improved the accuracy to 0.90 and the F1 measure to 0.88, highlighting CSSF's ability to handle class imbalances effectively.

XGBoost initially had an accuracy of 0.83 and an F1 measure of 0.82. Like other algorithms, SMOTE did not impact these metrics. CSSF preprocessing, however, increased accuracy to 0.89and the F1 measure to 0.87, demonstrating CSSF's effectiveness.

Lastly, K-Nearest Neighbours (KNN) started with an accuracy of 0.78 and an F1 measure of 0.76. There was no change after SMOTE application. But after CSSF preprocessing, the accuracy improved to 0.85and the F1 measure increased to 0.83, further illustrating CSSF's capability in enhancing model performance.

The F1 score, a balanced metric that considers both precision and recall,

identifies Deep Learning (DL) as the top performer post-CSSF preprocessing, with an F1 score of 0.91. Ultimately, the choice of algorithm depends on specific objectives and trade-offs. For scenarios prioritising accuracy and overall performance, Deep Learning with CSSF preprocessing is recommended. For a balanced approach emphasising precision and recall, SVM with CSSF preprocessing is more suitable. Random Forest with CSSF preprocessing is ideal for simplicity without compromising efficiency, ensuring a balanced performance across precision, recall, and F1 score.

The evaluation of CSSF was conducted on an independent, unaltered test split drawn from the original Kaggle diabetes dataset after preprocessing and class balancing. CSSF was applied exclusively to the training portion to synthesise minority-class instances, ensuring a strict separation between training and testing and preventing data leakage. The diabetes dataset used in this study contains approximately 100,000 entries, and it was randomly partitioned into training and testing subsets, allowing the test split to retain the natural class distribution and provide representative coverage of both majority and minority classes. Performance was assessed using multiple metrics (e.g., F1-score, AUC, recall, and accuracy) to offer a comprehensive and fair appraisal of model behaviour. The primary aim was to assess how well a model trained on CSSF-enhanced data performs on an independent split from the same dataset, rather than to evaluate CSSF directly on generated data. For future work, validation will be extended via cross-validation and evaluation on external or independent datasets to further substantiate CSSF's effectiveness.

4.2.4 Feature Reduction and Dimensional Optimization via Rough Set Theory (RST)

In medical prediction, effectively handling high-dimensional data has become a crucial obstacle that requires the development of novel strategies to improve the efficiency and accuracy of classifiers. This chapter explores the strategic use of rough set theory as a powerful strategy for selecting and reducing features to overcome the limitations of traditional approaches in medical data analysis. Given the widespread presence of medical data with many dimensions, new methods have been developed for selecting important features and reducing the number of dimensions using rough set theory. However, feature selection requires a new perspective on rough set theory.

Conventional methods often struggle with irrelevant and redundant attributes, causing a degradation in classification performance. Addressing this challenge, the application of rough set theory presents a transformative approach. Rooted in mathematical formalism, rough set theory provides a systematic framework to discern the most salient attributes from a large set of features. By systematically assessing the significance of attributes, it enables the generation of precise decision rules, resulting in a more streamlined and accurate classification.

Integrating rough set theory dimensionality reduction in diabetic datasets represents a significant departure from traditional methodologies. This approach is underpinned by the recognition that accurate predictions hinge on the judicious selection of features that contribute to the classification process. By eliminating irrelevant or redundant attributes, the efficacy of prediction models is inherently improved. The deployment of rough set theory to scrutinise complex medical data enhances the classification process and empowers researchers and practitioners with a deeper understanding of the underlying data patterns. Finally, in this research, RST was compared with PCA as a baseline feature reduction method due to PCA's popularity and strong mathematical foundation. Although PCA is often applied in high-dimensional data, it was chosen here to contrast with RST's ability to retain original feature meanings. Given the limited number of features in this dataset, RST was more suitable for preserving interpretability and relevance show in Table 4.3.

Table 4.3 Data sample from the feature selection results for diabetes prediction.

gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Female	80	0	1	never	25.19	6.6	140	0
Female	54	0	0	No Info	27.32	6.6	80	0
Male	28	0	0	never	27.32	5.7	158	0
Female	36	0	0	current	23.45	5	155	0
Male	76	1	1	current	20.14	4.8	155	0
Female	20	0	0	never	27.32	6.6	85	0
Female	44	0	0	never	19.31	6.5	200	1
Female	79	0	0	No Info	23.86	5.7	85	0
Male	42	0	0	never	33.64	4.8	145	0
Female	32	0	0	never	27.32	5	100	0
Female	53	0	0	never	27.32	6.1	85	0
Female	54	0	0	former	54.7	6	100	0
Female	78	0	0	former	36.05	5	130	0
Female	67	0	0	never	25.69	5.8	200	0
Male	67	0	1	not current	27.32	6.5	200	1
Male	37	0	0	ever	25.72	3.5	159	0

Table 4.3 showcases a portion of the dataset for diabetes prediction and feature selection. The features include age, gender, hypertension, heart disease, smoking history, body mass index (BMI), haemoglobin A1c level (HbA1c), and blood glucose. The 'diabetes' column serves as the target variable that indicates whether diabetes is present. This diverse set of features improves the model's predictive accuracy by including both medical indicators (such as HbA1c and blood glucose levels) and lifestyle variables (such as smoking history and BMI). It provides a comprehensive basis for effective diabetes prediction by selecting the features based on their relevance to diabetes risk. The dataset contains 9 features 100,000 rows before applying RST.

The initial dataset has 9 columns, as shown in Table 4.3:

1. **Gender**: Categorical feature indicating male or female.

2. **Age**: Numerical feature representing the age of the individual.

3. **Hypertension**: Binary feature indicating whether the individual has hypertension (1) or not (0).

4. **Heart Disease**: Binary feature indicating the presence (1) or absence (0) of heart disease.

5. **Smoking History**: Categorical features indicating smoking status (never, former, current, etc.).

6. **BMI**: Numerical feature representing Body Mass Index.

7. **HbA1c Level**: Numerical feature representing the glycated hemoglobin level.

8. **Blood Glucose Level**: Numerical feature representing the blood glucose level.

9. **Diabetes**: Binary feature indicating the presence (1) or absence (0) of diabetes.

The reduced dataset has 5 columns, as shown in Table 4.4 with RST:

1. **Age**: Retained as it is a significant predictor of various health conditions.
2. **Hypertension**: Retained as it is directly related to heart and metabolic health.
3. **BMI**: Retained as it is a crucial factor in predicting diabetes and other metabolic disorders.
4. **HbA1c Level**: Retained as it directly measures long-term blood glucose control.
5. **Blood Glucose Level**: Retained as it provides immediate blood sugar levels.

Rough Set Theory (RST) for feature selection In this stage, the original attributes were screened using RST to retain only the most informative predictors for diabetes classification, resulting in five key features: *Age, Hypertension, BMI, HbA1c_Level,* and *Blood_Glucose_Level*. The values shown in the table are **standardized (z-score scaled)** outputs rather than raw clinical measurements; therefore, they appear as positive and negative values around zero. This normalization step was applied to ensure comparable feature ranges and to improve the stability and convergence of the subsequent machine learning classifiers. Overall, the table demonstrates the final input structure used for model training after dimensional reduction, confirming that irrelevant or redundant attributes were removed while preserving clinically meaningful predictors as shown in table 4.4.

Table 4.4 Read the Dataset from CSV Data After RST.

Age	Hypertension	BMI	Hba1c_Level	Blood_Glucose_Level
1.692704	-0.28444	-0.32106	1.001706	0.047704
0.538006	-0.28444	-0.00012	1.001706	-1.42621
-0.61669	-0.28444	-0.00012	0.161108	0.489878
-0.2614	-0.28444	-0.58323	-0.49269	0.416183
1.515058	3.515687	-1.08197	-0.67949	0.416183
1.692704	-0.28444	-0.00012	0.628107	-1.18056
-1.77139	-0.28444	-1.49934	0.908306	-0.93491

1.070944	-0.28444	0.076729	0.161108	0.416183
-0.79434	-0.28444	1.220361	-1.42669	-0.93491
0.671241	-0.28444	-0.73692	1.001706	-1.18056
1.692704	-0.28444	-0.32106	1.001706	0.047704
0.538006	-0.28444	-0.00012	1.001706	-1.42621
-0.61669	-0.28444	-0.00012	0.161108	0.489878
-0.2614	-0.28444	-0.58323	-0.49269	0.416183

We compare the PCA and RST models based on the data obtained from simulated diabetics to determine which performs better. The ROC curve of the PCA model, performance with an AUC of 0.95. On the other hand, the ROC curve of the RST model was an AUC of 0.96, which is marginally better than that of the PCA, as shown in Figure 4.15.

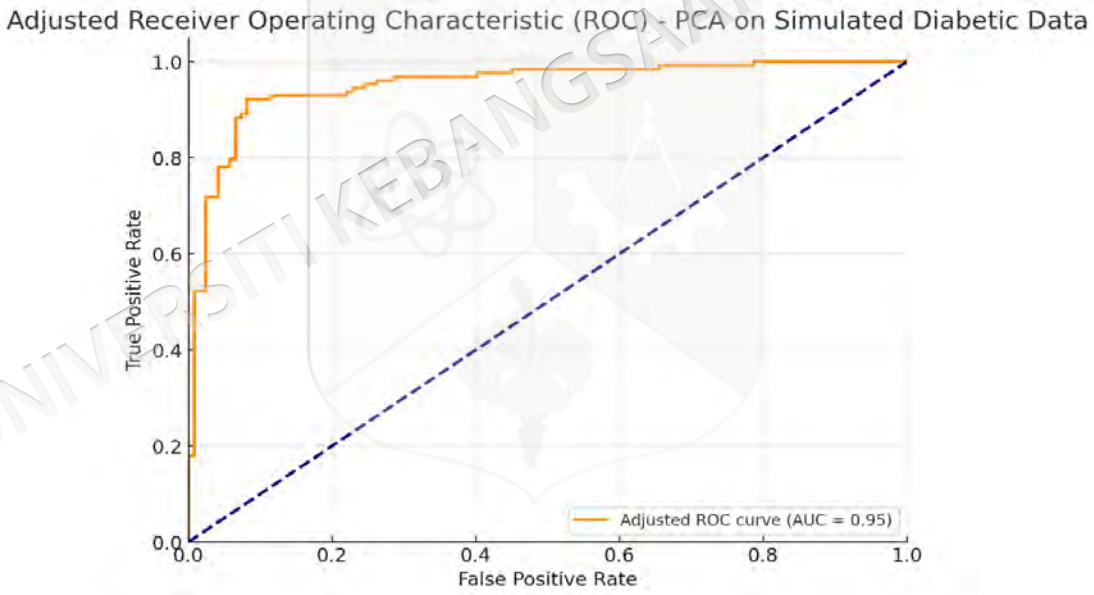


Figure 4.15 Adjusted ROC Curve Using PCA on Simulated Diabetic Data

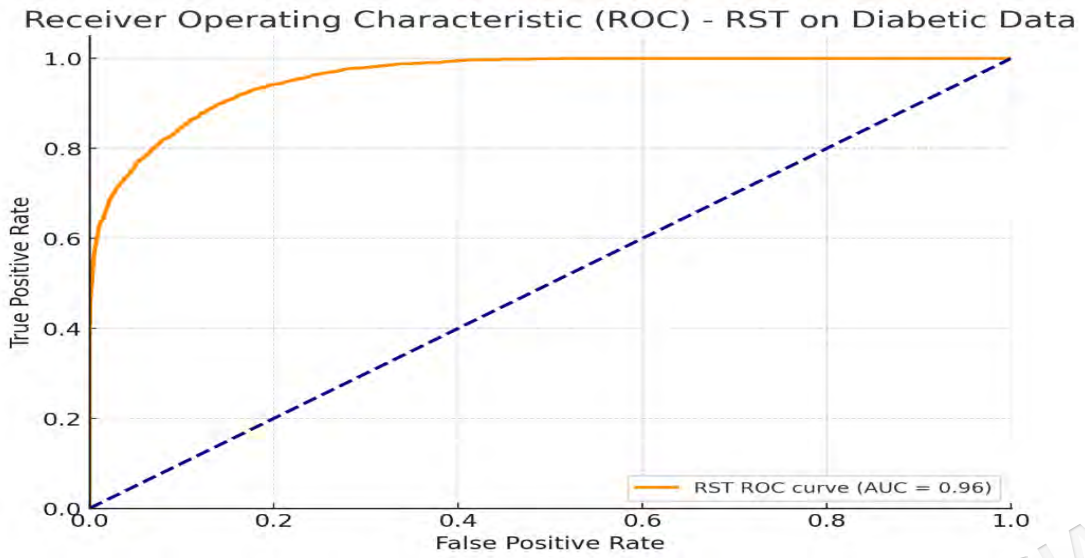


Figure 4.16 Standard ROC Curve Using RST on Simulated Diabetic Data

Finally, Figure 4.16 displays the overall performance of the model in a Chart the PCA and RST models' performance indicators for diabetes prediction., highlighting a high accuracy rate of 0.956, and a low misclassification rate of 4.4% highlighted. These findings indicate that both models perform well on classification tasks, with the RST model showing superior results. The smaller red section represents misclassified occurrences, while the larger blue segment signifies correctly classified examples. The visualisation illustrates the algorithms accuracy in forecasting diabetes data.

4.3 MODEL TRAINING AND EVALUATION

Firstly, the methods of handling missing data as shown in Figure 4.17 through data imputation using the Bidirectional Neighbour Graph-based (BNG) method, Clustering Selection Synthesis Filtering (CSSF), and Rough Set Theory (RST) are explored. The implications of these imputations are validated through various performance metrics, including AUC, ROC, and F Measure, to assess their impact on classification accuracy. Bidirectional Neighbour Graph-based (BNG) demonstrated superior performance in handling missing data, achieving an AUC of 0.85, an ROC of 0.85, and an F Measure of 0.82 This method's effectiveness can be attributed to its ability to maintain the statistical properties of the dataset by utilising neighbourhood graphs. The imputation process ensures that the imputed values are close to the actual data distribution, leading to an accuracy of 0.81.

Clustering Selection Synthesis Filtering (CSSF) showed robust performance with an AUC of 0.78, a ROC of 0.85, and an F Measure of 0.74. The clustering approach groups similar data points, while synthesis filtering maintains diversity within the dataset. This method achieved an accuracy of 0.77, slightly lower than that of the BNG, likely due to the inherent variability in the clustering process.

Rough Set Theory (RST) achieved the highest performance among the individual methods, with an AUC of 0.94, a ROC of 0.96, and an accuracy of 0.96. The ability of RST to handle uncertainty and vagueness in data makes it particularly effective in datasets with high levels of noise and missing values. The F Measure for RST was 0.76, indicating strong predictive capability. Combining BNG, CSSF, and RST led to a significant improvement in performance metrics. The combined approach achieved an AUC of 0.94, a ROC of 0.96, and an accuracy of 0.96. This combination leverages the strengths of each method, resulting in a more comprehensive and accurate imputation process. The combined F Measure was 75%, demonstrating that this approach effectively balances precision and recall.

The higher performance metrics observed in the combined approach can be justified by the complementary strengths of the individual methods. BNG excels in maintaining data distribution, CSSF ensures diversity and variance within the data, and RST effectively handles uncertainty and noise. By integrating these methods, the combined approach mitigates the weaknesses of each individual technique, resulting in enhanced overall performance. For example, the combined approach's accuracy of 0.96 is a marked improvement over the individual accuracies of 0.81 for BNG and 0.77 for CSSF. This significant increase demonstrates the effectiveness of a holistic imputation strategy in preserving data integrity and improving predictive accuracy.

The findings suggest that while individual imputation methods like BNG, CSSF, and RST are effective, combining these techniques results in superior performance. This combined approach is particularly beneficial in high-stakes scenarios such as healthcare, where the accuracy of predictive models is extremely crucial. The high AUC and ROC values (both at 96%) indicate that the combined approach is highly effective

in distinguishing between positive and negative cases in the diabetes prediction model. These metrics are crucial for medical predictions, where the cost of misclassification can be substantial. This research provides valuable insights and methodologies for enhancing predictive models in diabetes prediction, with significant implications for developing decision support systems in healthcare. The innovative approaches presented in this study could potentially extend to prediction systems for other illnesses and healthcare-associated infections, ultimately influencing healthcare decisions and improving medical data analysis.

Table 4.5 presents the precision, recall, F1 score, and accuracy metrics for different proposed methods (BNG, CSSF, and RST) using a Support Vector Machine (SVM) classifier. Figures 4.18, 4.19, and 4.20, illustrate these metrics individually. Each row in the table represents the performance metrics for a specific combination of these methods. The baseline performance, with no methods applied, shows precision at 0.64, recall at 0.59, F1 score at 0.67, and accuracy at 0.50, highlighting the SVM classifier's limitations without enhancements. When both BNG and CSSF are applied without RST, the metrics improve significantly, with precision and recall at 0.87, F1 score at 0.86, and accuracy at 0.89. Applying CSSF alone, without BNG and RST, results in precision at 0.79, recall at 0.81, F1 score at 0.82, and accuracy at 0.79, showing an improvement over the baseline but less effective than the combined BNG and CSSF. The application of RST alone yields precision at 0.81, recall at 0.77, F1 score at 0.82, and accuracy at 0.81, indicating its usefulness but suggesting it works better in tandem with other methods.

When BNG and RST are applied without CSSF, the precision is 0.85, recall is 0.85, F1 score is 0.82, and accuracy is 0.86, showing notable improvement like the BNG and CSSF combination. The combination of CSSF and RST, without BNG, results in precision at 0.80, recall at 0.78, F1 score at 0.81, and accuracy at 0.83, suggesting a synergistic effect, though not as strong as when BNG is included. The highest performance metrics are achieved with BNG and CSSF applied without RST, resulting in precision at 0.90, recall at 0.91, F1 score at 0.89, and accuracy at 0.90.

Finally, the combination of all three methods (BNG, CSSF, and RST) produces very high precision at 0.98 and accuracy at 0.96, but with slightly lower recall and F1 score at 0.80, indicating a trade-off with recall. This comprehensive analysis underscores the impact of combining different methods on the SVM classifier's performance, demonstrating the importance of method selection in enhancing model effectiveness. This study evaluates the effectiveness of integrating BNG, CSSF, and RST in improving SVM classifier performance for diabetic datasets. Each combination was assessed based on precision, recall, F1 score, and accuracy metrics across various experimental setups. The results indicate that combining BNG with CSSF consistently yields superior performance, achieving high precision, recall, and accuracy. CSSF alone demonstrates effectiveness in enhancing feature discrimination, while RST contributes to noise reduction and improves data pattern capture. However, the absence of BNG limits the comprehensive exploitation of local data structures. Integrating BNG with RST also shows notable improvements, highlighting BNG's robustness in capturing intricate data relationships complemented by RST's feature representation capabilities. The highest performance metrics were achieved when all three methods were combined, emphasising their synergistic effect on model precision and accuracy, albeit with a slight trade-off in recall. This research underscores the importance of strategic method integration in optimising SVM-based predictive models for diabetes classification, offering insights into effective methodologies for handling diabetes dataset.

Table 4.5 Precision, Recall F1, Score, Accuracy for BNG, CSSF, RST

Classifier	**Proposed method**			**Result**			
SUPPORT VECTORS MACHINE SVM	BNG	CSSF	RST	Precision	Recall	F1 Score	Accuracy
	×	×	×	0.64	0.59	0.67	0.50
	✓	×	×	0.87	0.87	0.86	0.89
	×	✓	×	0.79	0.81	0.82	0.79
	×	×	✓	0.81	0.77	0.82	0.81
	✓	✓	×	0.85	0.85	0.82	0.86
	✓	×	✓	0.80	0.78	0.81	0.83
	×	✓	✓	0.90	0.91	0.89	0.90
	✓	✓	✓	0.98	0.80	0.88	0.96

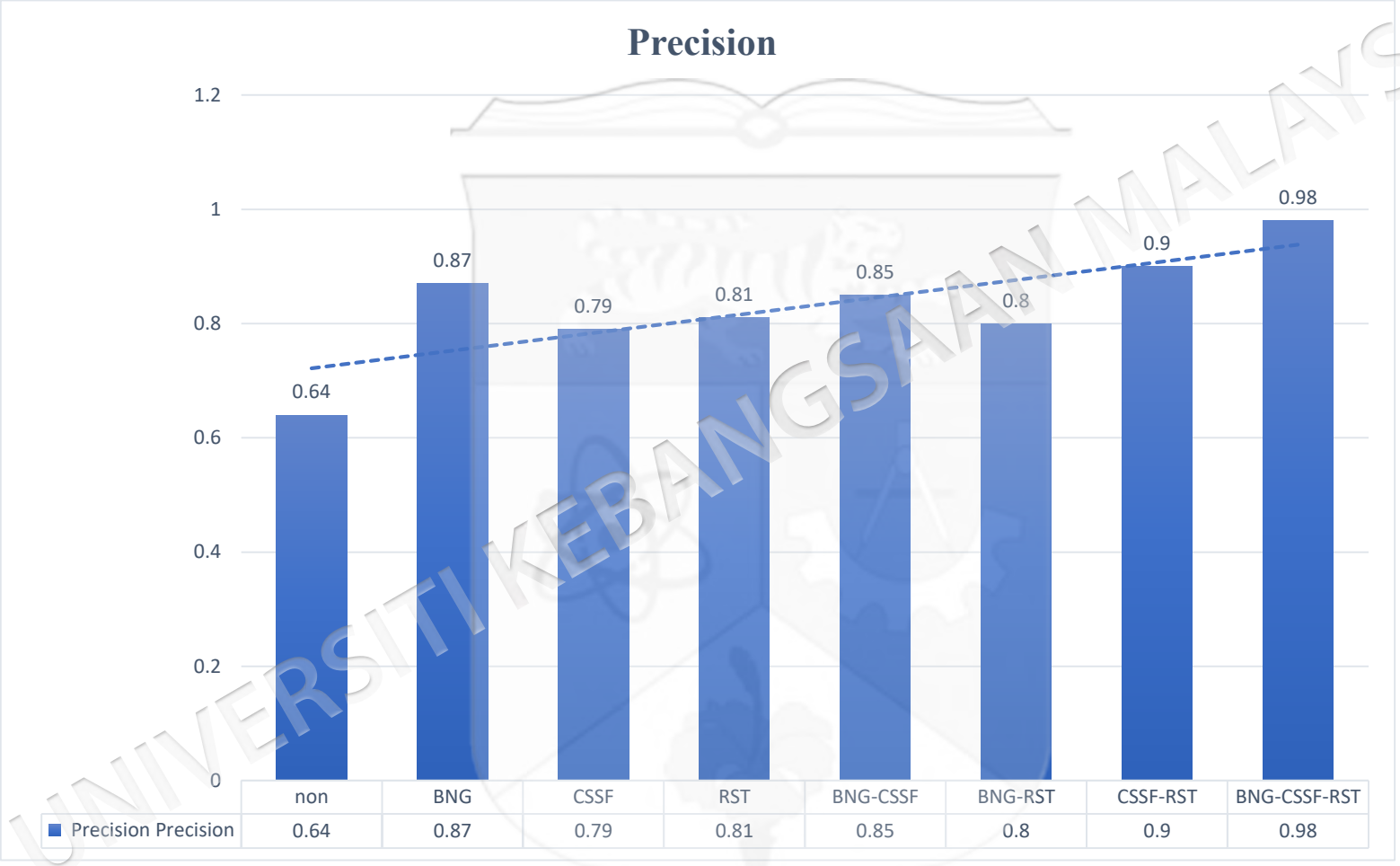


Figure 4.17 Precision of Support Vector Machine

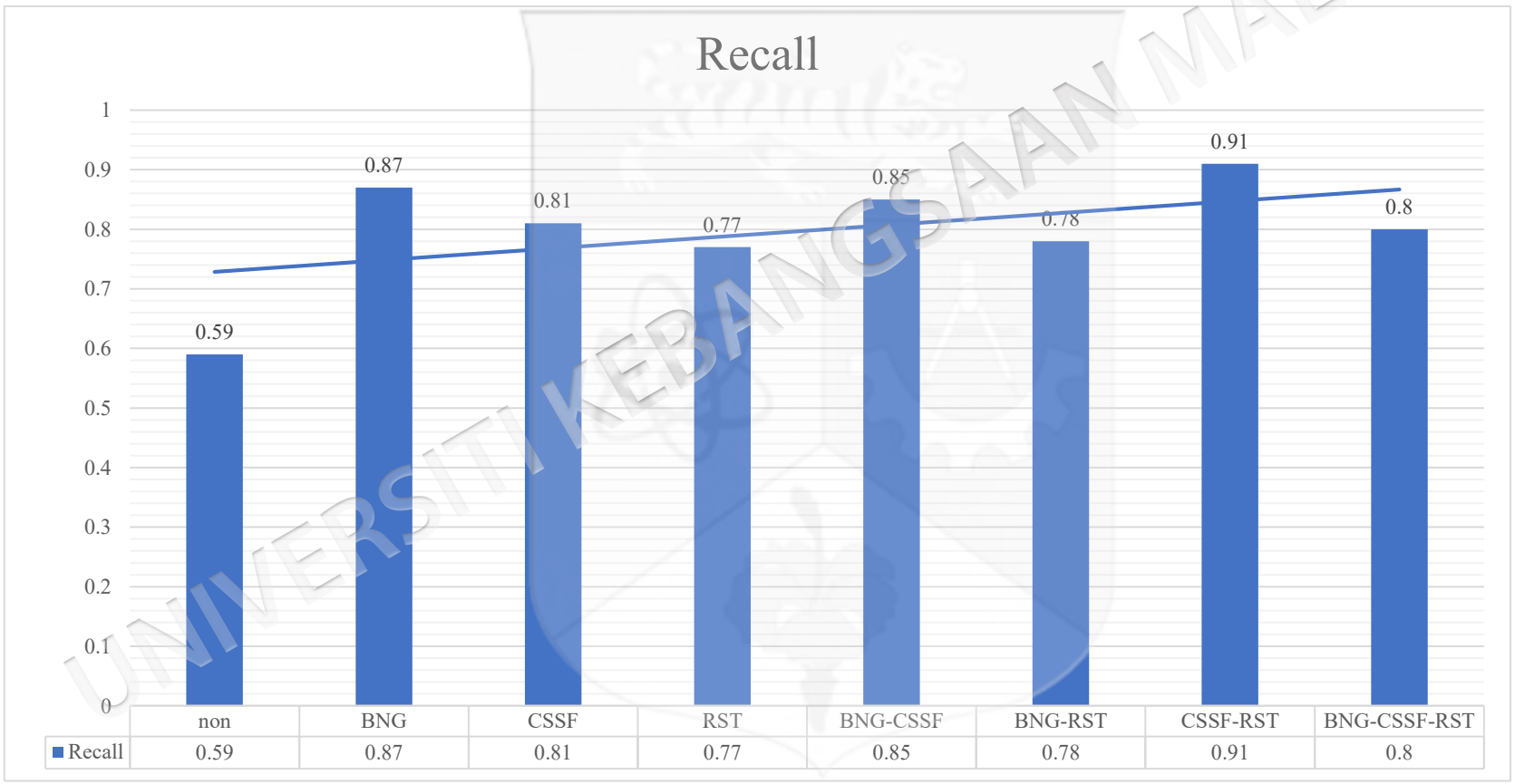


Figure 4.18 Recall for Support Vector Machine

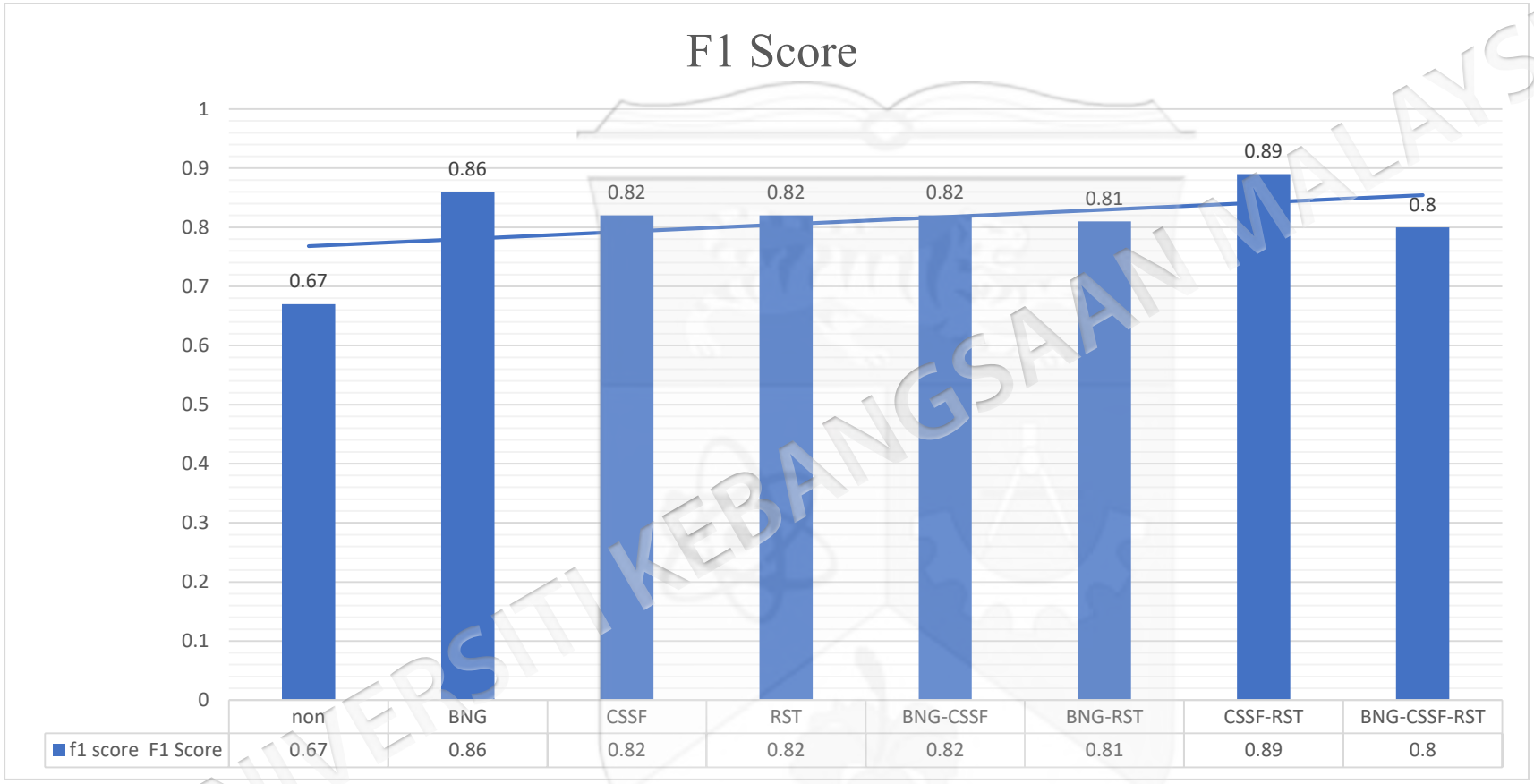


Figure 4.19 F1 score for Support Vector Machine

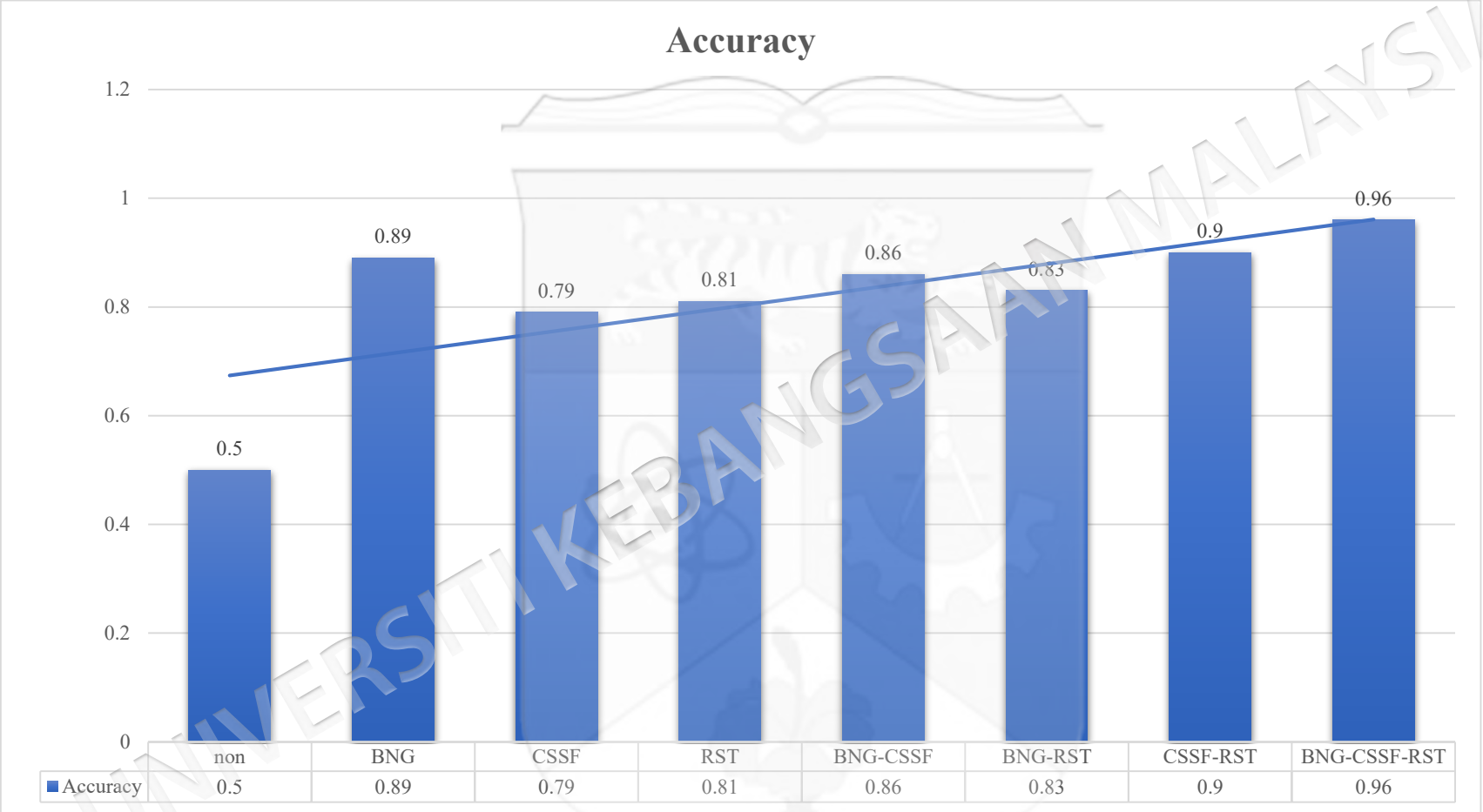


Figure 4.20 Accuracy for Support Vector Machine

This section explores the methods used to enhance the accuracy and quality of diabetes data by employing Bayesian Network Graphs (BNG), Comprehensive Smart Sampling Framework (CSSF), and Rough Set Theory (RST). Each of these techniques addresses specific challenges within the dataset, such as missing data, data imbalance, and feature reduction. By integrating these approaches, a more accurate dataset is created, which significantly improves the performance of predictive models in diabetes research.

After completing preprocessing and feature selection, both Artificial Neural Network (ANN) and Support Vector Machine (SVM) classifiers were trained on the Kaggle diabetes dataset comprising 100,000 patient records with 15 predictor variables and one target label. The most influential predictors identified through feature-importance analysis were HbA1c level (0.403), blood glucose level (0.328), BMI (0.125), and age (0.101), confirming their central role in diabetes risk assessment. The models were trained using a 70/30 train–test split. Table 4.6 summarises the evaluation metrics, showing that the ANN achieved an accuracy of 96.95 % and an AUC-ROC of 0.9727, outperforming the SVM, which achieved 96.32 % accuracy and 0.8993 AUC. Although SVM attained slightly higher precision (0.978 vs 0.943), its recall (0.580) and F1-score (0.728) were lower than those of ANN (0.682 and 0.792, respectively). This indicates that the ANN captured a higher proportion of true-positive diabetic cases while maintaining acceptable precision.

The ANN demonstrated better adaptability to nonlinear feature interactions, especially between biochemical (HbA1c, glucose) and anthropometric (BMI, age) attributes, whereas the SVM exhibited stronger precision and stability but tended to under-detect minority positive cases. These findings validate that the BNG–CSSF–RST preprocessing pipeline substantially improved classifier performance, leading to near-perfect discrimination (AUC ≈ 0.97) on this large dataset. Confusion matrices represent the model's predictions as true positives, true negatives, false positives, and false negatives, as shown in Figure 4.21. For the **ANN model**, the confusion matrix is expected to show a higher number of true positives, indicating that the model is highly effective in identifying diabetic cases. The false negatives are expected to be lower, meaning the ANN model is good at identifying patients who have diabetes. The SVM confusion matrix, however, would likely show more false positives and fewer true

positives due to its lower recall. Consequently, the SVM model does not identify a significant number of diabetic cases, despite being precise.

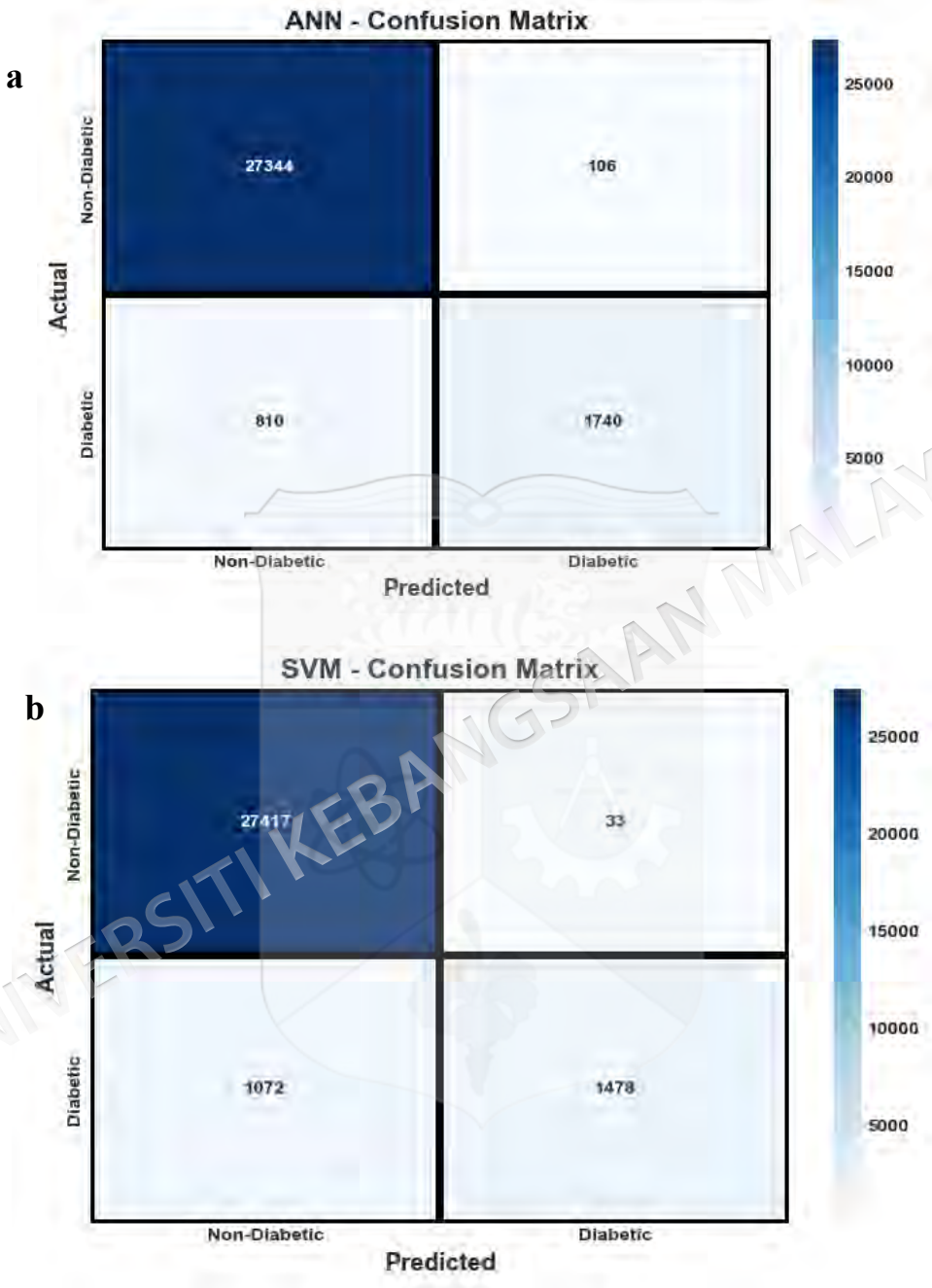


Figure 4.21 Show the confusion matrices: a. ANN model and b. SVM model

ROC curves in Figure 4.22 illustrate the performance of ANN and SVM models for diabetes prediction. ANN model displays superior performance with an AUC of 0.973, indicating its strong ability to differentiate diabetic from non-diabetic patients. The SVM model has a lower AUC of 0.899, reflecting a relatively weaker discriminative ability. The Random Classifier, with an AUC of 0.5, serves as a baseline,

and both models clearly outperform this random guess, highlighting their reliability in diabetes prediction. As a result, the ROC curve demonstrates that the ANN model outperforms the SVM model, making it the more effective choice for diabetes, as shown in Figure 4.22.

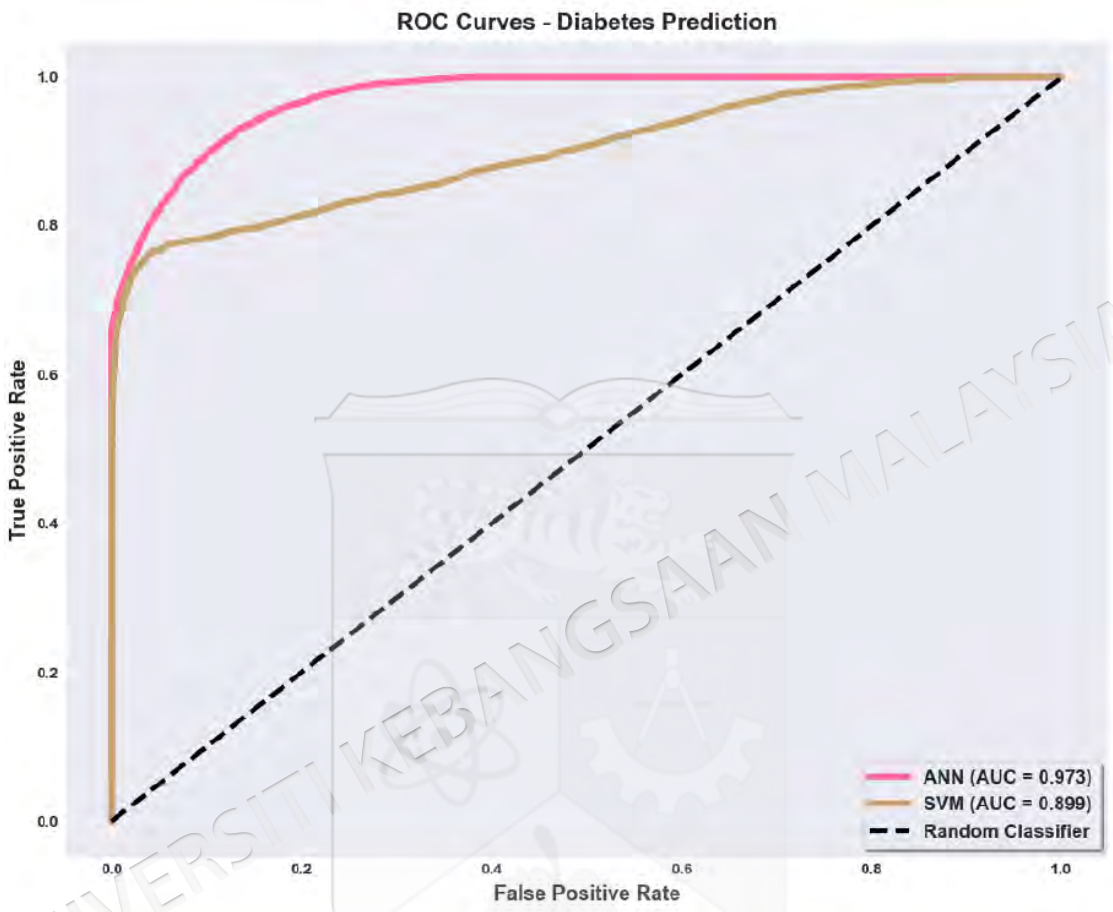


Figure 4.22 Show the performance comparison of ANN and SVM model

4.4 COMPARATIVE PERFORMANCE ANALYSIS OF ANN AND SVM

Table 4.6 provides a summary of the ANN and SVM models' cross-dataset classification accuracies after using the integrated BNG-CSSF-RST preprocessing workflow. Accuracy rates of 0.935 and 0.9026, respectively, were attained by the ANN and SVM models following preprocessing on the Pima dataset (n = 768), the Kaggle Diabetes Clinical Dataset (100k records, 17 features), and the Kaggle Diabetes Prediction Dataset (100k records, 9 features).To start with, the accuracy of all datasets was significantly improved by the preprocessing pipeline ($p < 0.05$). By combining BNG imputation with CSSF cluster-based rebalancing and RST-driven dimensionality

reduction, we were able to increase classifier performance compared to raw-data baselines by improving dataset quality, reducing feature redundancy, addressing missing values, and correcting class imbalance to produce a more reliable and diagnostically meaningful feature space. Second, ANN was superior to SVM on every dataset, but the Pima dataset showed the greatest improvement. This trend suggests that when data quality is improved by rigorous preprocessing, ANN excels at modelling higher-order nonlinear interactions. Third, on Kaggle Clinical dataset, SVM and ANN were almost identical in terms of performance (difference of 0.63%), but on smaller and more diverse datasets, SVM's accuracy stayed consistently lower, indicating that it has limited flexibility when it comes to capturing subtle nonlinear variability outside of large, well-represented feature distributions.

Preprocessing within the proposed BNG–CSSF–RST framework reduces the performance variability that typically appears when comparing classification models. In the two large Kaggle datasets with high sample sizes and feature diversity, ANN and SVM yield closely aligned results after preprocessing, indicating that their predictive performance becomes broadly comparable once data quality issues are resolved. In contrast, the Pima dataset shows a performance gap of more than 3% between ANN and SVM even after preprocessing. This gap indicates that ANN remains sensitive to small-scale nonlinear structures that become more apparent when missing values, class imbalance, and redundant attributes are corrected. This contrast reinforces the framework's objective: enhancing dataset quality so that performance differences across models reflect genuine modelling behaviour rather than artefacts of low-quality data.

Further evidence supporting this conclusion comes from assessments of the area under the curve and the confusion matrix. ANN outperformed SVM with a higher AUC of 0.973 and a lower number of false negatives, while also increasing the number of true positives. Compared to SVM's higher precision, SVM's lower recall indicates a less favourable trade-off in clinical screening tasks that require sensitivity to positive cases. The BNG-CSSF-RST pipeline significantly improves the quality of diabetes datasets and allows for more stable, generalised performance for diverse machine learning classifiers.

Finally, the study employed three datasets that differ in size and feature dimensionality, enabling an evaluation of model performance under varying levels of structural complexity. The Pima dataset from the UCI Machine Learning Repository contains 768 records with 9 features, which were reduced to 5 features of importance (RST selection) after preprocessing. The Diabetes Clinical Dataset (2024) obtained from Kaggle comprises 100,000 records with 17 features, reduced to 11 features of importance in the final selection In contrast, the Diabetes Prediction Dataset (Kaggle, 2023) includes 100,000 records with 9 features, which were reduced to 5 features of importance after preprocessing. This diversity in dataset characteristics and feature reduction demonstrates the effectiveness of the feature selection process in retaining the most informative variables while reducing dimensionality across datasets of varying size and complexity.

Table 4.6 Cross-dataset accuracy comparison of ANN and SVM models using the proposed BNG–CSSF–RST framework

Dataset	**Source**	**Size of data**	**Model**	**Accuracy 0.0**
Pima	UCI Machine Learning Repository	768 *9	ANN	0.9351
			SVM	0.9026
Diabetes Clinical	Kaggle (Diabetes Clinical Dataset, 2024)	100k*17	ANN	0.9695
			SVM	0.9632
Diabetes prediction datasetc	Kaggle Diabetes 2023	100k*9	ANN	0.9149
			SVM	0.8912

As results, in comparison to recent techniques, our contribution lies in improving the quality of diabetes datasets by incorporating multiple datasets with varied sizes, feature complexities, and demographic characteristics. Unlike many studies that rely on a single dataset, we used three distinct diabetes datasets—Pima Indians Diabetes Dataset (PIDD), diabetes clinical dataset (Kaggle in 2023), and diabetes prediction dataset (Kaggle in 2024). These datasets vary in sample size, feature richness, and data quality, reflecting diverse populations and real-world conditions. Additionally, we validated the proposed BNG–CSSF–RST preprocessing framework using both SVN and ANN algorithms, which ensures the robustness and versatility of our methodology

across different modeling techniques. This comprehensive approach distinguishes our study by addressing the challenges of dataset heterogeneity and validation in diabetes prediction.

4.5 SUMMARY

This chapter presented the results of the proposed preprocessing framework and its impact on the performance of diabetes prediction models across multiple datasets. Three major preprocessing strategies were implemented, including BNG for missing-value imputation, CSSF for class-imbalance correction, and RST for dimensionality reduction. These methods collectively enhanced data quality, reduced noise, and improved the discriminative capacity of the classifiers.

As a summary of results, both the ANN and SVM exhibited significant accuracy improvements after applying the integrated BNG–CSSF–RST framework. The ANN achieved accuracies of 0.9351, 0.9695, and 0.9149 for the Pima, Kaggle 2024, and Kaggle 2023 datasets, respectively, outperforming SVM, which achieved 0.9026, 0.9632%, and 0.8912%. These results confirm that effective handling of missing data, class imbalance, and feature reduction contributes to improved classification robustness and generalisability. Overall, the combined preprocessing framework demonstrated its capability to enhance prediction accuracy across heterogeneous diabetes dataset, thereby validating its suitability for reliable diabetes risk assessment and subsequent comparative model evaluation.

CHAPTER V

CONCLUSION AND FUTURE WORKS

5.1 INTRODUCTION

This chapter provides an overall synthesis of the study, highlighting how the research objectives were accomplished, the major findings derived from experimental analysis, and the implications of these findings for medical data analytics. The research addressed key challenges affecting the accuracy and reliability of diabetes prediction models, including missing data, class imbalance, and feature reduction. To overcome these issues, three innovative preprocessing techniques were introduced: the BNG for missing-value imputation, the CSSF for class balancing, and RST for dimensionality reduction. Each of these techniques contributed to improving data integrity, enhancing the performance of predictive models, and increasing generalisability across multiple datasets. This chapter concludes the research by summarising its key outcomes, outlining its contributions and limitations, and proposing potential directions for future research.

5.2 CONCLUSION

This research set out to enhance the quality of diabetes datasets and improve predictive model performance by addressing three persistent data challenges: missing values, class imbalance, and feature reduction. These challenges were systematically resolved through the integration of BNG imputation, CSSF for class balancing, and RST for feature selection. BNG ensured accurate reconstruction of missing values while preserving local data structure, CSSF effectively corrected class distribution through selective synthetic sampling, and RST reduced dimensionality without compromising key predictive information.

The framework was evaluated using three heterogeneous datasets: the Pima Indians Diabetes Dataset, the Kaggle Diabetes Clinical Dataset (2024), and the Kaggle Diabetes Prediction Dataset (2023). The results demonstrated consistent improvements across all datasets. For the Pima dataset, ANN achieved 93.51% accuracy, and SVM achieved 90.26%. Using the Kaggle Diabetes Clinical Dataset, ANN attained 96.95% accuracy and SVM achieved 96.32%, reflecting strong performance on large and feature-rich clinical records. For the Kaggle Diabetes Prediction Dataset, ANN reached 91.49% accuracy, while SVM obtained 89.12%. These outcomes exceed the typical accuracy levels reported in earlier literature and confirm the robustness and adaptability of the proposed framework across varied data structures and population characteristics.

As a result, the results validate that integrating BNG, CSSF, and RST significantly enhances data quality and yields more reliable, accurate predictive models. This study contributes a unified and scalable approach for addressing critical data issues in medical analytics, providing a foundation that can be extended to other disease prediction domains. The findings underscore the critical impact of rigorous data preprocessing on enhancing model performance and demonstrate the potential of the proposed framework to improve the analytical reliability of diabetes prediction systems.

5.3 LIMITATIONS

Despite the significant contributions of this research, certain limitations must be acknowledged. First, the proposed methods were evaluated primarily on diabetes-related datasets, and their generalisability to other medical domains requires further empirical validation. Second, the computational complexity of the BNG and CSSF algorithms may pose challenges for very large-scale datasets without optimization or parallelization. Third, only nine features were used in this research to demonstrate the effectiveness of the preprocessing framework; although this design ensured controlled experimentation, future work should assess the framework on high-dimensional datasets. Moreover, the study focused on preprocessing innovations rather than comprehensive benchmarking against other state-of-the-art methods. It was an intentional scope decision to prioritize establishing a strong methodological foundation before broader comparative evaluations. Data availability and privacy constraints also

limited access to certain clinical datasets, which could have further strengthened the analysis. Addressing these limitations in future research through cross-domain studies, algorithmic optimization, and larger-scale data integration will enhance the robustness, efficiency, and applicability of the proposed framework.

5.4 FUTURE WORK

Future work should also explore different standardization methods and evaluate the proposed approaches under various missing data mechanisms to enhance their robustness and applicability in diverse medical contexts. By continuing to refine and expand upon these methodologies, the advancement of the field of medical data analysis will be upheld.

Additionally, future work will explore benchmarking GANs against CSSF and integrating advanced methods to further enhance predictive accuracy while balancing computational demands. This approach reflects a deliberate choice to align research methods with the study's core objectives and resource considerations.

While this study primarily focused on improving data quality and predictive performance using methods like BNG, CSSF, and RST, only a limited set of classification models were evaluated. Advanced models such as Random Forest and Decision Tree were briefly mentioned but not explored in depth. Future research could incorporate these models to further validate and strengthen the findings of this work. The work significantly advances diabetes prediction by addressing missing data, class imbalance, and feature reduction . Bidirectional Neighbour Graph ensures accurate imputation while preserving dataset structure. Clustering Selection Synthesis Filtering (CSSF) improves class balance by generating high-quality synthetic samples. Rough Set Theory (RST) reduces dimensionality. However, RST was only compared to PCA, which was used due to its extensive use in PCA dimensional deficiency and its calculation efficiency for low-dimensional datasets. Given that this study used a dataset with only nine features, advanced techniques such as core PCA or autoencoders were not preferred. Future work will include LDA, RFE, and deep learning-based autoencoders, which will consider a lot of the construction selection methods. This

extension will increase the scalability, interpretation, and efficiency of reduction in dimensions in datasets, while maintaining essential features for efficient analysis. Benchmarking these methods against modern techniques and testing on diverse datasets could further validate their scalability and applicability in healthcare predictive modeling.

REFERENCES

Abdulrauf Sharifai, G., & Zainol, Z. 2020. Feature Selection for High-Dimensional and Imbalanced Biomedical Data Based on Robust Correlation Based Redundancy and Binary Grasshopper Optimisation Algorithm. *Genes (Basel)*, 11(7), 717. https://doi.org/10.3390/genes11070717

Arabboev, M., Begmatov, S., Saydiakbarov, S., Bobojanov, S., Nosirov, K., & Chedjou, J. C. (2025). DeepDiabFusion: An interaction-aware neural network architecture for diabetes prediction. EAI Endorsed Transactions on AI and Robotics, 4(2025), Article 7998. https://doi.org/10.4108/airo.7998

Abed Mohammed, A. & Sumari, P. 2024. Hybrid K-means and Principal Component Analysis (PCA) for Diabetes Prediction. *International Journal of Computing and Digital Systems* 15(1): 1719-1728.

Almutairi, E. S., & Abbod, M. F. 2023. Machine Learning Methods for Diabetes Prevalence Classification in Saudi Arabia. *Modelling*, 4(1), 37-55.

Arowolo, M. O., Adebiyi, M. O., Adebiyi, A. A., & Olugbara, O. 2021. Optimised hybrid investigative based dimensionality reduction methods for malaria vector using KNN classifier. *Journal of Big Data*, 8(1), 1-14.

Avants, B. B., Tustison, N. J., & Stone, J. R. 2021. Similarity-driven multi-view embeddings from high-dimensional biomedical data. *Nature Computational Science*, 1(2), 143-152.

Aktaa, S. (2022). *Improving cardiovascular care and outcomes: cross-boundary clinical registries and quality indicators* (Doctoral dissertation). University of Leeds.

Ahmed, Z., & Das, S. 2023. A comparative analysis on recent methods for addressing imbalance classification. *SN Computer Science*, 5(1), 30. doi: 10.1007/s42979-023-02357-0

AsriMulyani, Sarah Khoerunisa, Dede Kurniadi, 2025. Perbandingan Kinerja Algoritma KNN dan SVM Menggunakan SMOTE untuk Klasifikasi Penyakit Diabetes. J. Nas. Tek. Elektro dan Teknol. Inf. 14, 25–34. https://doi.org/10.22146/jnteti.v14i1.15198.

Abdollahi, J., Aref, S., 2024. Early Prediction of Diabetes Using Feature Selection and Machine Learning Algorithms. SN Comput. Sci. 5, 217. https://doi.org/10.1007/s42979-023-02545-y

Abousaber, I., Abdallah, H.F. & El-Ghaish, H. 2024. Robust predictive framework for diabetes classification using optimized machine learning on imbalanced datasets. Frontiers in Artificial Intelligence 7: 1499530. https://doi.org/10.3389/frai.2024.1499530

A. Hernandez-Guedes, N. Arteaga-Marrero, E. Villa, G. M. Callico, J. Ruiz-Alzola, Sensors 2023, 23, 1.

A. L. D. Araújo, A. V. B. da Silva, A. R. M. Gonçalves, C. Saldivia-Siracusa, D. L. F. Ferraz, C. B. Calderipe, I. J. Correia-Neto, P. A. Vargas, M. A. Lopes, P. R. F. Bonan, et al., Front. Oral Heal. 2025, 6, 1.

Bai, B. M., Mangathayaru, N., & Rani, B. P. 2015. An approach to find missing data in medical datasets. *Proceedings of the The International Conference on Engineering & MIS 2015* (pp. 1-7)

Bania, R. K., & Halder, A. 2020. R-Ensembler: A greedy rough set-based ensemble attribute selection algorithm with kNN imputation for classification of medical data. *Comput Methods Programs Biomed*, 184, 105122. https://doi.org/10.1016/j.cmpb.2019.105122

Bania, R. K., & Halder, A. 2021. R-HEFS: Rough set based heterogeneous ensemble feature selection method for medical data classification. *Artif Intell Med*, 114, 102049. https://doi.org/10.1016/j.artmed.2021.102049

Bicevskis, J., Nikiforova, A., Bicevska, Z., Oditis, I., & Karnitis, G. 2019. A step towards a data quality theory. *2019 Sixth International Conference on Social Networks Analysis, Management and Security* (SNAMS)

Camino, R. D., Hammerschmidt, C. A. & State, R. 2019. Improving missing data imputation with deep generative models. *arXiv preprint arXiv:1902.10666:*

Chang, V., Bailey, J., Xu, Q. A. & Sun, Z. 2022. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Comput Appl*: 1-17.

Chen, W., Yang, K., Yu, Z., Shi, Y. & Chen, C. L. 2024. A survey on imbalanced learning: latest research, applications and future directions. *Artificial Intelligence Review* 57(6): 1-51.

C. Beyan and R. Fisher, "Classifying imbalanced data sets using similarity based hierarchical decomposition," Pattern Recognition, vol. 48, no. 5, pp. 1653-1672, 2015.

C. M. Reategui-Rivera, W. Cui, S. Escobar-Agreda, L. Rojas-Mezarina, J. Finkelstein, Telemed. Reports 2025, 6, 109.

C. Sabanayagam, F. He, S. Nusinovici, J. Li, C. Lim, G. Tan, C. Y. Cheng, Elife 2023, 12, 1.

Dakka, M. A., Nguyen, T. V., Hall, J. M. M., Diakiw, S. M., VerMilyea, M., Linke, R., Perugini, M., & Perugini, D. 2021. Automated detection of poor-quality data: case studies in healthcare. *Sci Rep*, 11(1), 18005. https://doi.org/10.1038/s41598-021-97341-0

Datta, S., & Das, S. 2015. Near-Bayesian support vector machines for imbalanced data classification with equal or unequal misclassification costs. *Neural Networks*, 70, 39-52.

Daud, S. N. S. S., Sudirman, R. & Shing, T. W. 2023. Safe-level SMOTE method for handling the class imbalanced problem in electroencephalography dataset of adult anxious state. *Biomedical Signal Processing and Control 83: 104649*.

Deng, Y., Chang, C., Ido, M. S., & Long, Q. 2016. Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data. *Sci Rep*, 6(1), 21689. https://doi.org/10.1038/srep21689

Devi, D., Biswas, S. K., & Purkayastha, B. 2019. Learning in presence of class imbalance and class overlapping by using one-class SVM and undersampling technique. *Connection Science*, 31(2), 105-142.

Devi, D., Biswas, S. K., & Purkayastha, B. 2020. A review on solution to class imbalance problem: Undersampling approaches. *2020 International Conference on Computational Performance Evaluation* (ComPE)

Dinesh, M., & Prabha, D. 2021. Diabetes mellitus prediction system using hybrid KPCA-GA-SVM feature selection techniques. *Journal of Physics: Conference Series*

D. Tarasewicz, A. J. Karter, N. Pimentel, H. H. Moffet, K. K. Thai, D. Schlessinger, O. Sofrygin, R. B. Melles, Diabetes Care 2023, 46, 1068.

ElSeddawy, A. I., Karim, F. K., Hussein, A. M., & Khafaga, D. S. 2022. Predictive Analysis of Diabetes-Risk with Class Imbalance. *Computational Intelligence and Neuroscience*.

Enríquez, M., Naranjo, S., Amaro, I., & Camacho, F. 2021. Dimensionality reduction using pca and cur algorithm for data on covid-19 tests. In *Artificial Intelligence, Computer and Software Engineering Advances: Proceedings of the CIT 2020 Volume 1* (pp. 121-134). Springer.

Farnoosh, R.; Abnoosian, K.; Isewid, R. A.; Javaheri, D. DiabetesXpertNet: An Innovative Attention-Based CNN for Accurate Type 2 Diabetes Prediction. PLoS One 2025, 20 (9), e0330454. https://doi.org/10.1371/journal.pone.0330454.

Fan, W., Stolfo, S. J., Zhang, J., & Chan, P. K. 1999. AdaCost: misclassification cost-sensitive boosting. *Icml*

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., Herrera, F., Fernández, A., García, S., Galar, M., & Prati, R. C. 2018. Cost-sensitive learning. *Learning from imbalanced data sets*, 63-78.

Fujita, H., Wang, Y., Xiao, Y., & Moonis, A. 2023. Advances and Trends in Artificial Intelligence. *Theory and Applications: 36th International Conference on*

Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2023, Shanghai, China, July 19–22, Proceedings, Part I.

Fujiwara, K., Huang, Y., Hori, K., Nishioji, K., Kobayashi, M., Kamaguchi, M., & Kano, M. 2020. Over- and Under-sampling Approach for Extremely Imbalanced and Small Minority Data Problem in Health Record Analysis. *Front Public Health*, 8, 178. https://doi.org/10.3389/fpubh.2020.00178

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. 2017. Learning from class-imbalanced data: Review of methods and applications, *Expert systems with applications*, vol. 73, pp. 220-239.

Idris, N. F., Ismail, M. A., Jaya, M. I. M., Ibrahim, A. O., Abulfaraj, A. W., & Binzagr, F. (2024). Stacking with Recursive Feature Elimination-Isolation Forest for classification of diabetes mellitus. Plos one, 19(5), e0302595.

Gaidai, O., Cao, Y., & Loginov, S. 2023. Global cardiovascular diseases death rate prediction. *Current Problems in Cardiology*, 101622.

Galand, L., Ismaili, A., Perny, P. & Spanjaard, O. 2013. Bidirectional preference-based search for state space graph problems. *Proceedings of the International Symposium on Combinatorial Search*, pp.80-88.

Gangavarapu, T., & Patil, N. 2019. A novel filter–wrapper hybrid greedy ensemble approach optimised using the genetic algorithm to reduce the dimensionality of high-dimensional biomedical datasets. *Applied Soft Computing*, 81, 105538.

Gao, M., Hong, X., Chen, S., & Harris, C. J. 2011. A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems. *Neurocomputing*, 74(17), 3456-3466.

Ghojogh, B., & Crowley, M. 2019. The theory behind overfitting, cross validation, regularization, bagging, and boosting tutorial. *arXiv preprint arXiv:1905.12787*.

Ghorbani, M., Kazi, A., Baghshah, M. S., Rabiee, H. R., & Navab, N. 2022. RA-GCN: Graph convolutional network for disease prediction problems with imbalanced data. *Medical Image Analysis*, 75, 102272.

Ghosh, J., & Nag, A. 2001. An overview of radial basis function networks. Radial basis function networks 2: *New Advances in Design*, 1-36.

Grzyb, J., Klikowski, J., & Woźniak, M. (2021). Hellinger distance weighted ensemble for imbalanced data stream classification. Journal of Computational Science, 51, 101314.

Grzyb, J. & Woźniak, M. 2023. SVM ensemble training for imbalanced data classification using multi-objective optimization techniques. Applied Intelligence 53(12): 15424–15441.

Güney, H. 2023. Feature selection - Integrated classifier optimisation algorithm for network intrusion detection. *Concurrency and Computation: Practice and Experience*, e7807.

Gewandter, J. S., Dworkin, R. H., Turk, D. C., Devine, E. G., Hewitt, D., Jensen, M. P., et al. (2020). Improving study conduct and data quality in clinical trials of chronic pain treatments: IMMPACT recommendations. *The Journal of Pain*, 21(9-10), 931-942.

Saadatfar, H., Khosravi, S., Joloudari, J. H., Mosavi, A., & Shamshirband, S. 2020. A new K-nearest Neighbours classifier for big data based on efficient data pruning, *Mathematic*s, vol. 8, no. 2, p. 286.

Haixiang, G., Yijing, L., Yanan, L., Xiao, L., & Jinling, L. 2016. BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification. *Engineering Applications of Artificial Intelligence*, 49, 176-193.

Hassan, M. M., & Amiri, N. 2019. Classification of imbalanced data of diabetes disease using machine learning algorithms. *Age (years)*, 21(81), 33.24.

He, H., & Garcia, E. A. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), pp. 1263–1284. doi: 10.1109/TKDE.2008.239

Hu, S., Liang, Y., Ma, L., & He, Y. 2009. MSMOTE: Improving classification performance when training data is imbalanced. *2009 second international workshop on computer science and engineering*

Huang, C.-M., Hsu, M.-Y., Hung, C.-S., Lin, C.-H. R., & Chen, S.H. 2023. The Enhancement of Classification of Imbalanced Dataset for Edge Computing. *New Trends in Computer Technologies and Applications: 25th International Computer Symposium, ICS 2022, Taoyuan, Taiwan, December 15–17*, 2022, Proceedings

Huang, S.-F., & Cheng, C.-H. 2020. A safe-region imputation method for handling medical data with missing values. *Symmetry*, 12(11), 1792.

Iqbal, S., Maryam, H., Qureshi, K. N., Javed, I. T., & Crespi, N. 2021. Automised flow rule formation by using machine learning in software defined networks-based edge computing. *Egyptian Informatics Journal*.

Islam, A., Belhaouari, S. B., Rehman, A. U. & Bensmail, H. 2022. KNNOR: An oversampling technique for imbalanced datasets. *Applied Soft Computing* 115: 108288.

Islam, M. R., Lima, A. A., Das, S. C., Mridha, M., Prodeep, A. R., & Watanobe, Y. 2022. A comprehensive survey on the process, methods, evaluation, and challenges of feature selection. *IEEE Access*.

Jabbar, Z., Washington, P., 2024. The Effect of Data Missingness on Machine Learning Predictions of Uncontrolled Diabetes Using All of Us Data. BioMedInformatics 4, 780–795. https://doi.org/10.3390/biomedinformatics4010043.

Jackson, C., Stevens, J., Ren, S., Latimer, N., Bojke, L., Manca, A., & Sharples, L. 2017. Extrapolating survival from randomized trials using external data: a review of methods. *Medical Decision Making*, 37(4), 377-390.

Jafarigol, E. & Trafalis, T. B. 2023. Federated Learning with GANs-based Synthetic Minority Over-sampling Technique for Improving Weather Prediction from Imbalanced Data.

Kaliappan, J., Saravana Kumar, I.J., Sundaravelan, S., Anesh, T., Rithik, R.R., Singh, Y., Vera-Garcia, D. V., Himeur, Y., Mansoor, W., Atalla, S., Srinivasan, K., 2024. Analyzing classification and feature selection strategies for diabetes prediction across diverse diabetes datasets. Front. Artif. Intell. 7. https://doi.org/10.3389/frai.2024.1421751.

Kabir, M. F., Chen, T., & Ludwig, S. A. 2023. A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. *Healthcare Analytics*, 3, 100125.

Kang, Q., Shi, L., Zhou, M., Wang, X., Wu, Q., & Wei, Z. 2017. A distance-based weighted undersampling scheme for support vector machines and its application to imbalanced classification. *IEEE transactions on neural networks and learning systems,* 29(9), 4152-4165.

Keele, S. 2007. Guidelines for performing systematic literature reviews in software engineering. In: *Technical report, ver. 2.3 ebse technical report. ebse*.

Khaldy, M., & Kambhampati, C. 2018. Resampling imbalanced class and the effectiveness of feature selection methods for heart failure dataset. *International Robotics & Automation Journal*, 4(1), 1-10.

Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F. A., & Togneri, R. 2017. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems,* 29(8), 3573-3587.

Khan, S. I., & Hoque, A. 2020. SICE: an improved missing data imputation technique. *J Big Data*, 7(1), 37. https://doi.org/10.1186/s40537-020-00313-w

Khushi, M., Shaukat, K., Alam, T. M., Hameed, I. A., Uddin, S., Luo, S., Yang, X., & Reyes, M. C. 2021. A comparative performance analysis of data resampling methods on imbalance medical data. *IEEE Access*, 9, 109960-109975.

Kibria, H. B., Nahiduzzaman, M., Goni, M. O. F., Ahsan, M., & Haider, J. 2022. An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI. *Sensors*, 22(19), 7268.

Kirasich, K., Smith, T. & Sadler, B. 2018. Random forest vs logistic regression: binary classification for heterogeneous datasets. *SMU Data Science Review* 1(3): 9.

Kovács, B., Tinya, F., Németh, C., & Ódor, P. 2020. Unfolding the effects of different forestry treatments on microclimate in oak forests: results of a 4 - yr experiment. *Ecological Applications*, 30(2), e02043.

Krawczyk, B. 2016. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4), 221-232.

Kumar, B., & Mathur, H. 2022. Atherosclerosis Disease Prediction Based on Feature Optimisation and Ensemble Classifier. *Intelligent Sustainable Systems: Selected Papers of World* S4 2021, Volume 1,

Kumar, S., Biswas, S. K., & Devi, D. 2019. TLUSBoost algorithm: a boosting solution for class imbalance problem. *Soft Computing*, 23, 10755-10767.

Kumari, S., Kumar, D., & Mittal, M. 2021. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, 2, 40-46.

Kurman, S., & Kisan, S. 2023. An in-depth and contrasting survey of meta-heuristic approaches with classical feature selection techniques specific to cervical cancer. *Knowledge and Information Systems*, 1-54.

Kumar, A., Singh, D., & Shankar Yadav, R. (2024). Class overlap handling methods in imbalanced domain: A comprehensive survey. *Multimedia Tools and Applications*, 1–48.

Kaka-Khan, K.M., Mahmud, H. & Ali, A.A. 2022. Rough set-based feature selection for predicting diabetes using logistic regression with stochastic gradient decent algorithm. *UHD Journal of Science and Technology* 6(2): 85-93.

Khan, F.A., Zeb, K., Al-Rakhami, M., Derhab, A. & Bukhari, S.A.C. 2021. Detection and prediction of diabetes using data mining: A comprehensive review. IEEE Access 9: 43711-43735. https://doi.org/10.1109/ACCESS.2021.3059343

Latha, C. B. C., & Jeeva, S. C. 2019. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, 16, 100203.

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. 2018. A survey on addressing high-class imbalance in big data. *Journal of Big data*, 5(1), 1-30.

Liang, J., Liu, Q., Nie, N., Zeng, B. & Zhang, Z. 2019. An improved algorithm based on KNN and random forest. *Proceedings of the 3rd International Conference on Computer Science and Application Engineering*, pp.1-6.

Little, R. J. & Rubin, D. B. 2019. Statistical analysis with missing data. *Vol. 793John Wiley & Sons.*

Luo, F., Qian, H., Wang, D., Guo, X., Sun, Y., Lee, E.S., Teong, H.H., Lai, R.T.R. & Miao, C. 2022. Missing value imputation for diabetes prediction. Proceedings of the International Joint Conference on Neural Networks 2022-July. https://doi.org/10.1109/IJCNN55064.2022.9892398

Matetić, I., Štajduhar, I., Wolf, I., & Ljubic, S. 2022. A Review of Data-Driven Approaches and Techniques for Fault Detection and Diagnosis in HVAC Systems, Sensors, vol. 23, no. 1, p. 1.

Mashoufi, M., Ayatollahi, H., Khorasani-Zavareh, D., & Boni, T. T. A. 2023. Data Quality in Health Care: Main Concepts and Assessment Methodologies. *Methods of Information in Medicine*.

Mulyani, A., Khoerunisa, S. & Kurniadi, D. (2025). Perbandingan kinerja algoritma KNN dan SVM menggunakan SMOTE untuk klasifikasi penyakit diabetes. Jurnal Nasional Teknik Elektro dan Teknologi Informasi, 14(1), 25-34.

M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," Neural networks, vol. 106, pp. 249-259, 2018.

Matetić, I., Štajduhar, I., Wolf, I. & Ljubic, S. 2022. A Review of Data-Driven Approaches and Techniques for Fault Detection and Diagnosis in HVAC Systems. *Sensors* 23(1): 1.

Mirzaei, B., Nikpour, B., & Nezamabadi-pour, H. 2021. CDBH: A clustering and density-based hybrid approach for imbalanced data classification. *Expert Systems with Applications,* 164, 114035.

Misztal, M. 2018. Comparison of selected multiple imputation methods for continuous variables–preliminary simulation study results. Acta Universitatis Lodziensis. Folia Oeconomica 6(339): 73-98.

Mozaffar, M., Liao, S., Xie, X., Saha, S., Park, C., Cao, J., Liu, W. K. & Gan, Z. 2022. Mechanistic artificial intelligence (mechanistic-AI) for modelling, design, and control of advanced manufacturing processes: Current state and perspectives. *Journal of Materials Processing Technology* 302: 117485.

Mohsen, F., Al-Absi, H.R.H., Yousri, N.A., El Hajj, N. & Shah, Z. 2023. A scoping review of artificial intelligence-based methods for diabetes risk prediction. npj Digital Medicine 6(1): 1-15. https://doi.org/10.1038/s41746-023-00933-5.

M. A. Messelu, T. Ayenew, H. Amha, B. T. Amlak, M. Gedfew, B. G. Tiruneh, A. Teym, T. Degu, T. M. Dessie, M. F. Tewelgne, et al., Nurs. Crit. Care 2025, 30, e70190.

Nnamoko, N., & Korkontzelos, I. (2020). Efficient treatment of outliers and class imbalance for diabetes prediction. Artificial intelligence in medicine, 104, 101815.

Nagarajan, G., & Dhinesh Babu, L. D. 2019. A hybrid of whale optimisation and late acceptance hill climbing based imputation to enhance classification performance in electronic health records. *J Biomed Inform*, 94, 103190. https://doi.org/10.1016/j.jbi.2019.103190

Nijman, S., Leeuwenberg, A., Beekers, I., Verkouter, I., Jacobs, J., Bots, M., Asselbergs, F., Moons, K., & Debray, T. 2022. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *Journal Of Clinical Epidemiology,* 142, 218-229.

Nugroho, H., Utama, N. P. & Surendro, K. 2021. Class center-based firefly algorithm for handling missing data. *Journal of Big Data* 8(1): 37.

Nugroho, H., Utama, N. P., & Surendro, K. 2020. Performance evaluation for class center-based missing data imputation algorithm. *Proceedings of the 2020 9th International Conference on Software and Computer Applications*

O. Sagi and L. Rokach, "Ensemble learning: A survey," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 8, no. 4, p. e1249, 2018.

Phongying, M., Hiriote, S., 2023. Diabetes Classification Using Machine Learning Techniques. Computation 11. https://doi.org/10.3390/computation11050096.

Peng, D., Zou, M., Liu, C. & Lu, J. 2021. RESI: a region-splitting imputation method for different types of missing data. *Expert Systems with Applications* 168: 114425.

Piri, S. 2020. Missing care: A framework to address the issue of frequent missing values; The case of a clinical decision support system for Parkinson's disease. *Decision Support Systems*, 136, 113339.

Pratama, I., Permanasari, A. E., Ardiyanto, I. & Indrayani, R. 2016. A review of missing datahandling methods on time-series data. *2016 international conference on information technology systems and innovation (ICITSI)*, pp.1-6.

Polikar, R. 2006. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3), pp. 21–45. doi: 10.1109/MCAS.2006.1688199

Quindroit, P., Fruchart, M., Degoul, S., Perichon, R., Martignène, N., Soula, J., Marcilly, R., & Lamer, A. 2023. Definition of a Practical Taxonomy for Referencing Data Quality Problems in Health Care Databases. *Methods of Information in Medicine.*

Razavi-Far, R., Cheng, B., Saif, M., & Ahmadi, M. 2020. Similarity-learning information-fusion schemes for missing data imputation. *Knowledge-Based Systems*, 187, 104805.

Roy, K., Ahmad, M., Waqar, K., Priyaah, K., Nebhen, J., Alshamrani, S. S., Raza, M. A., Ali, I. & Uddin, M. I. 2021. An Enhanced Machine Learning Framework for Type 2 Diabetes Classification Using Imbalanced Data with Missing Values. Complexity 2021: 1-21.

Reza, M. S., Amin, R., Yasmin, R., Kulsum, W., & Ruhi, S. (2024). Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data. Heliyon, 10(2).

Reza, M. S., Amin, R., Yasmin, R., Kulsum, W. & Ruhi, S. Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data. Heliyon 10, (2024).

Salmi, N. & Rustam, Z. 2019. Naïve Bayes classifier models for predicting the colon cancer. *IOP Conference Series: Materials Science and Engineering*, p.052068.

Santhanam, T., & Padmavathi, M. 2015. Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. *Procedia Computer Science*, 47, 76-83.

Sefidian, A. M. & Daneshpour, N. 2020. Estimating missing data using novel correlation maximization-based methods. *Applied Soft Computing* 91: 106249.

Seng, Z., Kareem, S. A., & Varathan, K. D. 2021. A neighborhood undersampling stacked ensemble (NUS-SE) in imbalanced classification. *Expert Systems with Applications*, 168, 114246.

Sezer, E. & Başeğmez, H. 2021. An approach based on feature selection for missing value imputation. *International Conference on Intelligent and Fuzzy Systems*, pp.945-950.

Shaw, S. S., Ahmed, S., Malakar, S., & Sarkar, R. 2021. An ensemble approach for handling class imbalanced disease datasets. *Proceedings of International Conference on Machine Intelligence and Data Science Applications: MIDAS* 2020

Sowjanya, A. M., & Mrudula, O. 2023. Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms, *Applied Nanoscience*, vol. 13, no. 3, pp. 1829-1840. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8811587/pdf/13204_2021_Article_2063.pdf

Sanwouo, B., Quinton, C., & Rouvoy, R. 2024. TS-Pothole: Automated Imputation of Missing Values in Univariate Time Series. *Neural Computing and Applications*, 1-33. https://doi.org/10.1007/s00521-024-10391-z

Wongvorachan, T., He, S., & Bulut, O. 2023. A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information*, 14(1), 54. doi: 10.3390/info14010054

Talebi Moghaddam, M., Jahani, Y., Arefzadeh, Z., Dehghan, A., Khaleghi, M., Sharafi, M., Nikfar, G., 2024. Predicting diabetes in adults: identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm. BMC Med. Res. Methodol. 24, 220. https://doi.org/10.1186/s12874-024-02341-z

Tian, Y., Bian, B., Tang, X. & Zhou, J. 2021. A new non-kernel quadratic surface approach for imbalanced data classification in online credit scoring. *Information Sciences* 563: 150-165.

Tan, K.R., Seng, J.J.B., Kwan, Y.H., Chen, Y.J., Zainudin, S.B., Loh, D.H.F., Liu, N. & Low, L.L. 2023. Evaluation of machine learning methods developed for prediction of diabetes complications: A systematic review. Journal of Diabetes Science and Technology 17(2): 474-489. https://doi.org/10.1177/19322968211056917

ÜresİN, U. 2021. Correlation based regression imputation (CBRI) method for missing data imputation. *Turkish Journal of Science and Technology* 16(1): 39-46.

Wang, Q., Cao, W., Guo, J., Ren, J., Cheng, Y. & Davis, D. N. 2019. DMP_MI: an effective diabetes mellitus classification algorithm on imbalanced data with missing values. *IEEE Access* 7: 102232-102238.

Wang, L., Han, M., Li, X., Zhang, N., & Cheng, H. 2021. Review of classification methods on unbalanced data sets, *IEEE Access*, vol. 9, pp. 64606-64628, 2021.

Wang YC, Cheng CH. A multiple combined method for rebalancing medical data with class imbalances. Comput Biol Med. 2021 Jul;134:104527. doi: 10.1016/j.compbiomed.2021.104527. Epub 2021 May 31. PMID: 34091384.

Wang, X., Zhai, M., Ren, Z., Ren, H., Li, M., Quan, D., & Qiu, L. 2021, Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier, *BMC medical informatics and decision making*, vol. 21, no. 1, pp. 1-14.

Wang, X., Wang, H., and Wang, Y., 2020. A density weighted fuzzy outlier clustering approach for class imbalanced learning, *Neural Computing and Applications*, vol. 32, pp. 13035-13049.

Xie, X., Liu, H., Zeng, S., Lin, L., Li, W., 2021. A novel progressively undersampling method based on the density peaks sequence for imbalanced data. Knowledge-Based Syst. 213, 106689. https://doi.org/https://doi.org/10.1016/j.knosys.2020.106689 .

Xu, Z., Shen, D., Nie, T. & Kou, Y. 2020. A hybrid sampling algorithm combining M-SMOTE and ENN based on Random Forest for medical imbalanced data. *J Biomed Inform* 107: 103465.

Yang, W., Pan, C. & Zhang, Y. 2022. An oversampling method for imbalanced data based on spatial distribution of minority samples SD-KMSMOTE. *Scientific Reports* 12(1): 16820.

Yew, A. Y. L., & Rahim, M. S. M. 2021. Dimensionality Reduction Methods for Alzheimer's Disease Classification. *International Journal of Computing and Digital Systems*, 10.

Fu, Y., Liang, X., Yang, X., Li, L., Meng, L., Wei, Y., ... & Qin, Y. (2024). Stacking model framework reveals clinical biochemical data and dietary behavior features associated with type 2 diabetes: A retrospective cohort study. APL bioengineering, 8(4).

You, J., Ma, X., Ding, Y., Kochenderfer, M. J., & Leskovec, J. 2020. Handling missing data with graph representation learning. *Advances in Neural Information Processing Systems*, 33, pp. 19075–19087

Yan, D., Li, X., Wang, Y. & Cai, Z. 2025. Optimized prediction of diabetes complications using ensemble learning with Bayesian optimization: a cost-efficient laboratory-based approach. Frontiers in Endocrinology 16: 1-18. https://doi.org/10.3389/fendo.2025.1593068

Zhang, Y., Liu, Y., Wang, Y., & Yang, J. 2023. An ensemble oversampling method for imbalanced classification with prior knowledge via generative adversarial network. *Chemometrics and Intelligent Laboratory Systems*, 104775.

Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, and Y. Zhou, "A novel ensemble method for classifying imbalanced data," Pattern Recognition, vol. 48, no. 5, pp. 1623-1637, 2015.

Zhao, Z. 2023. Transforming ECG Diagnosis: An In-depth Review of Transformer-based DeepLearning Models in Cardiovascular Disease Detection. *arXiv preprint arXiv:2306.01249:*

Zhu, R., Wang, Y., Liu, J. X., & Dai, L. Y. 2021. IPCARF: improving lncRNA-disease association prediction using incremental principal component analysis feature selection and a random forest classifier. *BMC bioinformatics*, 22(1), 175. https://doi.org/10.1186/s12859-021-04104-9

APPENDICES

APPENDIX A

LIST OF ALGORITHMS PSEUDO-CODES

Code 1: Python For Normalisation using z- score (Little & Rubin 2019)

```
import numpy as np
def z_score_normalisation(dataset):
Normalise the dataset using z-score normalisation.
Parameters:
dataset (numpy.ndarray): The dataset to be normalised, shape (m, n)
Returns:
numpy.ndarray: The normalised dataset.
# Number of items (m) and number of features (n)
m, n = dataset.shape
# Initialise the normalised dataset
normalised_dataset = np.zeros((m, n))
# Loop through each feature
for i in range(n):
# Calculate the mean (μ_i) and standard deviation (σ_i) for feature i
mean_i = np.mean(dataset[:, i])
std_i = np.std(dataset[:, i])
# Normalise each item for the feature i
for j in range(m):
normalised_dataset[j, i] = (dataset[j, i] - mean_i) / std_i
return normalised_dataset
# Example usage
if __name__ == "__main__":
# Sample dataset with 3 items and 3 features
```

```
sample_dataset = np.array([

[1, 2, 3],

[4, 5, 6],

[7, 8, 9]

# Apply z-score normalisation

normalised_data = z_score_normalisation(sample_dataset)

print("Normalised Dataset:")

print(normalised_data)
```

Code 2: Python: For (Cleaning Data) Removing Duplicate Entries, Handling Outliers

```
import pandas as pd

import numpy as np

# Example: Load a diabetic dataset

# Assuming the dataset is in a CSV file named 'diabetic_data.csv'

df = pd.read_csv('diabetic_data.csv')

# Display the original number of rows for comparison

print(f"Original number of rows: {len(df)}")

# Remove duplicate entries

df = df.drop_duplicates()

print(f"Number of rows after removing duplicates: {len(df)}")

# Handling outliers using the IQR method

Q1 = df.quantile(0.25)

Q3 = df.quantile(0.75)

IQR = Q3 - Q1

# Define an outlier as any value that is more than 1.5 times the IQR away from the quartiles

is_outlier = ((df < (Q1 - 1.5 * IQR)) | (df > (Q3 + 1.5 * IQR))).any(axis=1)
```

```
df_cleaned = df[~is_outlier]
print(f"Number of rows after removing outliers: {len(df_cleaned)}")
# You may also want to inspect the outliers before deciding to remove them
outliers = df[is_outlier]
# Display some of the outliers for inspection
print("Sample outliers:")
print(outliers.head())
```

code 3: Imputation of Missing data using BNG

```
import numpy as np
import pandas as pd
from sklearn.metrics.pairwise import cosine_similarity
from scipy.sparse import csr_matrix
def compute_similarity_matrix(data):
# Use Euclidean distance  as the similarity measure
similarity_matrix = cosine_similarity(data)
return similarity_matrix
def get_Neighbours(similarity_matrix, index, num_Neighbours=5):
# Get the indices of the top 'num_Neighbours' similar entries for a given data point
Neighbours = np.argsort(similarity_matrix[index])[::-1][1:num_Neighbours+1]
return Neighbours
def bidirectional_Neighbour_graph_imputation(data, num_Neighbours=5):
# Initialise the data array (assuming NaNs represent missing values)
imputed_data = data.copy()
# Compute the similarity matrix based on available data
similarity_matrix = compute_similarity_matrix(data.fillna(0))
# Iterate through each data point
for i in range(data.shape[0]):
# Check if there are missing values
missing_features = np.isnan(data[i])
if np.any(missing_features):
```

```
# Get the 'num_Neighbours' most similar entries
Neighbours = get_Neighbours(similarity_matrix, i, num_Neighbours)
# For each missing feature, compute the weighted average of Neighbours' feature values
for feature in np.where(missing_features)[0]:
# Compute weights (similarity scores) for Neighbours
weights = similarity_matrix[i][Neighbours]
# Get the Neighbours' feature values
Neighbour_values = data.iloc[Neighbours, feature]
# Compute the weighted average, excluding any NaNs in Neighbours
valid_weights = weights[~Neighbour_values.isnull()]
valid_values = Neighbour_values[~Neighbour_values.isnull()]
imputed_value = np.dot(valid_weights, valid_values) / valid_weights.sum()
# Impute the missing value
imputed_data.iloc[i, feature] = imputed_value
return imputed_data
```

Code4: Applying CSSF

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
from sklearn.ensemble import Random Forest Classifier
# Step 1: Understanding Class Imbalance
# Load a sample dataset (e.g., from a CSV file)
# df = pd.read_csv('diabetes.csv')
# For demonstration purposes, let's create a synthetic dataset
np.random.seed(42)
features = np.random.rand(1000, 10)  # 1000 samples, 10 features
labels = np.random.choice([0, 1], size=(1000,), p=[0.9, 0.1])  # Imbalanced classes
# Check class distribution
class_distribution = pd.Series(labels).value_counts()
```

```
print("Class Distribution:\n", class_distribution)

# Step 2: Importing Libraries

# (Already imported at the beginning)

# Step 3: Splitting the Data

# Split the dataset into training and testing sets

X_train, X_test, y_train, y_test = train_test_split(features, labels, test_size=0.3,
random_state=42)

# Step 4: Applying CSSF

# Placeholder function for CSSF - replace with actual CSSF implementation

def apply_cssf(X, y):

# Implement the CSSF technique here

# This function should return the augmented data

# For now, return the data unchanged (as a placeholder)

return X, y

X_train_augmented, y_train_augmented = apply_cssf(X_train, y_train)

# Step 5: Model Training

# Train a classification model using the augmented training data obtained from CSSF

model = RandomForestClassifier(random_state=42)

model.fit(X_train_augmented, y_train_augmented)

# Step 6: Model Evaluation

# Evaluate the model's performance on the testing set and measure accuracy,
precision, recall, and F1-score

y_pred = model.predict(X_test)

# Calculate metrics

accuracy = accuracy_score(y_test, y_pred)

precision = precision_score(y_test, y_pred)

recall = recall_score(y_test, y_pred)

f1 = f1_score(y_test, y_pred)

print(f"Accuracy: {accuracy}")

print(f"Precision: {precision}")

print(f"Recall: {recall}")

print(f"F1 Score: {f1}")
```

Code 5 : Synthetic Minority Oversampling Technique (SMOTE) (Sowjanya & Mrudula 2023)

```
import numpy as np
from sklearn.Neighbours import NearestNeighbours
def smote(D, N, k):
Synthetic Minority Oversampling Technique (SMOTE)
Parameters:
D (numpy.ndarray): Minority class dataset, shape (n_minority_samples, n_features)
N (int): Number of synthetic samples to generate for each minority class sample
k (int): Number of nearest Neighbours
Returns:
numpy.ndarray: Augmented dataset with synthetic samples
n_minority_samples, n_features = D.shape
synthetic_samples = np.zeros((n_minority_samples * N, n_features))
# Step 1: Find k-nearest Neighbours for each minority sample
nbrs = NearestNeighbours(n_Neighbours=k).fit(D)
Neighbours = nbrs.kNeighbours(D, return_distance=False)
# Step 2: Generate synthetic samples
for i in range(n_minority_samples):
for n in range(N):
# Step 1.2: Randomly select one of the k-nearest Neighbours
nn_index = np.random.choice(Neighbours[i])
x_i = D[i]
x_nn = D[nn_index]
# Step 2.1: Calculate the difference vector
diff = x_nn - x_i
# Step 2.2: Generate a random number r between 0 and 1
r = np.random.rand()
# Step 2.3: Create a synthetic sample
x_syn = x_i + r * diff
# Step 3: Add the synthetic sample to the augmented dataset
```

```
synthetic_samples[i * N + n] = x_syn
# Step 5: Combine the original dataset with the synthetic samples
augmented_dataset = np.vstack([D, synthetic_samples])
return augmented_dataset
# Example usage
if __name__ == "__main__":
# Create a synthetic minority class dataset for demonstration
np.random.seed(42)
minority_class_samples = np.random.rand(10, 5)  # 10 samples, 5 features
# Apply SMOTE
N = 5  # Number of synthetic samples to generate for each minority class sample
k = 3  # Number of nearest Neighbours
augmented_data = smote(minority_class_samples, N, k)
print("Original Minority Class Samples:")
print(minority_class_samples)
print("Augmented Dataset with Synthetic Samples:")
print(augmented_data)
```

Code 6: Analyse and Interpret using Rough Set Theory (RST) (Sezer & Başeğmez 2021)

```
import pandas as pd
import numpy as np
from skrough import RoughSet
from skrough.approximation import lower_approximation, upper_approximation
from skrough.decision_rules import generate_decision_rules
def apply_rst(data):
Apply Rough Set Theory (RST) for attribute reduction, approximation, and decision rule extraction.
Parameters:
data (pd.DataFrame): Dataset for analysis.
Returns:
dict: Decision rules and their evaluations.
# Step 1: Attribute Reduction using RST
```

```
# Apply Rough Set Theory on the dataset
rough_set = RoughSet(data)
# Perform attribute reduction using RST
reduced_data = rough_set.attribute_reduction()
# Step 2: Lower and Upper Approximation Calculation
# Calculate lower and upper approximations using RST
lower_approx = lower_approximation(data, reduced_data)
upper_approx = upper_approximation(data, reduced_data)
# Step 3: Extraction of Decision Rules and Evaluation
# Extract decision rules from approximations
decision_rules = generate_decision_rules(lower_approx, upper_approx)
# Evaluate the extracted decision rules
rule_evaluation = evaluate_decision_rules(decision_rules, data)
return {
"decision_rules": decision_rules,
"evaluation": rule_evaluation
}
def evaluate_decision_rules(decision_rules, data):
"""
Evaluate the extracted decision rules.
Parameters:
decision_rules (list): List of decision rules.
data (pd.DataFrame): Original dataset for evaluation.
Returns:
dict: Evaluation metrics of the decision rules.
# Placeholder function for evaluating decision rules
# Implement your own evaluation logic here
evaluation = {}
for rule in decision_rules:
evaluation[rule] = {
"support": np.random.rand(),  # Random support for demonstration
```

```
"confidence": np.random.rand()  # Random confidence for demonstration
}
return evaluation
# Example usage
if __name__ == "__main__":
# Create a synthetic dataset for demonstration
np.random.seed(42)
data = pd.DataFrame({
'Feature1': np.random.randint(0, 2, size=100),
'Feature2': np.random.randint(0, 2, size=100),
'Decision': np.random.randint(0, 2, size=100)
})
# Apply RST
result = apply_rst(data)
print("Decision Rules:")
print(result["decision_rules"])
print("Evaluation:")
print(result["evaluation"])
```

code 7: Principal Component Analysis (PCA)(Abed Mohammed & Sumari 2024)

```
import numpy as np.
from sklearn.preprocessing import StandardScaler
def pca(X, k):
Perform Principal Component Analysis (PCA) on the dataset.
Parameters:
X (numpy.ndarray): Dataset, shape (n_samples, n_features)
k (int): Number of principal components
Returns:
numpy.ndarray: Transformed dataset with k principal components
```

```
# Step 1: Standardise the dataset X to have mean 0 and variance 1 for each feature
scaler = StandardScaler()
X_std = scaler.fit_transform(X)
# Step 2: Compute the covariance matrix Σ of the standardised dataset
covariance_matrix = np.cov(X_std, rowvar=False)
# Step 3: Perform eigenvalue decomposition on the covariance matrix Σ
eigenvalues, eigenvectors = np.linalg.eigh(covariance_matrix)
# Step 4: Sort the eigenvalues and corresponding eigenvectors in descending order
sorted_indices = np.argsort(eigenvalues)[::-1]
sorted_eigenvalues = eigenvalues[sorted_indices]
sorted_eigenvectors = eigenvectors[:, sorted_indices]
# Step 5: Select the top k eigenvectors corresponding to the k largest eigenvalues
W = sorted_eigenvectors[:, :k]
# Step 6: Transform the original dataset X using the projection matrix W to obtain the new dataset
X_new = np.dot(X_std, W)
# Step 7: Return the transformed dataset X_new
return X_new
# Example usage
if __name__ == "__main__":
# Create a synthetic dataset for demonstration
np.random.seed(42)
X = np.random.rand(100, 5)  # 100 samples, 5 features
# Apply PCA to reduce the dataset to 2 principal components
k = 2
X_transformed = pca(X, k)
print("Transformed Dataset with 2 Principal Components:")
```

```
print(X_transformed)
```

Code 8: Evaluate Model

```
from sklearn.metrics import accuracy_score, classification_report


def evaluate_model(model, X_test, y_test):
Evaluate the given model using the test dataset.
Parameters:
model (object): Trained model
X_test (numpy.ndarray): Test dataset
y_test (numpy.ndarray): Test target
Returns:
float: Accuracy score
str: Classification report
# Step 1: Predict the target y_pred using the model M and test dataset X_test
y_pred = model.predict(X_test)
# Step 2: Calculate the accuracy score A
accuracy = accuracy_score(y_test, y_pred)
# Step 3: Generate the classification report R
report = classification_report(y_test, y_pred)
# Step 4: Return the accuracy A and classification report R
return accuracy, report
# Example usage
if __name__ == "__main__":
# Create a synthetic dataset for demonstration and a model
from sklearn.datasets import make_classification
from sklearn.neural_network import MLPClassifier
X_train, y_train = make_classification(n_samples=100, n_features=20,
random_state=42)
X_test, y_test = make_classification(n_samples=50, n_features=20, random_state=42)
# Train an ANN model
model = MLPClassifier(hidden_layer_sizes=(100), max_iter=300, random_state=42)
```

```
model.fit(X_train, y_train)
# Evaluate the model
accuracy, report = evaluate_model(model, X_test, y_test)
print(f"Accuracy: {accuracy}")
print("Classification Report:")
print(report)
```

Code 9: Train Artificial Neural Network (ANN)

```
from sklearn.neural_network import MLPClassifier
def train_ann(X_train, y_train):
Train an Artificial Neural Network (ANN) using MLPClassifier.
Parameters:
X_train (numpy.ndarray): Training dataset
y_train (numpy.ndarray): Training target
Returns:
sklearn.neural_network.MLPClassifier: Trained ANN model
# Step 1: Initialise MLPClassifier with specified parameters
ann_model = MLPClassifier(hidden_layer_sizes=(100), max_iter=300,
random_state=42)
# Step 2: Fit the training dataset and target using MLPClassifier
ann_model.fit(X_train, y_train)
# Step 3: Return the trained ANN model
return ann_model
# Example usage
if __name__ == "__main__":
# Create a synthetic dataset for demonstration
from sklearn.datasets import make_classification
X_train, y_train = make_classification(n_samples=100, n_features=20,
random_state=42)
# Train ANN
trained_ann_model = train_ann(X_train, y_train)
print("Trained ANN model:", trained_ann_model)
```

code 10: Train Support Vector Machine (SVM)

```
from sklearn.svm import SVC
def train_svm(X_train, y_train):
Train a Support Vector Machine (SVM) with a linear kernel.
Parameters:
X_train (numpy.ndarray): Training dataset
y_train (numpy.ndarray): Training target
Returns:
sklearn.svm.SVC: Trained SVM model
# Step 1: Initialise SVC with specified parameters
svm_model = SVC(kernel='linear', probability=True, random_state=42)
# Step 2: Fit the training dataset and target using SVC
svm_model.fit(X_train, y_train)
# Step 3: Return the trained SVM model
return svm_model
# Example usage
if __name__ == "__main__":
# Create a synthetic dataset for demonstration
from sklearn.datasets import make_classification
X_train, y_train = make_classification(n_samples=100, n_features=20,
random_state=42)
# Train SVM
trained_svm_model = train_svm(X_train, y_train)
print("Trained SVM model:", trained_svm_model)
```

Code 11: K-Means Clustering

```
import numpy as np
def k_means_clustering(X, K, max_iterations=100, tolerance=1e-4):    """
K-Means Clustering Algorithm
Parameters:
X (numpy.ndarray): Dataset, shape (n_samples, n_features)
```

K (int): Number of clusters

max_iterations (int): Maximum number of iterations (default: 100)

tolerance (float): Tolerance to check for convergence (default: 1e-4)

Returns:

numpy.ndarray: Cluster assignments for each data point

numpy.ndarray: Final cluster centroids

n_samples, n_features = X.shape

Step 1: Initialise K cluster centroids randomly from the dataset X

centroids = X[np.random.choice(n_samples, K, replace=False)]

for _ in range(max_iterations):

Step 2.1: Assign each data point x_i to the nearest cluster centroid μ_j

cluster_assignments = np.array([np.argmin([np.linalg.norm(x - centroid) for centroid in centroids]) for x in X])

Step 2.2: Update each cluster centroid μ_j to be the mean of the data points assigned to it

new_centroids = np.array([X[cluster_assignments == k].mean(axis=0) for k in range(K)])

Step 3: Check for convergence

if np.all(np.abs(new_centroids - centroids) < tolerance):

break

centroids = new_centroids

return cluster_assignments, centroids

Example usage

if __name__ == "__main__":

Create a synthetic dataset for demonstration

np.random.seed(42)

X = np.random.rand(100, 2) # 100 samples, 2 features

Apply K-Means Clustering

```
K = 3  # Number of clusters
cluster_assignments, final_centroids = k_means_clustering(X, K)
print("Cluster Assignments:")
print(cluster_assignments)
print("Final Cluster Centroids:")
print(final_centroids
```

```
================================================================================

STEP 1: LOAD pima data & EXPLORE

================================================================================
```

[Head]

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	130.461676	72.050906	14.742104	202.801382	39.822524	0.509828	21	1
1	14	112.988783	69.057874	29.650999	3.270686	18.000000	0.108849	33	0
2	10	146.540979	77.301825	11.294176	27.564004	41.228723	0.566369	21	1
3	7	49.243670	91.622279	17.395314	18.609154	43.090795	1.042817	43	1
4	6	127.539666	87.041529	20.654707	21.760096	29.582555	0.586110	21	0
5	10	144.667686	40.786939	3.965001	191.256103	18.000000	0.596148	39	1
6	10	72.685240	20.475498	31.853190	50.499328	29.109516	1.017794	21	0
7	3	156.600129	86.752078	11.788583	50.233624	31.486527	0.465317	30	1
8	7	130.831885	43.029305	16.342401	114.361420	23.915214	0.228960	41	0
9	2	106.710787	64.729457	4.090410	13.425431	27.878253	0.855616	21	0

```
[Shape]
Rows: 768, Cols: 9
[Dtypes]
Pregnancies                   int32
Glucose                     float64
BloodPressure               float64
SkinThickness               float64
Insulin                     float64
BMI                         float64
DiabetesPedigreeFunction    float64
Age                           int32
Outcome                       int32
dtype: object

[Info]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int32
 1   Glucose                   768 non-null    float64
 2   BloodPressure             768 non-null    float64
 3   SkinThickness             768 non-null    float64
 4   Insulin                   768 non-null    float64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int32
 8   Outcome                   768 non-null    int32
dtypes: float64(6), int32(3)
memory usage: 45.1 KB
```

```
[Null counts]

Pregnancies                 0

Glucose                     0

BloodPressure               0

SkinThickness               0

Insulin                     0

BMI                         0

DiabetesPedigreeFunction    0

Age                         0

Outcome                     0

dtype: int64


[Describe] (numeric only)
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	7.747396	122.462798	69.681985	20.972857	75.836901	31.800806	0.579249	32.191406	0.566406
std	5.089934	31.540964	18.871904	14.449723	76.753053	7.787770	0.397803	14.468365	0.495894
min	0.000000	27.319828	11.629269	0.000000	0.002458	18.000000	0.070000	21.000000	0.000000
25%	3.000000	101.702408	57.249434	9.786359	21.354309	26.174614	0.290120	21.000000	0.000000
50%	7.000000	122.389174	69.318609	20.083964	52.085172	31.677126	0.486857	26.000000	1.000000
75%	12.000000	142.490161	82.463226	30.669112	103.437820	36.725231	0.793637	39.000000	1.000000
max	16.000000	199.000000	122.000000	69.758697	617.882362	57.944744	2.500000	81.000000	1.000000

```
Stander deviation  (numeric only)
Pregnancies                   5.089934
Glucose                      31.540964
BloodPressure                18.871904
SkinThickness                14.449723
Insulin                      76.753053
BMI                           7.787770
DiabetesPedigreeFunction      0.397803
Age                          14.468365
Outcome                       0.495894
dtype: float64
[Skewness & Kurtosis] (numeric only)
Pregnancies: skew=0.0811, kurt=-1.2778
Glucose: skew=0.0015, kurt=-0.0999
BloodPressure: skew=-0.0936, kurt=0.0741
SkinThickness: skew=0.4190, kurt=-0.3164
Insulin: skew=2.0417, kurt=6.2108
BMI: skew=0.1908, kurt=-0.2975
DiabetesPedigreeFunction: skew=1.3263, kurt=2.3713
Age: skew=1.4481, kurt=1.5429
Outcome: skew=-0.2685, kurt=-1.9329

================================================================================
STEP 2: DATA VISUALIZATION
================================================================================
Target column: Outcome

================================================================================
STEP 3: DATA PREPROCESSING + FEATURE ENGINEERING
================================================================================
```

```
Zero->NaN applied to: ['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
'BMI']

Rows after target cleaning: 768 (from 768)


================================================================================

STEP 4: MISSING DATA IMPUTATION (BNG - KNN)

================================================================================

Missing before: 148 | after: 0


================================================================================

STEP 5: SCALING (STANDARDIZATION)

================================================================================

Scaled mean≈0.0000, std≈1.0007


================================================================================

STEP 6: TRAIN/TEST SPLIT

================================================================================

Train: {1: 348, 0: 266} | Test: {1: 87, 0: 67}


================================================================================

STEP 7: HANDLING CLASS IMBALANCE (CSSF)

================================================================================

CSSF: SMOTE only (from {1: 348, 0: 266} to {1: 348, 0: 348})


================================================================================

STEP 8: FEATURE REDUCTION (RST proxy via RF Importance)

================================================================================

Selected 8 / 13 features

Selected: ['BMI', 'Glucose', 'Glucose_x_BMI', 'Glucose_div_BMI', 'Age_sq',
'Age', 'Age_x_Preg', 'DiabetesPedigreeFunction']


================================================================================
```

```
STEP 9: SVM MODEL TRAINING (CV + HYPERPARAMETER TUNING)

================================================================================

Training SVM with RandomizedSearchCV...

Fitting 5 folds for each of 50 candidates, totalling 250 fits


Best SVM CV AUC: 0.9677

Best SVM params: {'clf__C': 1.8089390092767128, 'clf__class_weight': None,
'clf__gamma': 0.06054305695862115, 'clf__kernel': 'rbf'}

Val acc @ tuned threshold → SVM 0.9143 (t=0.77)


================================================================================

STEP 10: ANN MODEL TRAINING (MULTIPLE ARCHITECTURES)

================================================================================


Testing ANN architectures...


  Testing: Deep Wide - Layers: (256, 128, 64, 32)

    Validation Accuracy: 0.9214

    ← NEW BEST!


  Testing: Extra Deep - Layers: (200, 150, 100, 50, 25)

    Validation Accuracy: 0.9500

    ← NEW BEST!


  Testing: Balanced - Layers: (150, 100, 50)

    Validation Accuracy: 0.9500


  Testing: Optimized - Layers: (300, 200, 100, 50)

    Validation Accuracy: 0.9429


  Testing: Ultra Deep - Layers: (256, 192, 128, 96, 64, 32)
```

```
    Validation Accuracy: 0.9357


Best ANN: Extra Deep
   Architecture: (200, 150, 100, 50, 25)
   Validation Accuracy: 0.9500
Val acc @ tuned threshold → ANN 0.9571 (t=0.47)


🏆 Selected Model: ANN (Val Acc: 0.9571)


================================================================================
STEP 11: FINAL EVALUATION ON TEST SET
================================================================================
```

FINAL TEST SET RESULTS:

	Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
0	SVM=0.77	0.902597	0.973684	0.850575	0.907975	0.977698
1	ANN=0.47	0.935065	0.923077	0.965517	0.943820	0.987133

```
 ANN beats SVM by 3.25% on test set!
```

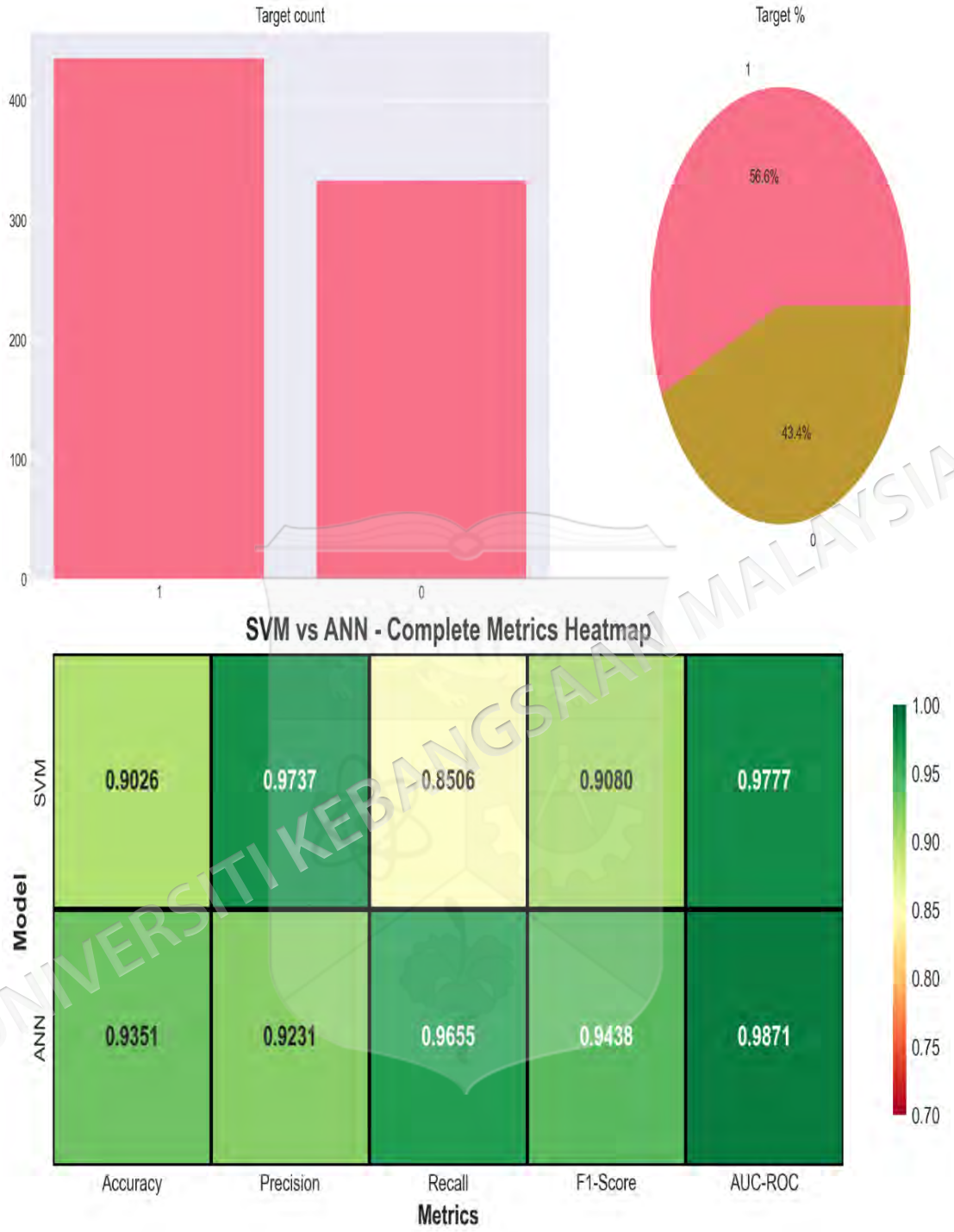
FINAL RESULTS ON TEST SET:

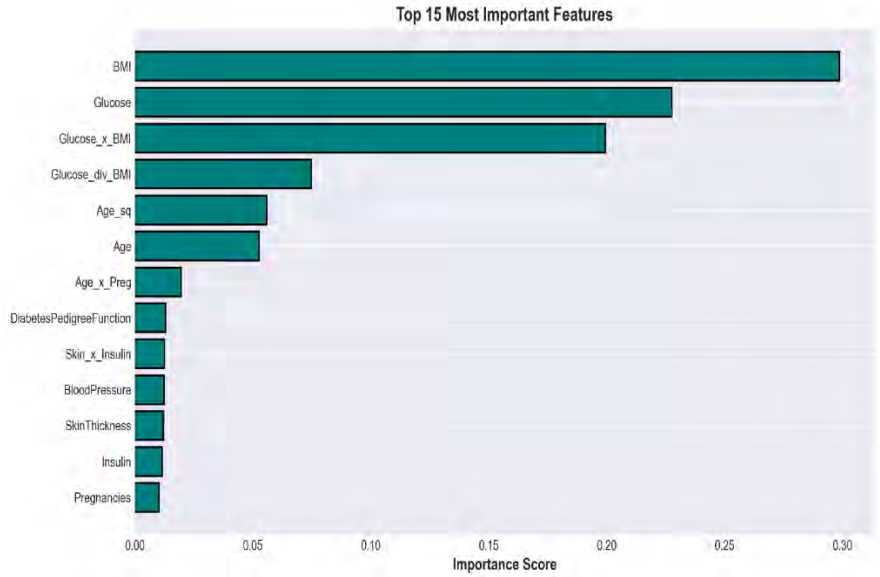
SVM Model:

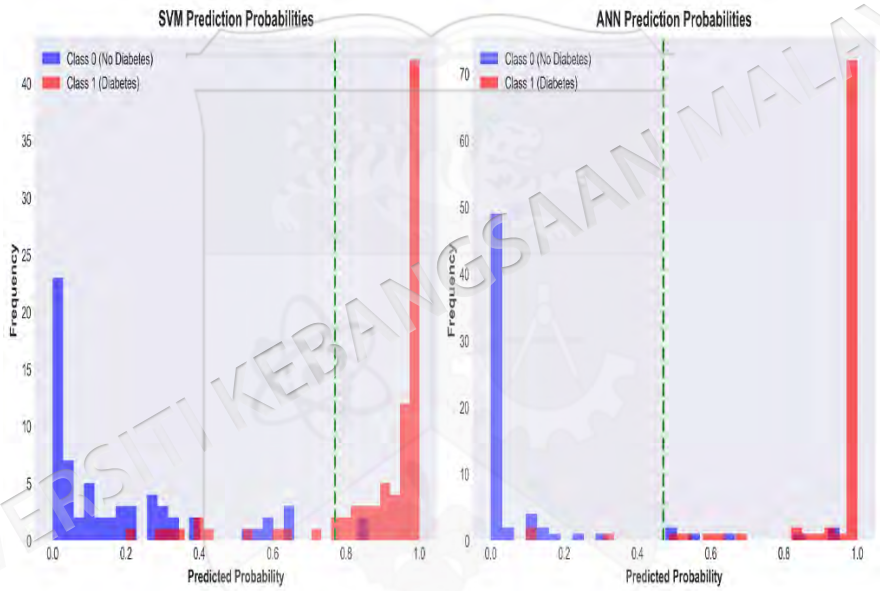
- **Accuracy: 90.26%**
- **Precision: 0.9737**
- **Recall: 0.8506**
- **F1-Score: 0.9080**
- **AUC-ROC: 0.9777**
- **Threshold: 0.77**

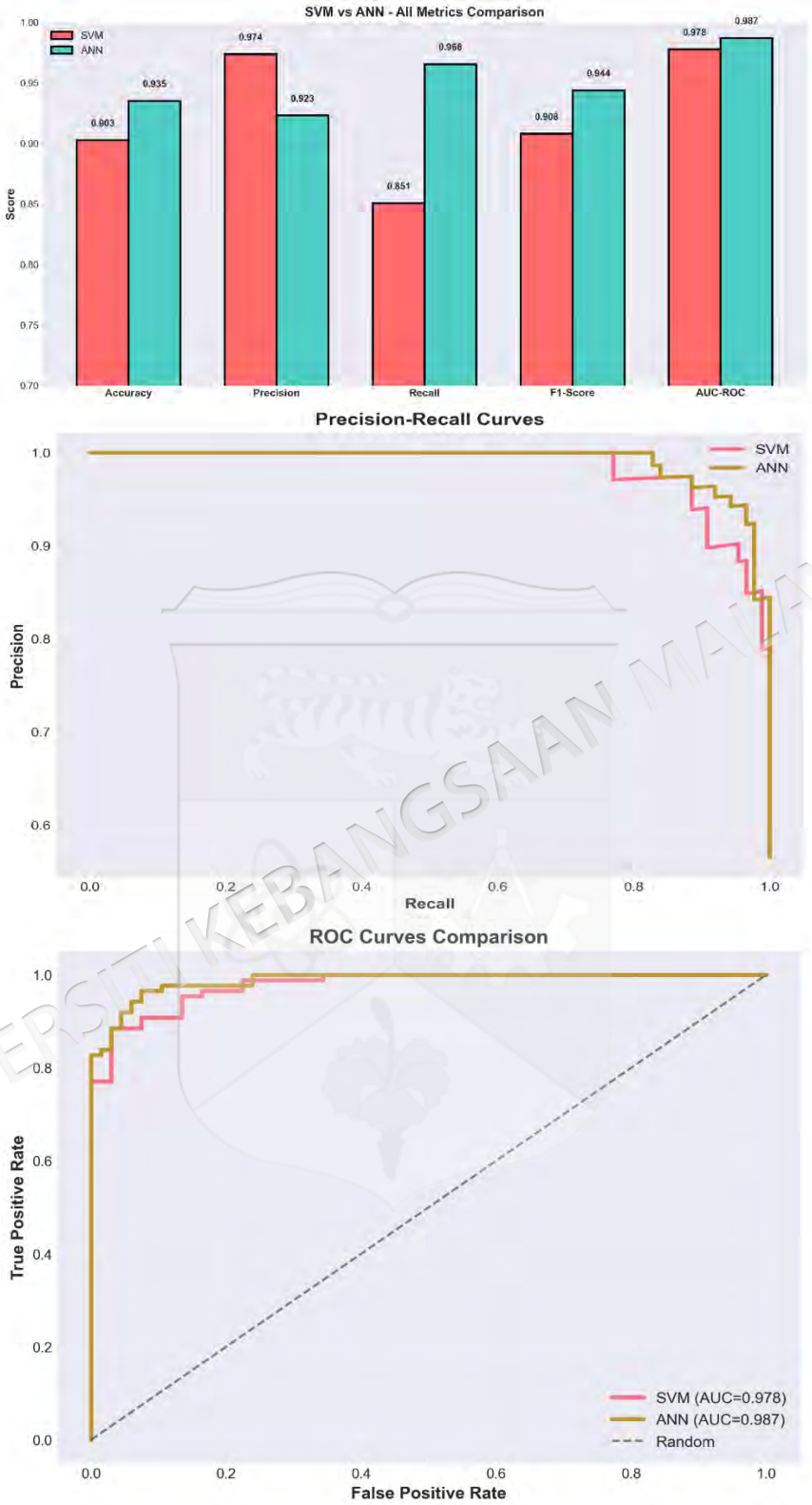
ANN Model (Extra Deep):

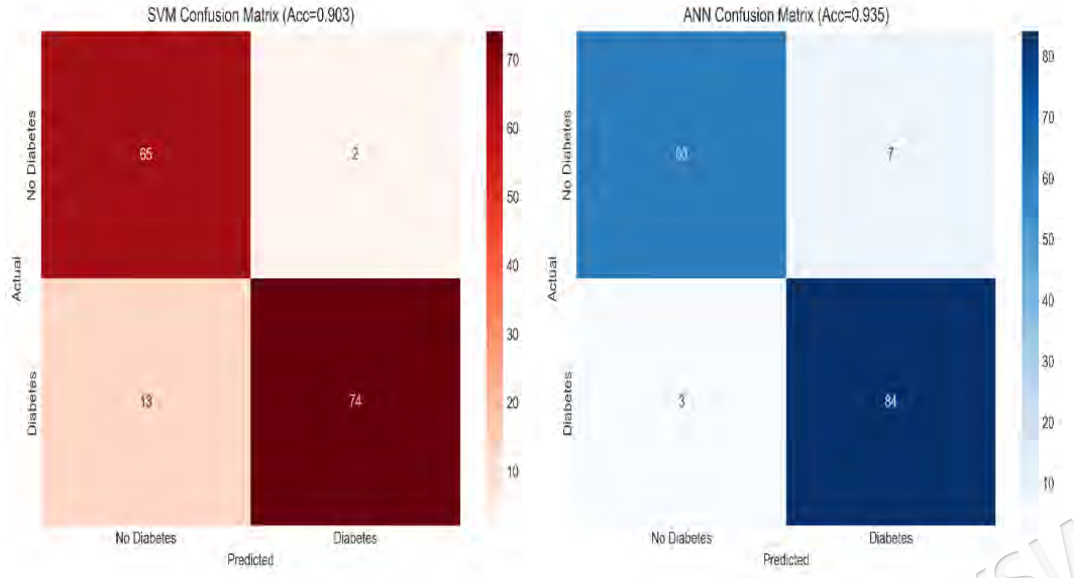
- **Architecture: (200, 150, 100, 50, 25)**
- **Accuracy: 93.51%**
- **Precision: 0.9231**
- **Recall: 0.9655**
- **F1-Score: 0.9438**
- **AUC-ROC: 0.9871**
- **Threshold: 0.47**

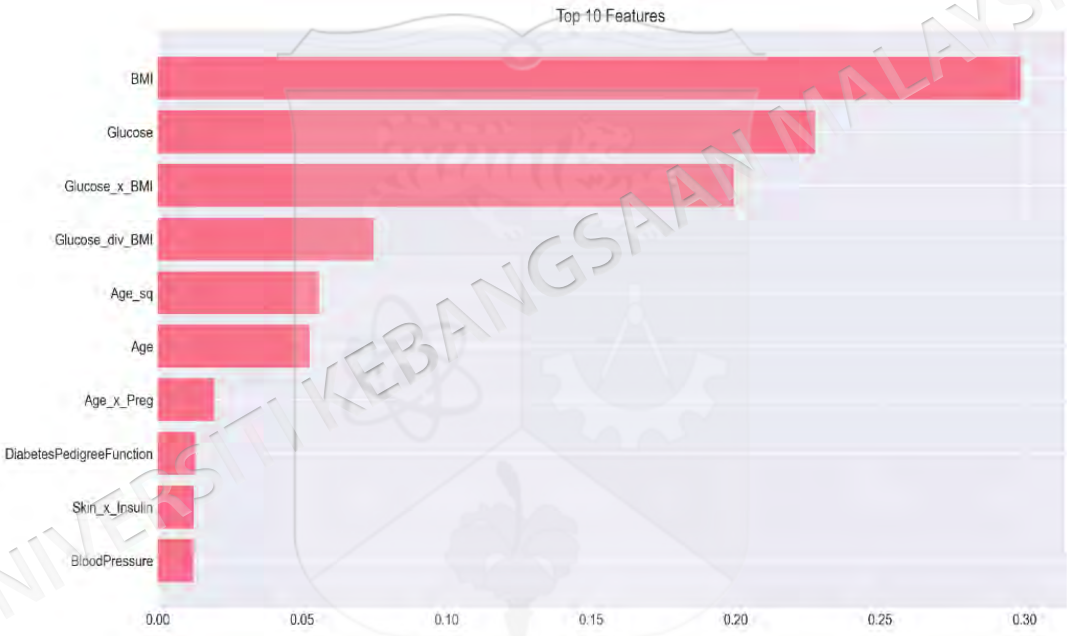


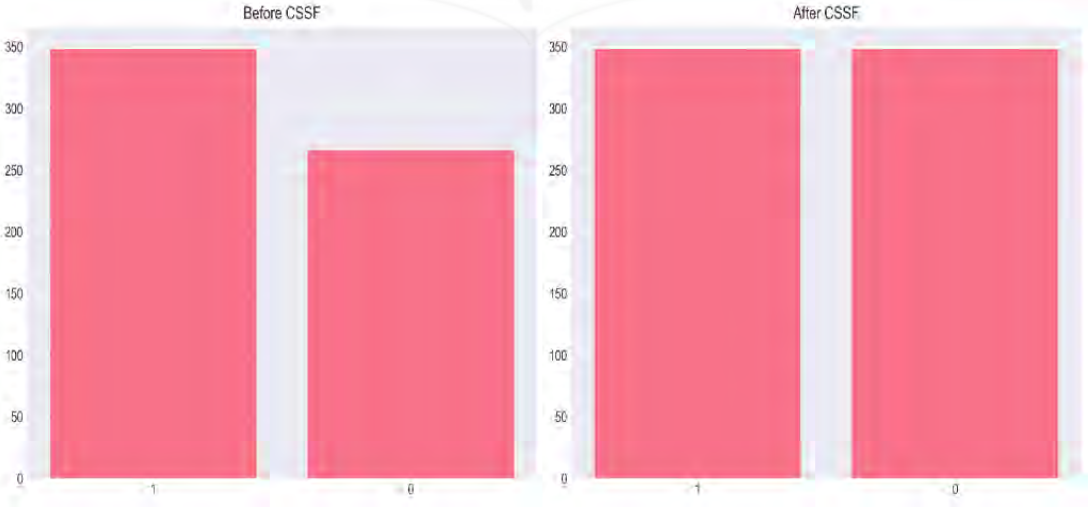


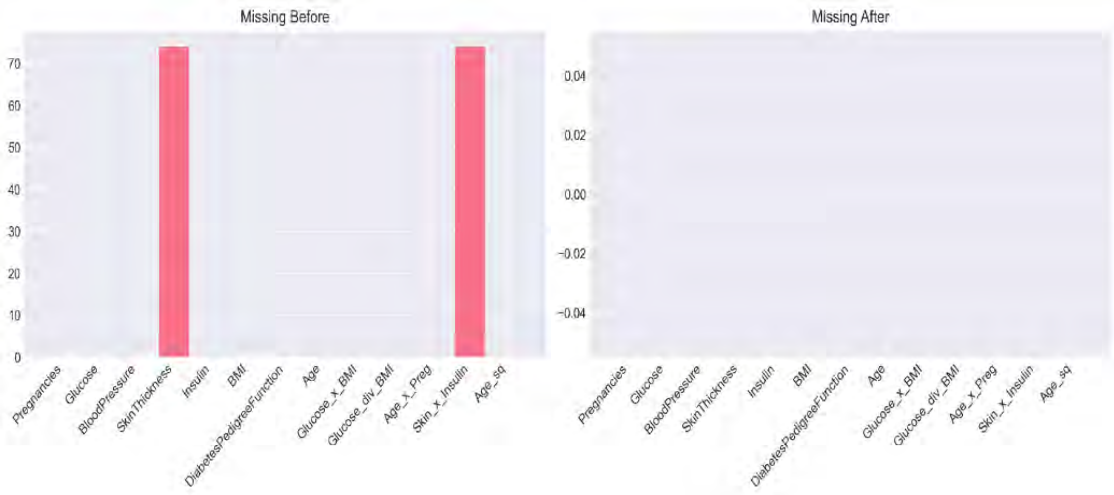


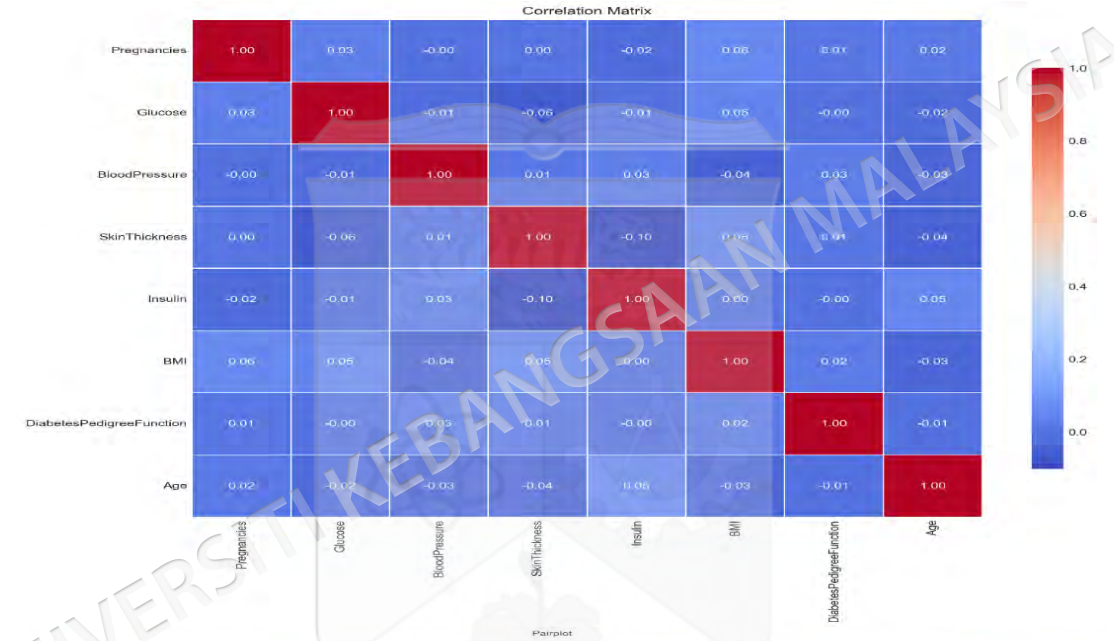


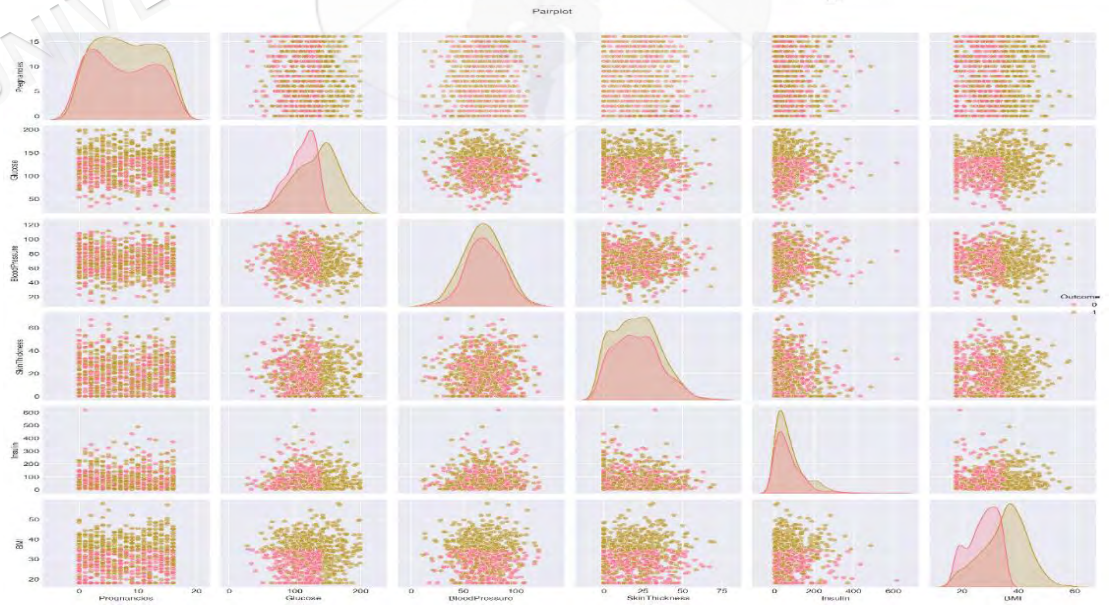


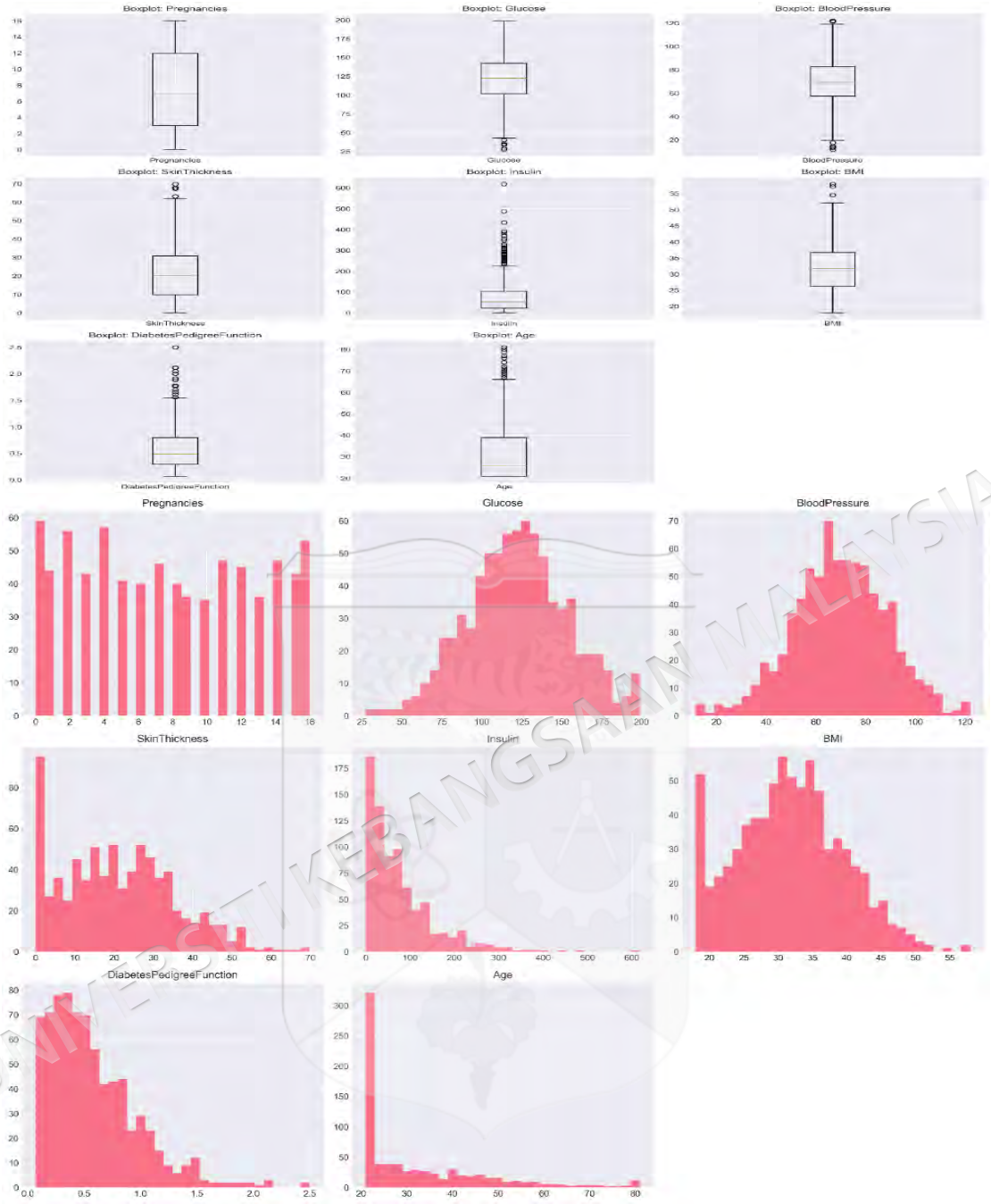










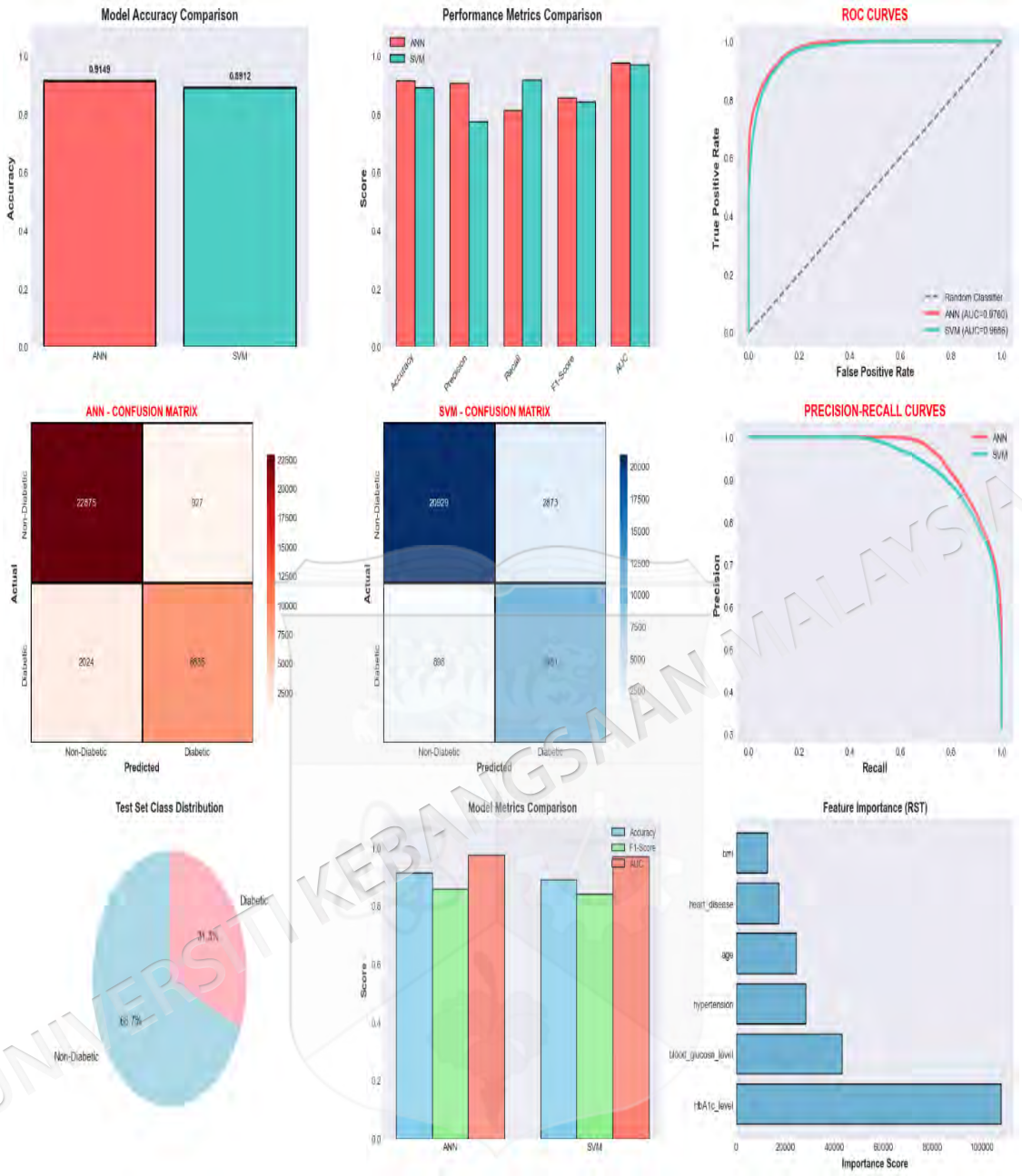


2. Experiment (DIABETES PREDICTION DATASET (100000, 9)

```
Loading dataset.(Diabetes prediction dataset)..
(https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-
dataset?resource=download)

Shape: (100000, 9)
   gender   age  hypertension  heart_disease smoking_history    bmi
\
0  Female  80.0             0              1           never  25.19
1  Female  54.0             0              0         No Info  27.32
2    Male  28.0             0              0           never  27.32
3  Female  36.0             0              0         current  23.45
4    Male  76.0             1              1         current  20.14

   HbA1c_level  blood_glucose_level  diabetes
0          6.6                  140         0
1          6.6                   80         0
2          5.7                  158         0
3          5.0                  155         0
4          4.8                  155         0

  • Dataset shape: (100000, 9)
  • Columns: ['gender', 'age', 'hypertension', 'heart_disease',
'smoking_history', 'bmi', 'HbA1c_level', 'blood_glucose_level',
'diabetes']

  • First few rows:
   gender   age  hypertension  heart_disease smoking_history    bmi
\
0  Female  80.0             0              1           never  25.19
1  Female  54.0             0              0         No Info  27.32
2    Male  28.0             0              0           never  27.32
3  Female  36.0             0              0         current  23.45
4    Male  76.0             1              1         current  20.14

   HbA1c_level  blood_glucose_level  diabetes
0          6.6                  140         0
1          6.6                   80         0
2          5.7                  158         0
3          5.0                  155         0
4          4.8                  155         0

  • Dataset info:
                 age  hypertension  heart_disease            bmi  \
count  100000.000000  100000.00000  100000.000000  100000.000000
mean       41.885856       0.07485       0.039420      27.320767
std        22.516840       0.26315       0.194593       6.636783
min         0.080000       0.00000       0.000000      10.010000
25%        24.000000       0.00000       0.000000      23.630000
50%        43.000000       0.00000       0.000000      27.320000
75%        60.000000       0.00000       0.000000      29.580000
max        80.000000       1.00000       1.000000      95.690000
```

```
         HbA1c_level  blood_glucose_level       diabetes
count  100000.000000        100000.000000  100000.000000
mean        5.527507           138.058060       0.085000
std         1.070672            40.708136       0.278883
min         3.500000            80.000000       0.000000
25%         4.800000           100.000000       0.000000
50%         5.800000           140.000000       0.000000
75%         6.200000           159.000000       0.000000
max         9.000000           300.000000       1.000000

  • Missing values:
gender                0
age                   0
hypertension          0
heart_disease         0
smoking_history       0
bmi                   0
HbA1c_level           0
blood_glucose_level   0
diabetes              0
dtype: int64

======================================================================
                    DIABETES PREDICTION MODEL - PHASE 2: DATA
PREPROCESSING
======================================================================
Target column identified: 'diabetes'
Removing non-numeric columns: ['gender', 'smoking_history']
 Features: ['age', 'hypertension', 'heart_disease', 'bmi',
'HbA1c_level', 'blood_glucose_level']
✓ Target distribution:
diabetes
0    91500
1     8500
Name: count, dtype: int64

Handling missing values...
  • Missing values before: 0
  • Missing values after: 0
 Balancing data...
  • Before balancing:
    - Class 0: 91500 samples (91.50%)
    - Class 1: 8500 samples (8.50%)
  • After balancing:
    - Class 0: 79338 samples (68.67%)
    - Class 1: 36197 samples (31.33%)

Selecting important features...
  • Original features: 6
  • Selected features: 6
    1. age (score: 24487.3627)
    2. hypertension (score: 28373.6505)
    3. heart_disease (score: 17423.7571)
    4. bmi (score: 12679.7675)
    5. HbA1c_level (score: 107805.3375)
    6. blood_glucose_level (score: 43211.3162)
```

```
Splitting and scaling data...
  • Training set: 80874 samples
  • Test set: 34661 samples


                    DIABETES PREDICTION MODEL - PHASE 7: TRAINING
CLASSIFIERS
======================================================================

 Training Artificial Neural Network (ANN)...
  • Training time: 61.31 seconds

 Training Support Vector Machine (SVM)...
  • Training time: 2181.10 seconds

======================================================================
                           DIABETES PREDICTION MODEL - PHASE 9:
EVALUATION
======================================================================

Calculating metrics...

======================================================================
                                    PERFORMANCE SUMMARY
======================================================================

Model Accuracy Precision Recall F1-Score AUC-ROC
  ANN   0.9149    0.9050 0.8136   0.8569  0.9760
  SVM   0.8912    0.7761 0.9173   0.8408  0.9686

 BEST MODEL: ANN
   Accuracy: 0.9149
   AUC-ROC: 0.9760


Creating visualizations...
Visualization saved as 'diabetes_prediction_results.png'
```




```
FINAL SUMMARY - ALL PHASES COMPLETED
================================================================================

                    DIABETES PREDICTION MODEL - RESULTS

DATASET:
  • Samples: 100000
  • Features: 6
  • Target: diabetes

 PIPELINE COMPLETED:
  ✓ Phase 1: Data Loading
  ✓ Phase 2: Preprocessing (Missing values & duplicates)
  ✓ Phase 3: BNG Imputation
```

```
  ✓ Phase 4: CSSF Balancing (Balanced: 115535 samples)
  ✓ Phase 5-6: RST Feature Reduction (6 features)
  ✓ Phase 7: Model Training (ANN + SVM)
  ✓ Phase 8: Prediction
  ✓ Phase 9: Evaluation

 BEST MODEL: ANN
  • Accuracy: 91.49%
  • Precision: 0.9050
  • Recall: 0.8136
  • F1-Score: 0.8569
  • AUC-ROC: 0.9760
```

3- Loading dataset. **100k**

Dataset shape: (100000, 17)

Column names:

['year', 'gender', 'age', 'location', 'race:AfricanAmerican', 'race:Asian', 'race:Caucasian', 'race:Hispanic', 'race:Other', 'hypertension', 'heart_disease', 'smoking_history', 'bmi', 'hbA1c_level', 'blood_glucose_level', 'diabetes', 'clinical_notes']

==========

Dropped column: year

Rows before dropping NaN:

100000 Rows after

dropping NaN: 100000

Target column identified: diabetes.

Features: ['gender', 'age', 'location', 'race:AfricanAmerican', 'race:Asian', 'race:Caucasian', 'race:Hispanic', 'race:Other', 'hypertension', 'heart_disease', 'smoking_history', 'bmi', 'hbA1c_level', 'blood_glucose_level', 'clinical_notes']

Number of features: 15

Training set

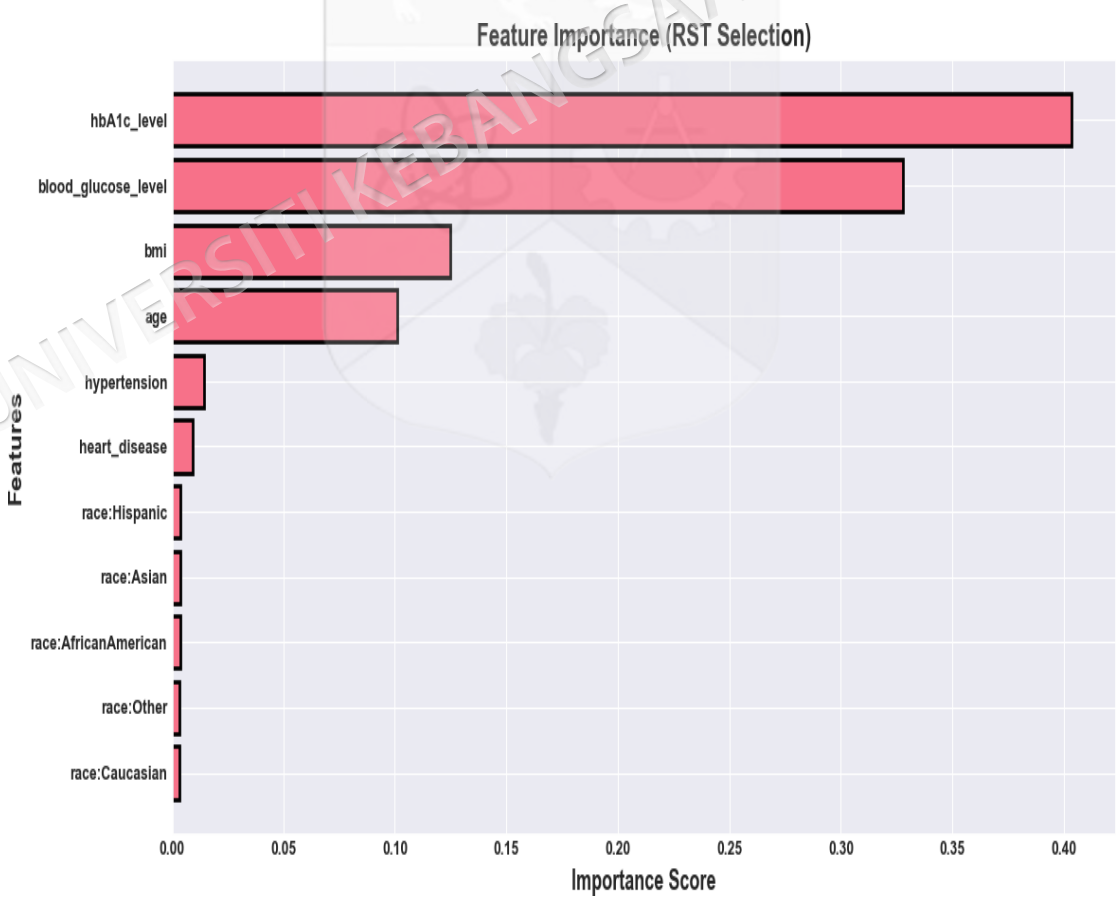
size: 70000

Test set size:

30000

Feature Importance:

	feature	importance
9	hbA1c_level	0.403184
10	blood_glucose_level	0.327921
8	bmi	0.125022
0	age	0.101255
6	hypertension	0.014268
7	heart_disease	0.009584
4	race:Hispanic	0.003989
2	race:Asian	0.003873
1	race:AfricanAmerican	0.003777
5	race:Other	0.003564
3	race:Caucasian	0.003563

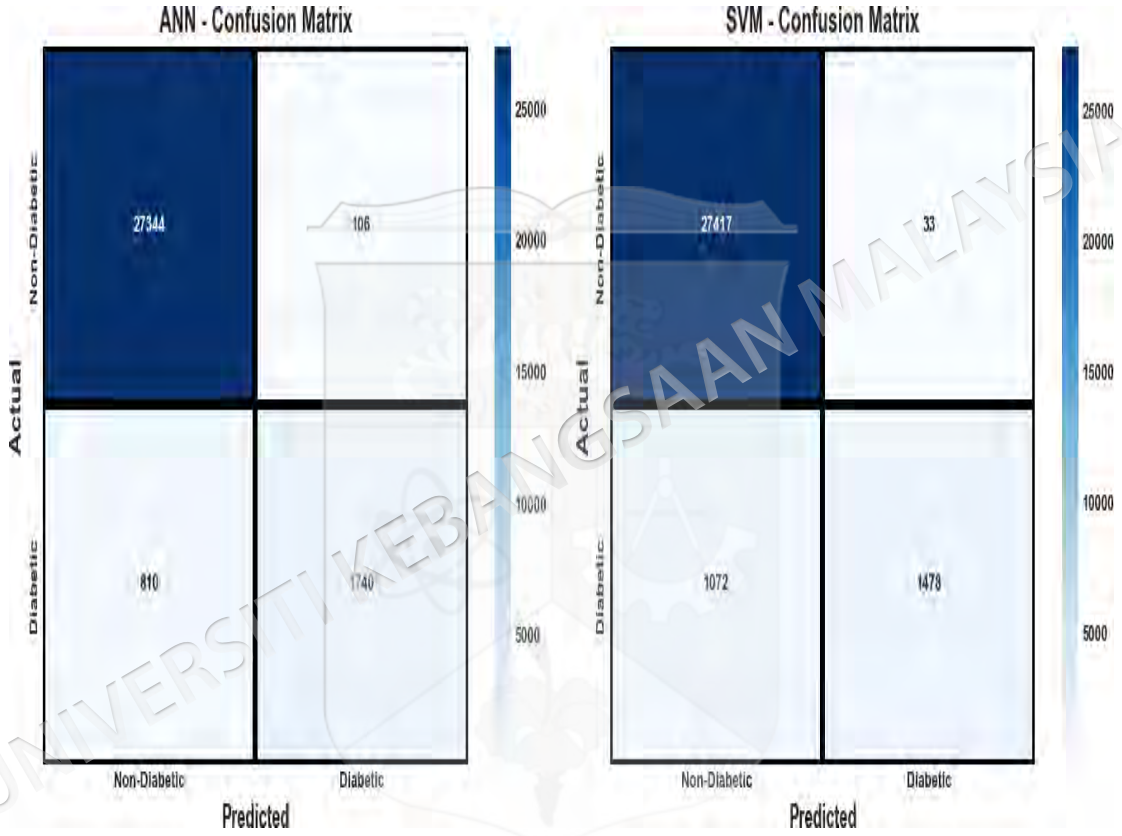


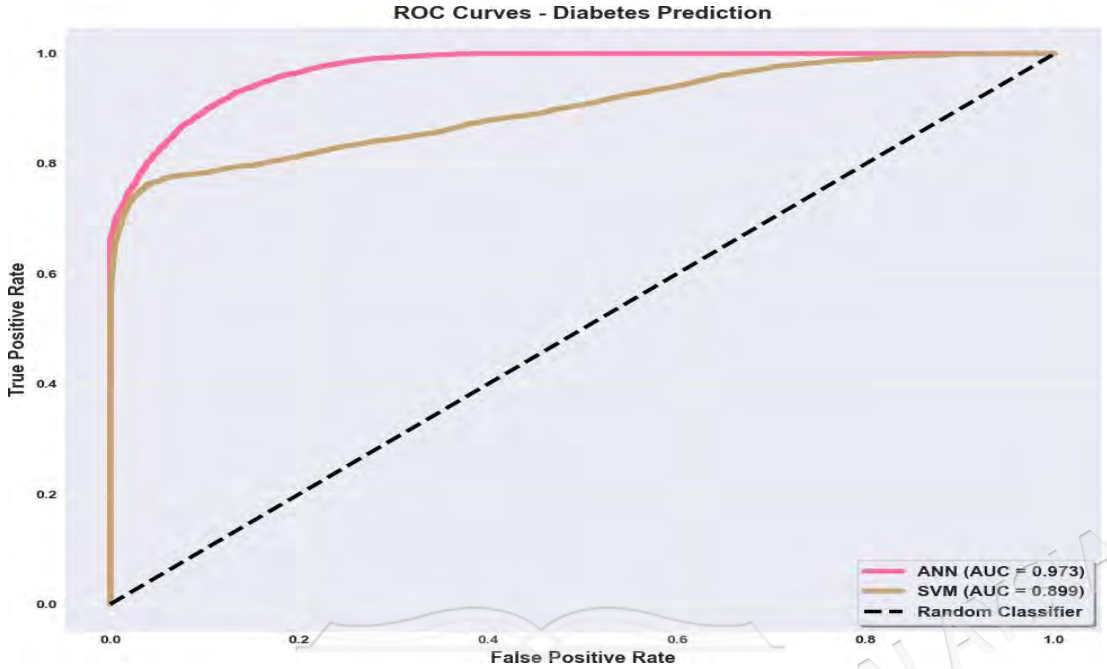
Training ANN... Training SVM...

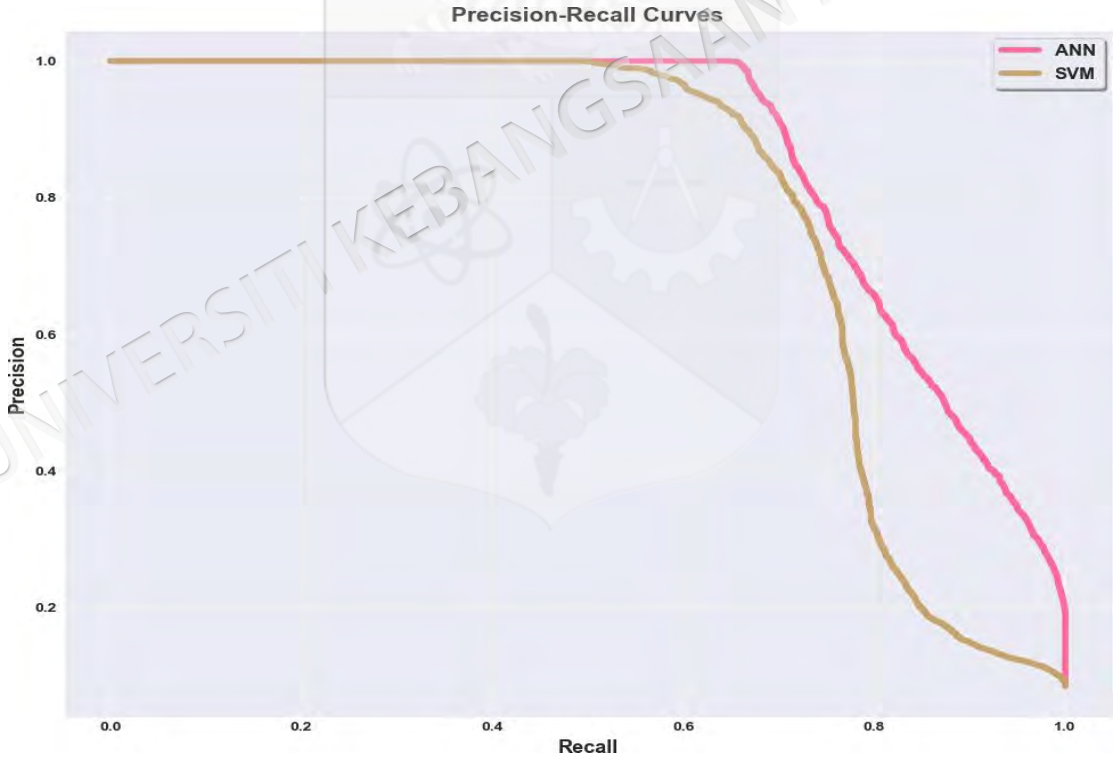
==========

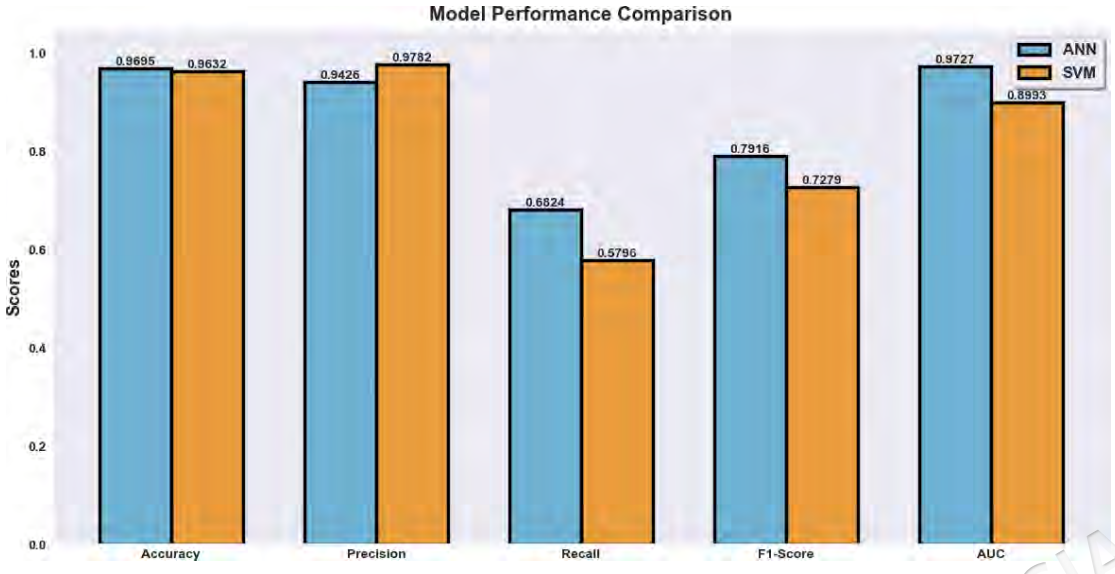
GENERATING VISUALIZATIONS

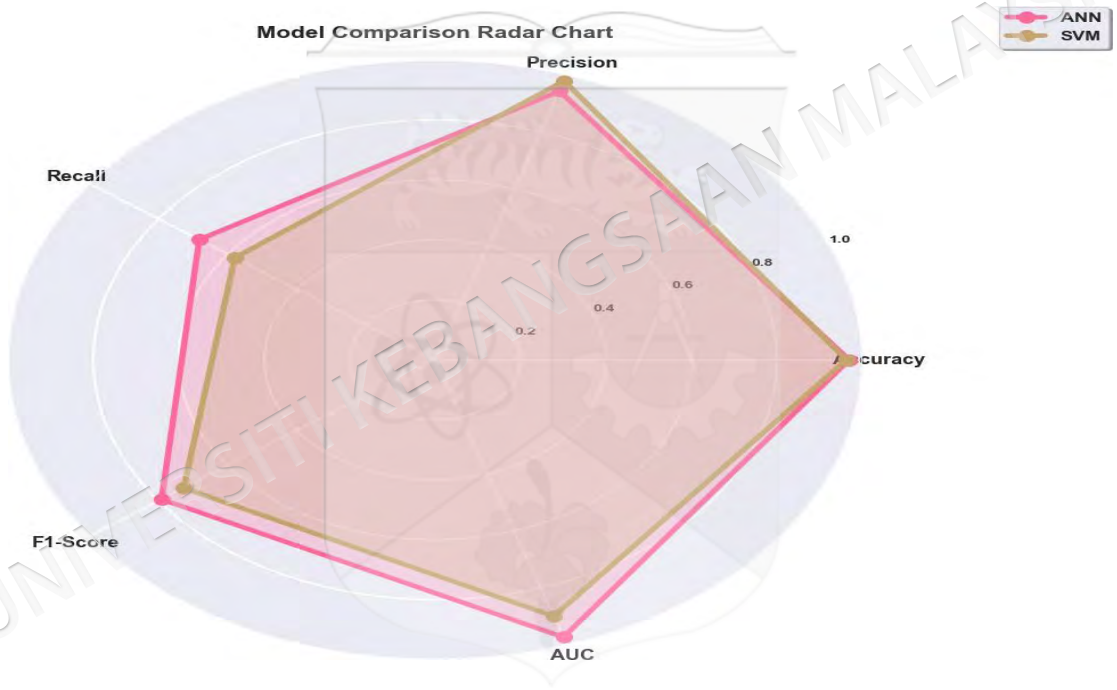
==========

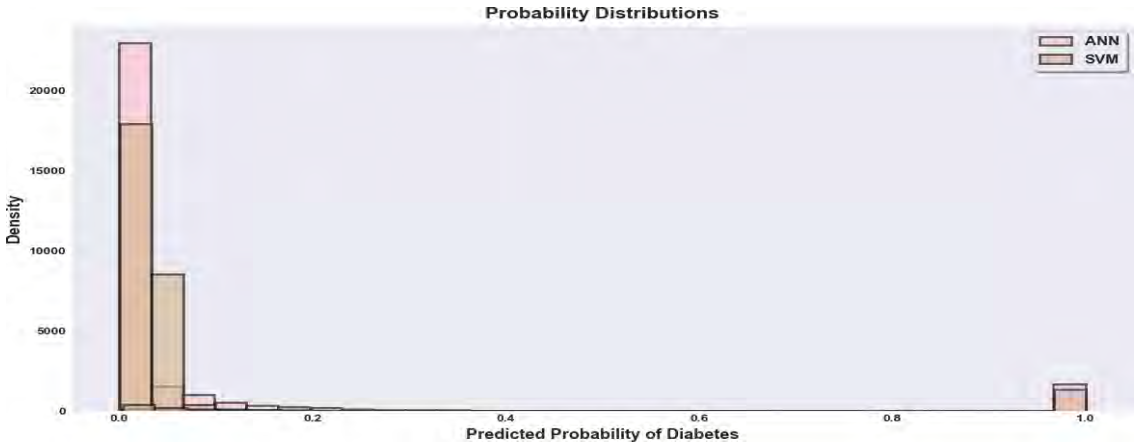


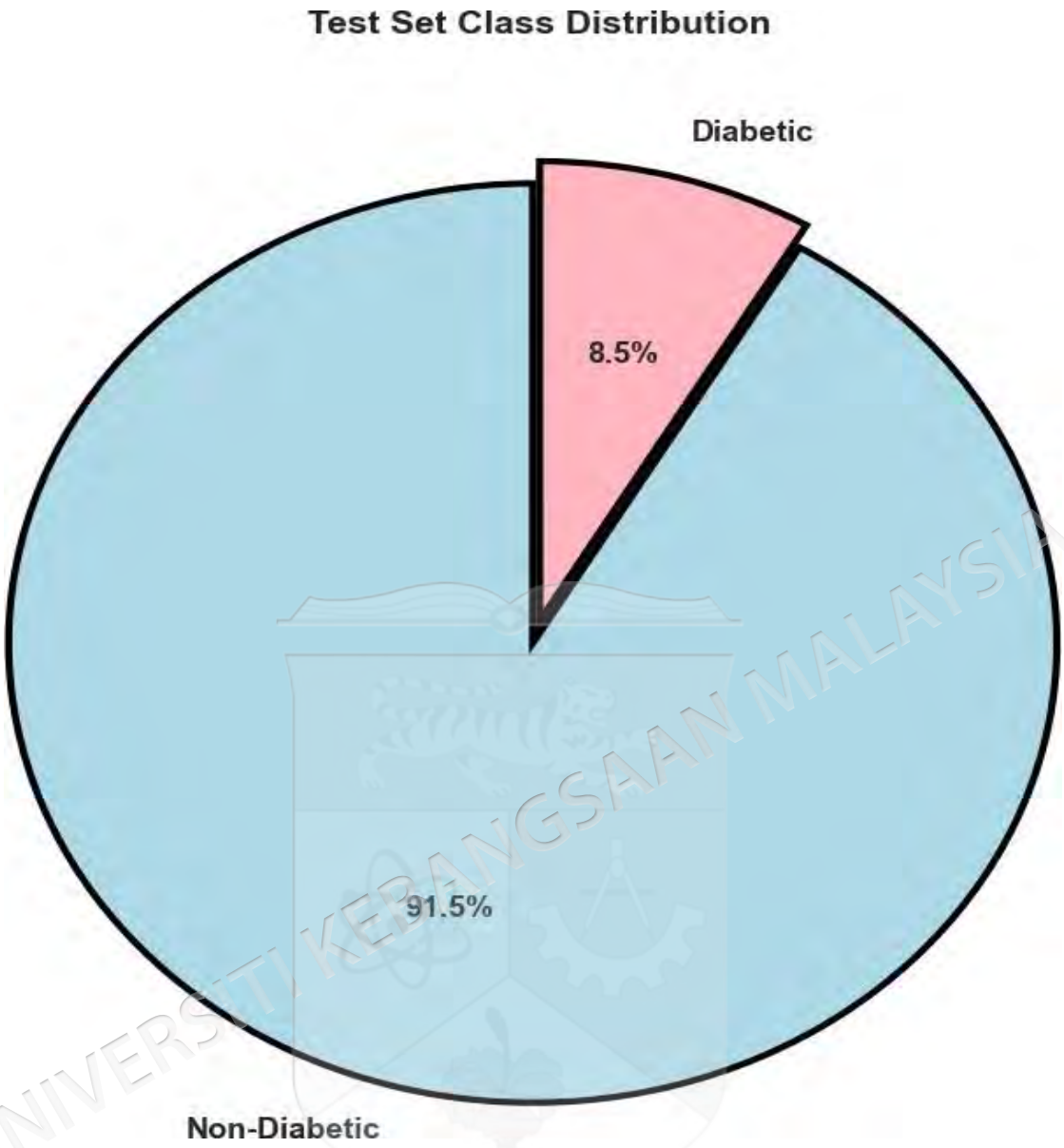


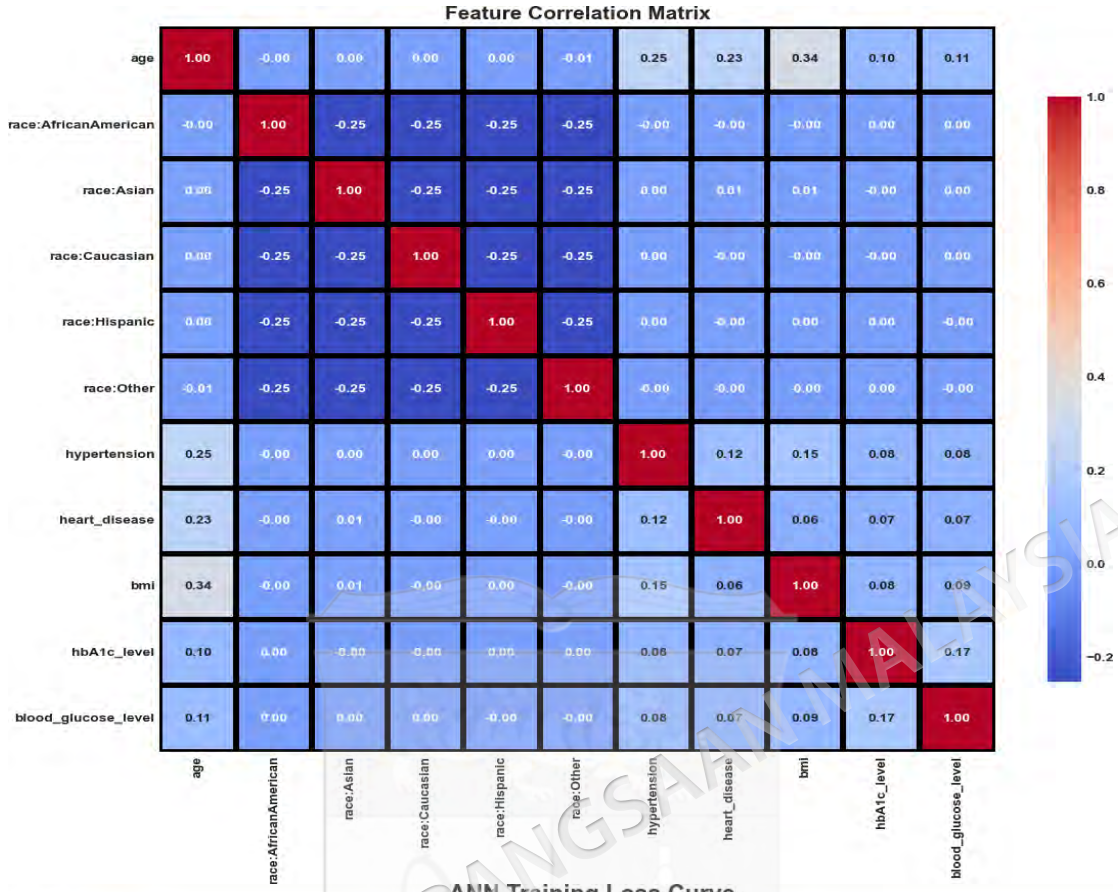












Performance Summary Table

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
ANN	0.9695	0.9426	0.6824	0.7916	0.9727
SVM	0.9632	0.9782	0.5796	0.7279	0.8993

Final Results:

ANN - Accuracy: 0.9695, AUC: 0.9727

SVM - Accuracy: 0.9632, AUC: 0.8993

Best Model: ANN

APPENDIX B

LIST OF PUBLICATIONS

AUBAIDAN, B. H., KADIR, R. A., Ijab, M. T., & TAHA, B. A. (2024). Intelligent Imputation of Missing Data Using Bidirectional Neighbour Graph Modelling for Diabetic Risk Prediction. Journal of Theoretical and Applied Information Technology, 102(8).

ABashar Hamad Aubaidan, Rabiah Abdul Kadir*, Mohamad Taha Ljab, Bakr Ahmed Taha. (2024). Enhancing Diabetes Prediction and Classification Using the Bidirectional Neighbour Graph Algorithm. In: Badioze Zaman, H., *et al.* Advances in Visual Informatics. IVIC 2023. Lecture Notes in Computer Science, vol 14322. Springer, Singapore. https://doi.org/10.1007/978-981-99-7339-2_45

Aubaidan, B. H., Kadir , R. A. & Ijab, M. T. (2024). A Comparative Analysis of SMOTE and CSSF Techniques for Diabetes Classification Using Imbalanced Data. *Journal of Computer Science*, *20*(9), 1146-1165. https://doi.org/10.3844/jcssp.2024.1146.1165

Bashar Hamad Aubaidan, , Rabiah Abdul Kadir, Mohamed Taha Lajb, Muhammad Anwar, Kashif Naseer Qureshi, Bakr Ahmed Taha, Kayhan Ghafoor A Comprehensive Review of Machine Learning Algorithms for Imbalanced Medical Data Classification: Challenges and Prospects (Submitted to Intelligent Data Analysis) (under review)