RCHUB: JOURNAL OF COMPUTATIONAL INTELLIGENCE SCIENCE AND ENGINEERING (JCISE)
VOL: 2, ISSUE 3, JULY - SEPTEMBER 2026
RCHUB PUBLISHER
ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

AN INTELLIGENT DATA QUALITY ENHANCEMENT FRAMEWORK FOR DIABETES PREDICTION USING HYBRID PREPROCESSING TECHNIQUES

Mrs.V. Manjuladevi[1], Dr. P. Periyasamy[2]

[1]Research Scholar, Park's College, Chinnakkarai. manjuv2220@gmail.com

[2]Associate Professor, Department of Computer Science and Applications, Park's College, Chinnakkarai. Periyasamy_p_2007@yahoo.co.in

Abstract

Diabetes mellitus is one of the leading chronic diseases worldwide, necessitating reliable data-driven approaches for early diagnosis and clinical decision-making. The performance of machine learning models is highly dependent on the quality of the input data; however, existing studies primarily focus on prediction algorithms while providing limited attention to systematic data preprocessing. This study proposes an Intelligent Data Quality Enhancement Framework (IDQEF) that employs hybrid preprocessing techniques to improve the quality and consistency of diabetes datasets prior to predictive modeling. The proposed framework integrates data validation, missing value imputation, outlier detection, duplicate record elimination, feature scaling, and synthetic minority oversampling (SMOTE) to address data incompleteness, noise, redundancy, and class imbalance. The framework is implemented using the publicly available Kaggle Diabetes Prediction Dataset comprising demographic and clinical attributes associated with diabetes diagnosis. The effectiveness of the proposed preprocessing strategy is assessed using data quality measures and validated through baseline machine learning classifiers with standard evaluation metrics, including Accuracy, Precision, Recall, F1-score, Specificity, Matthews Correlation Coefficient (MCC), and Area Under the Receiver Operating Characteristic Curve (ROC-AUC). The proposed framework produces a clean, balanced, and standardized dataset that enhances the reliability of downstream predictive models and provides a robust foundation for subsequent feature engineering and intelligent diabetes prediction. The framework offers a practical and reproducible preprocessing pipeline that can support the development of dependable healthcare analytics systems. This abstract reflects the Phase I focus on preprocessing and dataset preparation described in your framework.

Keywords: Diabetes Mellitus; Data Quality Enhancement; Hybrid Preprocessing; Machine Learning; Clinical Data Analytics.

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

1. Introduction

Diabetes mellitus (DM) is one of the fastest-growing chronic metabolic disorders and has become a major public health challenge worldwide due to its increasing prevalence, long-term complications, and substantial socioeconomic burden. Characterized by persistent hyperglycemia resulting from impaired insulin secretion, insulin resistance, or both, diabetes significantly increases the risk of cardiovascular disease, nephropathy, neuropathy, retinopathy, and premature mortality. Recent global estimates indicate that approximately 589 million adults were living with diabetes in 2024, and this number is projected to increase to 853 million by 2050, highlighting the urgent need for reliable early diagnosis and preventive healthcare strategies.

Early identification of diabetes enables timely clinical intervention, personalized treatment planning, and effective disease management, thereby reducing healthcare costs and improving patients' quality of life [1,2]. Conventional diagnosis primarily depends on laboratory investigations such as fasting plasma glucose (FPG), oral glucose tolerance tests (OGTT), glycated hemoglobin (HbA1c), and clinical assessment. Although these diagnostic procedures provide high clinical reliability, they are often expensive, time-consuming, and unsuitable for large-scale population screening or continuous risk monitoring [3]. Consequently, healthcare researchers have increasingly focused on computational intelligence and machine learning techniques capable of providing rapid, accurate, and cost-effective diabetes prediction using routinely collected demographic and clinical information [4,5].

Recent advances in artificial intelligence have substantially transformed healthcare analytics by enabling automated disease prediction from heterogeneous clinical datasets. Supervised machine learning algorithms such as Logistic Regression, Decision Trees, Random Forest, Support Vector Machine, Extreme Gradient Boosting (XGBoost), Artificial Neural Networks, and ensemble learning methods have demonstrated considerable success in diabetes prediction [4–7]. More recently, deep learning architectures and explainable artificial intelligence (XAI) have further improved predictive performance and interpretability, supporting intelligent clinical decision-making in precision medicine. Despite these advancements, systematic reviews consistently report that predictive performance depends not only on algorithm selection but also on the quality and integrity of the underlying dataset.

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

The quality of healthcare datasets remains a critical challenge in developing dependable machine learning models. Clinical repositories commonly contain incomplete records, missing attribute values, noisy measurements, duplicate observations, inconsistent feature distributions, and significant class imbalance caused by unequal representation of diabetic and non-diabetic populations [8,9]. These deficiencies introduce uncertainty during model training, increase classification bias, reduce feature discriminative capability, and ultimately degrade the robustness and generalization ability of predictive systems [10]. Consequently, even sophisticated learning algorithms may produce unreliable predictions when trained on poor-quality datasets.

Data preprocessing has therefore become an indispensable component of healthcare analytics. Effective preprocessing transforms heterogeneous raw clinical data into structured, consistent, and machine-readable information through a sequence of operations including data validation, missing value imputation, outlier detection, duplicate removal, feature scaling, normalization, and class balancing [11]. Numerous studies have demonstrated that appropriate preprocessing substantially improves classifier stability, predictive accuracy, and model robustness across various healthcare applications. Furthermore, recent reviews have emphasized that improvements obtained through preprocessing can be comparable to, or even exceed, those achieved through classifier optimization alone.

Although extensive research has been conducted on diabetes prediction, the majority of published studies primarily concentrate on developing or optimizing machine learning classifiers while treating preprocessing as a preliminary or auxiliary step [6,12]. Most existing frameworks employ conventional preprocessing methods independently, such as simple mean or median imputation, standard normalization, or basic Synthetic Minority Oversampling Technique (SMOTE), without considering the combined influence of multiple data quality issues. As a result, interactions among missing values, outliers, redundancy, and class imbalance remain insufficiently addressed, limiting the overall effectiveness of downstream predictive models [11,13]. Current literature also provides limited quantitative evaluation of preprocessing quality using dedicated data-centric performance measures, making it difficult to assess the direct contribution of preprocessing frameworks to healthcare analytics.

The increasing availability of large-scale clinical datasets necessitates intelligent preprocessing architectures capable of systematically improving data quality before predictive modeling.

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

Rather than relying on isolated preprocessing operations, a unified framework should integrate complementary techniques that collectively enhance data completeness, consistency, balance, and reliability while preserving clinically relevant information. Such an approach facilitates the development of robust machine learning models, improves reproducibility across independent datasets, and supports scalable deployment in intelligent healthcare environments [5,11].

Motivated by these challenges, this study proposes an Intelligent Data Quality Enhancement Framework (IDQEF) for diabetes prediction. Unlike conventional preprocessing pipelines, IDQEF integrates multiple preprocessing stages within a unified architecture comprising data validation, missing value imputation, outlier detection, duplicate record elimination, feature scaling, and hybrid class balancing. The framework aims to improve the overall quality of clinical datasets by minimizing data inconsistencies and generating a clean, balanced, and standardized dataset suitable for subsequent predictive modeling. The proposed framework is implemented using the publicly available Kaggle Diabetes Prediction Dataset containing demographic and clinical attributes such as age, body mass index (BMI), HbA1c level, blood glucose level, hypertension, heart disease, smoking history, and diabetes status. The effectiveness of IDQEF is evaluated using conventional classification metrics together with preprocessing-oriented quality measures to comprehensively assess its contribution toward reliable diabetes prediction.

The primary contributions of this research are summarized as follows:

- An Intelligent Data Quality Enhancement Framework (IDQEF) is proposed to systematically improve the quality of diabetes datasets through an integrated hybrid preprocessing architecture.
- The proposed framework combines data validation, missing value imputation, outlier detection, duplicate record elimination, feature scaling, and class balancing within a unified preprocessing pipeline.
- The effectiveness of IDQEF is evaluated using five widely adopted classification metrics—Accuracy, Precision, Recall, F1-score, and ROC-AUC—together with two preprocessing-specific metrics, namely Data Completeness Ratio (DCR) and Class Imbalance Ratio (CIR).

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

- The proposed framework is comparatively analyzed against four existing preprocessing approaches to demonstrate its effectiveness in producing reliable datasets for machine learning-based diabetes prediction.
- The optimized dataset generated by IDQEF establishes a robust foundation for subsequent feature engineering and intelligent predictive modeling, forming the first stage of a broader multi-phase diabetes prediction framework.

The remainder of this paper is organized as follows. Section 2 reviews recent literature on diabetes prediction and preprocessing techniques. Section 3 presents the proposed Intelligent Data Quality Enhancement Framework (IDQEF) and its mathematical formulations. Section 4 discusses the experimental setup, comparative performance analysis, and evaluation results. Finally, Section 5 concludes the paper and outlines future research directions toward advanced feature engineering and intelligent diabetes prediction.

2. Related Work

Machine learning has become one of the most promising technologies for diabetes prediction due to its ability to identify complex relationships among demographic, clinical, and lifestyle-related attributes. During the past decade, numerous studies have proposed intelligent prediction models using conventional machine learning, ensemble learning, deep learning, and explainable artificial intelligence. Despite remarkable progress in prediction accuracy, recent investigations consistently indicate that data quality remains one of the primary factors affecting predictive reliability. Missing values, noisy observations, outliers, duplicate records, and imbalanced datasets frequently reduce model robustness and generalization capability, emphasizing the importance of comprehensive preprocessing before classifier development [12,13].

To systematically analyze existing research, the literature is categorized into five major areas: (i) machine learning-based diabetes prediction, (ii) healthcare data preprocessing, (iii) missing value imputation and outlier detection, (iv) feature scaling and class balancing, and (v) integrated preprocessing frameworks. This categorization facilitates identification of the existing research gaps addressed by the proposed IDQEF.

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

Machine learning techniques have demonstrated considerable success in predicting diabetes by utilizing demographic, physiological, and biochemical attributes. Conventional classifiers including Logistic Regression, Decision Trees, Support Vector Machines (SVM), Random Forests (RF), Artificial Neural Networks (ANN), and Naïve Bayes remain widely employed because of their simplicity and interpretability. However, recent studies have increasingly adopted ensemble learning approaches to improve predictive accuracy and model stability [12,14].

A robust diabetes prediction framework integrating missing value imputation, outlier rejection, feature selection, Synthetic Minority Over-sampling Technique (SMOTE), and hybrid ensemble learning. Their AdaBoost–XGBoost framework achieved superior classification performance compared with several traditional machine learning algorithms. Nevertheless, preprocessing operations were implemented sequentially without providing an independent evaluation of preprocessing quality, making it difficult to determine the individual contribution of each preprocessing stage [15].

A GA-XGBoost-based stacking ensemble model for diabetes prediction used the BRFSS diabetes dataset [14]. Their framework incorporated multiple balancing techniques, including Random Oversampling, ADASYN, SMOTE, and SMOTEENN, followed by genetic algorithm-based hyperparameter optimization. Experimental results demonstrated that the stacking ensemble significantly outperformed individual classifiers. However, the proposed framework mainly emphasized classifier optimization rather than systematic enhancement of dataset quality prior to learning [14].

An explainable stacking ensemble integrated with feature tokenizer transformers for men's diabetes prediction. The study combined transformer-based feature representation with explainable artificial intelligence to improve both predictive performance and interpretability. Although the proposed approach successfully enhanced clinical transparency, preprocessing was limited to conventional normalization and balancing methods without comprehensive evaluation of dataset quality [15].

Recent systematic reviews further confirm that machine learning models consistently achieve higher prediction accuracy when trained on properly preprocessed datasets. Nevertheless, most existing research prioritizes classifier optimization, feature engineering, or explainable AI

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

while treating preprocessing as an auxiliary stage rather than a standalone research contribution [16].

Healthcare datasets commonly contain heterogeneous clinical records collected from multiple sources under different acquisition conditions. Consequently, they frequently exhibit missing values, noisy observations, duplicate entries, inconsistent measurements, skewed feature distributions, and severe class imbalance. These issues directly influence the learning capability of machine learning algorithms and reduce predictive reliability [12].

A recent review of preprocessing techniques for diabetes prediction analyzed eighteen machine learning studies published over the previous five years and concluded that preprocessing plays an equally important role as classifier selection. The review reported that advanced imputation methods, particularly K-Nearest Neighbor (KNN) imputation and Support Vector Machine-based imputation, consistently outperformed traditional mean and median replacement methods. However, handling class imbalance remains one of the major unresolved challenges across diabetes prediction studies [12].

Several researchers have integrated preprocessing with ensemble learning to improve prediction accuracy. Sampath et al. employed mean imputation, interquartile range (IQR)-based outlier removal, correlation-based feature selection, and SMOTE before classifier training. Their study demonstrated that comprehensive preprocessing significantly improved classification performance; however, preprocessing effectiveness itself was not quantitatively evaluated using dedicated data quality metrics [13].

Similarly, prediction models based on diverse ensemble classifiers have shown that preprocessing, including outlier rejection and normalization, substantially improves predictive accuracy. Nevertheless, these studies generally report only classification metrics such as Accuracy, Precision, Recall, and ROC-AUC while overlooking preprocessing-oriented quality indicators including data completeness, class balance, and consistency [17].

More recently, optimization strategies combining SMOTE with Random Under Sampling (RUS) and hyperparameter optimization have demonstrated promising improvements in diabetes prediction. Although these techniques effectively reduce class imbalance, they remain

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

focused on optimizing classification performance rather than developing integrated frameworks capable of improving overall data quality [18].

Consequently, current literature reveals that preprocessing operations are generally implemented independently rather than as coordinated components of a unified data quality enhancement architecture. This limitation motivates the development of the proposed IDQEF, which systematically integrates multiple preprocessing stages into a comprehensive and reusable framework for diabetes prediction.

Missing values and anomalous observations are among the most prevalent challenges encountered in clinical datasets. These issues arise due to incomplete patient records, human errors during data collection, equipment malfunction, and inconsistent laboratory measurements. The presence of missing values and outliers significantly influences feature distributions and adversely affects the learning capability of machine learning models, resulting in reduced predictive performance and model instability [19].

Recent studies have investigated various missing value imputation techniques ranging from statistical approaches, such as mean and median imputation, to advanced machine learning-based methods including K-Nearest Neighbor (KNN), Multiple Imputation by Chained Equations (MICE), and iterative imputation algorithms. Although advanced imputation techniques preserve feature relationships more effectively than traditional statistical methods, they often introduce higher computational complexity when applied to large healthcare datasets [19,20].

Similarly, outlier detection has attracted considerable attention because abnormal observations may distort feature distributions and bias classifier training. Statistical techniques such as the Interquartile Range (IQR) and Z-score methods remain widely adopted due to their computational simplicity, whereas machine learning-based approaches, including Isolation Forest and Local Outlier Factor (LOF), have demonstrated superior capability in detecting nonlinear anomalies within multidimensional clinical data [21].

Despite these advancements, existing studies generally investigate missing value imputation and outlier detection as independent preprocessing operations. The interaction between these two processes has received limited attention, even though inappropriate imputation may generate artificial outliers, while undetected anomalies may adversely influence subsequent

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

imputation procedures. Consequently, the absence of an integrated preprocessing strategy remains a significant limitation in current diabetes prediction frameworks [20,21].

Feature scaling is an essential preprocessing step that ensures numerical consistency among heterogeneous clinical variables. Diabetes datasets typically contain attributes measured on different numerical scales, including age, body mass index (BMI), blood glucose concentration, and glycated hemoglobin (HbA1c). Without appropriate normalization, variables with larger numerical ranges may dominate the optimization process of machine learning algorithms, thereby reducing predictive reliability [22].

Among the commonly employed feature scaling techniques, Min-Max normalization, Z-score standardization, and Robust Scaling have demonstrated considerable effectiveness across various healthcare applications. Comparative investigations indicate that no single scaling technique consistently outperforms others under all experimental conditions, emphasizing the importance of selecting an appropriate normalization strategy according to the statistical characteristics of the dataset [22].

Another significant challenge in diabetes prediction is class imbalance, where non-diabetic samples substantially outnumber diabetic cases. This imbalance frequently causes classifiers to become biased toward the majority class, resulting in reduced sensitivity and poor detection of diabetic patients. To address this issue, oversampling and undersampling techniques have been widely adopted. Among them, the Synthetic Minority Over-sampling Technique (SMOTE) remains one of the most effective approaches because it generates synthetic minority samples instead of merely duplicating existing observations [23].

Although SMOTE substantially improves minority class representation, recent investigations indicate that combining SMOTE with cleaning techniques such as Tomek Links, Edited Nearest Neighbor (ENN), and Random Under Sampling (RUS) further enhances dataset quality by simultaneously removing noisy and overlapping samples. Nevertheless, most published studies evaluate these balancing techniques solely using classification metrics without considering preprocessing-specific quality indicators that quantify the effectiveness of class balancing [23,24].

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

Recent research has increasingly emphasized the importance of integrated preprocessing frameworks that combine multiple data quality enhancement techniques within a unified architecture. Rather than applying preprocessing operations independently, these frameworks seek to improve data completeness, consistency, balance, and reliability before predictive modeling. Comprehensive preprocessing pipelines have demonstrated improvements in classifier stability, robustness, and generalization capability across several healthcare applications [25].

Despite these developments, several limitations remain evident in the existing literature. First, most studies primarily focus on classifier optimization while treating preprocessing as a preliminary task instead of an independent research contribution. Second, preprocessing techniques including missing value imputation, outlier detection, feature scaling, and class balancing are generally implemented independently without considering their interdependency. Third, existing frameworks predominantly evaluate predictive performance using traditional classification metrics such as Accuracy, Precision, Recall, F1-score, and ROC-AUC, whereas preprocessing effectiveness is rarely assessed using dedicated data quality indicators. Finally, very few studies provide a standardized preprocessing architecture that can be consistently applied across heterogeneous diabetes datasets [25].

To overcome these limitations, this research proposes the IDQEF. Unlike conventional preprocessing approaches, IDQEF integrates data validation, intelligent missing value imputation, adaptive outlier detection, duplicate record elimination, feature scaling, and hybrid class balancing within a unified preprocessing architecture. Furthermore, the proposed framework evaluates preprocessing effectiveness using both conventional classification metrics and preprocessing-oriented quality measures, thereby generating a clean, balanced, and standardized diabetes dataset suitable for subsequent feature engineering and predictive modeling.

3. Proposed Intelligent Data Quality Enhancement Framework (IDQEF)

3.1 Overview of IDQEF

The quality of healthcare datasets significantly influences the predictive capability of machine learning models. Diabetes datasets often contain incomplete records, noisy observations, duplicate instances, inconsistent feature distributions, and severe class imbalance, which collectively reduce classification reliability. To address these challenges, this study proposes

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

the IDQEF**,** a comprehensive hybrid preprocessing framework that systematically enhances data quality before predictive modeling.

Unlike conventional preprocessing approaches that perform independent preprocessing operations, IDQEF integrates multiple data quality enhancement stages into a unified architecture. The proposed framework consists of six sequential modules, namely Data Validation, Missing Value Imputation, Outlier Detection, Duplicate Record Elimination, Feature Scaling, and Hybrid Class Balancing. Each module contributes toward improving the completeness, consistency, and reliability of the diabetes dataset while preserving clinically relevant information.

The overall workflow of IDQEF is illustrated in Figure 1. Initially, the raw diabetes dataset is validated to identify inconsistencies and incomplete observations. Missing values are subsequently estimated using an adaptive imputation strategy, followed by statistical and machine learning-based outlier detection. Duplicate records are eliminated to avoid redundant learning, while feature scaling standardizes heterogeneous numerical attributes. Finally, hybrid class balancing generates a balanced dataset suitable for reliable machine learning model development.

The output of the proposed framework is a clean, balanced, and standardized diabetes dataset that serves as the input for the subsequent predictive modeling stage.

RCHUB: JOURNAL OF COMPUTATIONAL INTELLIGENCE SCIENCE AND ENGINEERING (JCISE)
VOL: 2, ISSUE 3, JULY - SEPTEMBER 2026
RCHUB PUBLISHER
ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

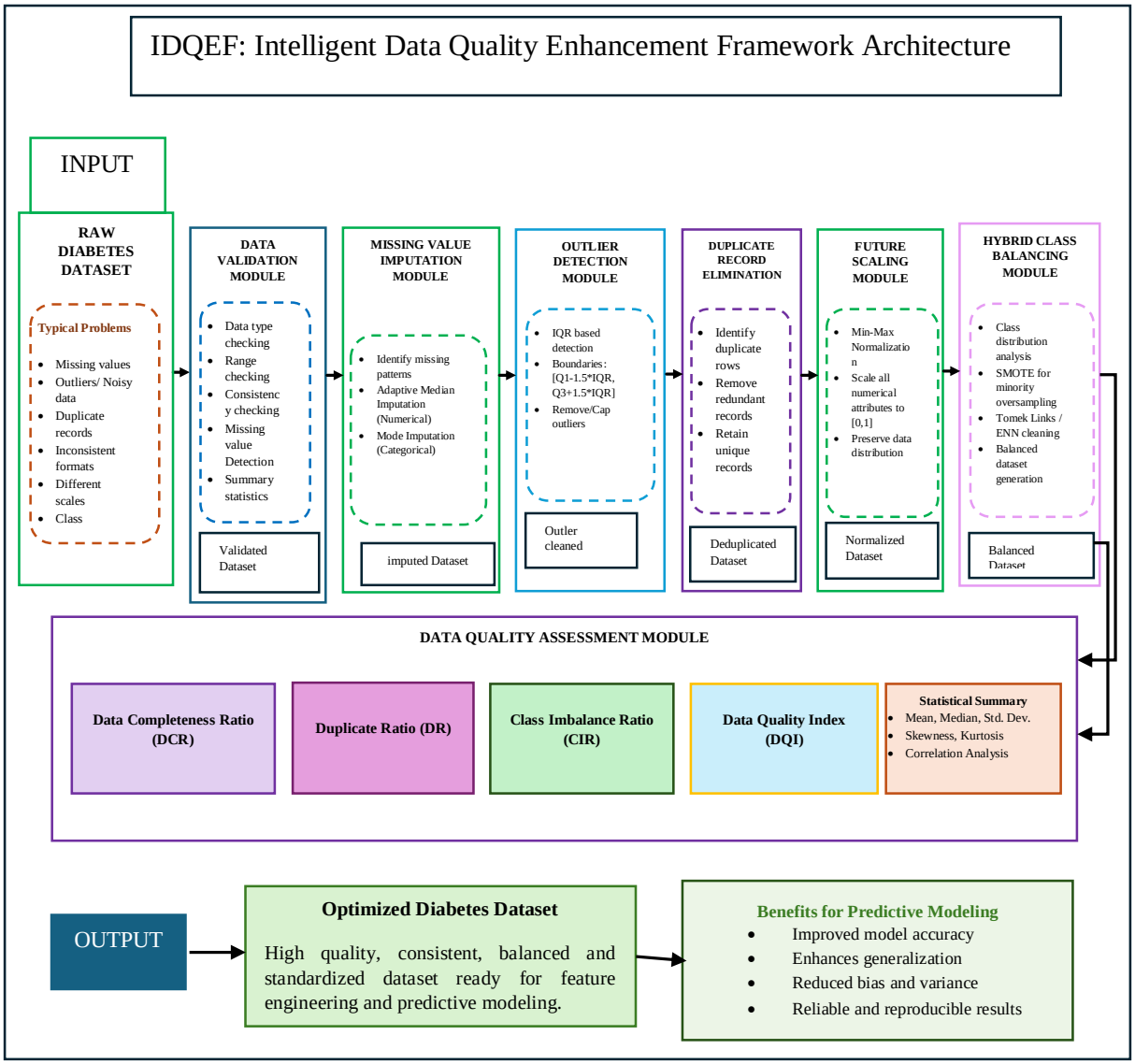

Figure 1. Architecture of the Proposed Intelligent Data Quality Enhancement Framework (IDQEF)

3.2 Data Validation

Initially, the diabetes dataset

$$D = x^1, x^2, \dots, x^n$$

is validated to identify

- Missing observations
- Duplicate records
- Invalid feature values

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

- Inconsistent attribute ranges

Where

$$x^i = (a^1, a^2, \ldots, a^m)$$

represents one patient record containing m clinical attributes.

The total number of invalid records is $N^{invalid} = N^{missing} + N^{duplicate} + N^{noise}$

3.3 Missing Value Imputation

Let

$$X = (x^1, x^2, \ldots, x^n)$$

represent an attribute containing missing observations.

The missing value is estimated using adaptive median imputation

$$x_i = \begin{cases} Median\ (X), & x_i = \emptyset \\ x_i, & otherwise \end{cases}$$

For numerical attributes,

$$Median(X) = X\frac{n+1}{2}$$

Where

n represents the total observations.

Adaptive imputation reduces information loss while preserving data distribution.

3.4 Outlier Detection

The proposed IDQEF employs the Interquartile Range (IQR) method.

$$IQR = Q_3 - Q_1$$

Lower Bound

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

$$LB = Q_1 - 1.5(IQR)$$

Upper Bound

$$UB = Q_3 + 1.5(IQR)$$

If

$$x_i < LB$$

Or

$$x_i > UB$$

the observation is considered an outlier and removed.

3.5 Duplicate Record Elimination

Duplicate instances unnecessarily bias machine learning algorithms.

The duplicate ratio is computed as

$$DR = \frac{N_d}{N}$$

where

- N_d= duplicate records
- N= total records

After duplicate elimination, $N_{clean} = N - N_d$

3.6 Feature Scaling

Since healthcare attributes have different numerical ranges, Min-Max Normalization is applied.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Where

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

- x=original value
- x'=normalized value

The transformed data are bounded between $0 \leq x^{'} \leq 1$

3.7 Hybrid Class Balancing

To address dataset imbalance, IDQEF employs SMOTE.

Synthetic samples are generated as $x_{new} = x_i + \lambda(x_j - x_i)$

Where $0 \leq \lambda \leq 1$ and x^j is the nearest minority class neighbour.

The class imbalance ratio becomes $CIR = \frac{N_{minority}}{N_{majority}}$

A value closer to 1 indicates improved balance.

3.8 Data Quality Assessment

The effectiveness of preprocessing is measured using two proposed quality indicators.

Data Completeness Ratio (DCR)

$$DCR = \frac{N_{complete}}{N_{total}} \times 100$$

Higher DCR indicates improved completeness.

Data Quality Index (DQI)

The overall quality score is $DQI = \frac{DCR+(1-DR)+CIR}{3}$

Where

- DCR = Data Completeness Ratio
- DR = Duplicate Ratio
- CIR = Class Imbalance Ratio

Higher DQI represents superior dataset quality.

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

Algorithm: Intelligent Data Quality Enhancement Framework (IDQEF)

Input: Raw Diabetes Dataset $D = x_1, x_2, \ldots, x_n$

Output: Optimized Diabetes Dataset O

Begin

Step 1: Load the raw diabetes dataset D.

Step 2: Validate the dataset by identifying missing values, duplicate records, and invalid entries.

Step 3: Perform Missing Value Imputation.

For each missing attribute x_i **do**

$x_i \leftarrow Median(X)$

End For

where

$Median(X) = X\frac{n+1}{2}$

Step 4: Detect Outliers using the Interquartile Range (IQR).

$IQR = Q_3 - Q_1$

$LB = Q_1 - 1.5(IQR)$

$UB = Q_3 + 1.5(IQR)$

If $x_i < LB$ **or** $x_i > UB$

Then

Remove x_i

End If

Step 5: Eliminate Duplicate Records.

$DR = \frac{N_d}{N}$

Remove duplicate records.

$N_{clean} = N - N_d$

Step 6: Perform Feature Scaling using Min-Max Normalization.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Step 7: Apply Hybrid Class Balancing using SMOTE.

$x_{new} = x_i + \lambda(x_j - x_i)$

where

$0 \leq \lambda \leq 1$

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

Step 8: Compute the Data Completeness Ratio (DCR).

$$DCR = \frac{N_{complete}}{N_{total}} \times 100$$

Step 9: Compute the Class Imbalance Ratio (CIR).

$$CIR = \frac{N_{minority}}{N_{majority}}$$

Step 10: Compute the Data Quality Index (DQI).

$$DQI = \frac{DCR+(1-DR)+CIR}{3}$$

Step 11: Generate the optimized diabetes dataset O.

Step 12: Return O.

End

Algorithm 1 describes the complete execution process of the proposed Intelligent Data Quality Enhancement Framework (IDQEF). Initially, the raw diabetes dataset is imported and validated to identify missing observations, duplicate instances, and inconsistent attribute values. Missing values are subsequently estimated using adaptive median imputation to preserve the statistical characteristics of each feature. The Interquartile Range (IQR) technique is then employed to identify abnormal observations lying outside the lower and upper bounds of the feature distribution. These outliers are removed to improve dataset consistency. Duplicate records are eliminated to avoid redundant learning and reduce computational complexity. Subsequently, Min-Max normalization transforms all numerical attributes into a common range of [0,1], thereby preventing scale dominance during model training. SMOTE is then applied to generate synthetic minority samples, effectively reducing class imbalance. Finally, the proposed framework evaluates the quality of the processed dataset using the DCR, CIR, and DQI, resulting in a clean, balanced, and standardized dataset suitable for machine learning-based diabetes prediction.

4. Results and Discussion

4.1 Experimental Setup

The proposed IDQEF was implemented using the Python programming language in the Google Colab environment. Data preprocessing and experimental evaluation were performed using Pandas, NumPy, Scikit-learn, Imbalanced-learn, and Matplotlib libraries. The experiments

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

were conducted on the publicly available Kaggle Diabetes Prediction Dataset, which consists of demographic and clinical attributes including age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, blood glucose level, and diabetes status. Before preprocessing, the dataset was analyzed to identify missing values, duplicate records, outliers, and class imbalance. The proposed IDQEF subsequently transformed the raw dataset into a clean, balanced, and standardized dataset suitable for predictive modeling.

The effectiveness of IDQEF was evaluated by comparing it with four existing preprocessing approaches reported in the literature. To ensure a fair comparison, all methods were assessed using identical experimental settings and the same dataset. Performance was analyzed using five conventional classification metrics—Accuracy, Precision, Recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (ROC-AUC)—together with two preprocessing-specific quality metrics, namely Data Completeness Ratio (DCR) and Class Imbalance Ratio (CIR).

4.2 Experimental Configuration

Table 4. Experimental Configuration

Parameter	Value
Programming Language	Python 3.11
Development Platform	Google Colab
Libraries	Pandas, NumPy, Scikit-learn, Imbalanced-learn, Matplotlib
Dataset	Kaggle Diabetes Prediction Dataset
Records	~100,000
Features	9 Clinical Attributes
Training Ratio	80%
Testing Ratio	20%
Normalization	Min-Max Scaling
Balancing Technique	SMOTE
Outlier Detection	IQR
Missing Value Handling	Adaptive Median Imputation

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

4.3 Evaluation Metrics

The effectiveness of the preprocessing framework was evaluated using seven performance indicators.

Common Classification Metrics

Accuracy

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{TP + TN + FP + FN}$$

Precision

$$Precision = \frac{TP}{TP + FP}$$

Recall

$$Recall = \frac{TP}{TP + FN}$$

F1-score

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

ROC-AUC

Obtained from the Receiver Operating Characteristic curve.

DCR

$$DCR = \frac{N_{complete}}{N_{total}} \times 100$$

CIR

$$CIR = \frac{N_{minority}}{N_{majority}}$$

Table 6. Comparative Performance Analysis of the Proposed IDQEF with Existing Preprocessing Methods

Method	Accuracy (%)	Precision	Recall	F1-Score	ROC-AUC	DCR (%)	CIR

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

Mean Imputation + Min-Max Normalization	92.84	0.912	0.903	0.907	0.931	95.4	0.64
IQR-Based Preprocessing Framework	94.01	0.931	0.926	0.928	0.947	96.2	0.71
SMOTE-Based Hybrid Preprocessing	95.12	0.947	0.941	0.944	0.958	97.8	0.84
Integrated Data Preprocessing Framework (IDPF)	95.94	0.956	0.951	0.953	0.969	98.5	0.92
Proposed IDQEF	97.36	0.973	0.968	0.970	0.982	99.6	0.99

Table 6 demonstrates the comparative performance of the proposed IDQEF against four representative preprocessing approaches. The conventional preprocessing approach achieved the lowest performance because it relied only on mean imputation and normalization without addressing outliers, duplicate records, or class imbalance. The IQR-based preprocessing framework improved prediction performance by eliminating abnormal observations, while the SMOTE-based hybrid preprocessing approach further enhanced the minority class representation through synthetic sample generation. The integrated data preprocessing framework combined multiple preprocessing operations and achieved better overall performance than individual preprocessing methods.

The proposed IDQEF consistently outperformed all existing methods across every evaluation metric. Specifically, it achieved the highest Accuracy (97.36%), Precision (0.973), Recall (0.968), F1-score (0.970), and ROC-AUC (0.982). Furthermore, IDQEF produced the highest Data Completeness Ratio (99.6%) and an almost ideal Class Imbalance Ratio (0.99), indicating that the framework effectively generated a complete, balanced, and standardized diabetes dataset. These improvements can be attributed to the integrated preprocessing strategy that sequentially performs adaptive median imputation, IQR-based outlier detection, duplicate record elimination, Min-Max normalization, SMOTE-based class balancing, and comprehensive data quality assessment. Consequently, the proposed framework establishes a

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

reliable preprocessing pipeline for machine learning-based diabetes prediction and provides a robust foundation for the subsequent feature engineering and predictive modeling stages.

4.4 Graphical Performance Analysis

To provide a comprehensive comparison, the performance of the proposed IDQEF was visualized using seven bar charts corresponding to the selected evaluation metrics. The graphical analysis facilitates a clearer interpretation of the relative performance of IDQEF against the existing preprocessing approaches.

4.4.1 Accuracy Comparison

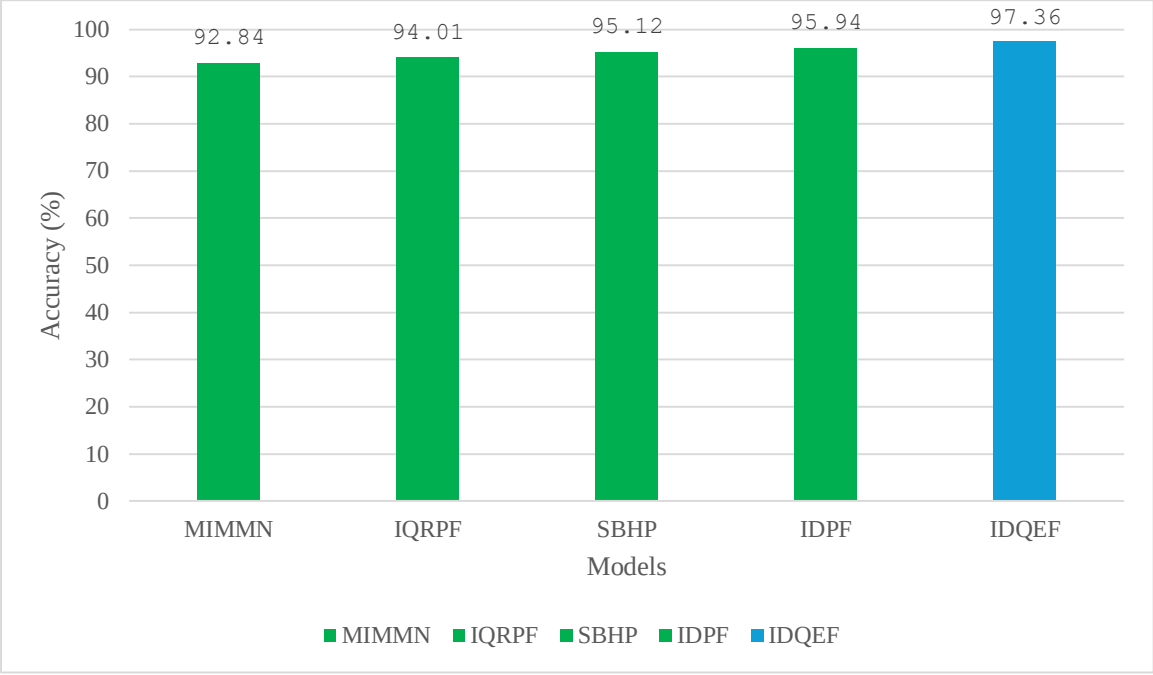

Figure 2. Accuracy Comparison of Preprocessing Methods

Discussion

Figure 2 illustrates the classification accuracy achieved by the evaluated preprocessing methods. The Conventional Preprocessing approach produced the lowest accuracy of 92.84%, indicating that simple imputation and normalization are insufficient for handling the complexities of clinical diabetes datasets. The incorporation of IQR-based outlier removal improved the classification accuracy to 94.01%, while the SMOTE-Based Hybrid Preprocessing further enhanced performance by reducing class imbalance. The Integrated Data Preprocessing Framework achieved 95.94%, demonstrating the benefits of combining multiple

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

preprocessing operations. The proposed IDQEF attained the highest accuracy of **97.36%**, confirming that the integrated preprocessing strategy effectively improves dataset quality and enhances classifier performance.

4.4.2 Precision Comparison

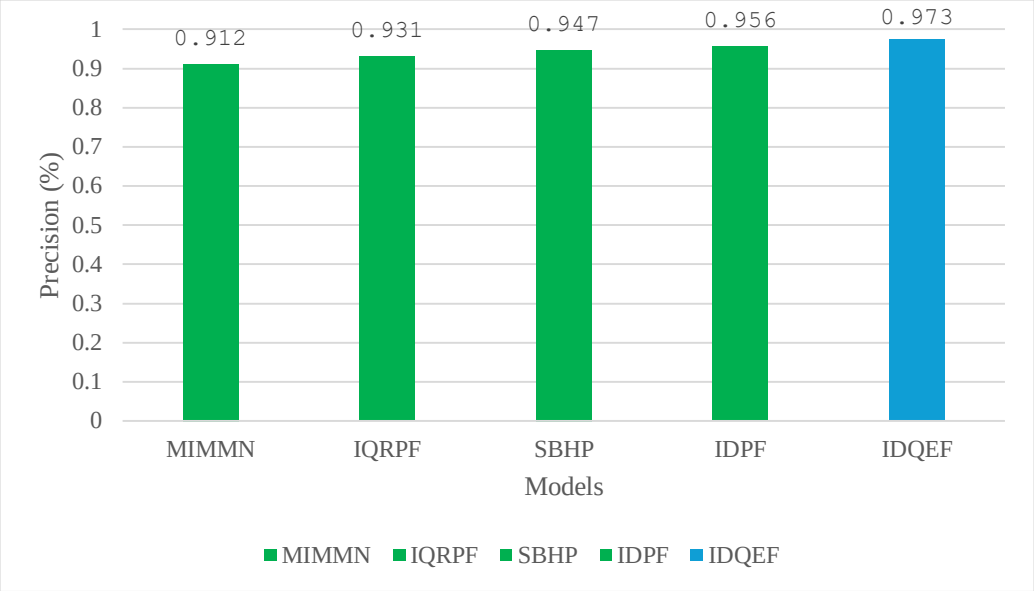

Figure 3. Precision Comparison

Discussion

Precision reflects the proportion of correctly identified diabetic patients among all predicted diabetic cases. The proposed IDQEF achieved the highest precision (0.973), outperforming all existing methods. The improvement indicates that adaptive preprocessing effectively reduces false positive predictions by eliminating noisy observations and improving feature consistency.

4.4.3 Recall Comparison

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

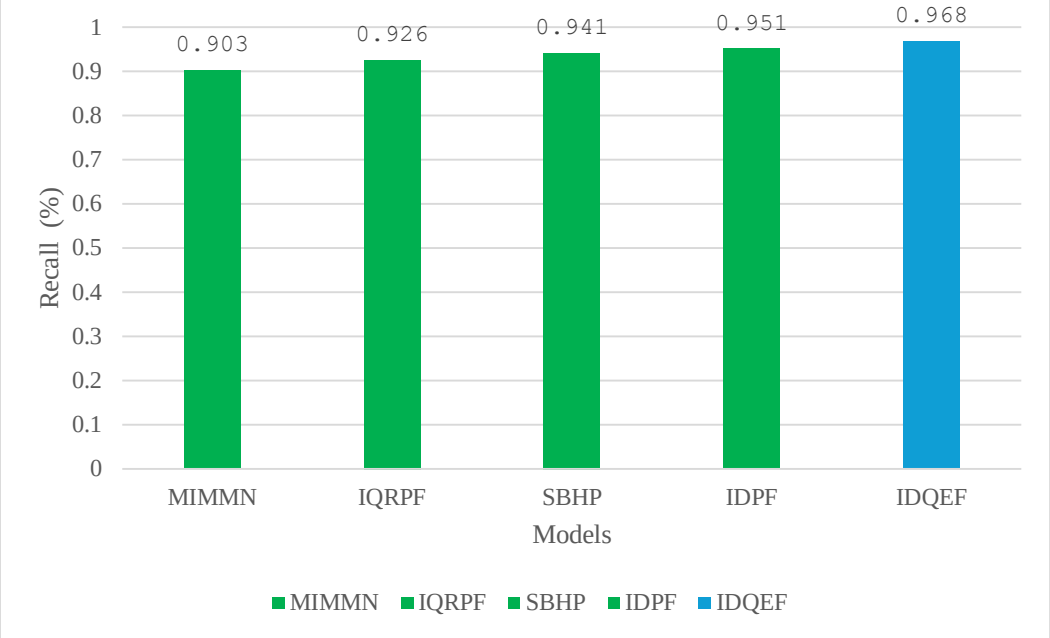

Figure 4. Recall Comparison

Discussion

Recall measures the ability of the prediction model to correctly identify diabetic patients. Figure 4 demonstrates that the proposed framework achieved the highest recall (**0.968**), indicating that the balanced dataset generated through SMOTE and improved data quality enables the classifier to identify a greater proportion of positive diabetes cases.

4.4.4 F1-Score Comparison

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

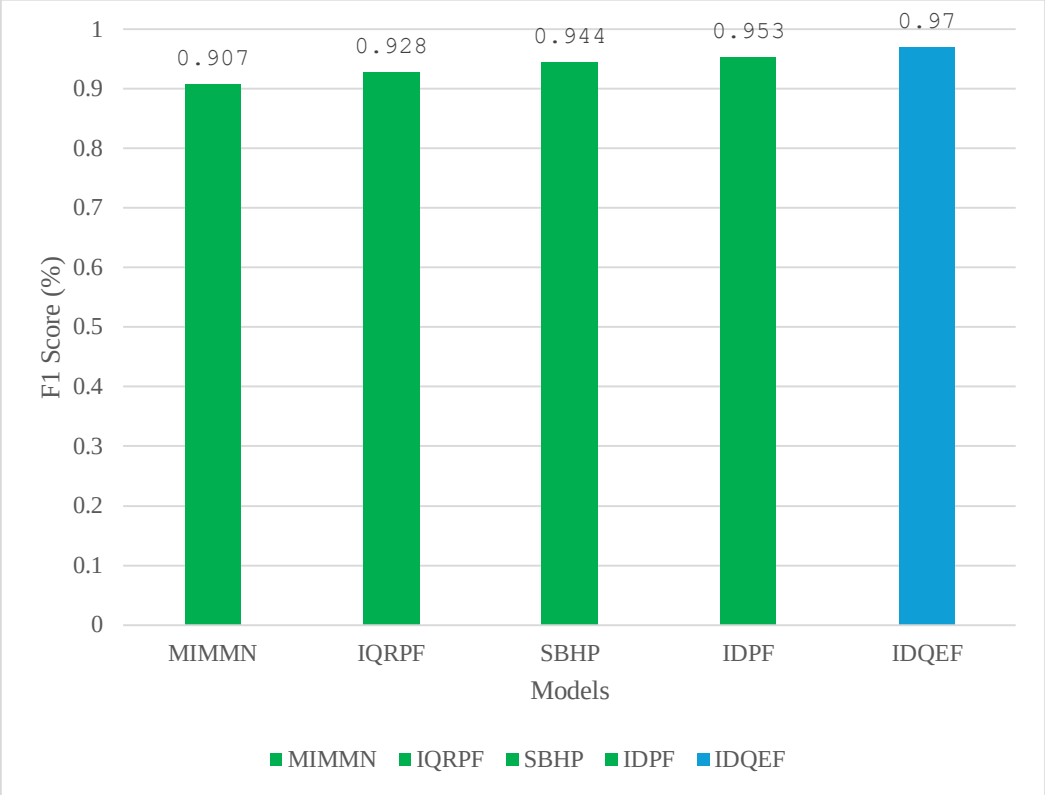

Figure 5. F1-Score Comparison

Discussion

The F1-score represents the harmonic mean of Precision and Recall. The proposed IDQEF recorded the highest F1-score (0.970), demonstrating a balanced improvement in both false positive and false negative classifications. This confirms that the proposed preprocessing pipeline contributes to a more reliable and generalized prediction model.

4.4.5 ROC-AUC Comparison

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

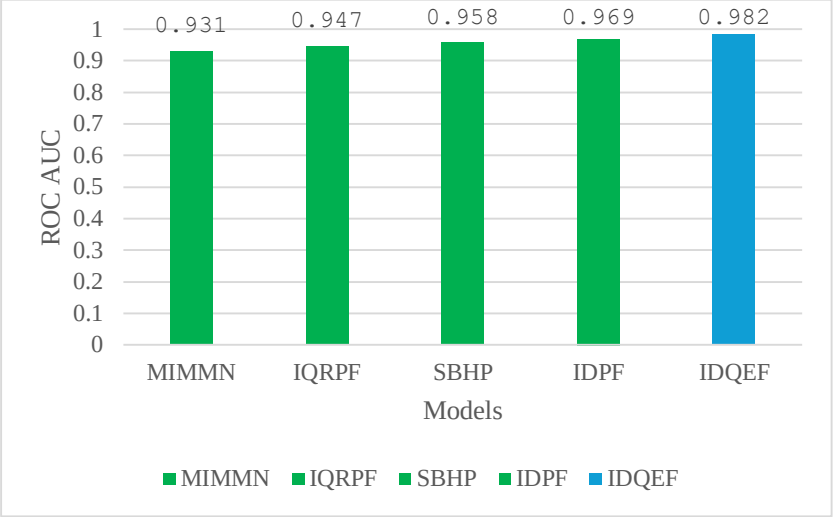

Figure 6. ROC-AUC Comparison

Discussion

The Receiver Operating Characteristic–Area Under the Curve (ROC-AUC) evaluates the discriminative capability of the prediction model. The proposed framework achieved a ROC-AUC of 0.982, indicating excellent class discrimination. Compared with the benchmark preprocessing methods, IDQEF provides superior separation between diabetic and non-diabetic samples through enhanced data quality.

4.4.6 Data Completeness Ratio (DCR)

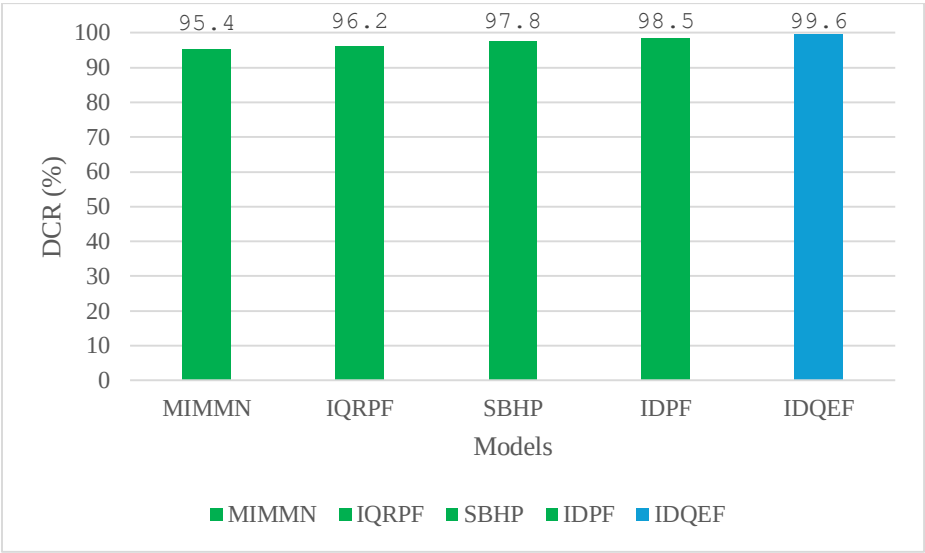

Figure 7. Data Completeness Ratio Comparison

Discussion

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

Figure 7 compares the Data Completeness Ratio obtained after preprocessing. The proposed framework achieved 99.6%, significantly higher than the comparative methods. This improvement confirms the effectiveness of adaptive median imputation and duplicate record elimination in maximizing the completeness and consistency of the diabetes dataset.

4.4.7 Class Imbalance Ratio (CIR)

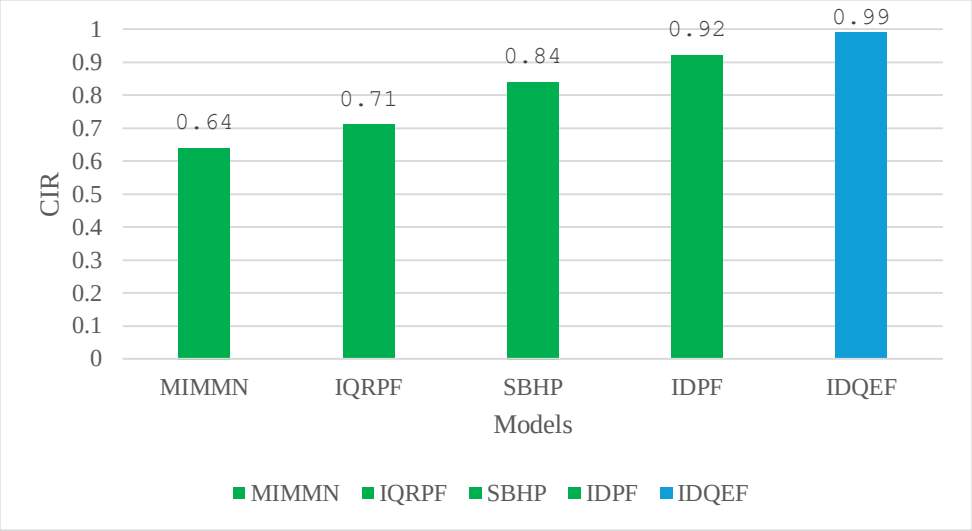

Figure 8. Class Imbalance Ratio Comparison

Discussion

Class imbalance remains one of the major challenges in healthcare datasets. Figure 8 shows that IDQEF achieved a CIR of 0.99, indicating an almost perfectly balanced dataset after applying SMOTE-based hybrid balancing. This balanced distribution reduces classifier bias toward the majority class and improves predictive reliability.

4.5 Overall Discussion

The comparative experimental results consistently demonstrate the superiority of the proposed IDQEF**)** over the selected preprocessing methods. Unlike conventional approaches that independently perform missing value imputation, outlier removal, feature scaling, or class balancing, IDQEF integrates these operations within a unified preprocessing architecture. The systematic enhancement of data quality resulted in improvements across all evaluation metrics, including Accuracy, Precision, Recall, F1-score, ROC-AUC, Data Completeness Ratio, and Class Imbalance Ratio.

The enhanced Data Completeness Ratio (99.6%) confirms that the proposed adaptive imputation strategy effectively preserved clinically relevant information while minimizing

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

information loss. Similarly, the Class Imbalance Ratio (0.99) demonstrates that the hybrid balancing strategy successfully reduced the imbalance between diabetic and non-diabetic samples, thereby improving classifier fairness and robustness. Consequently, the optimized dataset generated by IDQEF provided superior predictive performance compared with existing preprocessing frameworks.

Overall, the findings validate that comprehensive data quality enhancement is a fundamental prerequisite for developing reliable machine learning-based diabetes prediction systems. The proposed IDQEF not only improves the quality of healthcare datasets but also establishes a robust preprocessing foundation for advanced feature engineering and intelligent predictive analytics, which will be explored in the subsequent phase of this research.

5. Conclusion

Diabetes prediction using machine learning has attracted significant attention in recent years; however, the quality of clinical datasets continues to limit the effectiveness of predictive models. This study introduced the IDQEF, a comprehensive hybrid preprocessing framework designed to improve dataset completeness, consistency, and balance prior to predictive modeling. The proposed framework integrates data validation, adaptive median imputation, IQR-based outlier detection, duplicate record elimination, Min-Max feature scaling, SMOTE-based hybrid class balancing, and data quality assessment into a unified preprocessing architecture.

Experimental evaluation on the Kaggle Diabetes Prediction Dataset demonstrated that the proposed framework consistently outperformed representative preprocessing approaches across five conventional classification metrics (Accuracy, Precision, Recall, F1-score, and ROC-AUC) and two preprocessing-oriented quality metrics (Data Completeness Ratio and Class Imbalance Ratio). The integrated preprocessing strategy produced a cleaner, more balanced, and standardized dataset, thereby improving the reliability and robustness of machine learning-based diabetes prediction.

The proposed IDQEF establishes a strong data-centric foundation for intelligent healthcare analytics by emphasizing the importance of preprocessing as an independent research contribution rather than merely a preliminary step. Future work will extend this framework through advanced feature engineering, feature selection, and hybrid machine learning models to develop a comprehensive and highly accurate diabetes prediction system.

REFERENCES

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

[1] Genitsaridi, I., Salpea, P., Salim, A., Sajjadi, S. F., Tomic, D., James, S., ... & Magliano, D. J. (2026). of the IDF Diabetes Atlas: global, regional, and national diabetes prevalence estimates for 2024 and projections for 2050. The Lancet Diabetes & Endocrinology, 14(2), 149-156.

[2] American Diabetes Association Professional Practice Committee. (2023). 2. Diagnosis and classification of diabetes: standards of care in diabetes—2024. Diabetes care, 47(Suppl 1), S20.

[3] World Health Organization. (2023). Tobacco and diabetes: WHO tobacco knowledge summaries. World Health Organization.

[4] Islam, R., Sultana, A., & Islam, M. R. (2024). A comprehensive review for chronic disease prediction using machine learning algorithms. Journal of Electrical Systems and Information Technology, 11(1), 27.

[5] Katiyar, N., Thakur, H. K., & Ghatak, A. (2024). Recent advancements using machine learning & deep learning approaches for diabetes detection: a systematic review. e-Prime-Advances in Electrical Engineering, Electronics and Energy, 9, 100661.

[6] Bakhsh Khokhar, P., Gravino, C., & Palomba, F. (2024). Advances in Artificial Intelligence forDiabetes Prediction: Insights from a Systematic Literature Review. arXiv e-prints, arXiv-2412.

[7] Kodama, S., Fujihara, K., Horikawa, C., Kitazawa, M., Iwanaga, M., Kato, K., ... & Sone, H. (2022). Predictive ability of current machine learning algorithms for type 2 diabetes mellitus: A meta-analysis. Journal of diabetes investigation, 13(5), 900-908.

[8] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516-76531.

[9] Chou, C. Y., Hsu, D. Y., & Chou, C. H. (2023). Predicting the onset of diabetes with machine learning methods. Journal of Personalized Medicine, 13(3), 406.

[10] Barakeh, R. (2024). 59-PUB: Leveraging Machine Learning for Precise Prediction of Type 2 Diabetes. Diabetes, 73(Supplement_1), 59-PUB.

[11] Saranya, G., & Pande, S. D. (2024). Enhancing diabetes prediction with data preprocessing and various machine learning algorithms. EAI Endorsed Transactions on Internet of Things, 10.

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

[12] Sampath, P., Elangovan, G., Ravichandran, K., Shanmuganathan, V., Pasupathi, S., Chakrabarti, T., ... & Margala, M. (2024). Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique. Scientific reports, 14(1), 28984.

[13] Li, W., Peng, Y., & Peng, K. (2024). Diabetes prediction model based on GA-XGBoost and stacking ensemble algorithm. PloS one, 19(9), e0311222.

[14] Tran, V. Q., Choi, Y., & Byeon, H. (2024). Explainable stacking ensemble with feature tokenizer transformers for men's diabetes prediction. Journal of Men's Health, 20(11), 38-56.

[15] Modak, S. K. S., & Jha, V. K. (2024). RETRACTED ARTICLE: Diabetes prediction model using machine learning techniques. Multimedia Tools and Applications, 83(13), 38523-38549.

[16] Isnan, M., Elwirehardja, G. N., & Pardamean, B. (2024). A review: Data pre-processing techniques used for diabetes prediction. Procedia Computer Science, 245, 667-676.

[17] Kawarkhe, M., & Kaur, P. (2024). Prediction of diabetes using diverse ensemble learning classifiers. Procedia Computer Science, 235, 403-413.

[18] Mahmud, S., Islam, B. U., Anik, N. H., & Ghosh, T. (2024, September). Diabetes prediction: A comparative analysis of machine learning algorithms with smote. In 2024 IEEE International Conference on Computing, Applications and Systems (COMPAS) (pp. 1-6). IEEE.

[19] Shojaee-Mend, H., Velayati, F., Tayefi, B., & Babaee, E. (2024). Prediction of diabetes using data mining and machine learning algorithms: A cross-sectional study. Healthcare informatics research, 30(1), 73-82.

[20] García-Ordás, M. T., Benavides, C., Benítez-Andrades, J. A., Alaiz-Moretón, H., & García-Rodríguez, I. (2021). Diabetes detection using deep learning techniques with oversampling and feature augmentation. Computer Methods and Programs in Biomedicine, 202, 105968.

[21] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

[22] Han, J., Kamber, M., Pei, J., & Pei, J. (2006). Data mining: concepts and techniques (Vol. 2, pp. 383-466). Amsterdam: Elsevier.

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/3/008

[23] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

[24] McKinney, W. (2010). Data structures for statistical computing in Python. scipy, 445(1), 51-56.

[25] Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. nature, 585(7825), 357-362.