ISSN 2723-6471

A Framework for Diabetes Detection Using Machine Learning and Data Preprocessing

Ahmad Adel Abu-Shareha[1], Haneen Qutaishat[2], Asma Al-Khayat[3]

[1,2,3]*Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Al-Ahliyya Amman University, Amman,*

(Received: August 15, 2024; Revised: September 25, 2024; Accepted: October 03, 2024; Available online: October 15, 2024)

Abstract

People with diabetes are at an increased risk of developing other complications, such as heart disease and nerve damage. Therefore, diabetes prediction is crucial to reduce the severe consequences of this disease. This study proposed a comprehensive framework for diabetes prediction to maximize the information from available diabetes datasets, which include historical records, laboratory tests, and demographic data. The proposed framework implements a data imputation technique for filling in missing values and adopts feature selection methods to remove less important features for better diabetes classification. An oversampling technique and a parameter tuning approach were used to increase the samples and fine-tune the parameters for training the machine learning algorithms. Various machine learning algorithms, including Neural Networks, Logistic Regression, Support Vector Machines, and Random Forest, were used for the prediction. These algorithms were evaluated using both train-test split and cross-validation techniques. The experiments were conducted on the Pima Indian Diabetes dataset using various evaluation metrics, including accuracy, precision, recall, and F-measure. The results showed that the Random Forest algorithm, particularly when fine-tuned with Grid Search Cross Validation, outperformed other algorithms, achieving an impressive accuracy of 0.99. This demonstrates the robustness and effectiveness of the proposed framework, which outperformed the accuracy of state-of-the-art approaches.

Keywords: Diabetes, Classification, Machine Learning, Data Preprocessing, Performance Measures

1. Introduction

The consequences of diabetes are profound, affecting multiple organs and leading to severe complications such as blood vessel damage, angina, stroke, eye damage, and hearing impairment. This chronic condition significantly contributes to an increase in mortality rates [1]. Between 2000 and 2019, there was a 3% increase in age-standardized mortality rates due to diabetes, with lower-middle-income countries experiencing a 13% increase in mortality rates from this metabolic disorder. As such, in 2019, diabetes directly caused 1.5 million deaths, 48% of which were individuals under 70 years old. Moreover, diabetes was a factor in 460,000 kidney disease deaths and was associated with approximately 20% of cardiovascular deaths due to high blood glucose levels [2]. The situation nowadays seems worse, with over 422 million people worldwide currently living with diabetes. According to the International Diabetes Federation (IDF) reports, diabetes is becoming increasingly severe worldwide.

The statistics reveal global occurrences of 10.5%, with nearly half (44.7%) undiagnosed. Projections indicate that by 2045, approximately 783 million adults will have diabetes, meaning one in eight adults will suffer from this condition. This represents a 46% increase, significantly outpacing the estimated population growth of 20% during the same period [3]. With the occurrence of diabetes notably higher in developed countries and expected to rise to 5.4% by 2025, diabetes remains a critical public health challenge on an international scale [4]. These statistics highlight the urgent need for a robust predictive model to predict and mitigate the risk factors leading to diabetes, emphasizing the criticality of early intervention and personalized healthcare strategies. The statistics above confirm the pressing necessity to improve and automate the diabetes diagnosis and prognosis processes. While diabetes is a complex and multifactorial condition, early detection and intervention can significantly reduce the risk of complications and improve health outcomes for people with diabetes [5]. In recent years, artificial intelligence algorithms have been used to ease the complexity of diabetes diagnosis for diabetes prediction. Machine learning and data mining techniques have emerged as powerful tools for diabetes prediction and analyzing large datasets containing clinical, demographic, and laboratory

***Corresponding author: Ahmad Adel Abu-Shareha (aabushareha@yahoo.com)**
DOI: https://doi.org/10.47738/jads.v5i4.363

test results [6]. The diabetes diagnosis is a supervised classification problem based on given datasets. However, the success of the machine learning techniques depends heavily on the quality of the data used for model training and development. Poorly processed or incomplete data can lead to biased or inaccurate predictions, limiting the utility of predictive models in clinical practice. As such, data preprocessing steps, are crucial in developing accurate and reliable predictive of diabetes mellitus [7]. Data preprocessing techniques, including data cleaning, transformation, and feature selection, are essential for optimizing the performance of predictive models and improving their interpretability and generalizability [8].

Although various approaches have been developed for diabetes prediction comprising various preprocessing techniques and machine learning algorithms, the accuracy of the prediction remains inadequate [9]. The problem formed around analyzing and understanding the processed data and utilizing suitable methods and methods' combinations to improve the performance. Besides, the structure and the flow of these methods and techniques depend on the input data and its characteristics [10]. This paper proposes a framework for diabetes prediction, aiming to construct a robust and easily interpretable predictive model for better classification of diabetes. The proposed technique utilizes the most influenced and responsible factors along with regular factors like Glucose, BMI, Age, Insulin, etc., which contribute significantly to diabetes. This paper focuses on the importance of data preprocessing in diabetes prediction based on the characteristics of the data while implementing various machine learning algorithms to achieve the best accuracy. The proposed approach addresses the advantages and limitations of the Pima Indian Diabetes dataset (PIMA). The rest of this paper is organized as follows: Section 2 reviews the related work on diabetes detection. Section 3 presents the proposed framework for diabetes detection. Section 4 presents the results. Finally, the conclusion is given in Section 5.

2. Literature Review and Hypothesis Development

Supervised machine learning learns the relationship between input and output variables (s). In diabetes prediction, this learning process aims to identify the patterns between laboratory test results and health records in medical datasets and the output of being diabetic or non-diabetic. The ability of these techniques to learn depends on the algorithm and the input data. Various preprocessing steps of data mining are required to improve the learning process's performance. Various prediction models were proposed in the literature, using multiple data mining techniques, machine learning algorithms, and their integrations. This review primarily focuses on two key aspects: the machine learning algorithms and the preprocessing techniques utilized for early detection of diabetes.

An early approach for diabetes prediction was proposed by Dogantekin, et al. [11] using Linear Discriminant Analysis (LDA) to identify and select the significant features and Adaptive Network-based Fuzzy Inference System (ANFIS) as a classifier. The experiments were conducted using the PIMA dataset and evaluated using sensitivity and specificity, classification accuracy, and confusion matrix metrics. The results were obtained by splitting the dataset into training and testing sets with a percentage of 90/10. The results of the proposed model based on the ANFIS were accurate at 84.61%. Zangooei, et al. [12] used non-dominated sorting Genetic Algorithm-II (NSGA-II), a multi-objective evolutionary algorithm, to identify mapping points (MPs) for rounding real-values to integers in the preprocessing stage. Two ML algorithms were used: the Support Vector Regression (SVR) and the Support Vector Machine (SVM). Additionally, NSGA-II was used to optimize SVR kernel parameters, enhancing the model's performance. The prediction framework was tested on multiple datasets, including Liver Disorder, Breast Cancer, Hepatitis, and the PIMA dataset. The results showed that SVR achieved 86.13% and 84.61% accuracy for the SVM using the PIMA dataset and obtained through cross-validation.

Naz and Ahuja [13] proposed a diabetes prediction framework focusing on splitting the data to achieve the best accuracy. Various sampling (i.e., data splitting) techniques were inspected, such as linear, shuffled, stratified, and automatic sampling. Various classifiers were used; these are Neural Networks (NN), Naïve Bayes (NB), Decision Tree (DT), and Deep Learning (DL). The experiments were conducted on the PIMA dataset, using the best sampling technique, the shuffled sampling with an 80/20 percentage split for training and validation. The results showed that the accuracy ranged from 90% to 98%. Notably, DL demonstrated the highest accuracy, achieving an impressive rate of 98.07%, suggesting its potential as a predictive tool for healthcare professionals. Guldogan, et al. [14] conducted a study to evaluate the performance of two NN models, Multilayer Perceptron (MLP) and Radial Based Function (RBF),

for diabetes prediction based on the PIMA dataset. First, the data was analyzed using median summarization, and the normality of the distribution was assessed through the Kolmogorov-Smirnov test. Additionally, the Mann-Whitney U test was used for further analysis. The results were obtained through a 60/40 percentage split for training and testing. The results were 78.1% for MPL, while RBF achieved 76.8%. Feature's significances were assessed, and it was found that for MLP, glucose, BMI, and pregnancy are essential, whereas for RBF, glucose, skin thickness, and insulin.

Khanam and Foo [15] explored diabetes prediction using machine learning and deep learning techniques on the PIMA dataset. The dataset was split using cross-validation and an 85/15 percentage split. Preprocessing steps involved outlier removal, filling missing values with the corresponding mean value, feature scaling by normalizing the data in the range [0- 1], and feature selection using Pearson's correlation. DT, K-Nearest Neighbors (KNN), Random Forest (RF), NB, Adaboost (AB), LR, and SVM were used for the classification task. All the algorithms resulted in an accuracy greater than 70%, with LR and SVM achieving 77%–78%, while NN demonstrated the highest accuracy of 88.6%. Saxena, et al. [16] proposed a prediction framework with multiple preprocessing steps. Outlier removal and missing value imputation using the mean were implemented. Feature selection methods using correlation attribute analysis, information gain, and principal component analysis (PCA) were implemented, followed by hyper-parameter optimization. NN, DT, RF, and KNN were utilized for the classification. The evaluation was conducted on the PIMA dataset through cross-validation. Notably, the RF classifier achieved the highest accuracy of 79.8%, outperforming other models with accuracies of 77.60% for NN, 76.07% for DT, and 78.58% for KNN.

Chang, et al. [17] proposed a diabetes prediction framework for diagnosing type 2 diabetes integrated with an Internet of Medical Things (IoMT) framework. The prediction framework fills in missing values using the median, targeting the dataset's invalid zeros. Besides, feature selection techniques using PCA, k-means clustering, and importance ranking were implemented. Three supervised ML models, NB, RF, and DT, were used. The experiments on the PIMA dataset showed that the NB with refined feature selection (glucose, BMI, and age) achieved an accuracy of 79.13%. With a broader feature set, RF attained an accuracy of 79.57% and used median-based imputation to fill in missing values. The results were obtained through a 70/30 percentage split for training and testing. Reza, et al. [18] explored various preprocessing techniques to improve classification accuracy. For normalization, z-score and interquartile range (IQR) analysis was used. Median imputation was used to fill in the missing values. Synthetic Minority Over-sampling Technique (SMOTE) for oversampling was used. Various classifiers were used: DT, RF, SVM, and NN. The experiments used the PIMA dataset with 70/30 splitting and 5-fold cross-validation. The results showed that the utilized techniques achieved accuracies ranging from 77.10% to 96.91%, with the NN yielding the highest accuracy.

Various other approaches were implemented using different datasets. Tarokh [19] proposed a diabetes prediction framework with various machine learning algorithms and feature selection. First, statistical analysis of the features was conducted using t-tests for continuous variables and Chi-square tests for categorical variables. Feature selection via Logistic regression is then implemented. NB, DT, AB, and RF were used for classification. The experiments were conducted on the National Health and Nutrition Examination Survey (NHANES) dataset, which consists of 14 features and 6561 samples, with 657 diabetic and 5904 non-diabetic samples. The results extracted with cross-validation showed that RF achieved the highest accuracy of 94.25%, while the NB classifier exhibited the lowest accuracy of 86.70% using 10-fold cross-validation. Wu, et al. [20] developed a predictive framework for diabetes prediction based on a dataset from the Chinese population via the Dryad Digital Repository (DDR) website. Their preprocessing steps involved removing missing values. EXtreme Gradient Boosting (XGBoost) was used for the classification through 50/50 percentage split. the XGBoost achieved Area Under Curve (AUC) of 95.50% on test data. Table 1 summarizes the literature on diabetes prediction. As noted, DL obtained the highest accuracy (98.07%) in the reviewed framework for predicting diabetes onset.

Table 1. Comparative Analysis of Diabetes Diagnosis using PIDD

Reference	Algorithm	Preprocessing	Sampling	Accuracy
Dogantekin, et al. [11]	LDA, ANFIS	LDA for feature selection	90/10	ANFIS: 84.61%
Zangooei, et al. [12]	SVR and SVM	NSGA-II for optimization	CV	SVM: 86.13%
Naz and Ahuja [13]	NN, NB, DT, and DL	Sampling techniques	80/20	DL: 98.07%
Guldogan, et al. [14]	NN: MLP and RBF	Medians and normality tests	60/40	MLP: 78.1%

 ISSN 2723-6471

Khanam and Foo [15]	DT, KNN, RF, NB, AB, LR, SVM and NN	Outlier removal, mean imputation, normalization, Feature selection	CV	NN: 88.6%
Saxena, et al. [16]	NN, DT, RF and KNN	Feature selection, outlier removal, mean imputation, hyper-parameter optimization	CV	RF: 79.8%
Chang, et al. [17]	NB, RF, and J48 DT	Median imputation, PCA, k-means, and importance ranking	70/30	RF: 79.57%
Reza, et al. [18]	DT, RF, SVM, and NN	Z-score, IQR, median imputation, and SMOTE	CV 70/30	NN: 96.91% En: 96.64%
Tarokh [19]	NB, DT, AB, and RF	Feature selection (Logistic regression)	CV	RF: 94.25% (NHANES)
Wu, et al. [20]	GBoost	Missing value removal	50/50	AUC: 95.50% (DDR)

As noted, besides the machine learning algorithms, various data mining techniques were implemented to improve the results of the diabetes prediction. The trends for these approaches can be drawn from a fundamental approach using machine learning and feature selection to tuning and oversampling, as shown in figure 1. Overall, preprocessing steps, oversampling, and sampling techniques prove their influences on the results of diabetes prediction. It was also noted that RF and GBoost algorithms achieve good results in this domain.

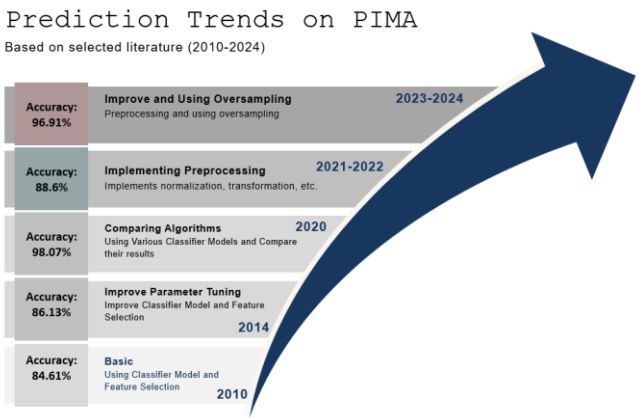

Figure 1. Diabetes Prediction Trends Overtime

3. The Proposed Framework

The main objective of the proposed framework is to maximize the information that can be discovered from the data and improve their quality to optimize the performance of machine learning for diabetes detection. Specifically, the framework processes the PIMA dataset, which has been studied extensively in the literature. This dataset's rich and varied features, including historical records, laboratory tests, and demographic data, provide a robust foundation for developing predictive models. The PIMA dataset is also widely used as a benchmark in the machine learning and medical research communities. This benchmark dataset ensures that the developed approaches and frameworks can be compared against a vast array of research, facilitating the validation and improvement of predictive models.

The PIMA dataset holds significant value, providing detailed medical records and demographic information specific to the Pima Indian population, a group known to have a higher prevalence of diabetes. Thus, the data allows for investigating the risk factors and patterns within this community. Additionally, the dataset includes a comprehensive range of features such as age, BMI, blood pressure, insulin levels, and glucose concentration, which are essential for identifying predictors of diabetes. The richness and variety of the data enable the development and testing of predictive models that can accurately identify at-risk individuals. Given the need for diabetes prediction and the value of the PIMA dataset, a framework for diabetes prediction is developed, as given in figure 2.

 ISSN 2723-6471

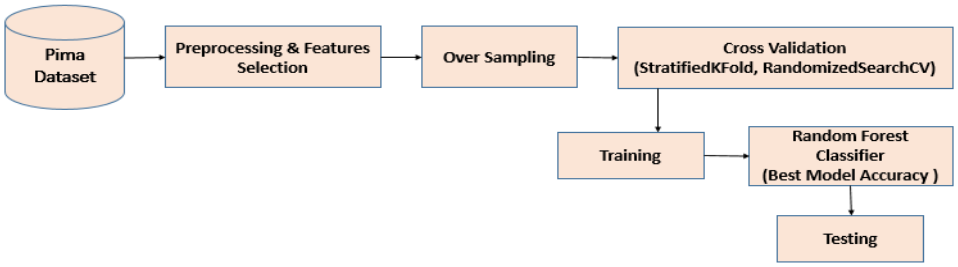

Figure 2. The Framework for Diabetes Prediction

3.1. Dataset Description

The PIMA dataset is collected by the National Institute of Diabetes and Digestive and Kidney Diseases. The selection criteria for instances in the dataset were strict, comprising females of at least 21 years old of Pima Indian heritage. The dataset comprises 768 samples and various medical predictor variables, including the number of pregnancies, BMI, insulin level, age, glucose levels, diastolic blood pressure, and skin thickness, alongside the target variable (Diabetic/Non-Diabetic). A detailed description of the variables of the PIMA dataset is given in table 2. Figure 3 shows the distribution of classes within the dataset, showing the number of diabetic and non-diabetic cases.

Table 2. Dataset Column Information

Column Name	Description
Pregnancies	Number of times the patient has been pregnant
Glucose	Plasma glucose concentration after 2 hours in an oral glucose tolerance test
Blood Pressure	Diastolic blood pressure (mm Hg)
Skin Thickness	Triceps skinfold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body mass index (weight in kg/(height in m)^2)
Diabetes Pedigree Function	A function which scores the likelihood of diabetes based on family history
Age	Age (years)
Outcome	Class variable (0 if non-diabetic, 1 if diabetic)

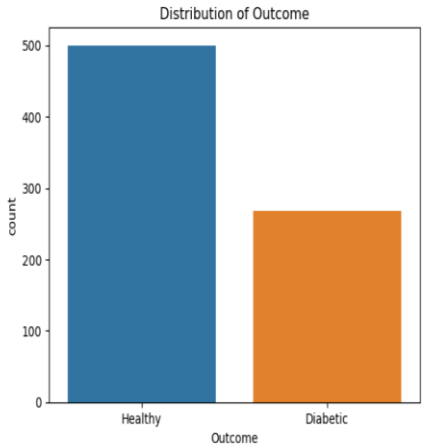

Figure 3. Distribution of Classes for PIMA Dataset

3.2. Dataset Analysis

A statistical summarization of the data is implemented using a Box plot, as illustrated in figure 4, which provides a clear visualization of the distribution of each feature within the PIMA dataset. The box plots highlight the median, quartiles, and potential outliers for each variable, offering valuable insights into the central tendency and variability of

 ISSN 2723-6471

the data. For instance, outliers in certain features, such as insulin levels, suggest careful data preprocessing is necessary to ensure robust model performance. This visual representation helps identify any skewness, spread, and anomalies within the dataset, guiding the implementation of appropriate preprocessing steps to enhance the quality and predictive power of the dataset for diabetes detection tasks. The correlation between these variables (including the outcome) is calculated and visualized in figure 5.

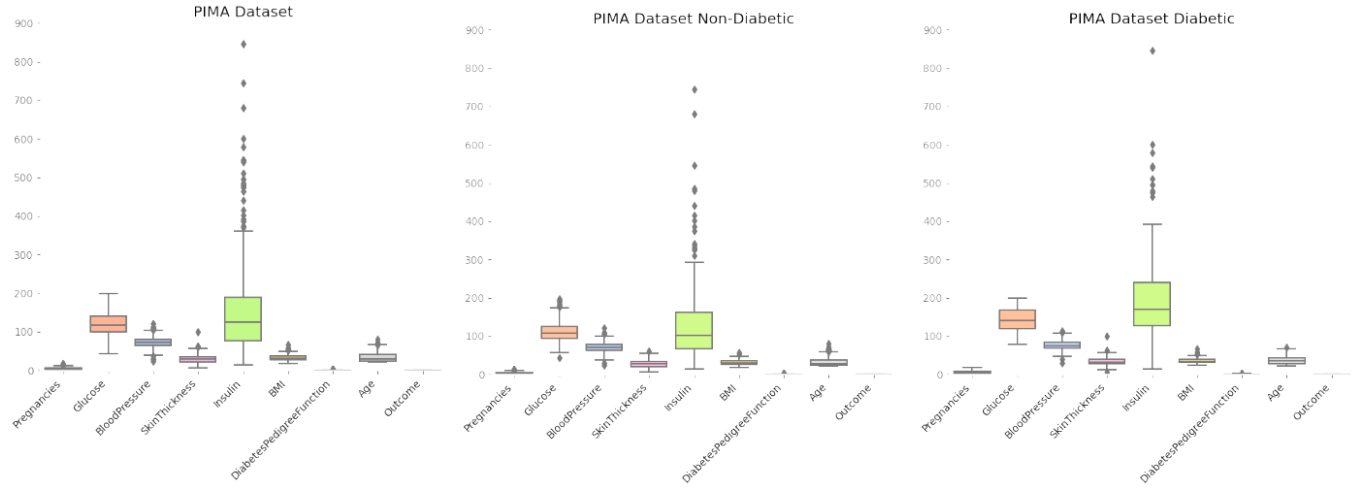

Figure 4. Box Plots of the Values Distributions in PIMA's Features

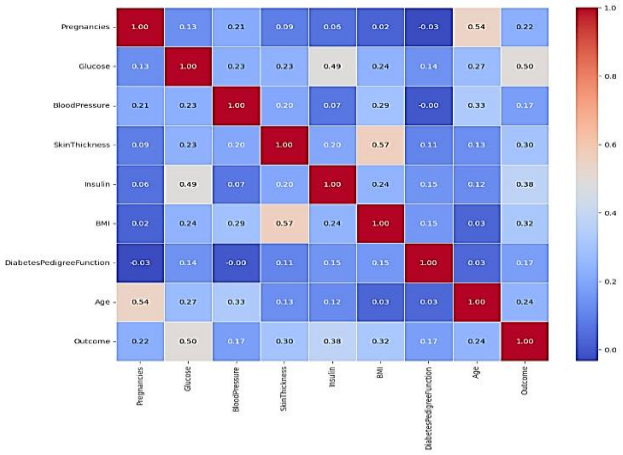

Figure 5. The Correlation between PIMA's Features

This correlation matrix provides a comprehensive overview of the relationships between different features in the PIMA dataset. Strong correlations indicate which features significantly impact the outcome, which is crucial for feature selection and model development. For instance, the correlation matrix reveals that glucose concentration has a strong positive correlation with the diabetes outcome, suggesting its importance as a predictor. Conversely, other features such as Blood Pressure and Diabetes Pedigree Function show weaker or more complex relationships with the outcome, highlighting the need for careful consideration in the modelling process.

The missing values are also counted as summarized in table 3 and presented in percentage in figure 6. Missing values analysis is critical for understanding the completeness of the dataset and guiding the implementation of appropriate data imputation techniques. Table 3 provides a detailed count of missing values for each feature, revealing that certain variable, such as Insulin levels and Skin Thickness, have a higher proportion of missing data than others. Figure 6 visualizes these percentages, clearly indicating the extent of missing values across different features. Due to the high percentage of missing values, the proposed framework employs an imputation technique for filling these missing values, thereby preserving the integrity of the dataset and improving the overall quality of the data used for training the machine learning algorithms.

 ISSN 2723-6471

Table 3. Summary of Missing Values in PIMA Dataset

Variable	Missing Values
Glucose	5
Blood Pressure	35
Skin Thickness	227
Insulin	374
BMI	11
Pregnancies	111
Diabetes Pedigree Function	0

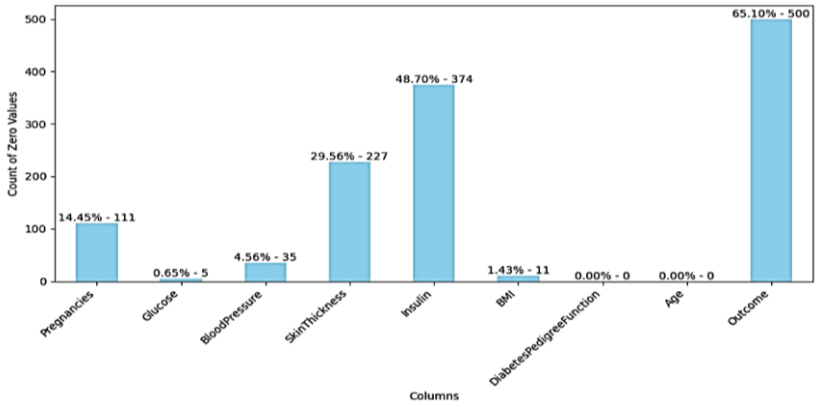

Figure 6. The Distribution of Missing Values

3.3. Data Preprocessing

Data preprocessing was conducted to ensure the quality and integrity of the dataset before model construction. The preprocessing pipeline included several steps, including filling in missing values using the imputation technique, oversampling, and feature selection. These preprocessing steps include handling missing values, addressing the class imbalance, selecting the most impactful features, optimizing the dataset, and laying a solid foundation for subsequent model development and analysis. The careful curation and preparation of the data are crucial for building robust and accurate machine-learning models. With a clean, balanced, and feature-rich dataset, the models are better equipped to learn and generalize from the data, ultimately leading to more reliable predictions.

First, when filling in missing values, the median of each column was calculated, and the missing values were identified and replaced with the corresponding median value. No samples were removed during this process, ensuring the dataset retained its full scope and diversity. This imputation strategy was chosen because the median is robust to outliers and provides a reliable central value for filling missing entries, as illustrated in figure 7. As noted in figure 7 compared to figure 4, median-based imputation helps maintain the overall distribution of each feature. By filling missing values with the median, the aim is to preserve the statistical properties of the dataset while avoiding the introduction of bias that could occur with other imputation methods.

 ISSN 2723-6471

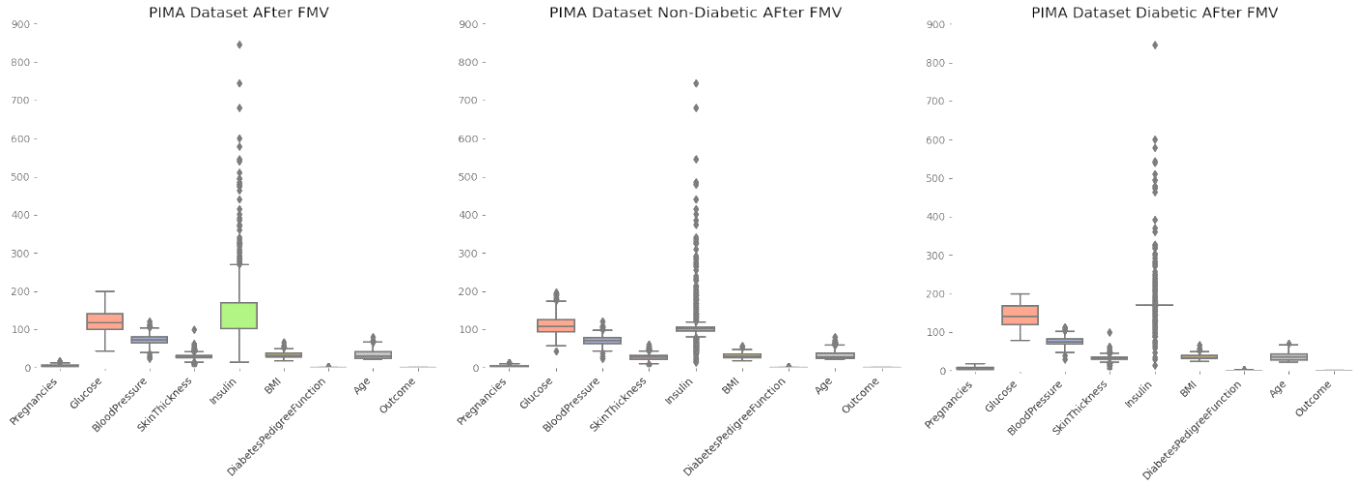

Figure 7. Box Plots of the PIMA's Features after Filling Missing Values

Feature selection uses a Random Forest model to identify the most important features strongly correlated with the target variable. All features above the threshold of 0.5 were selected, and the rest were filtered out. The selected features were Insulin, Glucose, Skin Thickness, BMI, Age, and Diabetes Pedigree Function, as illustrated in figure 8. These features were prioritized for their significant impact on predicting the target variable. Feature selection enhances the predictive power and efficiency of the framework.

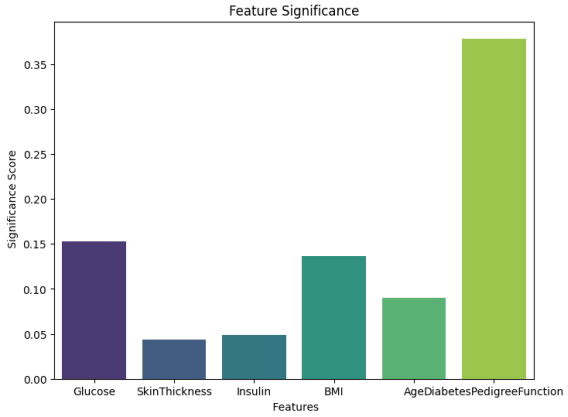

Figure 8. The Significance Values of Selected Features

To tackle class imbalance, SMOTE oversampling is used to increase samples in the minority class, resulting in a balanced dataset. Initially, we had 768 samples. After oversampling the dataset, the resulting data comprised 802 samples. This approach ensures that the machine learning models are trained on a more balanced representation of the classes, thereby enhancing model accuracy and robustness. Balancing the dataset is crucial for diabetes prediction because it helps mitigate the bias when one class is underrepresented. Without addressing the class imbalance, models tend to be biased towards the majority class, leading to poor performance in predicting the minority class, which, in this case, is crucial for identifying potential diabetes cases. Oversampling creates a dataset where both classes are equally represented, allowing the models to learn the characteristics of both classes effectively. However, SMOTE may generate unrealistic samples, particularly in cases where the minority class is sparse, potentially leading to overfitting. Additionally, SMOTE might introduce noise, as synthetic samples are interpolated between nearest neighbors, which may only sometimes reflect realistic patient profiles. To mitigate these issues, the distribution of the samples after oversampling is analyzed. As noted in figure 9 compared to figure 4, oversampling helps maintain the overall distribution of each feature.

 ISSN 2723-6471

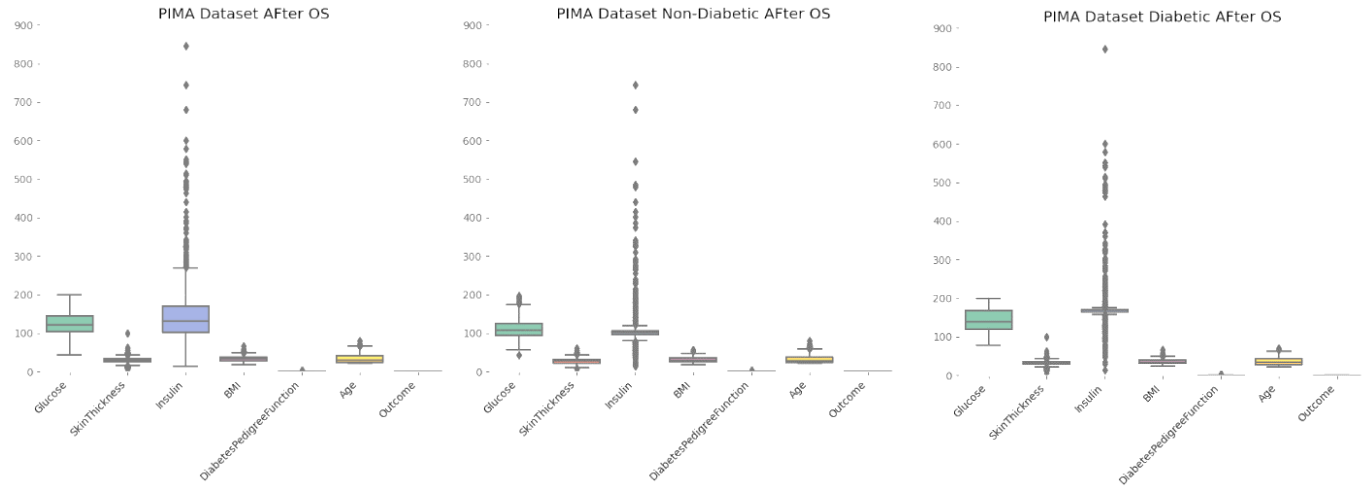

Figure 9. Box Plots of the PIMA's Features after SMOTE Oversampling

In summary, combining median imputation for missing values, SMOTE Over Sampler for class balance, and Random Forest for feature selection ensures the dataset is well-prepared for machine learning. These preprocessing techniques collectively enhance the quality and representativeness of the dataset, enabling the development of effective diabetes prediction models. The information obtained from the PIMA dataset is maximised by leveraging these methods.

3.4. Machine Learning Algorithms

This study uses different machine learning algorithms to predict whether patients have diabetes. These are NN, LR, SVM and RF. NN comprises layers of interconnected neurons, each processing input data through weighted connections and activation functions. For diabetes prediction in the PIMA dataset, a typical neural network architecture involves an input layer of six nodes, one hidden layer, and an output layer. The network learns by adjusting weights, w, and biases, b, during training, aiming to minimize a loss function l. After passing through the sigmoid activation functions, the network's output represents the probability of an individual having diabetes based on their input features.

LR calculates the probability of a binary outcome (diabetic or non-diabetic) based on a linear combination of input features. In the context of the Pima dataset, logistic regression estimates coefficients b for each feature xi, adjusting the intercept b0 to minimize the difference between the predicted probabilities and the actual binary outcomes in the training data. This is typically achieved using maximum likelihood estimation, where the goal is to find the set of coefficients that maximize the likelihood of the observed data.

SVM finds a hyperplane that best separates diabetic and non-diabetic instances in the feature space. In the Pima dataset, SVM maximizes the margin between support vectors, which are the closest points to the decision boundary, ensuring robust classification. This margin maximization helps improve the model's generalization ability by finding the optimal hyperplane that separates the two classes with the greatest distance, reducing the risk of misclassification.

RF builds multiple decision trees during training, each utilizing a random subset of features and data samples. In the PIMA dataset, this ensemble method aggregates predictions from individual decision trees to determine the final class (diabetic or non-diabetic).

3.5. Parameter Tuning

RF requires tuning the hyperparameters representing the number of trees and their depth, minimum number of leaves and minimum split. Optimizing such parameters significantly affects the results. As such, two methods to do this are Grid Search CV (GSCV) and Randomized Search CV (RSCV). GSCV is an exhaustive method that systematically tests the values of the hyper-parameter combinations to find the optimal settings. Every possible combination of hyper-parameters within a specified grid of values is evaluated using cross-validation to determine which combination yields the best accuracy. RSCV, on the other hand, explores a random subset of the hyper-parameter space. The optimization process is built on randomly selecting hyper-parameter values from specified distributions, allowing for a more efficient search than GSCV when the search space is large or continuous. A comparison between GSCV and RSVC is

 ISSN 2723-6471

given in table 4. GSCV and RSCV use internal cross-validation to evaluate each hyper-parameter combination, ensuring that the selected settings generalize well to unseen data.

Table 4. Comparison between GSCV vs. RSCV

Aspect	Grid Search CV	Randomized Search CV
Search Strategy	Exhaustive search	Random search
Parameter Space	All possible combinations	Randomly selected combinations
Time Complexity	Generally higher	Generally lower
Resource Efficiency	Less efficient	More efficient
Optimal Discovery	More likely	Less likely
Scalability	Less scalable	More scalable
Use Case	Small to medium-sized parameter spaces	Large parameter spaces
Implementation	Straightforward	More complex

4. Experimental Results

The experimental results were conducted using NN, RF, SVM, and LR machine learning algorithms, as well as GSCV and RSCV, to fine-tune the hyper-parameters of RF. All experiments were conducted using 10-fold cross-validation and an 80/20 percentage split. The evaluation is based on accuracy, precision, recall and f-measure.

4.1. Implementation

The experiments were conducted using Python 3, with a set of libraries: numpy, pandas, sci-kit-learn, imlearn, keras, matplotlib, and seaborn. The process flow is illustrated in figure 10.

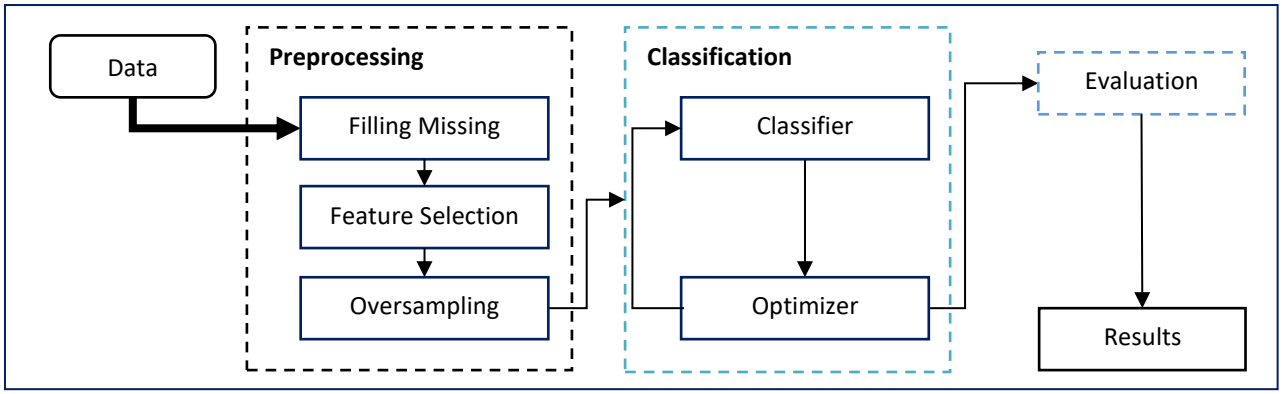

Figure 10. Implementation Process

4.2. Performance Metrics

Accuracy, precision, recall, and f-measure are calculated and used to assess the effectiveness of each model in predicting diabetes outcomes. These performance metrics are calculated based on the confusion matrix, which compares predicted and actual labels. Accuracy is the proportion of true results (both true positives and true negatives) to the total number of tested samples. Precision (also called positive predictive value) is the proportion of true positive results in the predicted positive instances. Precision indicates how many predicted positive instances are positive, reflecting the model's ability to avoid false positives. Recall (also called sensitivity or true positive rate) is the proportion of true positive results in the actual positive instances. Recall measures the model's ability to identify all positive instances, reflecting its ability to avoid false negatives. F-measure (or F1-score) is the harmonic mean of precision and recall, providing a metric that balances both concerns. These measures are calculated using Equation 1, Equation 2, Equation 3 and Equation 4.

$$Acc = tp + tn/(tp + tn + fp + fn) \quad (1)$$

 ISSN 2723-6471

$$\text{Precision} = \text{tp}/(\text{tp} + \text{fp}) \quad (2)$$

$$\text{R} = \text{tp}/(\text{tp} + \text{fn}) \quad (3)$$

$$\text{F1} = 2 * \text{Precision} * \text{Recall}/(\text{Precision} + \text{Recall}) \quad (4)$$

where tp is the number of true positives, tn is the number of true negatives, fp is the number of false positives, and fn is the number of false negatives.

4.3. Sampling and Data Splitting

K-fold cross-validation (CV) is a robust technique used to evaluate the performance of machine learning models, ensuring that the results are reliable and generalizable. The dataset is randomly divided into k equal-sized subsets, or "folds." The model is then trained and validated k times, using a different fold as the validation set and the remaining k-1 folds as the training set. This process helps ensure that every data point is used for training and validation, comprehensively assessing the model's performance. During each iteration, the model's performance is evaluated using the performance metrics, which are then averaged across all the iterations to obtain a final performance estimate. In this research study, 10-fold cross-validation was employed to evaluate the effectiveness of the machine learning algorithms.The percentage split method is a straightforward approach for evaluating the performance of machine learning algorithms. This method divides the dataset into two subsets: training and testing sets. Typically, a common split is 80% of the data for training and 20% for testing. The training set is used to train the model and learn the patterns and relationships in the data. The testing set, which the model has not seen during training, is used to evaluate the model's performance.

4.4. Results

The values of the utilized parameters for the implemented classifiers are listed in table 5, while the optimized parameters are listed in table 6. The results of the proposed framework are given in table 7 and figure 11. Table 7 presents the performance of different machine learning algorithms on the PIMA dataset using train-test split and cross-validation.

Table 5. The Parameters of the Utilized Classifier

Algorithm	Parameter	Value
NN	Optimizer	adam
	# Layers	3 (64, 32,1)
	Activation function(s)	ReLU, ReLU, Sigmoid
SVM	Kernel	Radial Basis Function
	Regularization Parameter (C)	1
	Gamma	1 (Scale)

Table 6. The Optimized Parameters of the Proposed Framework

Parameter	Range	Default	Random	Grid
Iteration		None	100	100
Folds		NA	10	10
Maximum Depth	1-50	Undetermined	41	17
# Estimators	1-200	100	78	79
Minimum leaf	1-20	1	1	2
Minimum Split	2-20	2	8	5

Table 7. Performance Results of the Proposed Framework

	Model	Accuracy	Precision	Recall	F1-Score
Percentage Split 80/20	NN	0.86	0.75	0.90	0.82
	LR	0.73	0.76	0.73	0.74

ISSN 2723-6471

	SVM	0.83	0.84	0.82	0.83
	RF	0.87	0.87	0.87	0.87
	RF+RSCV	0.88	0.87	0.87	0.87
	RF+GSCV	0.90	0.91	0.90	0.90
	NN	0.89	0.88	0.91	0.90
	LR	0.84	0.82	0.88	0.85
Cross Validation	SVM	0.91	0.89	0.94	0.92
	RF	0.88	0.88	0.88	0.88
	RF+RSCV	0.93	0.94	0.93	0.93
	RF+GSCV	0.99	0.99	0.99	0.99

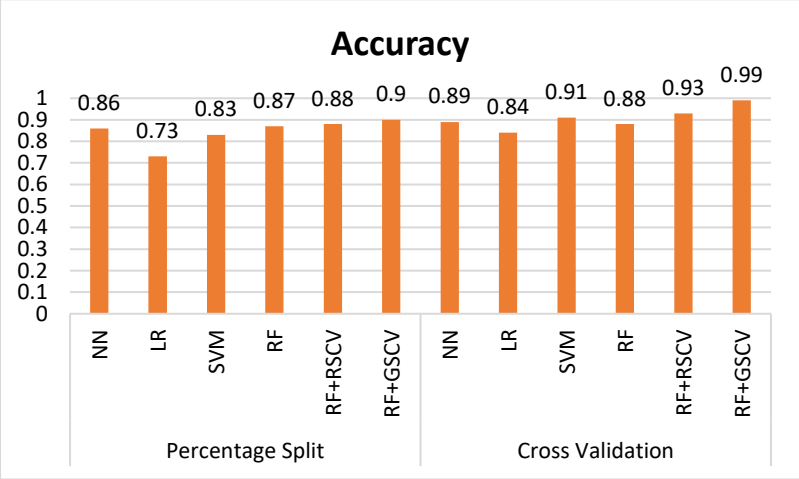

Figure 11. Accuracy of the Proposed Framework

The Random Forest model, particularly when optimized using GSCV, achieved the highest accuracy of 99% in cross-validation, surpassing performance compared to LR, SVM and NN. The significance of predictors such as BMI and glucose levels in forecasting diabetes risk was underscored by the interpretation of the results. Acknowledging limitations such as dataset size and inherent biases suggests caution when generalizing findings. Compared to the best results reported in the literature, the proposed work outperformed the results of Naz and Ahuja [13], which was 98.07 using DL, as reported in table 1.

5. Conclusion

In conclusion, this paper proposes a framework for diabetes prediction. The framework comprises preprocessing, oversampling, classification and parameter tuning methods. The results implemented on the PIMA dataset demonstrate the potential of the machine learning algorithms, notably RF with Grid Search CV, in enhancing the accuracy of predicting diabetes onset. The robust performance of these model highlights their utility in early detection and proactive management of diabetes. However, the study's limitations, including dataset constraints and inherent biases, warrant careful consideration in applying these findings broadly. Future research should focus on expanding datasets and refining modeling techniques to improve predictive accuracy and ensure broader applicability in clinical settings. The framework will also be enriched using other imputation methods, such as KNN and the mean value, to fill in the missing values. Besides, advanced preprocessing steps will be evaluated, including feature engineering, such as interaction terms and polynomial features. Finally, regularization techniques such as L1 and L2 will be tested and evaluated accordingly.

6. Declarations

6.1. Author Contributions

Conceptualization: A.A.A.-S., H.Q., and A.A.-K.; Methodology: H.Q. and A.A.-K.; Software: A.A.A.-S.; Validation: A.A.A.-S., H.Q., and A.A.-K.; Formal Analysis: A.A.A.-S., H.Q., and A.A.-K.; Investigation: A.A.A.-S.; Resources: A.A.-K.; Data Curation: A.A.-K.; Writing Original Draft Preparation: A.A.A.-S., H.Q., and A.A.-K.; Writing Review and Editing: H.Q., A.A.A.-S., and A.A.-K.; Visualization: A.A.A.-S.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] L. Z. Chee, S. Sivakumar, K. H. Lim, and A. A. Gopalai, "Gait acceleration-based diabetes detection using hybrid deep learning," *Biomedical Signal Processing and Control*, vol. 92, no. 105998, pp. 1-8, 2024.

[2] L. Xie, "Pima Indian Diabetes Database and Machine Learning Models for Diabetes Prediction," *Highlights in Science, Engineering and Technology*, vol. 88, no.1, pp. 97-103, 2024.

[3] S. Ramesh, H. Balaji, N.Ch.S.N Iyengar, and R. D. Caytiles, "Optimal predictive analytics of pima diabetics using deep learning", *International Journal of Database Theory and Application*, vol. 10, no. 9, pp. 47-62, 2017.

[4] H. King, R. E. Aubert, and W. H. Herman, "Global burden of diabetes, 1995–2025: prevalence, numerical estimates, and projections," *Diabetes care*, vol. 21, no. 9, pp. 1414-1431, 1998.

[5] T. A. Ojurongbe, H. A. Afolabi, A. Oyekale, K. A. Bashiru, O. Ayelagbe, O. Ojurongbe, S. A. Abbasi, and N. A. Adegoke, "Predictive model for early detection of type 2 diabetes using patients' clinical symptoms, demographic features, and knowledge of diabetes," *Health Science Reports*, vol. 7, no. 1, pp. 1-16, 2024.

[6] N. Nagarjuna and H. Lakshmi, "Predictive Modeling of Diabetes Mellitus Utilizing Machine Learning Techniques," *CVR Journal of Science and Technology*, vol. 26, no. 1, pp. 112-117, 2024.

[7] E. Barbierato and A. Gatti, "The challenges of machine learning: A critical review," *Electronics*, vol. 13, no. 2, pp. 1-30, 2024.

[8] N. G. Ramadhan, W. Maharani, and A. A. Gozali, "Chronic Diseases Prediction Using Machine Learning With Data Preprocessing Handling: A Critical Review," *IEEE Access*, vol. 12, no. 1, pp. 80698-80730, 2024.

[9] N. Nipa, M. H. Riyad, S. Satu, K. C. Howlader, and M. A. Moni, "Clinically adaptable machine learning model to identify early appreciable features of diabetes in Bangladesh," *Intelligent Medicine*, vol. 4, no. 1, pp. 22-32, 2024.

[10] P. Talari, N. Bharathiraja, G. Kaur, H. Alshahrani, M. S. Al Reshan, A. Sulaiman, A. Shaikh, "Hybrid feature selection and classification technique for early prediction and severity of diabetes type 2," *PLOS ONE*, vol. 19, no. 1, pp. 1-17, 2024.

[11] E. Dogantekin, A. Dogantekin, D. Avci, and L. Avci, "An intelligent diagnosis system for diabetes on linear discriminant analysis and adaptive network based fuzzy inference system: LDA-ANFIS," *Digital Signal Processing*, vol. 20, no. 4, pp.

1248-1255, 2010.

[12] M. H. Zangooei, J. Habibi, and R. Alizadehsani, "Disease Diagnosis with a hybrid method SVR using NSGA-II," *Neurocomputing*, vol. 136, no. 1, pp. 14-29, 2014.

[13] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," *Journal of Diabetes & Metabolic Disorders*, vol. 19, no.1, pp. 391-403, 2020.

[14] E. Guldogan, Z. Tunc, A. Acet, and C. Colak, "Performance evaluation of different artificial neural network models in the classification of type 2 diabetes mellitus," *The Journal of Cognitive Systems*, vol. 5, no. 1, pp. 23-32, 2020.

[15] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Express*, vol. 7, no. 4, pp. 432-439, 2021.

[16] R. Saxena, S. K. Sharma, M. Gupta, and G. Sampada, "A novel approach for feature selection and classification of diabetes mellitus: machine learning methods," C*omputational Intelligence and Neuroscience*, vol. 2022, no. 3820360, pp. 1-11, 2022.

[17] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Computing and Applications*, vol. 35, no. 22, pp. 16157-16173, 2023.

[18] M. S. Reza, R. Amin, R. Yasmin, W. Kulsum, and S. Ruhi, "Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data," *Heliyon*, vol. 10, no. 2, pp. 1-13, 2024.

[19] M. J. Tarokh, "Type 2 Diabetes Prediction Using Machine Learning Algorithms," *Jorjani Biomedicine Journal*, vol. 8, no. 3, pp. 4-18, 2020.

[20] Y. Wu et al., "Machine learning for predicting the 3-year risk of incident diabetes in Chinese adults," *Frontiers in Public Health*, vol. 9, no. 626331, pp. 1-12, 2021.