Ingénierie des Systèmes d'Information
Vol. 31, No. 1, January, 2026, pp. 193-203
Journal homepage: http://iieta.org/journals/isi

A Feasibility-Constrained Treble Opposite Algorithm for Missing Data Imputation in Clinical Diabetes Prediction

Mohamed Boussalem*, Hichem Haouassi, Abdelaali Bekhouche, Nabil Azizi

ICOSI Laboratory, Abbes Laghrour University, Khenchela 40004, Algeria

Corresponding Author Email: boussalem_mohamed@univ-khenchela.dz

https://doi.org/10.18280/isi.310118

Received: 13 September 2025
Revised: 15 November 2025
Accepted: 19 January 2026
Available online: 31 January 2026

Keywords:
diabetes diagnosis, missing data imputation, metaheuristic optimization, treble opposite algorithm, swarm intelligence, medical data analysis, machine learning

ABSTRACT

Diabetes is a prevalent chronic disease that poses a significant global health challenge. Machine learning techniques have demonstrated strong potential for early diabetes diagnosis; however, their performance is often compromised by missing values in clinical datasets. Missing data can distort statistical relationships, introduce bias, and reduce the reliability of predictive models. Therefore, developing robust imputation techniques is essential for improving the quality of medical data used in machine learning applications. This study proposes a novel missing data imputation framework based on the Treble Opposite Algorithm (TOA), a swarm-based metaheuristic optimization method designed to enhance the quality of incomplete clinical datasets. Two feasibility-constrained variants are introduced: General Feasibility-Constrained TOA (GFC-TOA) and Class-Specific Feasibility-Constrained TOA (CSFC-TOA). The first applies global feature constraints to ensure biologically plausible imputations, while the second incorporates class-specific constraints that preserve the statistical characteristics of diabetic and non-diabetic samples. The proposed methods were evaluated using multiple machine learning classifiers, including Support Vector Machine, Random Forest, Decision Tree, Naïve Bayes, and K-Nearest Neighbors, on the Pima Indians Diabetes Dataset. Experimental results demonstrate that the CSFC-TOA imputation strategy consistently outperforms both classical imputation methods and several state-of-the-art approaches, significantly improving classification accuracy, recall, precision, and F1-score. These findings indicate that incorporating feasibility constraints into metaheuristic optimization provides an effective strategy for missing data imputation and enhances the reliability of machine learning models for clinical diabetes prediction.

1. INTRODUCTION

Diabetes is one of the most widespread chronic diseases worldwide. Its etiology is multifactorial, involving a combination of genetic predisposition, sedentary lifestyle, poor dietary habits, and other contributing factors of environmental, metabolic, or physiological origin [1]. Diabetes is a chronic condition causing uncontrolled blood glucose (hyperglycemia), which can lead to serious complications such as heart disease, stroke, kidney failure, and nerve damage [2]. There are two main types of diabetes: Type 1, is characterized by the body's inability to produce insulin and accounts for about 10% of all diagnosed cases; and Type 2, occurs when the body either does not produce enough insulin or cannot use it effectively, representing approximately 90% of cases worldwide [3].

According to recent data from the International Diabetes Federation (IDF) in April 2025 [4], approximately 589 million individuals aged between 20 and 79 are currently affected by diabetes across the globe. This number is projected to rise to 853 million by 2050, with over 3.4 million deaths attributed to diabetes each year. Moreover, one in every eight adults faces a high risk of developing Type 2 diabetes, while an estimated 1.8 million children and adolescents under 20 suffer from Type 1 diabetes. In light of the alarming statistics and the significant impact of diabetes on both individual health and public health systems, early detection has become a critical necessity to enable timely diagnosis and effective treatment.

In this context, machine learning (ML) has emerged as a powerful tool for the early prediction and diagnosis of diabetes, thanks to its ability to learn complex patterns from multidimensional medical data. Numerous studies have explored ML techniques for diabetes prediction. For instance, early efforts compared neural networks with traditional methods such as logistic regression and linear perceptron [5], while others evaluated a variety of classifiers including SVM, KNN, Decision Trees, Naïve Bayes, and Random Forests [6, 7]. Notably, Abdulhadi and Al-Mousa [8] combined logistic regression with feature selection techniques to achieve up to 94.25% accuracy using national Health and nutrition examination survey data.

Despite these advances, the reliability of ML models in medical applications remains heavily dependent on the quality and completeness of input data [9]. In real-world clinical datasets, missing values are a pervasive issue. They often result from incomplete medical tests, human error during data

entry, or loss of information during data transfer [10]. These missing values can severely affect the performance of ML models by introducing bias, reducing statistical power, and distorting learned patterns [11]. Addressing this issue is therefore a critical step in the data preprocessing pipeline.

A wide range of imputation strategies have been developed to handle missing data. Simple techniques, such as replacing missing values with the mean or median, are computationally efficient but may over simplify the data structure and ignore feature correlations [12]. More advanced methods, including those based on machine learning and optimization, attempt to estimate missing values more accurately by exploiting underlying data relationships [13]. However, these methods often suffer from high computational complexity or sensitivity to noise, limiting their effectiveness in robust clinical prediction scenarios.

While various imputation techniques exist, many are either too simplistic to capture the complexity of medical data or too sensitive to noise and outliers. Moreover, few approaches are designed with optimization under feasibility constraints—a necessary condition in clinical datasets where imputed values must remain within physiologically valid ranges. Furthermore, metaheuristic algorithms have shown promise in solving complex optimization problems, yet their potential remains under explored in the specific context of robust imputation for diabetes prediction.

To bridge this gap, this paper proposes a novel missing data imputation method based on the Treble Opposite Algorithm (TOA), a recent swarm-based metaheuristic. We develop two feasibility-constrained variants GFC-TOA and CSFC-TOA, which are designed to generate realistic imputations while preserving the data structure. The effectiveness of these methods is evaluated using five machine learning classifiers on the Pima Indians Diabetes Dataset (PIDD). The results demonstrate that our proposed CSFC-TOA approach significantly outperforms both classical and state-of-the-art imputation methods, enhancing the accuracy and robustness of diabetes prediction models.

The remainder of this paper is organized as follows: Section 2 presents the related work on missing data imputation approaches for diabetes detection. Section 3 details the proposed TOA-based imputation method along with its two variants. Experimental results are reported in Section 4. In Section 5, we compare the performance of our two proposed approaches, followed by a comparative analysis with classical and state-of-the-art imputation techniques. Classifier performance is also evaluated and discussed. Finally, Section 6 concludes the paper and outlines directions for future research.

2. RELATED WORKS

Data preprocessing is crucial for improving the performance and reliability of diabetes prediction models. Among its critical tasks, handling missing data remains one of the most persistent challenges. Various techniques have been introduced in the literature to address this issue. Studies such as [9, 10, 14] impute missing entries by replacing them with the mean value of the corresponding feature within its respective class. This technique is intended to preserve the dataset's size, which is especially important for small samples; however, it may negatively affect the model's prediction performance. Another commonly adopted strategy is to remove records containing missing values [15, 16], which ensures working only with observed data but results in a smaller dataset. A simpler method involves substituting missing value with zeros [17], but this often distorts data distributions and undermines model performance. Although these conventional methods occasionally yield acceptable results, their inherent limitations have driven the exploration of more sophisticated approaches.

Two new data imputation algorithms are proposed to handle missing data issues [13]. The first, Random Forest with Mean (RFM) employs RF model to estimate missing values by training on complete records. During the imputation process, missing entries in non-target columns are temporarily replaced with their global mean values, after which the RF predictions substitute the missing values. The second method, Random Forest with Class' Mean (RFCM), adopts a similar procedure but initializes missing entries using class-specific means (diabetic or non-diabetic) instead of global means. Both imputation strategies were evaluated on the Pima Indian Diabetes dataset using several classifiers, including Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), and Decision Tree (DT). The highest accuracy was achieved with RF under the RFCM approach (87.53%), followed by DT with 84.07%.

While statistical and clustering-based approaches have shown moderate success, recent studies have explored metaheuristic strategies for more adaptive and data-driven imputation. The Salp Swarm Algorithm (SSA) is utilized as a metaheuristic approach for imputing missing values in the PIMA diabetes dataset, aiming to enhance diabetes prediction [18]. The imputation quality is evaluated using three classification models: SVM, KNN, and NB. Experimental findings indicate that KNN yields the highest average classification accuracy at 79.06%, while NB registers the lowest at 73.35%. SVM achieves a moderate performance with an accuracy of 79.42%. Overall, the results demonstrate that the ISSA surpasses conventional imputation techniques, such as mean substitution and zero-filling, in terms of predictive performance.

The Improved Grey Wolf Optimizer (GWO) is employed as a metaheuristic strategy to imputing missing values in the PIMA diabetes dataset [19], with the objective of enhancing classification performance. Both cross-validation and a data split of 70% for training and 30% for testing are employed for evaluation. Imputation effectiveness is assessed with three classifiers: SVM, KNN, and NB. Experimental results indicate that SVM achieves the highest accuracy at 79.53%, followed by KNN with 79.06%, while NB obtains the lowest performance with 74.28%. These findings highlight the superiority of the IGWO approach over traditional methods.

The authors [20] aim to enhance the reliability of diabetes prediction by first identifying factors correlated with the disease through a correlation analysis with the target variable, using the Pima Indian Diabetes dataset. Subsequently, various imputation techniques are applied to handle missing values, including CART (Classification and Regression Tree), GMM (Gaussian Mixture Models), and RFR (Random Forest Regressor). The impact of these imputation methods on predictive performance is evaluated across several multiple classifiers: DT, SVM, KNN, NB, and RF. Further improvements are achieved by optimizing model hyperparameters via grid search. The best performance is obtained using GMM-based imputation, where the RF classifier achieves the highest accuracy of 90.80%, followed

by DT with an accuracy of 86.93%.

In contrast to complex model-based imputation, some studies have focused on assessing classifier performance after applying simpler techniques such as KNN imputation. Leverages the PID dataset to assess the effectiveness of ten classification algorithms in detecting diabetes [21]. Among the evaluated models are LR, DT, RF, SVM, NB. The study focuses on comparing these classifiers after applying KNN for imputing missing values. Logistic Regression demonstrates superior performance, achieving the highest precision (72.9%), accuracy (77.73%), and F1-score (65.04%), highlighting its robustness in this predictive context.

Beyond graph-based methods, nature-inspired metaheuristic algorithms have also attracted attention for their potential in missing data imputation. Swarm-based metaheuristic algorithms have demonstrated remarkable effectiveness across various domains, including feature selection in sentiment analysis [22], word sense disambiguation [23], among others. Their success is largely driven by inherent properties such as self-learning, flexibility, adaptability, and versatility, which make bio-inspired and swarm-based approaches particularly powerful for solving complex optimization problems [24]. In the context of diabetes prediction, however, their application to missing data imputation remains limited. To the best of our knowledge, only two swarm-based imputation methods, ISSA [18] and IGWO [19] have been proposed for this purpose, as discussed earlier.

Building on the demonstrated potential of swarm-based metaheuristics and addressing the limited research on their use for missing data imputation in diabetes prediction, this study proposes a novel swarm-based metaheuristic framework designed specifically for this task. The approach aims to overcome the shortcomings of existing methods—such as limited adaptability to data patterns and suboptimal handling of complex feature relationships—while leveraging the self-learning and flexibility inherent in bio-inspired optimization. The following section presents the original optimization algorithm and details its customized adaptation for robust and efficient missing value imputation.

3. PROPOSED APPROACH

In this study, we present a novel missing-data imputation framework built upon the recently developed TOA [25], a metaheuristic optimization method designed to balance exploration and exploitation through a triple-opposite learning strategy. By leveraging TOA's adaptive search capabilities, our approach aims to generate accurate and robust estimates for missing values, thereby preserving data integrity and enhancing downstream prediction performance.

Figure 1 illustrates the overall pipeline of our proposed approach and the integration of the TOA-based imputation within the machine learning workflow for diabetes prediction. After loading the dataset, a structured pre-processing workflow was employed to enhance data quality and ensure compatibility with the learning algorithms. This process begins with the detection and precise localization of missing values (NaN), a crucial step that enables targeted and accurate imputation rather than applying unnecessary transformations to complete records.

The imputation procedure, which constitutes the core contribution of this study, is described in detail in Section 3.2. To ensure that all features contribute equally during model training, Min-Max scaling is applied to normalize all variables within the [0, 1] range, as show in Eq. (1).

$$Normalised_{value} = \frac{Original_{value} + Min}{Max + Min} \quad (1)$$

where, *Min* and *Max* denote, respectively, the lowest and highest observed values for the given feature.

Following normalization, the dataset was partitioned using 5-fold cross-validation, in which the data is split into five equal folds. In each iteration, four folds are used for training and one fold for testing, ensuring that every instance is used for both training and evaluation. This approach provides a robust and unbiased assessment of the predictive performance of the developed models.

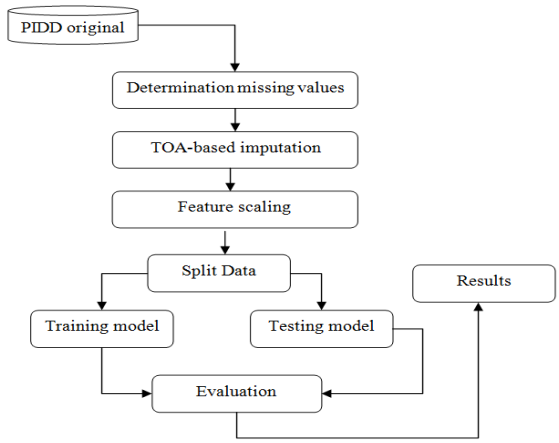

Figure 1. Diabetes prediction with TOA imputation
Note: TOA = Treble Opposite Algorithm.

3.1 Original Treble Opposite Algorithm

The TOA is a new metaheuristic optimization algorithm proposed in 2024 [25]. It belongs to the family of swarm-based metaheuristics, meaning it operates with a population of candidate solutions, as represented in Eq. (2), which evolve collectively to find the global optimum.

$$X = \{x_1, x_2, x_3, \dots, x_n\} \quad (2)$$

TOA begins by randomly initializing a population of solutions uniformly across the search space to ensure unbiased and well-distributed exploration. Each solution is defined within lower ($x_{l,j}$) and upper ($x_{u,j}$) bounds for dimension j, and generated using a uniform random number as in Eq. (3).

$$x_{i,j} = U(x_{l,j}, x_{u,j}) \quad (3)$$

The algorithm proceeds through three sequential search phases, each generating new candidate solutions and selecting the best among them.

Guided Search Toward Best Solution

This phase guides the search using the current global best solution x_b. Two candidate solutions are generated using Eq. (4) and Eq. (5): one moving toward x_b and the other away from it. The better of the two, x_{c11} or x_{c12} is selected as x_{c1}, as defined in Eq. (6).

$$x_{c11,j} = x_{i,j} + U(0,1).\left(x_{b,j} - 2x_{i,j}\right) \quad (4)$$

$$x_{c12,j} = x_{i,j} + U(0,1).\left(x_{i,j} - 2x_{b,j}\right) \quad (5)$$

$$x_{c1} = \begin{cases} x_{c11}, f(x_{c11}) < f(x_{c12}) \\ x_{c12}, else \end{cases} \quad (6)$$

Guided Search Using Midpoint

Two solutions (x_{s1} and x_{s2}) are randomly selected from the population (Eq. (7)), and their midpoint xt (Eq. (8)) serves as the reference. Two candidates are then generated: one moving toward x_t (Eq. (9)) and the other away from it (Eq. (10)). The better of the two, x_{c21} or x_{c22}, is selected as x_{c2} (Eq. (11)).

$$x_{s1}, x_{s2} = U(X) \quad (7)$$

$$x_{i,j} = \frac{x_{s1,j} + x_{s2,j}}{2} \quad (8)$$

$$x_{c21,j} = x_{i,j} + U(0,1).\left(x_{t,j} - 2x_{i,j}\right) \quad (9)$$

$$x_{c22,j} = x_{i,j} + U(0,1).\left(x_{i,j} - 2x_{t,j}\right) \quad (10)$$

$$x_{c2} = \begin{cases} x_{c21}, f(x_{c21}) < f(x_{c22}) \\ x_{c22}, else \end{cases} \quad (11)$$

Purely Random Search

This phase introduces random, directionless perturbations to maintain diversity and avoid premature convergence. Two random candidates, x_{c31} and x_{c32}, are generated using Eq. (12) and Eq. (13), respectively, and the better one is selected as x_{c3} (Eq. (14)).

$$x_{c31,j} = x_{i,j} + 0.1.U(-1,1).\left(1 - \frac{t}{t_m}\right).\left(x_{u,j} - x_{l,j}\right) \quad (12)$$

$$x_{c32,j} = x_{i,j} + U(-1,1).\left(1 - \frac{t}{t_m}\right).\left(x_{u,j} - x_{l,j}\right) \quad (13)$$

$$x_{c3} = \begin{cases} x_{c31}, f(x_{c31}) < f(x_{c32}) \\ x_{c32}, else \end{cases} \quad (14)$$

After each phase, the current solution is updated if the new candidate performs better (Eq. (15)), and the global best x_b is replaced if an even better solution is found (Eq. (16)). Solution quality is evaluated using the objective function f(x) where lower values indicate better solutions, as TOA is designed for minimization problems.

$$x_i' = \begin{cases} x_c, f(x_c) < f(x_i) \\ x_i, else \end{cases} \quad (15)$$

$$x_b' = \begin{cases} x_i, f(x_i) < f(x_b) \\ x_b, else \end{cases} \quad (16)$$

In TOA, the objective function f(x) measures the quality of a solution, with lower values indicating better performance, as the algorithm is primarily designed for minimization problems.

3.2 TOA-based imputation

We propose an adaptation of TOA to effectively address the challenge of missing data imputation in diabetes detection.

In this formulation, the initial population X, as defined in Eq. (17), consists of *n* candidate imputation vectors x_i, where each x_i each vector x_i represents a complete set of replacement values for the missing entries in the dataset.

$$X = \{x_1, x_2, x_3, \dots, x_n\} \quad (17)$$

The population is initialized using Eq. (18), where $x_{l,j}$ and $x_{u,j}$ denote the lower and upper bounds of the j^{th} dimension of x_i, as specified in the next sections.

$$x_{i,j} = U\left(x_{l,j}, x_{u,j}\right) \quad (18)$$

Here, the dimensionality of each candidate solution corresponds to the total number of missing values in the dataset.

The optimization objective is to impute these missing values so as to maximize the predictive performance of a classification model. Therefore, we redefine the fitness function in Eq. (19) as the classification accuracy obtained after imputing the dataset with a candidate solution x_i.

$$f(x_i) = Accuracy(x_i) \quad (19)$$

Unlike the original TOA, which is inherently designed for minimization problems, our adaptation is formulated for maximization. To accommodate this change, we reformulate the core update rules and selection criteria (originally given in Eqs. (6), (11), (14), (15), and (16)) into their maximization-oriented counterparts, shown in Eqs. (20)-(24). These modifications direct the search toward imputations that yield the highest classification accuracy.

$$x_{c1} = \begin{cases} x_{c11}, f(x_{c11}) > f(x_{c12}) \\ x_{c12}, else \end{cases} \quad (20)$$

$$x_{c2} = \begin{cases} x_{c21}, f(x_{c21}) > f(x_{c22}) \\ x_{c22}, else \end{cases} \quad (21)$$

$$x_{c3} = \begin{cases} x_{c31}, f(x_{c31}) > f(x_{c32}) \\ x_{c32}, else \end{cases} \quad (22)$$

$$x_i' = \begin{cases} x_c, f(x_c) > f(x_i) \\ x_i, else \end{cases} \quad (23)$$

$$x_b' = \begin{cases} x_i, f(x_i) > f(x_b) \\ x_b, else \end{cases} \quad (24)$$

Constraint Feasibility

To maintain the feasibility of candidate solutions, we apply the feasibility check defined in Eq. (25). For each element $x_{i,j}$, the algorithm ensures that it lies within the permissible range of its corresponding feature. If a value violates these bounds, it is replaced by a random value drawn from the valid interval. The pseudo-code of the TOA-based missing data imputation method is presented in Algorithm 1.

$$x_{i,j} = \begin{cases} x_{i,j}, x_{l,j} < x_{i,j} < x_{u,j} \\ U\left(x_{l,j}, x_{u,j}\right) else \end{cases} \quad (25)$$

We consider two types of feasibility constraints:

i. General Feasibility Constraint

The minimum and maximum values for each feature are

extracted from the entire dataset, ensuring that imputations remain within biologically and statistically plausible ranges.

ii. Class-Specific Feasibility Constraint

Feature bounds are computed per class (diabetic vs. non-diabetic), enabling the generation of imputations that better preserve class-specific distributions.

Accordingly, we evaluate two variants of the proposed method:

- TOA-based imputation under General Feasibility Constraints (GFC TOA imputation)
- TOA-based imputation under Class-Specific Feasibility Constraints (CSFC TOA imputation).

The pseudo-code of the proposed TOA-based missing data imputation approach is provided in Algorithm 1, and the comparative effects of GFC and CSFC on model performance are discussed in Section 4.

Algorithm 1. Treble Opposite Algorithm for missing data imputation
Input:
Dataset of diabetes PIDD
Number of particles (species) (NP)
Maximum number of iterations (T)
Output:
X_{best}: Best vector imputation.
Begin
for all x in X
Initialize $x_{i,j}$ using Eq. (18)
Update $x_{\mathbf{b}}$ using Eq. (24)
end for
for t=1:tm **do**
for all x in X
Perform 1st guided search using Eqs. (4), (5) and (20)
Feasibility verification using Eq. (25)
Update $\mathbf{x_i}$ using Eq. (23) and $\mathbf{x_b}$ using Eq. (24)
Perform 2end guided search using Eqs. (7) to (10) and (21)
Feasibility verification using Eq. (25)
Update $\mathbf{x_i}$ using Eq. (23) and $\mathbf{x_b}$ using Eq. (24)
Perform random search using (12), (13) and (22)
Feasibility verification using Eq. (25)
Update $\mathbf{x_i}$ using Eq. (23) and $\mathbf{x_b}$ using Eq. (24)
end for
end for
end

4. EXPERIMENTS AND RESULTS

This section details the experimental setup designed to assess the effectiveness of the proposed TOA-based imputation approaches. It begins with a description of the dataset employed in the study, followed by the evaluation metrics used to quantify both imputation and classification performance. The dataset is partitioned into validation, and test subsets, where the validation set is exclusively used for fitness evaluation during the imputation optimization process, while the test set remains completely unseen and is only used for the final assessment of classification performance. The parameter settings of the TOA are then presented. Finally, we report and analyze the results obtained with the proposed methods, comparing them against both conventional imputation techniques and recent state-of-the-art approaches.

4.1 PID dataset

This study employs the Pima Indians Diabetes Dataset (PIDD), a widely recognized benchmark in diabetes prediction research. The dataset is publicly accessible on Kaggle at: https://www.kaggle.com/uciml/pimaindians-diabetes-database.

The dataset comprises 768 patient records, with 500 non-diabetic and 268 diabetic cases. Each record is described by eight clinical attributes (Features). Capturing physiological and medical measurements relevant to diabetes risk assessment. A detailed description of these features is provided in Table 1.

Table 1. Dataset feature descriptions

Feature	Description
Pregnancies	Total number of times the patient has been pregnant
Glucose	Plasma glucose level (mg/dL) measured 2 hours after an oral glucose tolerance test
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skinfold thickness (mm), representing subcutaneous fat
Insulin	Serum insulin level (mu U/ml) measured 2 hours after glucose intake
BMI	Body Mass Index, calculated as weight (kg) divided by height squared (m^2)
DiabetesPedigreeFunction	Risk score for diabetes based on family history
Age	Patient's age in years
Outcome	Class label: 1 indicates diabetic, 0 indicates non-diabetic

Table 2 presents the statistical analysis of missing values and their distribution across the dataset. In total, 652 missing entries were identified, affecting five out of the eight clinical features. The highest proportions of missing data occur in Insulin (57.37%) and SkinThickness (34.81%), followed by BloodPressure, BMI, and Glucose. This uneven distribution underscores the critical role of the imputation phase in the preprocessing pipeline, as addressing these gaps effectively is essential to prevent bias, preserve statistical integrity, and enhance the robustness of subsequent diabetes prediction models.

Table 2. Missing values statistics in the dataset

Feature	Number of Missing Values	Rate (%)
Glucose	5	0.77
BloodPressure	35	5.36
SkinThickness	227	34.81
Insulin	374	57.37
BMI	11	1.69
Total Missing Values	652	100

4.2 Machine learning models used

To assess the effectiveness of the proposed imputation strategies, five widely recognized machine learning classifiers were employed. These models were chosen for their proven performance in medical diagnosis tasks, their diversity in

learning paradigms, and their ability to handle both linear and non-linear decision boundaries. The selected models include: Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF) and K-Nearest Neighbors (KNN). This variety ensures a comprehensive evaluation by covering probabilistic, discriminative, and ensemble-based learning approaches, allowing for a robust comparison of model performance after imputation.

4.3 Performance measure

In this study, model performance is evaluated using four widely adopted metrics: accuracy, precision, recall, and F1-score. These metrics are derived from the confusion matrix, a useful tool that summarizes how well a classification model performs. The confusion matrix in Table 3 compares predicted outcomes with actual results using four fundamental components: True Positives (TP), False Negatives (FN), True Negatives (TN), and False Positives (FP). Together, these measures provide a comprehensive assessment of the model's ability to correctly classify both positive and negative cases.

Table 3. Confusion matrix

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Accuracy

Accuracy measures the overall effectiveness of a model in correctly classifying instances. It is calculated as the ratio of correctly predicted observations to the total number of observations. In other words, it reflects how closely the model's predictions align with the true outcomes. The formula used to compute accuracy is provided in Eq. (26).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{26}$$

Precision

Quantifies the proportion of correctly predicted positive instances among all instances that were predicted as positive. In other words, it reflects the model's ability to avoid false positives. Precision is calculated using Eq. (27).

$$Precision = \frac{TP}{TP + FP} \tag{27}$$

Recall

Assesses the model's ability to correctly identify all relevant instances of the positive class. It is defined as the ratio of true positive predictions to the total number of actual positive cases. Recall is computed using Eq. (28).

$$Precision = \frac{TP}{TP + FN} \tag{28}$$

4.4 Results

To evaluate the proposed imputation strategies, we applied both General Feasibility Constraints (GFC-TOA) and Class-Specific Feasibility Constraints (CSFC-TOA) within the TOA framework. In the GFC-TOA approach, feature-level boundary values were extracted from the dataset and incorporated into Eq. (25) to ensure that generated imputations remained within valid ranges. Similarly, CSFC-TOA was implemented by extracting class-dependent feature boundaries for diabetic and non-diabetic classes, integrating these into the same equation to generate imputations that better reflect class-specific distributions.

The extracted constraint values, algorithmic behavior, and classification performances are presented and analyzed in the following subsections.

4.4.1 General feasibility constraints

Table 4 lists the GFC values derived from the PIDD dataset, focusing on features containing missing values. For each feature, the minimum and maximum observed values define the allowable range for generated imputations. Only solutions satisfying these constraints are considered valid.

Table 4. General feasibility constrains values

Feature	**Min**	**Max**
Glucose	44	199
BloodPressure	24	122
SkinThickness	7	99
Insulin	14	846
BMI	18.20	67.10

4.4.2 Class-specific feasibility constraints

Table 5 presents the class-specific constraint values. By analyzing the minimum and maximum values of each feature within diabetic and non-diabetic groups in PIMA dataset, we observe noticeable differences between the two classes. These differences are leveraged to guide the generation of feasible solutions according to the specific class (diabetic or non-diabetic).

Table 5. Class-specific feasibility constraints values

Feature	**Non-diabetic**		**Diabetic**	
	Min	**Max**	**Min**	**Max**
Glucose	44.0	197.0	78.0	199.0
BloodPressure	24.0	122.0	30.0	114.0
SkinThickness	7.0	60.0	7.0	99.0
Insulin	15.0	744.0	14.0	846.0
BMI	18.2	57.3	22.9	67.1

4.4.3 Algorithm's behaviour

The TOA algorithm is governed by two key parameters: the population size (NP, number of agents) and the number of iterations (T). To determine the optimal combination of NP and T and define the stopping criteria, experiments were conducted with population sizes of 10, 20, 30, 50, and 100 agents, each evaluated over 100 iterations. For each configuration, the fitness function (classification accuracy) was recorded for both GFC-TOA and CSFC-TOA, as shown in Figures 2 and 3.

In the case of GFC-TOA (Figure 2), the best performance was achieved with 100 agents at approximately 41 iterations, after which accuracy plateaued.

In Figure 3, which corresponds to CSFC-TOA, the highest accuracy was achieved with 100 agents at around 47 iterations, followed by stable performance.

Based on this analysis, the parameter configurations and stopping criteria for the two proposed algorithm variants were defined as follows. For GFC-TOA, the population size was set to NP = 100 with a maximum of T = 41 iterations, whereas for

CSFC-TOA, the population size was also fixed at NP = 100 with T = 47 iterations. The number of agents was fixed to 100 in both variants as a trade-off between search diversity and computational efficiency**,** providing sufficient exploration of the solution space while maintaining a reasonable computational cost, as confirmed by preliminary sensitivity analyses. In both cases, the optimization process terminates when the predefined maximum number of iterations is reached.

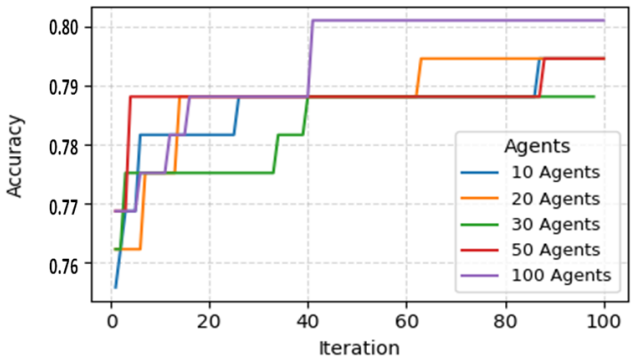

Figure 2. Accuracy vs. Iteration for different size of agents, case of: GFC-TOA based imputation
Note: GFC= General Feasibility Constraints; TOA = Treble Opposite Algorithm.

The two variants differ notably in convergence speed. CSFC-TOA starts with a higher initial accuracy (≈0.81) and converges rapidly within the first 20 iterations, regardless of population size. In contrast, GFC-TOA begins at a lower accuracy (≈0.76) and shows slower, more variable convergence depending on the number of agents.

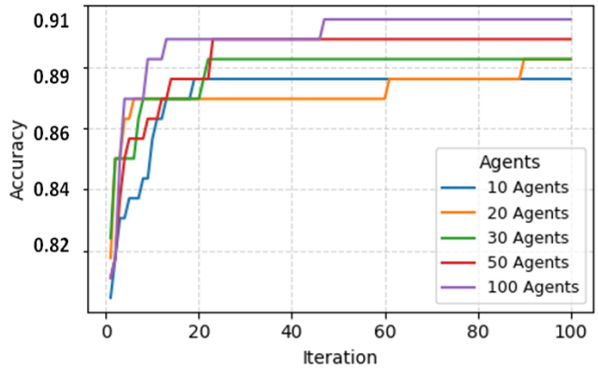

Figure 3. Accuracy vs. Iteration for different size of agents, case of: CSFC-TOA based imputation
Note: GFC= General Feasibility Constraints; TOA = Treble Opposite Algorithm.

4.4.4 Performances of GFC-TOA and CSFC-TOA based imputation

Both imputation variants, GFC-TOA and CSFC-TOA, were evaluated using five classification models. The selected classifiers include Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbors (KNN), chosen for their proven effectiveness in medical diagnosis tasks. Model performance was assessed based on Accuracy, Recall, Precision, F1-score, and AUC, with the results summarized in Table 6.

Table 6. Performances of proposed approaches

Classifier	Imputation	Accuracy	Recall	Precision	F1_score	AUC
DT	GFC-TOA	0.7720	0.6410	0.6840	0.6610	0.7420
	CSFC-TOA	0.8870	0.8140	0.8610	0.8330	0.8700
KNN	GFC-TOA	0.7900	0.6530	0.7220	0.6850	0.8080
	CSFC-TOA	0.8440	0.6940	0.8390	0.7570	0.8900
SVM	GFC-TOA	0.7900	0.5710	0.7760	0.6540	0.8400
	CSFC-TOA	0.8500	0.7540	0.8090	0.7790	0.8970
NB	GFC-TOA	0.7900	0.6790	0.7130	0.6950	0.8310
	CSFC-TOA	0.8220	0.7460	0.7510	0.7460	0.8830
RF	GFC-TOA	0.8030	0.6530	0.7520	0.6980	0.8430
	CSFC-TOA	0.9000	0.7950	0.9070	0.8470	0.9460

Note: KNN = K-Nearest Neighbors; RF = Random Forest; SVM = Support Vector Machine; DT = Decision Tree; NB = Naïve Bayes; GFC= General Feasibility Constraints; TOA = Treble Opposite Algorithm; CSFC= Class-Specific Feasibility Constraints.

5. COMPARATIVE STUDY OF TOA-BASED IMPUTATION PERFORMANCE

This section presents a comprehensive performance comparison of classifiers using our proposed imputation approaches. The evaluation includes intra-comparisons between the two TOA-based variants, as well as comparisons with classical imputation techniques and state-of-the-art methods. Finally, we analyze the relative performance of classifiers when using the CSFC-TOA imputation strategy.

5.1 Comparison of GFC-TOA vs. CSFC-TOA

The comparative analysis between the two imputation variants, GFC-TOA and CSFC-TOA, according to Table 6, demonstrates that CSFC-TOA consistently outperforms GFC-TOA across all classifiers and evaluation metrics. Quantitatively, CSFC-TOA achieves an average improvement of approximately 7.2% in accuracy, 12.1% in recall, 10.4% in precision, 11.4% in F1-score, and 8.4% in AUC over GFC-TOA. The most significant gains are observed for Decision Tree and Random Forest classifiers, where accuracy improvements reach up to 11.5% and 9.7%, respectively, accompanied by substantial increases in F1-score and AUC.

Figure 4 presents these average performance gains across all evaluation metrics, highlighting consistent improvements, particularly in recall and F1-score. These results indicate that incorporating class-specific constraints during the imputation process leads to more discriminative feature representations, reduces the noise caused by missing values, and enhances the predictive performance of machine learning models for diabetes detection. Consequently, CSFC-TOA can be

considered the dominant imputation strategy in this experimental setting.

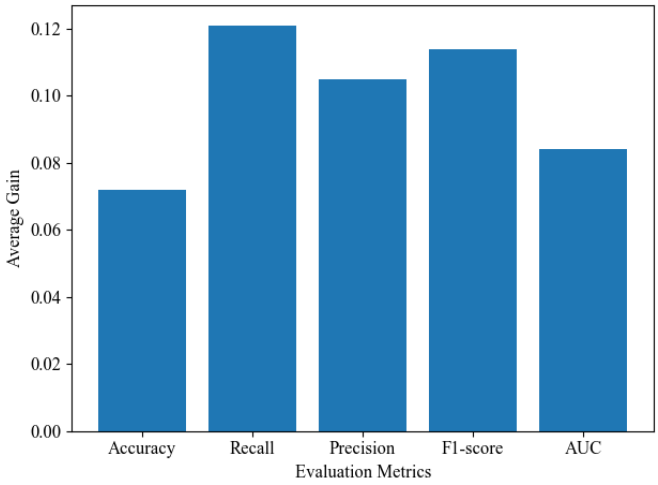

Figure 4. Average metric gains achieved by CSFC-TOA Over GFC-TOA

Note: GFC= General Feasibility Constraints; TOA = Treble Opposite Algorithm; CSFC= Class-Specific Feasibility Constraints.

5.2 Comparison of TOA-based imputation against classical methods

To rigorously evaluate the effectiveness of the proposed imputation strategies, we evaluated the performance of five widely used machine learning classifiers: NB, SVM, DT, RF and KNN, on the PIMA dataset. The dataset was preprocessed using both conventional imputation techniques (zero-filling, mean, median, and mode) and the proposed TOA-based strategies (GFC-TOA and CSFC-TOA). Performance results, obtained using 5-fold cross-validation, are summarized in Table 7 and measured in terms of Accuracy, Recall, Precision, F1-score, and ROC-AUC.

Table 7 compares the performance of classifiers KNN, RF, SVM, DT, and NB using different imputation methods: classical approaches **(**Mean, Median, Mode, Without**)** and the proposed variants GFC-TOA and CSFC-TOA. Classical methods show limited performance, with marginal differences in accuracy, F1-score, and AUC (<2%). The GFC-TOA approach improves results, but CSFC-TOA consistently outperforms all other methods**,** with significant gains, particularly for DT (+11.5% in accuracy) and RF (+9.7%), along with notable increases in F1-score and AUC.

On average, CSFC-TOA improves performance metrics by 7-12% in accuracy, 6-18% in recall, 3-15% in precision, 5-15% in F1-score, and 5-13% in AUC, confirming that incorporating class-specific constraints reduces noise from missing values and produces more discriminative feature representations. These results establish CSFC-TOA as the dominant imputation strategy for diabetes prediction Our two proposed imputation variants outperform classical methods, with CSFC-TOA achieving the highest performance. These results highlight the robustness of optimization-driven, class-aware strategies in improving data quality and enhancing predictive performance across diverse machine learning algorithms when handling missing values.

Table 7. Classification performances using different missing value imputation methods

Classifier	Imputation	Accuracy	Recall	Precision	F1-score	AUC
KNN	Without	0.732	0.556	0.639	0.592	0.691
	Mean	0.747	0.642	0.637	0.639	0.723
	Median	0.733	0.627	0.615	0.621	0.708
	Mode	0.720	0.567	0.607	0.586	0.685
	GFC-TOA	0.7900	0.6530	0.7220	0.6850	0.8080
	CSFC-TOA	0.8440	0.6940	0.8390	0.7570	0.8900
RF	Without	0.767	0.593	0.695	0.640	0.727
	Mean	0.749	0.582	0.661	0.618	0.710
	Median	0.760	0.597	0.680	0.635	0.723
	Mode	0.755	0.589	0.671	0.627	0.717
	GFC-TOA	0.8030	0.6530	0.7520	0.6980	0.8430
	CSFC-TOA	0.9000	0.7950	0.9070	0.8470	0.9460
SVM	Without	0.773	0.545	0.740	0.626	0.721
	Mean	0.777	0.556	0.749	0.635	0.726
	Median	0.773	0.549	0.743	0.628	0.721
	Mode	0.768	0.545	0.732	0.621	0.716
	GFC-TOA	0.7900	0.5710	0.7760	0.6540	0.8400
	CSFC-TOA	0.8500	0.7540	0.8090	0.7790	0.8970
DT	Without	0.700	0.578	0.573	0.575	0.672
	Mean	0.699	0.555	0.572	0.562	0.666
	Median	0.712	0.571	0.597	0.582	0.679
	Mode	0.741	0.608	0.640	0.622	0.710
	GFC-TOA	0.7720	0.6410	0.6840	0.6610	0.7420
	CSFC-TOA	0.8870	0.8140	0.8610	0.8330	0.8700
NB	Without	0.755	0.619	0.662	0.639	0.724
	Mean	0.750	0.597	0.665	0.627	0.714
	Median	0.746	0.593	0.660	0.621	0.711
	Mode	0.742	0.593	0.648	0.617	0.708
	GFC-TOA	0.7900	0.6790	0.7130	0.6950	0.8310
	CSFC-TOA	0.8220	0.7460	0.7510	0.7460	0.8830

Note: KNN = K-Nearest Neighbors; RF = Random Forest; SVM = Support Vector Machine; DT = Decision Tree; NB = Naïve Bayes; GFC= General Feasibility Constraints; TOA = Treble Opposite Algorithm; CSFC= Class-Specific Feasibility Constraints.

5.3 Comparison of TOA based imputation against state-of-the-art approaches

In this section, we compare the performance of our proposed imputation strategies with several state-of-the-art imputation methods reported in the literature. To ensure a scientifically sound and fair comparison, we selected studies that used the same PIMA dataset without modifying its structure through sample deletion, feature selection, or class balancing techniques. This allows us to isolate and rigorously assess the impact of the missing data imputation methods alone. The approaches considered for comparison are detailed in Section 2.

To ensure a fair and scientifically rigorous comparison with state-of-the-art imputation approaches, all methods were evaluated under a unified experimental framework using 5-fold cross-validation on the Pima Indians Diabetes Dataset. No feature selection or alternative datasets were used in any method, and all classifier parameters were kept consistent across experiments. This unified setup allows for confident and unbiased assessment of the relative performance of each method. Table 8 presents the classification accuracy of the state-of-the-art methods and compares them with the accuracy achieved by our proposed approach in its two variants.

A Friedman non-parametric test was conducted to evaluate the effectiveness of different imputation methods using classification accuracy from five classifiers (DT, KNN, NB, RF, and SVM) under the same cross-validation protocol. Classifiers were treated as blocks and imputation methods as treatments. Table 9 shows the average ranks of the imputation methods based on the Friedman test.

The Friedman test revealed statistically significant differences among the compared imputation methods ($p < 0.05$). CSFC-TOA obtained the best overall performance, achieving the lowest (best) average rank of 1.2, followed by GMM with an average rank of 1.8. RFR and RFCM formed a second performance tier with comparable average ranks of 4.2 and 4.25, respectively, while IGWO (4.66) and GFC-TOA (5.4) demonstrated moderate competitiveness. Classical imputation techniques, including CART (5.8), KNN (8.4), MF (8.5), and RFM (9.5), consistently ranked lower, indicating inferior performance across classifiers. Among recent swarm-based approaches, CSFC-TOA clearly outperformed IGWO and ISSA (average rank 7), confirming its superiority in terms of robustness and classification accuracy, while GFC-TOA provided stable yet less competitive improvements.

Overall, these findings confirm the effectiveness of the TOA-based imputation strategies for handling missing data. The consistent top performance of CSFC-TOA across a wide range of classifiers underscores its robustness and positions it as a promising, reliable imputation approach for predictive modeling tasks in medical and clinical applications.

Table 8. Accuracy of several state-of-the-art imputation methods

Classifier	Imputation Method										
	KNN	**MF**	**RFM**	**RFCM**	**CART**	**GMM**	**RFR**	**IGWO**	**ISSA**	**GFC-TOA**	**CSFC-TOA**
DT	0.7173	0.741	0.734	0.841	0.823	0.8693	0.8156	/	/	0.7720	0.8870
KNN	0.7435	/	/	/	0.771	0.8105	0.7908	0.791	0.784	0.7900	0.8440
NB	0.7460	0.676	0.646	0.816	0.778	0.8170	0.8105	0.743	0.733	0.7900	0.8220
RF	0.7357	0.76	0.748	0.875	0.817	0.9080	0.8565	/	/	0.8030	0.9000
SVM	0.7695	0.743	0.722	0.775	0.784	0.8039	0.7923	0.795	0.788	0.7900	0.8500
	[21]	[13]	[13]	[13]	[20]	[20]	[20]	[19]	[18]	Our work	Our work

Note: KNN = K-Nearest Neighbors; RFM= Random Forest with Mean; RFCM= Random Forest with Class' Mean; CART= Classification and Regression Tree; GMM= Gaussian Mixture Models; RFR= Random Forest Regressor; GFC= General Feasibility Constraints; TOA = Treble Opposite Algorithm; CSFC= Class-Specific Feasibility Constraints.

Table 9. Average ranks of the imputation methods based on the Friedman test

Imputation Method	CSFC-TOA	GMM	RFR	RFCM	IGWO	GFC-TOA	CART	ISSA	KNN	MF	RFM
Average Rank	1.2	1.8	4.2	4.25	4.66	5.4	5.8	7	8.4	8.5	9.5

Note: KNN = K-Nearest Neighbors; RFM= Random Forest with Mean; RFCM= Random Forest with Class' Mean; CART= Classification and Regression Tree; GMM= Gaussian Mixture Models; RFR= Random Forest Regressor; GFC= General Feasibility Constraints; TOA = Treble Opposite Algorithm; CSFC= Class-Specific Feasibility Constraints.

The comparative study demonstrates that the proposed CSFC-TOA imputation method consistently outperforms both classical techniques (Mean, Median, Mode, Without) and other state-of-the-art and metaheuristic approaches (GMM, RFR, RFCM, IGWO, ISSA) across multiple classifiers for diabetes prediction. Quantitative improvements are substantial, with CSFC-TOA achieving average gains of 7–12% in accuracy, 6–18% in recall, 3–15% in precision, 5–15% in F1-score, and 5–13% in AUC. Statistical validation using the Friedman test confirms these differences are significant ($p < 0.05$), with CSFC-TOA attaining the best average rank (1.2), followed by GMM (1.8), while classical methods consistently rank lower (8.4–9.5). These results highlight the effectiveness of class-specific constraint-based imputation in reducing noise from missing values, generating more discriminative feature representations, and significantly enhancing predictive performance. CSFC-TOA thus emerges as the dominant and most robust imputation strategy in this experimental setting.

6. CONCLUSION

This study introduced two novel imputation methods for handling missing values in diabetes diagnosis dataset. Both approaches are based on the TOA, a metaheuristic optimization technique designed to efficiently explore the search space. The first variant, GFC-TOA, employs general feasibility constraints to guide the imputation process but may generate values less aligned with class-specific characteristics. The enhanced version, CSFC-TOA, incorporates class-specific feasibility constraints, focusing the search on more relevant regions for each class. This refinement leads to more realistic mutations and improved predictive performance. A key strength of our approach is that it does not rely on any

external information and does not modify the dataset by removing or adding samples, except for imputing the missing values.

A notable advantage of the proposed methods is their independence from external data sources and the preservation of the original dataset structure, with modifications applied only to fill in missing values. Comparative experiments against both conventional techniques (mean, median) and advanced methods (ISSA, IGWO, GMM, RFM, RFCM) demonstrated that CSFC-TOA consistently achieved superior classification performance across diverse machine learning models, including SVM, DT, RF, and LR. On the Pima Indians Diabetes Dataset, the RF classifier yielded the highest accuracy for diabetes detection when paired with CSFC-TOA.

Despite these promising results, the method's evaluation on a single dataset limits the generalizability of the findings. Future research will explore the application of CSFC-TOA to a wider range of medical datasets, investigate strategies to reduce computational cost, and integrate the approach with deep learning architectures to further enhance imputation quality and extend its applicability to real-world clinical decision support systems.

REFERENCES

[1] Kohei, K.A.K.U. (2010). Pathophysiology of type 2 diabetes and its treatment policy. JMaJ, 53(1): 41-46.

[2] Balaji, R., Duraisamy, R., Kumar, M.P. (2019). Complications of diabetes mellitus: A review. Drug Invention Today, 12(1).

[3] Ozougwu, J.C., Obimba, K.C., Belonwu, C.D., Unakalamba, C.B. (2013). The pathogenesis and pathophysiology of type 1 and type 2 diabetes mellitus. Journal of Physiology and Pathophysiology, 4(4): 46-57. https://doi.org/10.5897/JPAP2013.0001

[4] Duncan, B.B., Magliano, D.J., Boyko, E.J. (2026). IDF diabetes atlas 11th edition 2025: Global prevalence and projections for 2050. Nephrology Dialysis Transplantation, 41(1): 7-9. https://doi.org/10.1093/ndt/gfaf177

[5] Yabo, M.M.I., Garko, A.B., Muslim, A.A., Suru, H.U. (2022). A review of diabetes datasets. Journal of Computer Sciences and Applications, 10(1): 6-15. https://doi.org/10.12691/jcsa-10-1-2

[6] Sarwar, M.A., Kamal, N., Hamid, W., Shah, M.A. (2018). Prediction of diabetes using machine learning algorithms in healthcare. In 2018 24th International Conference on Automation and Computing (ICAC), Newcastle Upon Tyne, UK, pp. 1-6. https://doi.org/10.23919/IConAC.2018.8748992

[7] Alam, T.M., Iqbal, M.A., Ali, Y., Wahab, A., Ijaz, S., Baig, T.I., Hussain, A., Malik, M.A., Raza, M.M., Ibrar, S., Abbas, Z. (2019). A model for early prediction of diabetes. Informatics in Medicine Unlocked, 16: 100204. https://doi.org/10.1016/j.imu.2019.100204

[8] Abdulhadi, N., Al-Mousa, A. (2021). Diabetes detection using machine learning classification methods. In 2021 International Conference on Information Technology (ICIT), Amman, Jordan, pp. 350-354. https://doi.org/10.1109/ICIT52682.2021.9491788

[9] 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. https://doi.org/10.1016/j.cmpb.2021.105968

[10] Rubaiat, S.Y., Rahman, M.M., Hasan, M.K. (2018). Important feature selection & accuracy comparisons of different machine learning models for early diabetes detection. In 2018 International Conference on Innovation in Engineering and Technology (ICIET), Dhaka, Bangladesh, pp. 1-6. https://doi.org/10.1109/CIET.2018.8660831

[11] Hasan, M.K., Alam, M.A., Roy, S., Dutta, A., Jawad, M.T., Das, S. (2021). Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). Informatics in Medicine Unlocked, 27: 100799. https://doi.org/10.1016/j.imu.2021.100799

[12] Little, R.J., Rubin, D.B. (2019). Statistical Analysis with Missing Data. John Wiley & Sons.

[13] Torkey, H., Ibrahim, E., Hemdan, E.E.D., El-Sayed, A., Shouman, M.A. (2022). Diabetes classification application with efficient missing and outliers data handling algorithms. Complex & Intelligent Systems, 8(1): 237-253. https://doi.org/10.1007/s40747-021-00349-2

[14] Abdulhadi, N., Al-Mousa, A. (2021). Diabetes detection using machine learning classification methods. In 2021 International Conference on Information Technology (ICIT), pp. 350-354.

[15] Rakshit, S., Manna, S., Biswas, S., Kundu, R., Gupta, P., Maitra, S., Barman, S. (2017). Prediction of diabetes type-II using a two-class neural network. In International Conference on Computational Intelligence, Communications, and Business Analytics, Springer Singapore, pp. 65-71. https://doi.org/10.1007/978-981-10-6430-2_6

[16] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics, 9: 515. http://doi.org/10.3389/fgene.2018.00515

[17] Kumar, S., Bhusan, B., Singh, D., Kumar Choubey, D. (2020). Classification of diabetes using deep learning. In 2020 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, pp. 0651-0655. https://doi.org/10.1109/ICCSP48568.2020.9182293

[18] Hassan, G.S., Ali, N.J., Abdulsahib, A.K., Mohammed, F.J., Gheni, H.M. (2023). A missing data imputation method based on salp swarm algorithm for diabetes disease. Bulletin of Electrical Engineering and Informatics, 12(3): 1700-1710. https://doi.org/10.11591/eei.v12i3.4528

[19] Ahmed, A., İnan, T. (2023). A missing data imputation method based on grey wolf algorithm for diabetes disease. Journal of Engineering Systems and Architecture, 7(1): 55-72. https://doi.org/10.53600/ajesa.1321182

[20] Pourrostami, H., Alavi, S.A., Hosseeini, A., Mousapour Mamoudan, M., Jolai, F., Aghsami, A. (2024). Investigating the impact of missing value imputation methods on the prediction of diabetes using machine learning. Journal of Industrial and Systems Engineering, 16(3): 30-62.

[21] Rahutomo, F., Putranto, T.A. (2025). Comparison of machine learning classification methods for early detection of diabetes. Revue d'Intelligence Artificielle,

39(1): 1-9. https://doi.org/10.18280/ria.390101
[22] Bekhouche, M., Haouassi, H., Bakhouche, A., Rahab, H., Mahdaoui, R. (2023). Improved binary crocodiles hunting strategy optimization for feature selection in sentiment analysis. Journal of Intelligent & Fuzzy Systems, 45(1): 369-389. https://doi.org/10.3233/JIFS-222192
[23] Haouassi, H., Bekhouche, A., Rahab, H., Mahdaoui, R., Chouhal, O. (2024). Discrete student psychology optimization algorithm for the word sense disambiguation problem. Arabian Journal for Science and Engineering, 49(3): 3487-3502. https://doi.org/10.1007/s13369-023-07993-5
[24] Valdez, F. (2021). Swarm intelligence: A review of optimization algorithms based on animal behavior. Recent Advances of Hybrid Intelligent Systems Based on Soft Computing, 273-298. https://doi.org/10.1007/978-3-030-58728-4_16
[25] Kusuma, P.D., Novianty, A. (2024). A new metaheuristic algorithm called treble opposite algorithm and its application to solve portfolio selection. Mathematical Modelling of Engineering Problems, 11(3): 807. https://doi.org/10.18280/mmep.110326