www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 16(4), Sep-Dec 2025, 381-398. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2025.v16i4.999

Missing data imputation with Harmony Search Algorithm

Edgar Alberto Oviedo-Salas[1], *Fausto Antonio Balderas-Jaramillo*[1]
[1]TecNM Instituto Tecnológico de Ciudad Madero, México.
eaos9407@gmail.com, fausto.bj@cdmadero.tecnm.mx

Abstract. Incomplete data poses a significant obstacle in data science and machine learning, influencing model outcomes and occurring commonly across domains such as health, nutrition, electricity, agriculture, chemistry and water resources. Missing data refers to the absence of information for one or more variables in a dataset. Accurate imputation is therefore crucial to ensure the reliability and validity of analyses and predictive models. This study proposes a Harmony Search Algorithm (HSA) to address missing-data imputation, emphasising its flexibility and adaptability. The approach seeks the best imputations by minimising error metrics such as MAE, MSE and RMSE. Computational tests indicate that HSA is a promising method for imputing missing data in a range of contexts.

Keywords: Data science, harmony search, imputation, machine learning

Article Info
Received February 25, 2025
Accepted July 6, 2025

1 Introduction

Incomplete data poses a significant obstacle in data science and is prevalent across various domains. These missing data represent the lack of information on one or more variables within the set. The problem is due to several factors, the main one being human error in recording information, the omission of complete records, and several failures in measuring devices, among others, that affect the information of a dataset. Ignoring the incomplete data leads to low-quality estimates with possible biases. State-of-the-art catalogs them into three categories: Missing Completely at Random MCAR, Missing At Random MAR, and Missing Not At Random MNAR. In the MCAR approach, the data does not depend on a category or its intrinsic value; the mechanism is random. Therefore, allows the use of several imputation techniques. While the Missing At Random MAR approach, the data depend on other values but not specifically on itself, allows the use of several imputation techniques. Finally, in the case of Missing Not At Random MNAR, the absence of data is directly linked to the values of the variable itself, which complicates imputation and requires the use of more advanced methods.

It is essential to recognize missing data approaches to analyze and apply various imputation techniques (Rubin et al., 1995; Memon et al., 2022; Wells et al., 2013; Dong & Peng, 2013; Rubin, 1976). Most machine learning algorithms cannot work with incomplete data sets. Therefore, the data set receives treatment in different ways, such as erasing incomplete information or using imputation techniques. The treatment consists of filling in a value using data predicted through imputation strategies. These methods utilize the existing values within the dataset to substitute the missing entries.

We will address the main techniques for managing incomplete data through imputation, such as Simple Imputation, the K-Nearest Neighbors method, Random Forest, and Multiple Imputation by Chained Equations. The Simple imputation method substitutes missing values by leveraging all non-missing entries in the dataset. It employs strategies such as using the mode, mean, or median values to fill in the gaps. However, when applied to high-dimensional datasets, simple imputation might introduce bias or lead to unrealistic outcomes (Jerez et al., 2010). The K-Nearest Neighbors method groups the available values into clusters and substitutes the missing entries with those obtained from the closest neighbors. The assessment of proximity uses the distance measures, including the Minkowski, Manhattan, Cosine, Jaccard, Hamming, and Euclidean distances (García-Laencina et al., 2009). The Random Forest is a collection of decision trees designed to improve prediction accuracy and minimize overfitting. The algorithm iteratively predicts missing values using observed data and refines the estimates at each step until it reaches satisfactory results (Stekhoven & Bühlmann, 2012). Finally, the Multiple Imputation by Chained Equations (MICE) uses the regression techniques to predict missing values, generating multiple imputations and producing several versions of the dataset. The approach combines the resulting versions to obtain robust estimates. The MICE model stands out for its great versatility in dealing with a wide variety of variable types (van Buuren & Groothuis-Oudshoorn, 2011).

This study explores the application of the Harmony Search Algorithm (HSA) as an alternative to impute missing values. HSA is a metaheuristic technique with inspiration in the musical process to obtain the perfect harmony. This method addresses optimization challenges and finds extensive applications across diverse domains, including Operations Research (Geem et al., 2001), Artificial Intelligence and Evolutionary Computing, Engineering (Goh et al., 2020; Zhang & Zhang, 2018), Data Science (Shuai et al., 2022), Health Research and Bioinformatics (Abdulkhaleq et al. 2022), Business (Abu Doush et al., 2022), among other areas (Siddique & Adeli, 2015).

2 State of the art

Little and Rubin (2002) developed a transformer neuronal network to impute incomplete data in energy demand profiles. Their methodology integrates the K-means algorithm to handle the replacement of missing values. The model was tested on two case studies involving residential houses located in Cornwall and Fintry, UK. Its performance was assessed across three typical scenarios: MCAR, MAR, and MNAR. The proposed method demonstrates notable advancements in both scenarios: in Cornwall, it achieves performances of 57.52% MCAR, 49.71% MAR, and 50.21% MNAR with 30% missing data, outperforming the linear approach. Similarly, in Fintry, the performance enhancements are 46.45% for MCAR, 48.71% for MAR, and 54.87% for MNAR. The authors emphasize that this methodology has profound implications for end users, energy providers, and policymakers, as it facilitates more precise data and better-informed decisions for energy demand analysis and strategic planning.

Ferri et al. (2023) analyze diverse imputation techniques and classification workflows aimed at repurposing highly incomplete Electronic Health Records (EHR) for Machine Learning (ML) applications. The study focuses on identifying imputation methods for managing missing data and assessing ML models that address significantly incomplete numeric values in EHR. The research evaluates 30 different combinations of imputation approaches and classification algorithms, utilizing Complete Blood Count, Census, and Life expectancy data in hospital admissions associated with COVID-19, specifically considering characteristics with a 95% missing data rate. These combinations incorporate methods like average imputation, K-Nearest Neighbors, Bayesian regression, Generative Adversarial Imputation Networks (GAIN), logistic regression, random forest, gradient boosting, and deep neural networks with multiple layers. Computational experiments demonstrates that the integration between transformation and encoding imputations using tree ensemble models surpasses advanced imputation strategies.

Memon et al. (2023) assess various imputation techniques for addressing incomplete categorical data, using health records sourced from Kawempe National Referral Hospital, Uganda. The study examines five techniques: Mode, KNN, Random Forest, Sequential Hot-Deck (SHD), and MICE. The computational results indicate that KNN attains the highest precision in estimating missing values within categorical data, especially when the missing data rate ranges from 5% to 50% MCAR. Random Forest ranks place, delivering superior results at lower percentages of missing data (10%, 15%, 20%). Conversely, SHD exhibits the lowest accuracy across all levels of incomplete data. The study concludes that KNN is the most effective method for imputing missing values across several categories.

Emmanuel et al. (2021) analyze two approaches for imputing missing values, utilizing KNN and Random Forest on the power plant fan dataset and the Iris dataset, with missing rates ranging from 5% to 20%. The computational experiments reveal that KNN imputation performs better than Random Forest in the Iris dataset, while Random Forest demonstrates superior performance over KNN in the power plant fan dataset. The study concludes that both KNN and RF are capable of effectively handling datasets with missing values.

Kabir et al. (2020) examine incomplete values within the municipal water network database from Calgary City, Canada. The study analyzes the effectiveness of single imputation methods, including mean, median, and linear regression, in comparison to advanced multiple imputation techniques such as IRMI, AMELIA, and sequential imputation for missing values (IMPSEQ). The findings from their computational tests reveal that IMPSEQ delivers superior performance compared to both the single imputation approaches and other multiple imputation strategies.

Aljuaid et al. (2017) evaluate five imputation methods for handling incomplete data: KNN, Hot-Deck, and C5.0. Their analysis is based on synthetic datasets of varying sizes, comparing classification accuracy between the original and imputed data. The findings reveal that these techniques perform effectively, even with higher levels of missing data, such as 10% in the credit card dataset and 25% in the adult dataset.

Elasra et al. (2022) apply multiple imputation techniques utilizing the Markov Chain Monte Carlo (MCMC) framework and the Gibbs sampling algorithm on longitudinal educational data. The study evaluates three iterations—2, 50, and 100—and demonstrates that 25 iterations are optimal for effectively imputing incomplete data and achieving a complete longitudinal dataset. The findings conclude that this approach efficiently imputes missing values, enabling more accurate estimation of educational production functions while minimizing biased results.

3 Methodology

To ensure the quality, consistency, and adequate preparation of the data for the imputation process of missing values and subsequent analysis, we implement a rigorous and standardized preprocessing process on all datasets used in this study: Breast Cancer, Salaries, Diabetes, and Wine Dataset, all sourced from the Kaggle platform. This process was essential for transforming the raw data into an optimal format for algorithmic modeling.

The initial phase involves cleaning and structuring each dataset, removing irrelevant columns to ensure the data starts from a complete state before the controlled introduction of missing values for experimentation. Subsequently, categorical variables were converted to numerical representations using direct mappings, while numerical variables were normalized using Min-Max scaling to transform their values to the range [0, 1]. This process uses Equation 1:

$$X_{normalized} \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

Where X is the original feature value, X_{min} and X_{max} are the minimum and maximum observed values for that feature, respectively. This transformation prevents features with wide ranges from dominating the learning of scale-sensitive algorithms.

As a central step for the evaluation, the procedure simulates missing values in a controlled manner by randomly introducing defined percentages of *NaNs* (10%, 20%, 30%, 40%, and 50%) into the predictor variables. This simulation was replicated 20 times with 1000 runs to generate varying sparsity conditions and thus robustly evaluate the performance of the Harmony Search Algorithm HSA against standard imputation techniques.

Performance evaluation uses widely recognized quantitative metrics, such as Mean Square Error MSE, Mean Absolute Error MAE, and Root Mean Square Error RMSE, which measure the difference between imputed values and known actual values. These metrics allow for a comprehensive evaluation of the precision and accuracy of imputations. Furthermore, to statistically validate the significance of the observed differences between HSA and other techniques, nonparametric tests such as the Wilcoxon test and Friedman analysis were applied, providing robustness and rigor to the comparative analysis. Finally, the combination of meticulous preprocessing, controlled simulation of missing values, evaluation using standard metrics, and statistical validation ensures the robustness and reliability of the results presented in this study.

3.1 Interpreting Distributions

Analyzing feature distributions in data sets on Breast Cancer, Salaries, Diabetes, and Wine reveals crucial patterns and properties for understanding their nature and preparing them appropriately for imputation and modeling.

Fig.1 and Fig. 2 displays the distribution of the Breast Cancer dataset and the Salaries dataset. Fig.1 presents the Cancer distribution with binary or discrete categorical variables that predominate, such as Menopause, Breast, Diagnosis Result, and Breast Quadrant, which display count distributions. The age variable presents a more continuous and closer to normal distribution. There is a considerable imbalance in the target variable, *Diagnosis Result*.

Fig.2 presents the distribution of the Salaries dataset, which uses numeric, categorical, and ordinal attributes, *experience_level*, *employment_type*, *remote_ratio*, and company_size, demonstrating clear concentrations across several attributes. The salary variables exhibit strong positive skewness, with a concentration of low values and a long tail toward high values. Furthermore, some categories, such as *employment_type* and *remote_ratio*, show extremely high frequencies in certain classes.

Analysis of the datasets reveals distinct challenges in imputation. The Cancer dataset requires methods that handle categorical variables and data imbalance. The Salaries dataset, on the other hand, requires robust algorithms that can handle a mix of variable types and strong value skew.

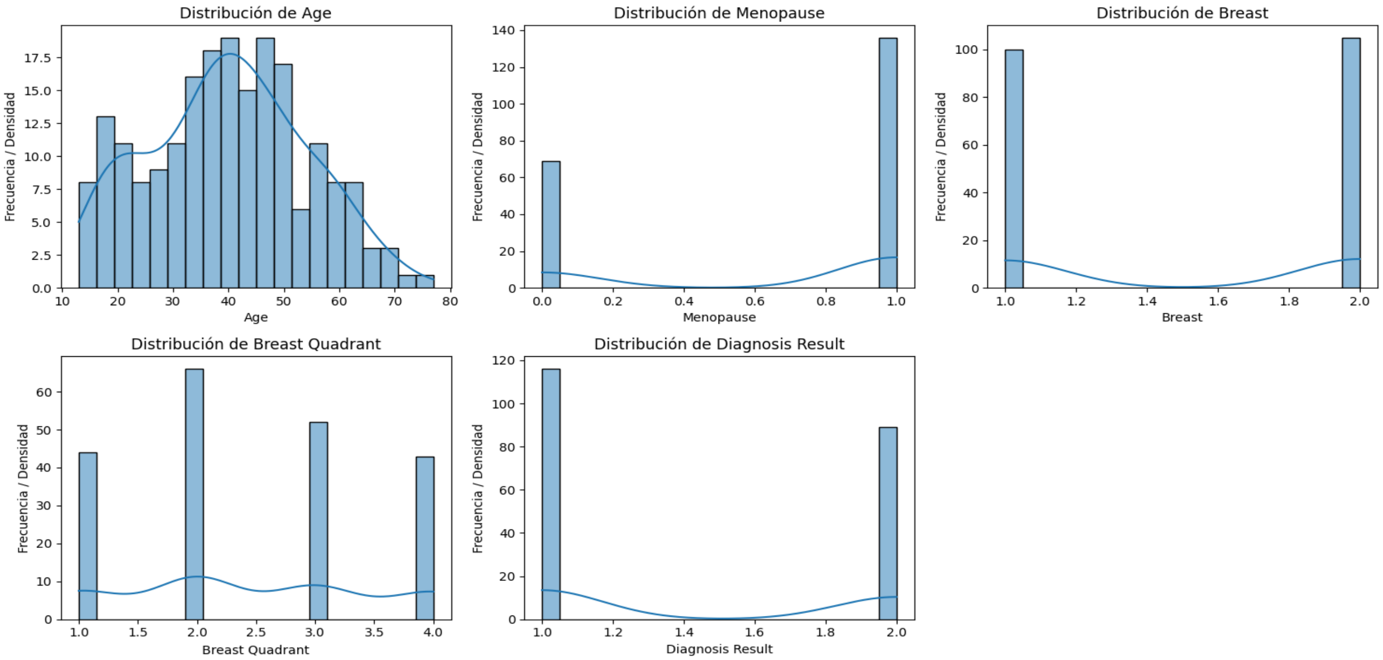

Fig. 1. Cancer distribution.

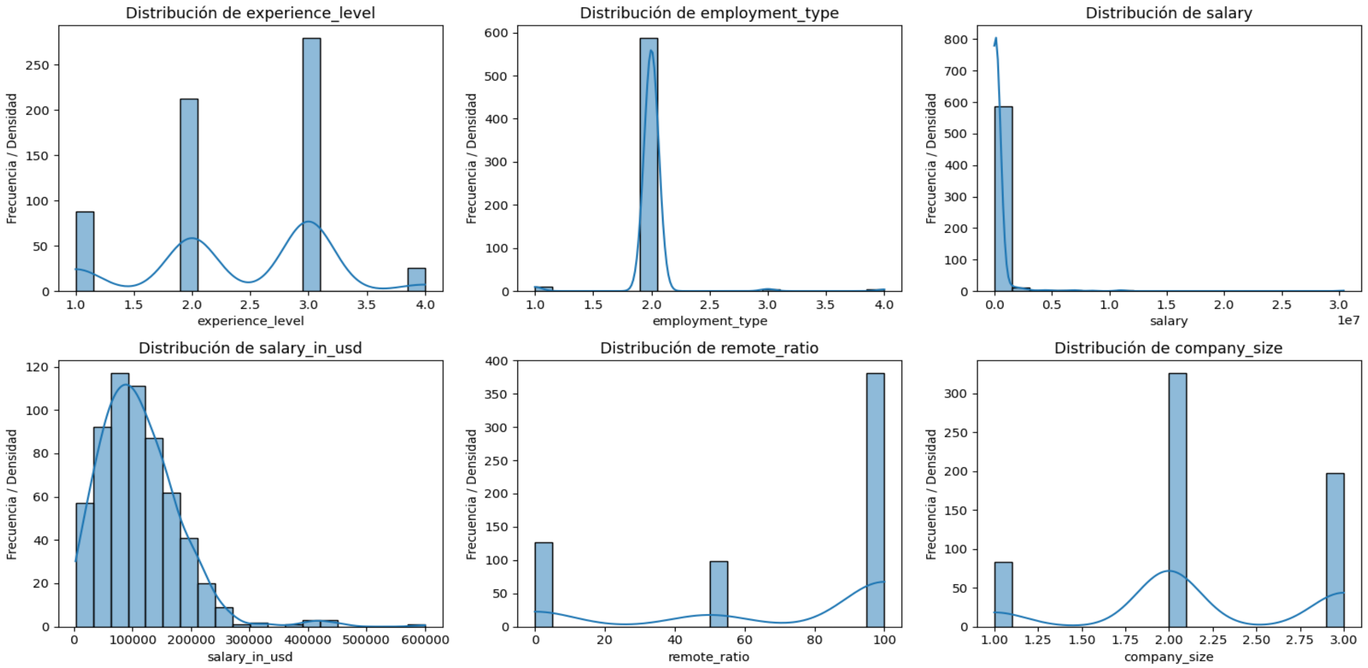

Fig. 2. Salary distribution

Fig.3 and Fig.4 presents the distributions of the Diabetes and Wine datasets. Fig.3 shows the distribution of the Diabetes dataset, which consists primarily of continuous numerical variables. Several critical values, such as blood pressure, skin thickness, insulin, and BMI, require special handling during imputation. Most other numerical variables are positively skewed, and the target variable also exhibits class imbalance.

Finally, Fig. 4 presents the Wine distributions, which are characterized by continuous numerical variables, many with skewed or bimodal distributions, such as Malic Acid, Flavonoids, and Proline. Variables such as Alcohol and Ash have near-normal distributions. The target variable *Customer_Segment* is categorical with multiple classes, which display an imbalance in the distribution of instances. In summary, the distributions confirm the mixed nature of the data, the presence of skewness, and imbalances common in real-world data. These observations inform decisions regarding normalization, imputation, and the selection of appropriate models for each set.

The analysis of the Diabetes and Wine datasets confirms the complexity of the real data. The Diabetes dataset contains continuous variables with positive skewness and class imbalance, while the Wine dataset exhibits skewed and bimodal distributions, as well as imbalance in its target variable. These characteristics demonstrate the necessity for robust imputation and normalization methods to ensure reliable results.

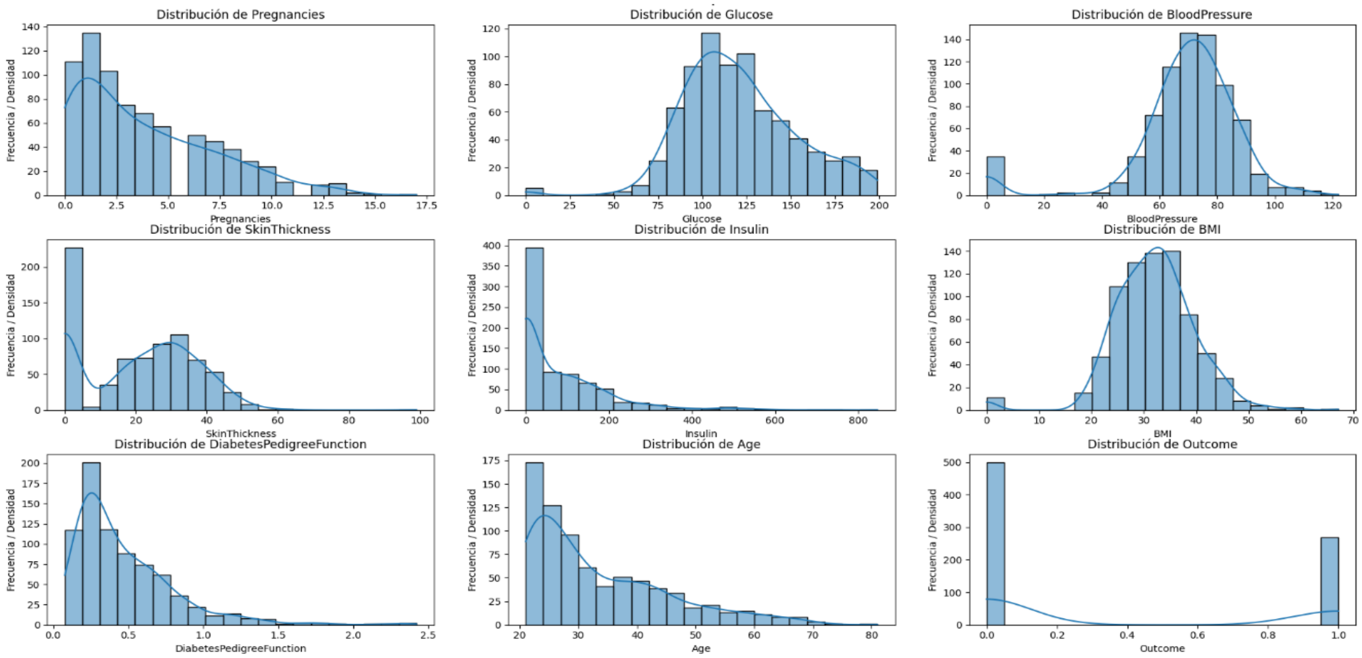

Fig. 3. Diabetes distribution

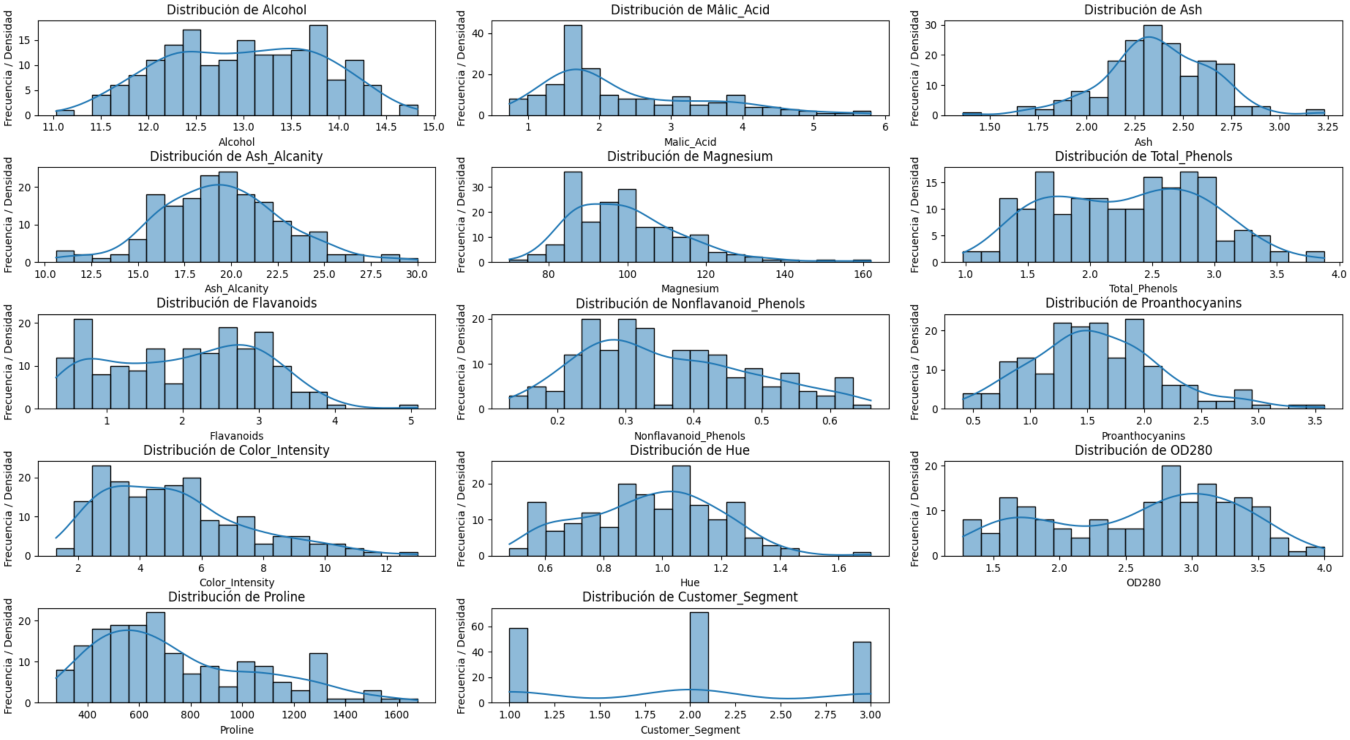

Fig. 4. Wine distribution

4 Harmony Search Algorithm Approach

Harmony search algorithm (HSA) is an optimization technique introduced by Zong Woo Geem et al. (2001). This method draws inspiration from musical harmony, modeling the process of generating and refining solutions after how musicians craft harmonious chords. The algorithm leverages a harmony memory to store and refine solutions, assessing new candidates based on the objective function. Its goal is to identify the optimal solution for various complex problems.

HSA is a metaheuristic technique inspired by musical improvisation processes, demonstrating its effectiveness in solving optimization problems in large, nonlinear search spaces. Its design allows for an efficient balance between exploration and exploitation of the solution space, which is particularly useful for issues such as missing data imputation, where the search for plausible values must consider multiple variables and constraints. This feature, coupled with the fact that, unlike other methods, HSA does not require large populations or complex operators, simplifies its implementation and reduces the computational burden, making it a flexible and adaptable option for imputation, where the process finds a combination of imputed values that minimizes reconstruction error without falling into a local optimum. Furthermore, its stochastic nature and its ability to explore different regions of the solution space make this method a promising alternative to traditional techniques and other heuristics, especially in datasets with varied characteristics and heterogeneous levels of missing values. For these reasons, we select the HSA method as the basis for developing the imputation method presented in this study.

Algorithm 1 shows our proposed procedure to impute missing data; the process generates the solutions with a defined number of iterations. The algorithm input includes initial variables such as Harmony Memory (HM), best score (best), current solution (Sol), dataset (Ds), current score (score), solution scores (Scores), and worst score (WorstScore). Algorithm parameters such as Harmony Memory Size (HMS), Harmony-Memory Considering Rate (HMCR), Pitch-Adjusting Rate (PAR), and Bandwidth (BW) are also used.

Algorithm 1 HSA Imputation General Procedure

```
Input:HM, Sol, Bestsol, Ds, Scores ← ∅ ▷ Empty Sets/Lists
best, score, WorstScore ← 0
Ds ← LoadPreparedDataSet()               ▷ Setup
InitializeParam (HMS, HMCR, PAR, BW)
InitializeHarmonyMemory (HM)
  for i = 1 to MaxIter do ▷ Main Iteration Loop
        Sol ← Generate Solution (HM, Ds, HMCR, PAR, BW)
        score ← objective function (Ds, Sol)
        Add score to Scores
        WorstScore ← Max (Scores)
       if score < WorstScore then
            HM [findworstidxF(Scores)] ← Sol
        end if
        if score < best then
            best ← score
            Bestsol ← Sol
        end if
    end for
Output: best, Bestsol
```

The method starts with the *LoadPreparedDataSet* function, which loads the dataset for preprocessing and amputation. The *InitializeParam* function initializes the main parameters to use in HSA, receiving values for these elements: HMCR and PAR, with values between 0 and 1; HMS, which determines the number of solutions to save in memory during the searching process; and BW, which adjusts the range for the selected values in the HM with a range from 0 to 100. *InitializeHarmonyMemory* generates an initial solution set, in which each solution is a copy of the original data and imputes missing values with random values within the ranges observed in the existing data. The Generate Solution function produces possible imputations for missing data that are close to the original data set, while minimizing the error between the imputed data and the original information. Finally, the objective function process evaluates and compares the accuracy of the solutions, measuring the error between the imputed values and the original values, guiding the algorithm to improve and identify the most precise imputations. The algorithm output consists of two values, *best* and *BestSol*, which correspond to the best current solution and the best solution, respectively.

4.1 Initialize Memory

Algorithm 2 shows the initialization procedure for the HSA harmony memory. The input for this algorithm is data (the dataset) and HMCR (Harmony Memory Considering Rate). The output is HM (Harmony Memory). This memory is a list of possible solutions for the missing data. The harmony memory is optimized iteratively, where each solution is a version of the original dataset with missing values imputed randomly, using the existing values of the columns with complete data. Using a range of minimum and maximum values ensures that the imputed values are reasonable in the context of the existing data, which is critical to maintaining the integrity of the dataset throughout the imputation process. Therefore, the process generates an initial harmony memory with candidate solutions for imputed values, allowing the algorithm to continue iterating and refining the solutions with a range of values in the dataset, making it a simple but effective approach to generating viable initial solutions.

Algorithm 2 HSA Initialize Memory Process

```
Input: data, HMCR
HM ← ∅                    ▷ Initialize Harmony Memory
for k = 1 to HMCR do
    sol ← data     ▷ Initialize solution with input data
    for j = 1 to data.Columns do
        if IsNull(data_j) then
            getMaxMin (minv, maxv, data_j)
            adjustmentF(minv, maxv)
            fill values ← addrandomvF(minv, maxv, data)
            sol_j ← addF(fill values)
        end if
    end for
    HM ← addF(sol) ▷ Add the solution to the Harmony Memory
end for
Return: HM
Output: HM
```

4.2 Imputation Process

The main contribution is the use of this approach to impute missing data due to its capacity to explore a wide range of possible solutions, which allows finding imputations that minimize the error concerning the original values. Specifically, the initial solutions are not generated randomly as in traditional HSA. The basis of the solutions is the observed values in the columns of the dataset, thus ensuring a realistic and consistent imputation with the existing data.

Algorithm 3 displays the improvisation process that imputes the missing data; the goal is to create a candidate new solution. The input for this algorithm is HM (Harmony Memory), data (the dataset), HMCR (Harmony Memory Considering Rate), PAR (Pitch-Adjusting Rate), and BW (Bandwidth). The procedure output is *Newsol* (the new solution generated). The solution uses the information from Harmony Memory, applying several random rules, combining the exploitation of existing solutions with the exploration of new allocations, controlled by the parameters Harmony Memory Considering Rate (HMCR) (See lines 8 and 12), Pitch Adjustment Rate (PAR), and Bandwidth (BW) (See line 16). The function loops through each column in the dataset to check if it contains missing values. If so, we perform a row-by-row iteration to impute null values.

For each missing value, the function evaluates using HMCR values if the random value is less than HMCR, where the process selects a value from harmony memory. The procedure imitates the memory of musicians in improvisation, in which we reuse the previous solutions that have been effective. Otherwise, if it is higher than HMCR, the method generates a random value within a defined range between the maximum ($minv$) and the minimum ($maxv$) values of the current column in the original data, allowing the exploration of new solutions by imputing values not present in previous solutions. Eq. 3 shows the condition of HMCR, where selected sol_{rc} represents the initial solution from Harmony Memory (HM) that iteratively loops through rows (r) and columns (c) $HM = \{hm_{01}, hm_{ij}, \ldots, hm_{rc}\}$.

$$New\_sol_{r,c} = \begin{cases} selected_{sol_{r.c}}, & rand(0,1) < HMRC \\ randF\ (minv, maxv), & Otherwise \end{cases} \tag{3}$$

$$New_{sol_{r,c}} = New_{sol_{r.c}} + rand(-BW, BW\ ), randF\ (0,1) < PAR \tag{4}$$

Upon obtaining the new value (either with the harmony memory or randomly), the function makes a new decision using the PAR value. This decision evaluates if a new random value is less than the PAR value, where the function adjusts the imputed value by adding or subtracting a small random value within a range defined by BW (see Eq. 4). This adjustment introduces variations in the selected values, emulating the fine-tuning in musical improvisation. The objective is to explore close solutions that can improve the overall quality of the imputation.

Algorithm 3 HSA Improvisation Process

```
Input: HM, data, HMCR, PAR, BW
New sol ← data ▷ Initialize new solution with input data
for c = 1 to data.Columns do
  if IsNull(data$_c$) then ▷Check nulls in original data columns
   for r = 1 to New_sol.Index do ▷Iteration elements/rows
     if IsNull(New_sol$_{r,c}$) then  ▷Is the current element null?
          if randomF (0, 1) < HMCR then  ▷HM Consideration
            New_sol$_{r,c}$ ← randomF(HM)  ▷ Assign random HM
          else      ▷ Improvise new value
            getMaxMin(minv, maxv, data$_c$)
            adjustmentF(minv, maxv)
            New sol$_{r,c}$ ← randomF(minv, maxv)
          end if
      end if
      if randomF(0, 1) < PAR then     ▷ Pitch Adjustment
        New_sol$_{r,c}$ ← New_sol$_{r,c}$ – randomF(-BW, BW)
      end if
    end for
  end if
end for
Return: New_sol
Output: New_sol
```

5 Experimental Results

5.1 Configuration and Instances

Computational tests were performed on a computer with an AMD Ryzen 7 4000 series processor at 2.90 GHz and 8 GB of RAM, running Windows 10, using PyCharm and Python as the programming language. Given the stochastic nature of the HSA algorithm, we perform 20 independent runs of 1000 iterations. We report the average values of the results to ensure the stability and reliability of the metrics. To evaluate the performance of the methods, we use the MAE, MSE, and RMSE metrics. The results are presented in graphs showing these metrics on the Y-axis and the percentages of missing data (10%, 20%, 30%, 40%, and 50%) on the X-axis. The instances and code are available at csalas07 (n.d.) GitHub repository.

For comparative imputation methods, we use the Scikit-learn library for standard implementations. The experts widely recognize it for its efficiency and robustness in machine learning. For methods as the Mean, Median, and Mode imputers, we use the SimpleImputer library, setting the strategy to mean, median, and most frequent, respectively. KNN procedure, employing the KNNImputer class, with five nearest neighbors for imputation. For the MICE imputer, the IterativeImputer class was used, with

$random\ state = 0$ to ensure reproducibility, and the other parameters using their library default values. Finally, the Random Forest imputer made use of the RandomForestRegressor class, using the library's default parameters for its configuration.
Assessing the generality of the proposed approach, we selected four Kaggle datasets widely used in the literature (Cancer, Salary, Diabetes, and Wine) with diverse characteristics in terms of domain, size, and type of variables. Each set comes from different domains, which contributes to demonstrating the versatility and robustness of the algorithm. The Cancer and diabetes represent biomedical problems with complex and sensitive data, where imputation quality is crucial. Furthermore, Salary and Wine allow the method to be evaluated in socioeconomic and agro-industrial contexts, broadening its application spectrum.

Simulating the presence of missing data, random imputation was applied to the original datasets, uniformly removing values from continuous variables according to predefined percentages MCAR. This technique reproduces common scenarios where missing data occur without a specific pattern, allowing an objective and consistent evaluation of the performance of imputation methods under realistic and controlled conditions. Table 1 summarizes the key properties of each dataset: name, number of attributes, number of instance elements (n), and number of classes.

Table 1. Instance set

Name	Attributes	n	Class
Cancer	9	1917	2
Diabetes	9	6912	2
Salary	6	3642	0
Wine	14	2492	1
Cancer	9	1917	2

5.2 Analysis of missing data for Cancer dataset

Fig. 5 presents the performance of the Mean, Median, Mode, KNN, MICE, Random Forest, and HSA imputation techniques on the Cancer dataset, evaluated using Mean Absolute Error MAE, Mean Squared Error MSE, and Root Mean Squared Error RMSE, respectively. Visually, the HSA method exhibits consistently competitive behavior across all three percentages of missing data (from 10% to 50%), maintaining relatively low and stable error values. HSA performs comparable to or better than several methods under various conditions, even approaching the performance of Random Forest, a state-of-the-art algorithm, in terms of MAE and MSE. In contrast, the Mean method generally shows the highest MAE, MSE, and RMSE values. Methods such as KNN and MICE exhibit fluctuations, with KNN showing a remarkable initial performance that deteriorates markedly as the percentage of missing data increases, and MICE exhibiting variability and error spikes at several points.

Table 2 presents the Friedman test, which measures performance through average ranks. The Harmony Search Algorithm HSA imputation method is a tool with intermediate potential. HSA's ranks consistently rank moderately or highly, indicating that, while it does not perform as well as the Mean method, it does outperform other methods, such as the Median and Mode, in some levels of data loss. The HSA ranking suggests that it is a more robust alternative than other simplistic approaches, although the Mean appears to be the most effective. Therefore, although HSA does not stand out as the best, its competitive performance positions it as a viable option and a starting point for consideration.

Table 2. Friedman Test Cancer Average Range

Loss (%)	HSA	KNN	MICE	Mean	Median	Mode	RF
10	5.75	3.67	3.15	1.75	4.85	5.05	3.77
20	4.3	4.10	3	1.5	4.95	5.25	4.90
30	4.9	3.65	2.5	1.05	5.05	5.95	4.90
40	4.75	3.77	3.15	1.00	5.10	6.05	4.17
50	4.75	3.52	2.92	1.05	5.05	5.75	4.95

Subsequently, Table 3 displays the Wilcoxon test; the analysis suggests that the HSA imputation method is a potentially valuable tool in the field of data analysis. The HSA relevance is notable when compared to conventional methods such as Mean and MICE, as the consistently low p-values demonstrate that HSA produces significantly different results, offering a unique perspective that could be crucial for several research objectives. While its results become statistically similar to those of Median, Mode, and, in various ranges, KNN and RF, this does not diminish its potential, which demonstrates that its performance is comparable to methods considered reliable. HSA's ability to generate results distinct from those of simple techniques, such as Mean, positions it as a sophisticated and potentially superior alternative for tasks where more robust imputation is required.

Table 3. Wilcoxon Test Cancer P-Value HSA vs Imputers

Loss (%)	KNN	MICE	Mean	Median	Mode	RF
10	0.0008507	0.0002098	0.0000019	0.6215134	0.5958195	0.0003948
20	0.4897449	0.0239506	0.0000629	0.1429062	0.1230927	0.3488102
30	0.1893482	0.0007076	0.0000019	0.1429062	0.0637226	0.7561665
40	0.2942524	0.0440540	0.0000019	0.0761531	0.0695801	0.2161674
50	0.0239506	0.0016899	0.0000019	0.0758514	0.0531693	0.7841263

In addition to directly evaluating accuracy of imputation, the impact of the HSA method on the performance of a subsequent classification task on the Cancer dataset with the imputed values was analyzed. Table 4 presents the key classification metrics (Accuracy, Recall, F1-Score, ROC AUC) for the HSA method at different percentages of missing data.

Table 4. Precision Test Cancer

Loss (%)	Accuracy	Recall	F1-Score	ROC AUC
10	0.9032	1	0.9211	0.9275
20	0.9032	0.9714	0.9189	0.9381
30	0.8871	0.9714	0.9067	0.9079
40	0.9032	1	0.9211	0.937
50	0.871	0.9429	0.8919	0.8958

The results show that HSA maintains high accuracy (Accuracy) in the classification task, ranging from 0.8710 to 0.9032 as the percentage of missing data increases from 10% to 50%. Recall remains consistently high (between 0.9429 and 1.0000), indicating an excellent ability to correctly identify positive cases. Similarly, the F1 score remains within a robust range (0.8919 to 0.9211), and the area under the ROC curve AUC also demonstrates high discriminatory power (0.8958 to 0.9381). These values suggest that, despite the presence of missing data and their subsequent imputation with HSA, the resulting dataset retains the information necessary for effective classification, which is crucial for practical applications.

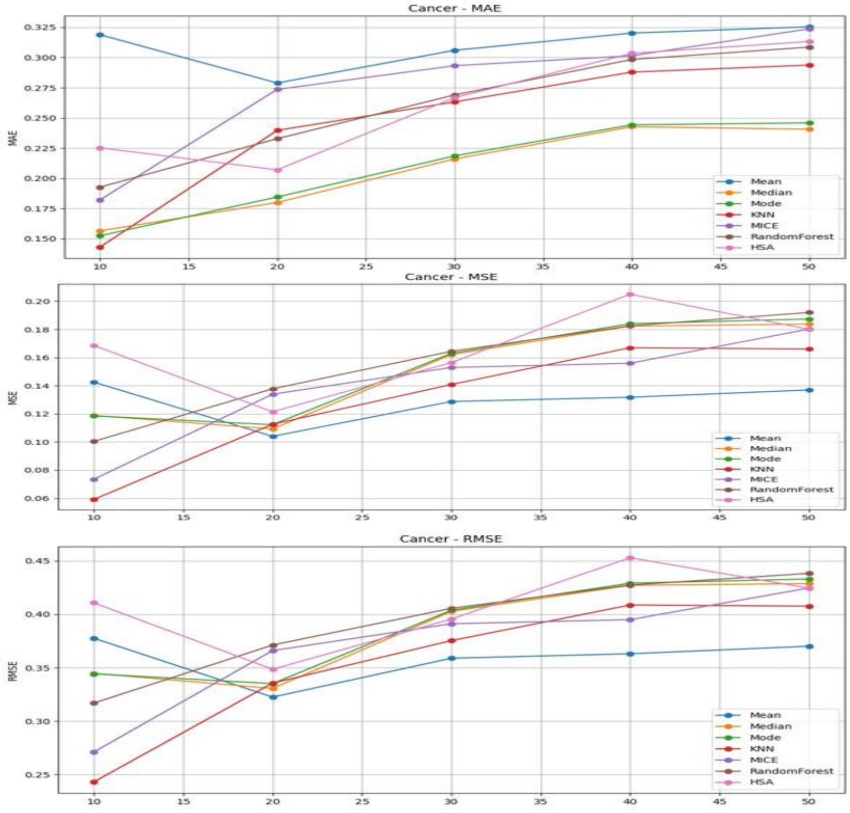

Fig. 5. Cancer comparison.

Analysis of the results for the Cancer set reveals that while HSA exhibits remark- able visual competitiveness and stability in imputation (Fig.5), statistical vali- dation (Table 2, 3) offers nuances. The Wilcoxon test indicates that, for RMSE, HSA is statistically inferior to KNN, MICE, and Mean, possibly due to the characteristics of this dataset favoring methods that capture more direct relationships or the robustness of the mean. However, no statistically significant difference was found between HSA and the Median, Mode, and Random Forest methods, validating HSA ability to compete with advanced algorithms. Crucially, HSA's performance on the downstream classification task (Table 4) consistently shows high levels of Accuracy, Recall, F1-Score, and ROC AUC, indicating that, despite the lower accuracies in direct imputation, HSA manages to pre- serve the structural quality of the data for effective classification. This suggests that HSA primary value lies in its ability to generate high-quality imputed datasets for downstream machine learning tasks, making it a robust alternative for applications where the ultimate goal is prediction or decision-making, especially in scenarios with higher data complexity.

5.3 Analysis of missing data for Diabetes dataset

Fig. 6 presents the performance of the imputation techniques Mean, Median, Mode, KNN, MICE, Random Forest, and HSA on the Diabetes dataset, evaluated using Mean Absolute Error MAE, Mean Squared Error MSE, and Root Mean Squared Error RMSE, respectively. Visually, the HSA method exhibits competitive and relatively stable behavior across all three percentages of missing data (from 10% to 50%). Its performance consistently remains among the lowest error methods at most points. For MAE, HSA ranks as the method with the lowest or near-lowest error at all percentages, demonstrating remarkable robustness. For MSE and RMSE, HSA also presents a solid performance, remaining among the two or three lowest error methods, frequently surpassed by Random Forest and MICE, but with very controlled error growth as missingness increases.
The Random Forest method generally shows excellent performance, often achieving the lowest MAE, MSE, and RMSE, with a very flat and stable error curve. MICE also present very competitive performance, staying close to Random Forest and HSA, and outperforming simpler methods in most scenarios. In contrast, the means, median, and mode methods consistently show higher error values than HSA, Random Forest, KNN, and MICE, and their errors increase as missing data increases. KNN shows intermediate performance, with a gradual increase in its error metrics.

Table 5 presents the Friedman test; the HSA method shows considerable potential, albeit with a critical dependence on the percentage of missing data. Initially, with only 10% missingness, HSA proves to be an undisputed leader with the lowest average rank (1.15), suggesting that it is the most effective method in scenarios with little data loss. However, this superior performance is volatile. As data loss increases, HSA's performance gradually deteriorates, yielding its leading position. At the highest levels of missingness, Random Forest (RF) emerges as the most robust and consistent method, outperforming all others. Therefore, HSA emerges as a potentially high-impact tool for scenarios with minimally incomplete data, although its reliability decreases significantly as the amount of missing data grows.

Table 5. Friedman Test Diabetes Average range

Loss (%)	HSA	KNN	MICE	Mean	Median	Mode	RF
10	1.15	3.25	2.4	4.8	6	7	3.4
20	3.55	3.15	2.05	4.65	5.65	6.95	2
30	5.55	2.85	1.55	4	5.6	6.8	1.65
40	6.35	2.65	1.7	4	5.05	6.6	1.65
50	6.45	2.65	1.7	4	5.05	6.5	1.65

Table 6 presents the Wilcoxon test, demonstrating that the Harmony Search Algorithm is a tool with significant and nuanced potential. Consequently, a significant statistical difference exists between the Mean and MICE methods, evidenced by consistently low p-values, which highlights HSA's ability to generate unique solutions. This finding validates HSA as a robust alternative. However, the picture becomes more complex when noting the absence of statistical differences with simpler methods, such as Median and Mode, in most scenarios. This similarity does not detract from HSA's merit, which suggests that its performance is comparable to that of established and reliable solutions.

In addition to directly evaluating accuracy of imputation, the impact of the HSA method on the performance of a subsequent classification task on the Diabetes dataset with the imputed values was analyzed. Table 7 presents the key classification metrics (Accuracy, Recall, F1-Score, ROC AUC) for the HSA method at different percentages of missing data.

Accuracy remained within acceptable ranges, reaching its highest point at 0.7706 for 20% missing data. Although a slight decrease in performance was detected for the metrics; Accuracy, Recall, F1-Score, and ROC AUC; as the percentage of missing data increased from 10% to 50%, the ROC AUC, by consistently remaining above 0.76 in all cases, indicates that the imputation performed by HSA preserves the underlying predictive power of the model. This suggests that HSA performs effective imputation that significantly contributes to maintaining the inherent structure of the data, which is essential for robust and accurate classification.

Table 6. Wilcoxon Test Diabetes P-Value HSA vs Imputers

Loss (%)	KNN	MICE	Mean	Median	Mode	RF
10	0.000001907	0.00029280	0.0000019	0.00000191	0.00000191	0.00000191
20	0.27892307	0.00271225	0.0019855	0.00036262	0.00000381	0.00485992
30	0.00000191	0.00000191	0.0000019	0.20244980	0.00013351	0.00008845
40	0.00008845	0.00000191	0.0000019	0.00012026	0.47490501	0.00000191
50	0.00008845	0.00000191	0.0000019	0.00001907	0.97021716	0.00000191

Table 7. Precision Test Diabetes

Loss (%)	Accuracy	Recall	F1-Score	ROC AUC
10	0.7619	0.7619	0.7554	0.7986
20	0.7706	0.7706	0.7643	0.7935
30	0.7489	0.7489	0.7364	0.7865
40	0.7446	0.7446	0.7311	0.7691
50	0.7186	0.7186	0.7021	0.7222

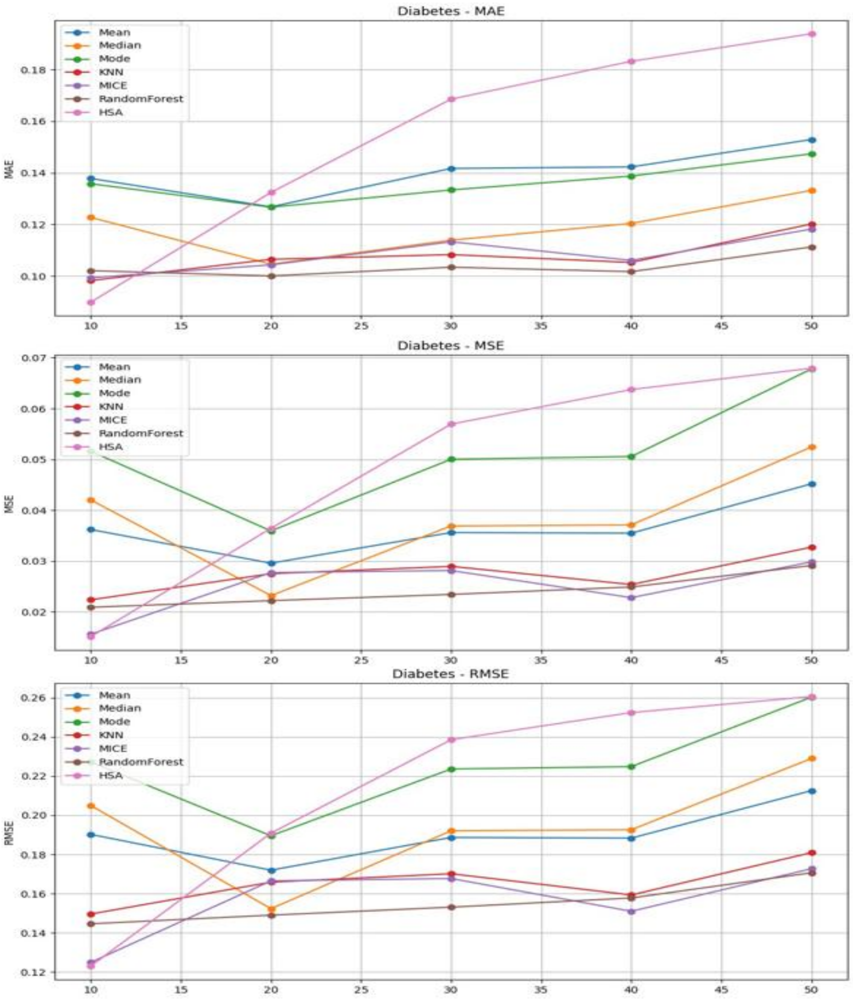

Fig. 6. Diabetes Comparison.

In the Diabetes dataset, the combined results of the Friedman test and Wilcoxon post-hoc analyses delineate a key performance pattern among the imputation techniques. HSA proved to be an exceptionally robust and reliable method in scenarios with a low proportion of missing data (10% RMSE), establishing itself as the gold standard under these high completeness conditions. Its consistent performance across this initial loss range positions it as an invaluable and highly promising tool for applications where source data quality is paramount. While its effectiveness decreased as loss increased, the robustness of MICE and Random Forest proved superior and more versatile under higher uncertainty scenarios. Both methods consistently maintained the lowest missingness ratios and exhibited significant superiority under higher loss conditions. This suggests that, for this dataset, the choice of the optimal method depends largely on the level of data completeness, making HSA a potential tool for near-complete data.

5.4 Analysis of missing data for Salary dataset

Fig. 7 presents the performance of the Mean, Median, Mode, KNN, MICE, Random Forest, and HSA imputation techniques on the Salary dataset, evaluated using Mean Absolute Error (MAE), Mean Squared Error MSE, and Root Mean Squared Error RMSE, respectively. Visually, the HSA method obtains the lowest MAE, MSE, and RMSE values at 10% missing data, suggesting superior initial performance. However, as the percentage of missing data increases (from 10% to 50%), HSA's error metrics; MAE, MSE, RMSE display a steady and steep upward trend, outperforming most other methods and ending with the highest or among the highest errors at 50%. In contrast, Random Forest and MICE prove more robust in handling high percentages of missing data in these metrics, maintaining comparatively low errors. The mean, median, mode, and KNN methods all show gradual increases in their errors, with the mean and median generally lagging behind the HSA by high percentages.

Table 8 presents the Friedman test of the Salary dataset. The test reveals Harmony Search Algorithm HSA as a tool with great potential, particularly in scenarios with a low amount of missing data. At 10% and 20% loss, HSA emerges as the clear leader, obtaining the lowest average ranks and demonstrating superior effectiveness to all other methods. This initial lead is a strong indicator of its value in contexts where accurate imputation does not require a high rate of data incompleteness. Although HSA's performance degrades as data loss increases, yielding the top spot to methods such as MICE and Random Forest RF, its ability to excel in low-loss conditions cements it as a highly relevant option. Therefore, HSA is an invaluable potential tool, especially for data imputation in data sets with minimal to moderate loss.

Table 8. Friedman Test Salary Average range

Loss (%)	HSA	KNN	MICE	Mean	Median	Mode	RF
10	1.00	4.05	3.35	3.65	6.20	6.20	3.55
20	1.05	3.95	2.77	3.98	6.20	6.50	3.55
30	2.85	3.60	1.90	3.30	6.25	6.35	3.75
40	3.55	3.90	1.85	2.70	6.40	6.60	3.00
50	5.70	3.22	1.35	2.70	6.12	5.97	2.92

Table 9 presents the Wilcoxon test of the Salary dataset. The test highlights the potential of the HSA as an adaptive tool whose relevance changes with the amount of missing data. Initially, at 10% and 20% loss, HSA markedly differentiates itself from all other methods, suggesting its ability to generate unique results in scenarios with few missing data. However, as the loss increases to 30% and 40%, HSA starts to converge with the performance of most state-of-the-art methods, such as KNN, MICE, and RF, becoming statistically indistinguishable from them, demonstrating its compatibility and ability to align with robust results. It is particularly notable that at the 50% level, HSA differentiates itself from all state-of-the-art methods but becomes indistinguishable from Median and Mode, suggesting a change in its behavior.

Table 9. Wilcoxon Test Salary P-Value HSA vs Imputers

Loss (%)	KNN	MICE	Mean	Median	Mode	RF
10	0.000001907	0.000001907	0.000001907	0.000001907	0.000001907	0.000001907
20	0.000001907	0.000001907	0.000001907	0.000001907	0.000001907	0.000003815
30	0.278797395	0.112563231	0.537866329	0.000026703	0.000026703	0.545875549
40	0.910817249	0.003182496	0.026641846	0.000001907	0.000001907	0.311794281
50	0.000013351	0.000003815	0.000103203	0.262680669	0.261098862	0.000001907

On the Salary dataset, the combined Friedman test and Wilcoxon results reveal a distinctive performance pattern. HSA demonstrated exceptionally robust and reliable performance in scenarios with a low proportion of missing data (10% RMSE with an average rank of 1.00 and 20% RMSE with 1.05), establishing itself as a potential and practically unmatched tool in these conditions of high data integrity. Its consistent performance at this initial loss range positions it as an invaluable and very promising option for applications where the quality of the source data is paramount. While its effective- ness decreased as the loss increased, reaching average ranks of 5.70 at the 50% loss level, the robustness of MICE and Random Forest proved superior and more versatile in scenarios of greater uncertainty. Both methods consistently maintained the lowest ranges; MICE with 1.35 at 50% and RF with 2.92 at 50%; and exhibited significant superiority under higher loss conditions, with MICE even significantly outperforming HSA above 40% RMSE. This suggests that, for this dataset, the choice of the optimal method depends largely on the level of data completeness, making HSA a potential tool for near-complete data.

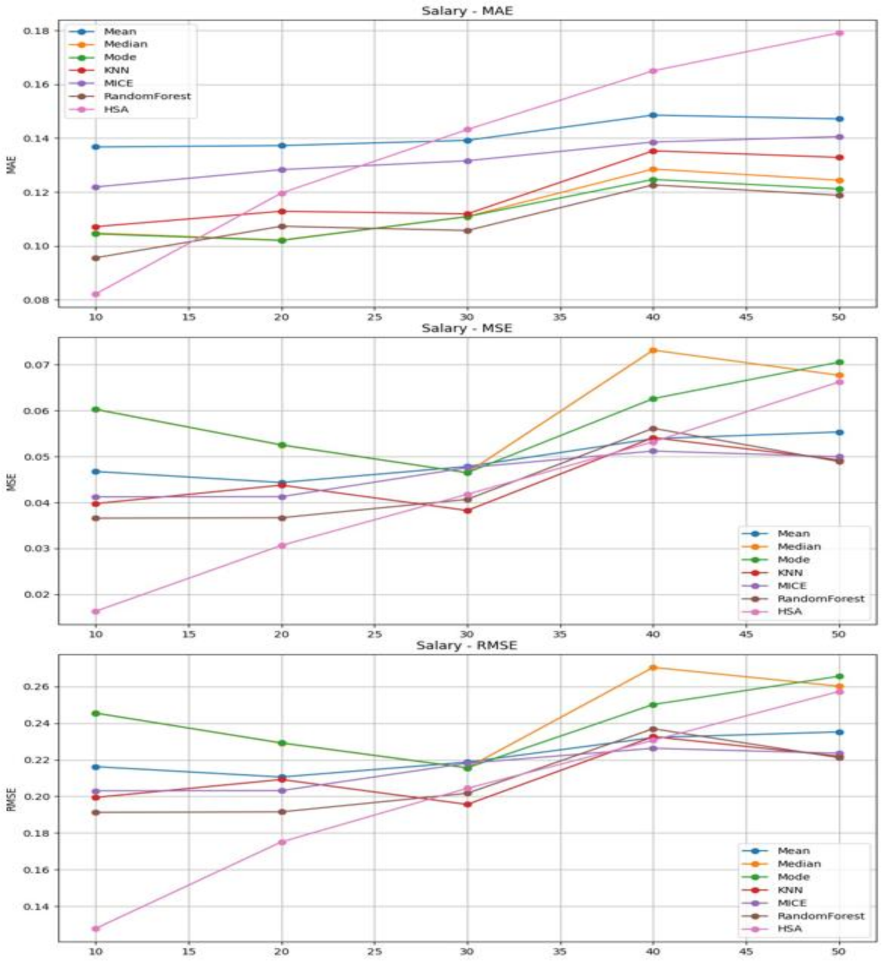

Fig. 7. Salary comparison.

5.5 Analysis of missing data for Wine dataset

Fig. 8 displays the performance of the imputation techniques (Mean, Median, Mode, KNN, MICE, Random Forest, and HSA) on the Wine dataset, evaluated using Mean Absolute Error (MAE), Mean Squared Error MSE, and Root Mean Squared Error RMSE, respectively. Visually, the HSA method exhibits competitive and relatively stable behavior across all three percentages of missing data (from 10% to 50%). Its performance consistently remains among the lowest error methods at most points. For MAE, HSA ranks among the two or three lowest error methods at most percentages, demonstrating good robustness. For MSE and RMSE, HSA also shows a solid performance, remaining among the best methods, with very controlled error growth as missingness increases.

The Random Forest method generally presents a solid performance, often achieving the lowest MAE, MSE, and RMSE, with a very flat and stable error curve, indicating its robustness to increasing missing data. MICE also present very competitive performance, remaining close to Random Forest and HSA, and outperforming simpler methods in most scenarios. In contrast, the

Mean, Median, and Mode methods consistently exhibit higher error values than HSA, Random Forest, KNN, and MICE, and their errors increase more steeply as the amount of missing data increases. KNN shows intermediate performance, with a gradual increase in its error metrics.

Table 10 presents the Friedman test for the Wine dataset; the test demonstrates that HSA is an imputation tool with great potential. At 10% and 20% loss, HSA establishes itself as the leading method, achieving the lowest average ranks and outperforming all its competitors. This initial performance underscores its exceptional effectiveness in scenarios with low to moderate amounts of missing data. Although its main rival, Random Forest RF, demonstrates greater robustness and outperforms it at higher loss levels, HSA's consistency at intermediate ranks and its initial lead consolidates it as a highly relevant option. The results suggest that HSA is a high-value alternative, capable of delivering outstanding performance in data imputation, especially in contexts with relatively high data integrity.

The Wilcoxon test in Table 11 describes the performance for the Wine dataset, demonstrating a changing dynamic of the approach. At low to moderate loss levels (10% to 40%), HSA consistently stands out as a distinctive method, producing results that are significantly different from almost all other imputers. This behavior underscores HSA's ability to generate unique and potentially superior solutions, validating its relevance in scenarios where accurate and differential imputation is needed. The situation evolves at 50% of missing data, where HSA becomes statistically indistinguishable from KNN and MICE, suggesting that, under conditions of high data loss, HSA's performance converges with that of other robust methods.

Analysis of the precision metrics for the Wine dataset (Table 12) reveals the remarkable robustness of the predictive model against data loss. In low-loss scenarios (10% and 20%), the model exhibits near-perfect performance, with Accuracy, Recall, and F1- Score values of 1.0000 and an ROC AUC that remains at 1.0000 and 0.9995 respectively. As the percentage of missing data increases to 30%, the metrics experience a slight reduction, but remain exceptionally high; Precision at 0.9815 and ROC AUC at 0.999. Even with significant losses of 40% and 50%, the model demonstrates surprising resilience, stabilizing its metrics above 0.94 for Accuracy, Recall, and F1-Score, and an ROC AUC above 0.98. This consistent high performance underscores the effectiveness of the imputation techniques applied on the Wine dataset, allowing the model to maintain robust predictive capability even when half of the data are missing.

Table 10. Friedman Test Wine Average range

Loss (%)	HSA	KNN	MICE	Mean	Median	Mode	RF
10	1.00	3.15	3.20	5.65	5.80	6.55	2.65
20	1.00	3.15	3.05	5.35	5.70	6.95	2.80
30	1.05	3.10	3.40	5.55	5.55	6.90	2.45
40	1.60	2.80	3.40	5.25	5.85	6.90	2.20
50	2.65	2.20	3.30	5.20	5.80	7.00	1.85

Table 11. Wilcoxon Test Wine P-Value HSA vs Imputers

Loss (%)	KNN	MICE	Mean	Median	Mode	RF
10	0.00000191	0.00000191	0.00000191	0.00000191	0.00000191	0.00008845
20	0.00000191	0.00000191	0.00000191	0.00008845	0.00000191	0.00000191
30	0.00000191	0.00000191	0.00008845	0.00000191	0.00000191	0.00012026
40	0.00070763	0.00089088	0.00008845	0.00000191	0.00000191	0.02395058
50	1.00000000	0.23051262	0.00000191	0.00000191	0.00000191	0.78412628

Table 12. Precision Test Wine

Loss (%)	Accuracy	Recall	F1-Score	ROC AUC
10	1.0000	1.0000	1.0000	1.0000
20	1.0000	1.0000	1.0000	0.9995
30	0.9815	0.9815	0.9814	0.999
40	0.9444	0.9444	0.9448	0.983
50	0.9444	0.9444	0.9449	0.9851

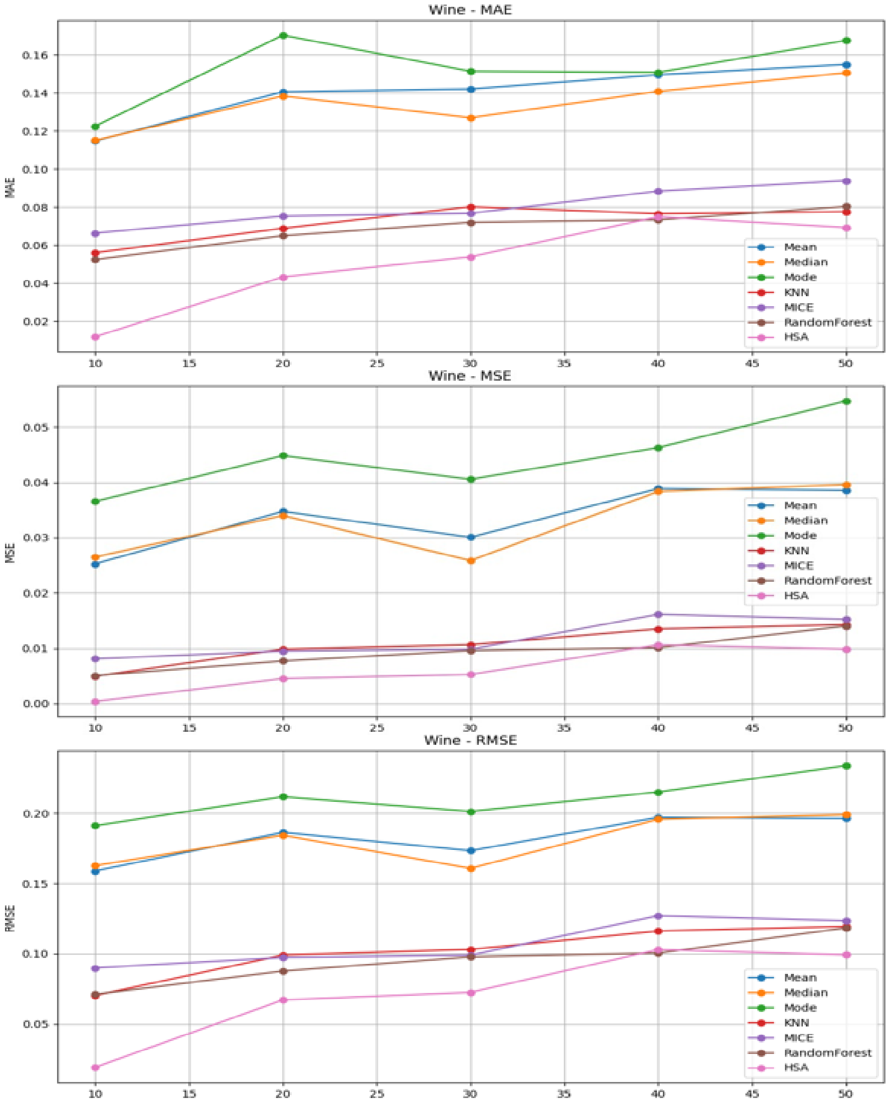

Fig. 8. Wine comparison.

Analysis of the Wine dataset, the Friedman test confirmed a statistically significant difference between the imputation techniques. HSA demonstrated exceptional performance, with average ranks of 1.00 at 10% and 20% loss, and 1.05 at 30%, proving to be a robust and highly reliable tool specifically in low-to-medium data loss and high integrity scenarios. Although its performance faltered at higher losses, Random Forest showed complementary robustness in those cases, reaching a rank of 1.85 at 50% loss. Thanks to the effectiveness of these imputations, the overall predictive model for Wine demonstrates astonishing resilience even with 50% missing data, it maintained extraordinarily high precision metrics; Accuracy and Recall at 0.9444, F1-Score at 0.9449, and ROC AUC at 0.9851, validating HSA high efficacy in preserving data quality.

5.6 Practical Importance

Finally, the robustness and accuracy of the Harmony Search algorithm in imputing missing data, demonstrated in diverse domains such as healthcare, finance, and product quality, transcends purely technical metrics. More complete and reliable data directly results in the construction of more accurate predictive and classificatory models. This, in turn, can lead to more accurate and timely diagnoses in the clinical setting, fairer and more efficient economic and human resource decisions in the financial sector, and more rigorous quality control and product characterization in industry. In this context, HSA emerges as a promising tool to address the challenge of incomplete data, potentially enabling more informed decision-making and greater positive impact in various real-world applications.

6 Conclusions

In this research, we propose a Harmony Search Algorithm to impute missing data in percentages of 10%, 20%, 30%, 40%, and 50%, comparing the proposal with other imputation techniques; Mean, Median, Mode, KNN, MICE, Random Forest.

For classification datasets such as cancer and diabetes, HSA has exhibited re- markable performance. In terms of reduction in imputation error RMSE, although it was occasionally slightly outperformed by more complex machine learning algorithms such as Random Forest or MICE, its ability to preserve the inherent structure of the data was consistently robust. This was evidenced by the preservation of classification metrics; Accuracy, F1-Score, ROC AUC in models trained on HAS imputed data, confirming its effectiveness in preserving crucial discriminative information for downstream predictive tasks. In these scenarios, HSA consistently emerged as a potential option to basic mean, median, or mode imputations.

On the other hand, on the Salary regression dataset, the results indicated that HSA did not always outperform decision tree-based methods or even simpler imputations. This could suggest that, for certain types of regression problems with more direct or less intricate relationships, the search strength of a general metaheuristic like HSA might not translate into a significant advantage over algorithms optimized for such data structures.

In summary, the Harmony Algorithm is established as a valuable metaheuristic for data imputation, demonstrating particular effectiveness in classification contexts where the structural integrity of information and predictive capacity are paramount. Its ability to explore complex search spaces and find estimates of missing values while maintaining data consistency positions it as a robust and viable tool in the field of data preprocessing.

7 Future Work

This research has thoroughly evaluated the performance of the Harmony Algorithm (HSA), a leading bio-inspired metaheuristic, in the task of imputing incomplete data. The findings provide a solid foundation for understanding its applicability to traditional and machine learning methods. Based on these results, several crucial avenues for future research are outlined that will further exploit the potential of HSA and metaheuristics in the field of imputation.

First, and in line with the interest in the applicability of advanced approaches, a priority area for future work is the comprehensive and systematic comparison of HSA with other evolutionary or bio-inspired metaheuristics algorithms and the current techniques from state-of-the-art. This direct comparison, conducted under controlled and replicable experimental conditions, would identify which design features of these heuristics are most advantageous for addressing the specific challenges of data imputation in diverse structures and types of missing values.

A second line of research will focus on the adaptive optimization of HSA internal parameters. Calibration of parameters such as the Harmony Recognition Rate HAR, Pitch Matching Rate PAR, and Number of Harmonies HM is critical for performance. Future studies could employ meta-optimization techniques or automatic parameter tuning algorithms to discover optimal configurations that maximize HSA imputation accuracy on different types of datasets. Furthermore, it is critical to evaluate HSA performance in more complex and demanding data scenarios. This includes imputation on very high-dimensional datasets, the presence of highly nonlinear relationships between features, and, crucially, the ability to handle various missing data patterns, such as Missing Not At Random MNAR, where most traditional methods typically face significant limitations. In these contexts, HSA strength as a metaheuristic, capable of exploring large search spaces and avoiding local optima, could prove to be a competitive advantage.
Finally, the development of hybrid strategies that combine HSA global search capabilities with elements of other imputation algorithms represents a promising avenue. Such hybridizations could enhance the robustness and accuracy of imputation. Furthermore, research into the scalability and computational efficiency of HSA, through its implementation in parallel or distributed processing architectures, will be vital for its effective application in massive volumes of data.

References

Abdulkhaleq, M. T., Rashid, T. A., Alsadoon, A., Hassan, B. A., Mohammadi, M., Abdullah, J. M., Chhabra, A., Ali, S. L., Othman, R. N., Hasan, H. A., Azad, S., Mahmood, N. A., Abdalrahman, S. S., Rasul, H. O., Bacanin, N., & Vimal, S. (2022). Harmony search: Current studies and uses on healthcare systems. In Artificial Intelligence in Medicine (Vol. 131, p. 102348). Elsevier. https://doi.org/10.1016/j.artmed.2022.102348.

Abu Doush, I., Al-Betar, M. A., Awadallah, M. A., Alyasseri, Z. A. A., Makhadmeh, S. N., & El-Abd, M. (2022). Island neighboring heuristics harmony search algorithm for flow shop scheduling with blocking. Swarm and Evolutionary Computation, 74, 101127. https://doi.org/10.1016/j.swevo.2022.101127.

Aljuaid, T., & Sasi, S. (2017, January 18). Proper imputation techniques for missing values in data sets. Proceedings of the 2016 International Conference on Data Science and Engineering, ICDSE 2016. https://doi.org/10.1109/ICDSE.2016.7823957.

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. In SpringerPlus (Vol. 2, Issue 1, pp. 1–17). SpringerOpen. https://doi.org/10.1186/2193-1801-2-222.

Elasra, A. (2022). Multiple Imputation of Missing Data in Educational Production Functions. Computation, 10(4), 49. https://doi.org/10.3390/computation10040049.

Ferri, P., Romero-Garcia, N., Badenes, R., Lora-Pablos, D., Morales, T. G., Gómez de la Cámara, A., García-Gómez, J. M., & Sáez, C. (2023). Extremely missing numerical data in Electronic Health Records for machine learning can be managed through simple imputation methods considering informative missingness: A comparative of solutions in a COVID-19 mortality case study. Computer Methods and Programs in Biomedicine, 242, 107803. https://doi.org/10.1016/j.cmpb.2023.107803.

García-Laencina, P. J., Sancho-Gómez, J. L., Figueiras-Vidal, A. R., & Verleysen, M. (2009). K nearest neighbours with mutual information for simultaneous classification and missing data imputation. Neurocomputing, 72(7–9), 1483–1493. https://doi.org/10.1016/j.neucom.2008.11.026.

Goh, R. Y., Lee, L. S., Seow, H. V., & Gopal, K. (2020). Hybrid harmony search-artificial intelligence models in credit scoring. Entropy, 22(9), 989. https://doi.org/10.3390/e22090989.

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial Intelligence in Medicine, 50(2), 105–115. https://doi.org/10.1016/j.artmed.2010.05.002.

Kabir, G., Tesfamariam, S., Hemsing, J., & Sadiq, R. (2020). Handling incomplete and missing data in water network database using imputation methods. Sustainable and Resilient Infrastructure, 5(6), 365–377. https://doi.org/10.1080/23789689.2019.1600960.

Little, R. J. A., & Rubin, D. B. (2002). Statistical Analysis with Missing Data. In Wiley (Ed.), Statistical Analysis with Missing Data (1st ed.). Wiley. https://doi.org/10.1002/9781119013563.

Memon, S. M. Z., Wamala, R., & Kabano, I. H. (2022). Missing Data Analysis Using Statistical and Machine Learning Methods in Facility-Based Maternal Health Records. SN Computer Science, 3(5), 1–15. https://doi.org/10.1007/s42979-022-01249-z.

Memon, S. M., Wamala, R., & Kabano, I. H. (2023). A comparison of imputation methods for categorical data. Informatics in Medicine Unlocked, 42, 101382. https://doi.org/10.1016/j.imu.2023.101382.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581–592. https://doi.org/10.1093/biomet/63.3.581.

Rubin, D. B., Stern, H. S., & Vehovar, V. (1995). Handling "Don't Know" survey responses: The case of the Slovenian plebiscite. Journal of the American Statistical Association, 90(431), 822–828. https://doi.org/10.1080/01621459.1995.10476580.

Shuai, L. L., Ye, J. H., & Ma, C. (2022). Missing Fault Data Processing Method Based On Improved Harmony Search Algorithm. ICNSC 2022 - Proceedings of 2022 IEEE International Conference on Networking, Sensing and Control: Autonomous Intelligent Systems. https://doi.org/10.1109/ICNSC55942.2022.10004095.

Siddique, N., & Adeli, H. (2015). Harmony Search Algorithm and its Variants. International Journal of Pattern Recognition and Artificial Intelligence, 29(8). https://doi.org/10.1142/S0218001415390012.

Stekhoven, D. J., & Bühlmann, P. (2012). Missforest-Non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45(3), 1–67. https://doi.org/10.18637/jss.v045.i03.

Wells, B. J., Nowacki, A. S., Chagin, K., & Kattan, M. W. (2013). Strategies for Handling Missing Data in Electronic Health Record Derived Data. EGEMs (Generating Evidence & Methods to Improve Patient Outcomes), 1(3), 7. https://doi.org/10.13063/2327-9214.1035.

Zhang, J., & Zhang, P. (2018). A study on harmony search algorithm and applications. Proceedings of the 30th Chinese Control and Decision Conference, CCDC 2018, 736–739. https://doi.org/10.1109/CCDC.2018.8407228.

Zong Woo Geem, Joong Hoon Kim, & Loganathan, G. V. (2001). A New Heuristic Optimization Algorithm: Harmony Search. SIMULATION, 76(2), 60–68. https://doi.org/10.1177/003754970107600201.