A missing data imputation method based on salp swarm algorithm for diabetes disease

Geehan Sabah Hassan[1], Noora Jamal Ali[2], Asma Khazaal Abdulsahib[3], Farah Jasim Mohammed[1], Hassan Muwafaq Gheni[4]
[1]College of Education for Women, University of Baghdad, Baghdad, Iraq
[2]Department of Electronic Technologies, Institute of Medical Technology Al-Mansour, Baghdad, Iraq
[3]College of Education, Ibn Rushd, University of Baghdad, Baghdad, Iraq
[4]Department of Computer Techniques Engineering, Al-Mustaqbal University College, Hillah, Iraq

Article Info

Article history:

Received Aug 6, 2022
Revised Aug 13, 2022
Accepted Aug 28, 2022

Keywords:

Classification
Diabetes disease
Machine learning
Missing values
Salp swarm algorithm

ABSTRACT

Most of the medical datasets suffer from missing data, due to the expense of some tests or human faults while recording these tests. This issue affects the performance of the machine learning models because the values of some features will be missing. Therefore, there is a need for a specific type of methods for imputing these missing data. In this research, the salp swarm algorithm (SSA) is used for generating and imputing the missing values in the pain in my ass (also known Pima) Indian diabetes disease (PIDD) dataset, the proposed algorithm is called (ISSA). The obtained results showed that the classification performance of three different classifiers which are support vector machine (SVM), K-nearest neighbour (KNN), and Naïve Bayesian classifier (NBC) have been enhanced as compared to the dataset before applying the proposed method. Moreover, the results indicated that issa was performed better than the statistical imputation techniques such as deleting the samples with missing values, replacing the missing values with zeros, mean, or random values.

Corresponding Author:

Geehan Sabah Hassan
College of Education for Women, University of Baghdad
Baghdad, Iraq
Email: Jihan.s@coeduw.uobaghdad.edu.iq

1. INTRODUCTION

During data mining (DM) processes, the quality of the considered data determines the quality of its outcome; hence, data pre-processing is an important step towards achieving clean and quality data and determines the success of the mining process. Data pre-processing is the major step in knowledge discovery in database (KDD) process as it decreases data complexity and gives better conditions to subsequent data analysis. Data pre-processing aids in understanding the nature of the data, thereby allowing accurate and efficient data analysis. The next important step of KDD is the data itself. The input data must be prepared in a suitable format and structure that will suit each DM task perfectly. Raw data is not expected to be perfect without pre-processing. Since good DM models usually require well-structured data, the data quality must be improved via thorough data cleansing. The data values must be correct and consistent as missing data is a major problem during DM processes, especially when occurring in large amounts; however, it is not all attributes (instances) with missing values can be removed from the sample [1]–[3]. The problem of data loss is particularly apparent in decision-making processes, especially in online applications where data must be used exactly as it was generated. As a result, computer intelligence techniques such as B neural networks and other pattern recognition approaches have been used in current decision-making procedures. However, decision-making processes cannot advance when some variables are not monitored, and the primary issue is that traditional computational intelligence algorithms cannot successfully handle input data with model views (MVs) or conduct regression or classification tasks [4], [5]. In most applications, finding a solution to the missing data problem is a tiresome effort, and this is not considered in most decision-making tasks. As a result, dealing with concerns relating to lost property necessitates quick and inefficient procedures. This raises the demand for computational and mental resources, such as procedures and theoretical frameworks that can lead to near-completion [6], [7]. Most of the time, inefficient tactics are used since there isn't enough time to identify better ways to cope with lost data at the time of observation, hence ineffective techniques like case deletion are used. Unfortunately, some widely used procedures cause more harm than good by producing biased and incorrect results. The remainder of the paper is laid out as follows: the section 2 introduces missing values and highlights the most important research on diabetic machine learning models. Section 3 outlines the proposed strategy. The pain in my ass (also known Pima) Indian diabetes disease (PIDD) data set and its analysis are presented in the fourth part. Furthermore, the proposed embedding algorithm is assessed. Finally, the fifth section summarizes the proposed algorithm's outcomes and offers suggestions for future research.

– Missing data in medical datasets

Thinking about how the data points were lost in the first place is the simplest technique to deal with lost data. The three processes of missing data are randomly missing, randomly missing, and unignorable [2], [3], [6], [7]. To begin, the phrase "totally missing completely at random" (MCAR) refers to the fact that the data that is missing is not logged at random. While missing in random (MAR) denotes the fact that some data points for specific observations in the data collection are not logged in a random manner. The non-ignorable state type implies that the missing data is dependent on the missing values rather than being random. One of the easy ways of handling MVs is to delete the attributes that contain them from the data set. However, this is not a good method when dealing with data that contains many records with MVs as it will result to bias during the inference. In the presence of MVs, data analysis is a difficult task as it will expose the analyst to serious problems; in fact, if handled in a non-professional manner, it can lead to bias during data analysis and cause ambiguous conclusions; it can also limit the generalizability of the study outcome [8], [9].

– Nature-inspired algorithms and salp swam algorithm

The ability of nature-inspired metaheuristics to provide solutions to modern optimization problems has attracted much research interest, especially their performance on, nomadic people (NP-hard) optimization problems, such as the travelling salesman problem and feature selection [10]–[13]. One of the nature-inspired metaheuristics commonly used in solving difficult optimizations tasks is the particle swarm optimization (PSO) which was first developed in 1995 by Eberhard and Kennedy [14]. The PSO was inspired by the swarm behavior of natural species, such as the flocking of birds and the schooling of fish. The PSO has found application in different optimization field where it has performed excellently. The firefly algorithm (FA) is another metaheuristic that has demonstrated good performance in may applications; it was developed by Yang, (2009). In these multiagent frameworks, the search mechanisms are governed by efficient local search, randomization, and optimal solution selection. However, the randomization normally uses uniform or Gaussian distribution. Different types of nature-inspired algorithms have been proposed during the past two decades. Most of them inspired from a biological organism or social life, such as artificial bee colony (ABC), ant colony optimizer (ACO), FA, grey wolf optimizer (GWO), ant lioner optimizer (ALO), and nomadic people optimizer (NPO) [15]–[20]. NIAs have played a great role for solving different types of optimization problems, medical case studies [21]–[23], engineering [24]–[36], energy [37]–[48], and information security [49]–[51]. Salp swarm algorithm (SSA) is a recent nature inspired algorithm, which is inspired from the cylindrical jellyfishes-like creatures which are belong to Salpidae [52]–[55]. These creatures are moved by pushing the water backward in order to move forward. The swarming behavior of these Salps inspired the authors to propose SSA for solving the difficult optimization problems. Figures 1 portray the main shape and salp chain in SSA. This chain is formulated mathematically by dividing the population into two dfferent groups leader and followers. Where the leader is responsible for leading the other followers to better positions. The position of the leader is updated as (1):

$$x_j^1 = \begin{cases} F_j + c_1\left(\left(ub_j - lb_j\right)c_2 + lb_j\right) & c_3 \geq 0 \\ F_j - c_1\left(\left(ub_j - lb_j\right)c_2 + lb_j\right) & c_3 < 0 \end{cases} \tag{1}$$

Where x_j^1 denotes the position of the leader salp in the search space, while ub and lb denote the upper and lower bounderies, finally c_1, c_2 and c_3 represent three random numbers.

2. METHOD

In this section, the proposed imputation algorithm based on SSA is presented. The section is divided into two sub-sections. In the first subsection, the proposed imputation algorithm in general, while the second subsection explains the SSA algorithm used in this study in details.

2.1. The proposed imputation algorithm

In the process of imputing or estimating the missing values in the targeted case study, the imputation algorithm based is designed for this purpose. The proposed algorithm consists of several stages as follows. Figure 1 shows the block diagram of the proposed algorithm.

a. Stage 1 dataset preparation. The proposed algorithm's preparation of the data set is the first step. It entails the following three processes for reading and preparing the data set: i) step 1 is to read the data set, ii) step 2 convert the data from its original (.xlsx) format to a comma separated values (.csv) file, which can be read by practically any current computer language, and iii) step 3 normalize the dataset to a constant range [0,1] using the minmax technique stated as in (2).

$$N_v = \frac{X_v - Min}{Max - Min} \tag{2}$$

Where N_v represents the normalized value, while X_v represents the original value. Min and Max denote the maximum and minimum values of a specific feature respectively.

b. Stage 2 the inputs. In this stage, the algorithmic parameters such as the size of the swarm, the maximum number of iterations, and other SSA controlling variables are entered.

c. Stage 3 determine the positions of the missing values. In order to fill the missing values, the positions of these values should be determined. In addition, the number of these missing values is determined as well. Based on the previous two information, the solution representation for each solution in the swarm is structured.

d. Stage 4 SSA implementation. In this stage, the SSA is executed to search for the best values, which replace the missing values in the dataset. The main steps of SSA are given in the next subsection.

e. Stage 5 evaluation. In this stage, the best solution obtained using SSA is evaluated in terms of classification accuracy, error rate, sensitivity, and specificity.

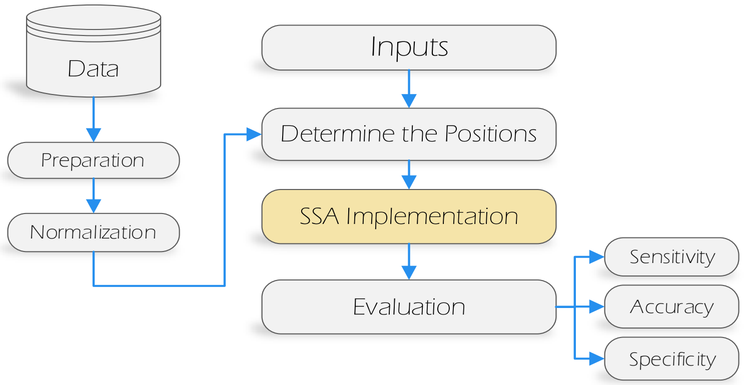

Figure 1. The block diagram of the proposed algorithm

2.2. SSA for missing values estimation

The SSA method is designed in such a way that the following stages are required to be completed. Although not all these steps are required, they help to implement the technique more efficiently.

a. In the parameter vector [S.S MaxItr], set the initial settings. The upper bound (UB) and lower bound (LB) values limit the search space. UB and LB values are assigned based on the case study, while swarm size (S.S) and maximum number of iterations are set according to different conditions.

b. Initialization: generate a random position for each solution in the swarm, via the uniform distribution method as (3):

$$X_i = (UB - LB) \times Rand(0,1) + LB \tag{3}$$

Where $Rand$ is a randomization method, which generates a random value in range [0,1].

c. Fitness function: in order to evaluate each generated–or estimated–solution via the classification accuracy (A). The accuracy is generated via K-fold cross validation where K is equal to 5. Three different classification models are used in this study, K-nearest neighbour (KNN), support vector machine (SVM), and Naïve bayesian classifier (NBC).

d. Position updating move the followers X_i towards another leaders X_i^1 with higher intensity) $I(F_i) < I(F_j)$ via (4):

$$x_j^i = \frac{1}{2}\left(x_j^i + x_j^{i-1}\right) \tag{4}$$

Where $i > 2$ and x_j^i denote the followers' positions in the j^{th} dimensions. Check the boundaries limits: check whether the values obtained in the new position of the solution is within the search space or not, as in (5):

$$x_j^i = \begin{cases} LB\ If\ x_j^i < LB \\ UB\ If\ F_i.x_j^i > UB \end{cases} \tag{5}$$

Then, x_j^i s evaluated using the fitness function explained in step c.

e. Sorting and ranking: after updating the positions of all fireflies, the swarm is sorted and ranked based on the fitness value. Obtain the leader (X_{Best}) value from the swarm (which will always be the topmost value after sorting). Compare every value of the X with itself.

f. Stop condition: the first and second steps are executed only one time, while the rest steps (c-f) are iterated for t times. Meaning that the algorithm checks t if it is still less than $MaxItr$–which has been identified in the first step–then go to step d. Otherwise, exit the loop and return the last F_{Best}.

2.3. SSA for missing values estimation

The data was first gathered by the national institute of diabetes and digestive and kidney diseases. During the investigation, the World Health Organization's (WHO) recommendations were followed. Females must be at least 21 years old and of Pima native American descent to participate in this study. This data set has already been used by multiple researchers to develop classification algorithms; thus, it was chosen for this study so that it could be compared to other current PID diagnosis investigations. This data set contains 768 examples, each with its own set of eight characteristics. Table 1 lists all the features in this data set, along with their numerical values.

The last value, a binary, was used for the classification task; it was partitioned into 2 classes which are "class zero (non-diabetic) and class one (diabetic)". The first 8 features in the dataset served as the input while the last value served as the ground truth. There are a total number of 268 diabetic cases (34.90%) in the dataset while non-diabetic cases accounted for 65.10% (500 cases). The missing data in most of medical case studies is a standard issue, for two main reasons. First, some of the medical tests are above the budget of the patients so they cannot afford them. Second, sometimes the values were not recorded correctly due to the time constraints. These missing values may affect on the classification performance. PIMA dataset is also associated with a large percentage of missing data as depicted in Table 2. All the features contain missing values, except the first feature where there are no missing values in it.

Table 1. The features set in the dataset

F	Name	Type
1	No. Of times pregnant	Numeric
2	Plasma glucose concentration	Numeric
3	Diastolic blood pressure	Numeric (mmH_g)
4	Triceps skin fold thickness	Numeric (mm)
5	2 hours serum insulin	Numeric ($\mu U/ml$)
6	Body mass index	Numeric (kg/m^2)
7	Diabetes pedigree function	Numeric
8	Age	Numeric (years)

Table 2. Information about missing values in the dataset

	Name	Type
1	No of times pregnant	-
2	Plasma glucose concentration	5
3	Diastolic blood pressure	35
4	Triceps skin fold thickness	227
5	2 hours serum insulin	374
6	Body mass index	11
7	Diabetes pedigree function	1
8	Age	63

3. RESULTS AND DISCUSSION

3.1. Experimental settings

To evaluate the performance of proposed imputation algorithm, a set of experiments should be implemented. The evaluation process consists of several experiments, each experiment consists different test settings. The imputation algorithm has been written and executed using MATLAB, version 2018b, and implemented in the environment of Windows 10 with CPU 2.6 GH-64 bit, and RAM 8 GB. On the other hand, the settings of the experiments depend mainly on the structural parameters, which are: number of iterations (ITR) and the number of solutions in the swarm (N). In order to validate the effect of these two parameters on the performance of the algorithm, several values of each one is implemented, as follows:

a. Case 1 based on N: changing the number of solutions has an impact on the performance of any nature optimization algorithm, sometimes, the large size of N enhances the performance, however this may affect on the speed of the algorithm. Therefore, to determine the best N as much as possible, several tests are performed, $N = \{10,15,20,30\}$.
b. Case 2 based on ITR. The number of iterations has another impact on the performance of the optimization algorithms. To determine the best possible ITR, several tests are performed where $ITR = \{25, 50, 100, 200\}$.
c. Case 3 based on classifier. As explained in the previous section, the fitness function of proposed imputation algorithm depends on three different classifiers. In other words, there three different versions of the proposed imputation algorithm, imputation SSA with KNN (SSA-KNN), imputation SSA with SVM (SSA-SVM), and imputation SSA with NBC (SSA-NBC).

The settings of the tests can be summarized in Table 3, each test was executed 10 run times. The obtained results of each test are:

a. Beginning accuracy ($B.Acc$). Represents the obtained accuracy based on the original dataset with missing values.
b. K-fold cross validation ($CV.Acc$). Represents the obtained accuracy using the proposed imputation algorithm.
c. Original holdout accuracy ($OR_c.Acc$). Represents the obtained accuracy based on different classifiers and the original dataset, when the dataset is divided into training set (65%) and testing set (35%).
d. Optimized holdout accuracy ($OP_c.Acc$). Represents the obtained accuracy based on different classifier and the enhanced dataset, when the enhanced dataset is divided into training set (65%) and testing set (35%).

Table 3. Tests settings

Test	N	ITR
T_1	10	25
T_2	10	50
T_3	10	100
T_4	10	200
T_5	15	25
T_6	15	50
T_7	15	100
T_8	15	200
T_9	20	25
T_{10}	20	50
T_{11}	20	100
T_{12}	20	200
T_{13}	30	25
T_{14}	30	50
T_{15}	30	100
T_{16}	30	200

3.2. Obtained results

3.2.1. Results obtained using KNN as a fitness function

In this part, KNN classification model is used for measuring the fitness of each solution in the swarm. The results of this experiments were obtained based on all $[T_1 - T_{16}]$ mentioned in Table 3, where each test has been implemented ten times. The average results of each test are summarized in Figures 2 and 3. Figure 2 illustrates the results obtained using cross validation of the original and the optimized dataset. While the Figure 3 illustrates the comparison results obtained using holdout results of three classifier.

It can be seen from the Figures 2 and 3 that the proposed imputation algorithm based on KNN model as a fitness function has enhanced the results. In other words, the proposed algorithm estimated and filled the missing values in PIDD dataset with values better for the prediction and classification process. In addition, it can be seen in the Figure 3 that KNN model has the best performance when it was used for the validation of the generated dataset, as compared to the others two classifiers. However, SVM has a very close performance to KNN, while the performance of NBC was the worst.

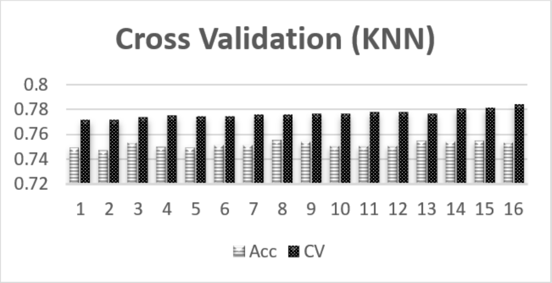

Figure 2. Comparison between average results of the obtained accuracies

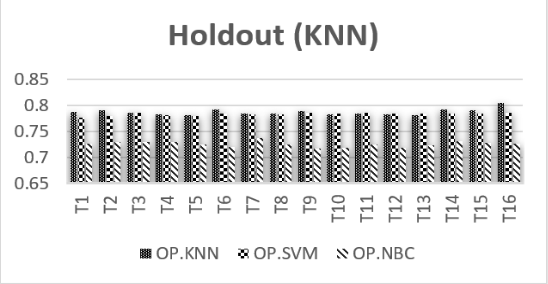

Figure 3. Comparison between the average accurizes using holdout

3.2.2. Results obtained using SVM as a fitness function

In this experiment, SVM classification model is used for evaluating the solutions in the swarm. The experiments have been validated based on the test mentioned in Table 3. Ten run times have been implemented, and the average of these runs for each test is presented in Figures 4 and 5. The figures showed different results as compared to the previous experiment, as the SVM in Figure 5 showed a superior performance. SVM was ranked first, while NBC ranked third and attained the worst performance just like the previous experiment. On the other hand, the comparison between the obtained results in this experiment were much better than the results obtained using the original dataset with missing values (see Figure 6).

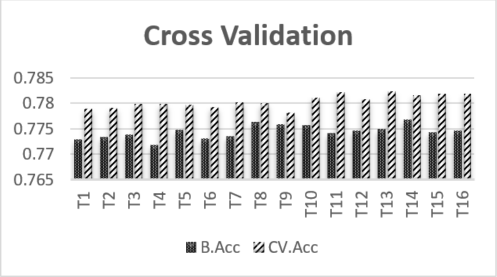

Figure 4. Comparison between average results of the obtained accuracies

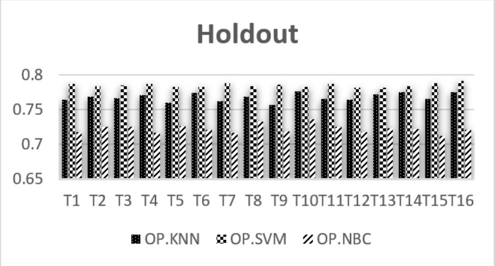

Figure 5. Comparison between the average accurizes using holdout

3.2.3. Results obtained using NBC as a fitness function

In the final experiment, NBC classifier is used for evaluating the generated datasets or the solutions in the swarm. The algorithm has been implemented ten run times based on the tests mentioned in Table 3. While the average of these runs is presented in the Figures 6 and 7.

Figures 6 and 7 show that NBC has the worst performance as compared to the other three classifiers. Moreover, the comparison between the accuracy obtained based on the dataset filled using the proposed imputation algorithm were better than the original dataset in all tests. Therefore, NBC enhances the performance of the proposed algorithm in general, but with worse results as compared to the other classifiers.

In the previous subsection, it was clear that the proposed SSA imputation algorithm based on all classifiers was able to handle the problem of the missing values in the PIDD dataset. Even the worst performance of NBC classifier was better than the best performance of all tests based on the original dataset. Moreover, there are three observations can be summarized as follows:

a. When KNN used as a fitness function, the holdout validation experiments showed that KNN classifier based on the 35% testing set was better than the other classifiers. However, KNN ranked the second position when SVM or NBC used as fitness functions. In general, SVM showed the best performance due to the sequential minimum optimization (SMO) algorithm for tuning the C and γ in the RBF kernel function.
b. All the results obtained using SVM and KNN were more than 77%, while the results obtained using NBC were in range [70% and 75%].
c. It can be seen from cross-validation experiments, that the results were better when the number of the solutions–or the swarm size–are increased (i.e., tests $T_{10} - T_{16}$). Meaning that the number of solutions has an obvious impact on the searching performance of FA. On the other hand, the number of iterations (ITR) has a less impact on SSA.

The evaluation measurements other than the classification accuracy (explained in section 4.2) are presented in the Table 4. In the previous subsections, the proposed SSA imputation algorithm based on different classifiers was evaluated. The evaluation process depended mainly on sixteen tests, and two validation methods: cross validation and holdout. In this section, the proposed imputation algorithm is benchmarked and compared against four well-known imputation approaches on PIDD dataset. These approaches are:

a. A_1: removing the entire row with the missing values or attributes. This approach leads to decrease the amount of training data which may affect on the classification process.
b. A_2: replacing the missing values with zeros. In some cases, this could be a good solution, however, the value of zero may also affect on the classification process when the classification model is trained based on modified data.
c. A_3: replacing the missing values by the average or mean of the other values of the attribute. In most cases, this approach is better than the previous approaches because the generated values depend mainly on the other values of the same attribute.
d. A_4: replacing the missing values by random values in the range [0,1]. However, this method may generate values effects on the classification models. In other words, the values may have some noise, or change the distribution of the samples.

The approaches above have been integrated with three classifiers used in this study and executed ten run times. Then, the best, the mean, the standard deviation was recorded. Table 5 presents the comparison of the four approaches against immunofloresensi assay (IFA)-KNN, IFA-SVM, and IFA-NBC. In addition, the mentioned approach, the classification accuracy of the dataset without implemented any imputation approach is also presented.

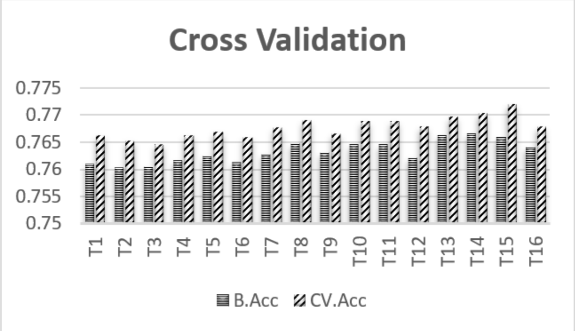

Figure 6. Comparison between average results of the obtained accuracies

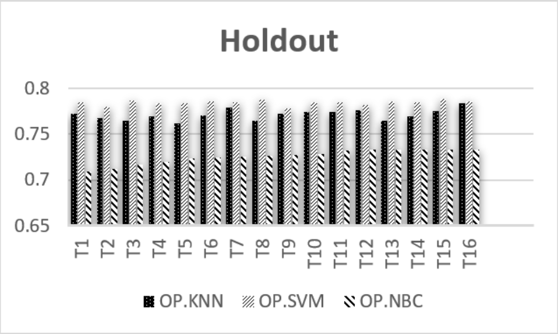

Figure 7. Comparison between the average accurizes using holdout

Table 4. Evaluation measurements

Algorithm	Sensitivity	Specificity	MSE
IFA-KNN	0.520583	0.862484	0.159723
IFA-SVM	0.532583	0.873494	0.154563
IFA-NBC	0.489166	0.762887	0.161783

Table 5. Comparison against other imputation approaches

Classifier	Approach	Best	Mean	Std. Dev
KNN	$Original$	0.75008	0.75008	0
	A_1	0.73641	0.73122	0.24782
	A_2	0.75421	0.75231	0.21412
	A_3	0.76822	0.76741	0.19321
	A_4	0.76025	0.75942	0.20411
	IFA	0.794153	0.78421	0.18695
SVM	$Original$	0.77935	0.77935	0
	A_1	0.75982	0.75611	0.22782
	A_2	0.76724	0.76514	0.21842
	A_3	0.77942	0.77862	0.20142
	A_4	0.77834	0.77285	0.19782
	IFA	0.790758	0.78793	0.002744
NBC	$Original$	0.70414	0.70414	0
	A_1	0.69842	0.69215	0.25413
	A_2	0.69624	0.69342	0.24821
	A_3	0.70128	0.70101	0.20421
	A_4	0.70431	0.70321	0.20142
	IFA	0.73348	0.72569	0.00754

It is obvious that the proposed imputation algorithm obtained the highest results as compared to the other approaches. A_1 with all classifiers attained the worst position, because in this approach the many samples were deleted from the dataset, which decreases the training set. The second approach A_2 had almost the same performance with slightly better results due to using zero as the value for all missing data. On the other hand, the third and fourth approaches A_3 and A_4 were better than the previous approaches because of filling the missing data with mean or random values. The generated values are better than using zero, or removing the sample with missing data, because at least these approaches filled them. Moreover, the best attained results were obtained using IFA-KNN, however, IFA-SVM has better average results. The standard deviation proofed that both of IFA-SVM and IFA-NBC are more stable than IFA-KNN.

4. CONCLUSION

The missing data or missing values is an issue with most of the medical datasets. It occurred for two main reasons: a) the expense of the medical tests and b) the fault of recording all the features for time constraints or human faults. Therefore, there is a need for a specific process for reparation these missing data, this process is called "Imputation". In this research, SSA is used as an imputation method. Three different classifiers are used for evaluating the generated missing values, these classifiers are: KNN, SVM, and NBC. The proposed imputation algorithm has been evaluated based two main experiments. First, using cross validation with 5 folds, while in the second experiment, the algorithm has been evaluated using holdout validation method, where the generated dataset was divided into training set (65%) and testing set (35%). The results showed that the proposed imputation algorithm could estimate the missing values in PIDD and enhanced the classification accuracy for all classifiers. SVM showed ranked the best, while NBC ranked the worst.

ACKNOWLEDGEMENTS

The authors are grateful to the Iraqi Ministry of Higher Education and Scientific Research (MOHESR) for technically supporting the current research.

REFERENCES

[1] Sowmya R and Suneetha K R, "Data mining with big data," in *2017 11th International Conference on Intelligent Systems and Control (ISCO)*, Jan. 2017, pp. 246–250, doi: 10.1109/ISCO.2017.7855990.

[2] K. Yothapakdee, S. Charoenkhum, and T. Boonnuk, "Improving the efficiency of machine learning models for predicting blood glucose levels and diabetes risk," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 27, no. 1, pp. 555–562, Jul. 2022, doi: 10.11591/ijeecs.v27.i1.pp555-562.

[3] W. Cardoso *et al.*, "Modeling of artificial neural networks for silicon prediction in the cast iron production process," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 2, pp. 530–538, Jun. 2022, doi: 10.11591/ijai.v11.i2.pp530-538.

[4] N. A. M. Aseri *et al.*, "Comparison of meta-heuristic algorithms for fuzzy modelling of Covid-19 illness' severity classification," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 1, pp. 50–64, Mar. 2022, doi: 10.11591/ijai.v11.i1.pp50-64.

[5] S. M. Mahdi and O. F. Lutfy, "Control of a servo-hydraulic system utilizing an extended wavelet functional link neural network based on sine cosine algorithms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 2, pp. 847–856, Feb. 2022, doi: 10.11591/ijeecs.v25.i2.pp847-856.

[6] U. A. Badawi, "Fish classification using extraction of appropriate feature set," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 3, pp. 2488–2500, 2022, doi: 10.11591/ijece.v12i3.pp2488-2500.

[7] W. D. Abdullah, A. A. Jasim, and L. R. Hazim, "Predictions and visualization for confirmed, recovered and deaths Covid-19 cases in Iraq," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 2, p. 1197, May 2022, doi: 10.11591/ijeecs.v26.i2.pp1197-1205.

[8] A. R. T. Donders, G. J. M. G. v. d. Heijden, T. Stijnen, and K. G. M. Moons, "Review: a gentle introduction to imputation of missing values," *Journal of Clinical Epidemiology*, vol. 59, no. 10, pp. 1087–1091, Oct. 2006, doi: 10.1016/j.jclinepi.2006.01.014.

[9] S. J. Choudhury and N. R. Pal, "Imputation of missing data with neural networks for classification," *Knowledge-Based Systems*, vol. 182, pp. 1–30, Oct. 2019, doi: 10.1016/j.knosys.2019.07.009.

[10] E. N. Sholikhah, N. A. Windarko, and B. Sumantri, "Tunicate swarm algorithm based maximum power point tracking for photovoltaic system under non-uniform irradiation," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 5, pp. 4559–4570, Oct. 2022, doi: 10.11591/ijece.v12i5.pp4559-4570.

[11] S. Abdulaziz, G. Atlam, G. Zaki, and E. Nabil, "Cuckoo search algorithm and particle swarm optimization based maximum power point tracking techniques," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 2, pp. 605–616, May 2022, doi: 10.11591/ijeecs.v26.i2.pp605-616.

[12] M. Alzaqebah *et al.*, "Hybrid feature selection method based on particle swarm optimization and adaptive local search method," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 3, pp. 2414–2422, Jun. 2021, doi: 10.11591/ijece.v11i3.pp2414-2422.

[13] A. A. A. -Arabo and R. Z. Alkawaz, "Implementation of combined new optimal cuckoo algorithm with a gray wolf algorithm to solve unconstrained optimization nonlinear problems," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 3, pp. 1582–1589, Sep. 2020, doi: 10.11591/ijeecs.v19.i3.pp1582-1589.

[14] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, 1995, pp. 39–43, doi: 10.1109/MHS.1995.494215.

[15] X.-S. Yang, "Firefly algorithms for multimodal optimization," in *SAGA 2009*, Stochastic., Berlin, Heidelberg: Springer, 2009, pp. 169–178.

[16] M. Dorigo, V. Maniezzo, and A. Colorni, "Ant system: optimization by a colony of cooperating agents," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 26, no. 1, pp. 29–41, Feb. 1996, doi: 10.1109/3477.484436.

[17] D. Karaboga, "An idea based on honey bee swarm for numerical optimization," *Computer Science*, pp. 1–10, 2005.

[18] S. Q. Salih and A. A. Alsewari, "A new algorithm for normal and large-scale optimization problems: nomadic people optimizer," *Neural Computing and Applications*, vol. 32, no. 14, pp. 10359–10386, Jul. 2020, doi: 10.1007/s00521-019-04575-1.

[19] S. Mirjalili, "The ant lion optimizer," *Advances in Engineering Software*, vol. 83, pp. 80–98, May 2015, doi: 10.1016/j.advengsoft.2015.01.010.

[20] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Advances in Engineering Software*, vol. 69, pp. 46–61, Mar. 2014, doi: 10.1016/j.advengsoft.2013.12.007.

[21] N. J. Ali Kadhim and J. Kadhim Abed, "Enhancing the prediction accuracy for cardiotocography (CTG) using firefly algorithm and naive bayesian classifier," *IOP Conference Series: Materials Science and Engineering*, vol. 745, no. 1, pp. 1–11, 2020, doi: 10.1088/1757-899X/745/1/012101.

[22] S. F. Jabar, "A classification model on tumor cancer disease based mutual information and firefly algorithm," *Periodicals of Engineering and Natural Sciences (PEN)*, vol. 7, no. 3, pp. 1152–1162, Sep. 2019, doi: 10.21533/pen.v7i3.656.

[23] N. Sureja, B. Chawda, and A. Vasant, "A novel salp swarm clustering algorithm for prediction of the heart diseases," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 1, pp. 265–272, Jan. 2022, doi: 10.11591/ijeecs.v25.i1.pp265-272.

[24] A. Malik *et al.*, "Pan evaporation estimation in Uttarakhand and Uttar Pradesh States, India: validity of an integrative data intelligence model," *Atmosphere*, vol. 11, no. 6, pp. 1–25, May 2020, doi: 10.3390/atmos11060553.

[25] A. Malik *et al.*, "The implementation of a hybrid model for hilly sub-watershed prioritization using morphometric variables: case study in India," *Water*, vol. 11, no. 6, pp. 1–19, May 2019, doi: 10.3390/w11061138.

[26] S. Q. Salih, A. A. Alsewari, and Z. M. Yaseen, "Pressure vessel design simulation," in *Proceedings of the 2019 8th International Conference on Software and Computer Applications*, Feb. 2019, pp. 120–124, doi: 10.1145/3316615.3316643.

[27] S. Q. Salih *et al.*, "Integrative stochastic model standardization with genetic algorithm for rainfall pattern forecasting in tropical and semi-arid environments," *Hydrological Sciences Journal*, vol. 65, no. 7, pp. 1145–1157, May 2020, doi: 10.1080/02626667.2020.1734813.

[28] M. H. Hassan, S. Kamel, S. Q. Salih, T. Khurshaid, and M. Ebeed, "Developing chaotic artificial ecosystem-based optimization algorithm for combined economic emission dispatch," *IEEE Access*, vol. 9, pp. 51146–51165, 2021, doi: 10.1109/ACCESS.2021.3066914.

[29] S. Q. Salih, M. Habib, I. Aljarah, H. Faris, and Z. M. Yaseen, "An evolutionary optimized artificial intelligence model for modeling scouring depth of submerged weir," *Engineering Applications of Artificial Intelligence*, vol. 96, pp. 1–13, Nov. 2020, doi: 10.1016/j.engappai.2020.104012.

[30] B. Ahmadi and R. Çağlar, "Determining the Pareto front of distributed generator and static VAR compensator units placement in distribution networks," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 4, pp. 3440–3453, Aug. 2022, doi: 10.11591/ijece.v12i4.pp3440-3453.

[31] M. H. M. Ali, M. M. S. Mohamed, N. M. Ahmed, and M. B. A. Zahran, "Comparison between P&O and SSO techniques based MPPT algorithm for photovoltaic systems," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 1, pp. 32–40, 2022, doi: 10.11591/ijece.v12i1.pp32-40.

[32] E. T. Yassen, M. Ayob, A. A. Jihad, and M. Z. A. Nazri, "A self-adaptation algorithm for quay crane scheduling at a container terminal," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 4, pp. 919–929, 2021, doi: 10.11591/IJAI.V10.I4.PP919-929.

[33] S. Nayak, S. K. Kar, and S. S. Dash, "Combined fuzzy PID regulator for frequency regulation of smart grid and conventional power systems," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 24, no. 1, pp. 12–21, Oct. 2021, doi: 10.11591/ijeecs.v24.i1.pp12-21.

[34] M. Al gabalawy, R. M. Hossam, S. A. Hussien, and N. S. Hosny, "Switched capacitor based multi-level boost inverter for smart grid applications," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 5, pp. 3772–3781, Oct. 2021, doi: 10.11591/ijece.v11i5.pp3772-3781.

[35] L. H. Abood and B. K. Oleiwi, "Design of fractional order PID controller for AVR system using whale optimization algorithm," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 3, pp. 1410–1418, Sep. 2021, doi: 10.11591/ijeecs.v23.i3.pp1410-1418.

[36] H. A. -Mawgoud, S. Kamel, S. Q. Salih, and A. S. Alghamdi, "Optimal integration of capacitor and PV in distribution network based on nomadic people optimizer," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 3, pp. 1237–1248, Sep. 2021, doi: 10.11591/ijeecs.v23.i3.pp1237-1248.

[37] A. Q. Mohammed, K. A. A. -Anbarri, and R. M. Hannun, "Optimal combination and sizing of a stand–alone hybrid energy system using a nomadic people optimizer," *IEEE Access*, vol. 8, pp. 200518–200540, 2020, doi: 10.1109/ACCESS.2020.3034554.

[38] U. Beyaztas, S. Q. Salih, K.-W. Chau, N. A. -Ansari, and Z. M. Yaseen, "Construction of functional data analysis modeling strategy for global solar radiation prediction: application of cross-station paradigm," *Engineering Applications of Computational Fluid Mechanics*, vol. 13, no. 1, pp. 1165–1181, Jan. 2019, doi: 10.1080/19942060.2019.1676314.

[39] P. N. Vinh, B. H. Dinh, V.-D. Phan, H. D. Nguyen, and T. T. Nguyen, "Minimize electricity generation cost for large scale wind-thermal systems considering prohibited operating zone and power reserve constraints," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 3, pp. 1905–1911, Jun. 2021, doi: 10.11591/ijece.v11i3.pp1905-1911.

[40] Z. M. Yaseen *et al.*, "Laundry wastewater treatment using a combination of sand filter, bio-char and teff straw media," *Scientific Reports*, vol. 9, no. 1, p. 18709, Dec. 2019, doi: 10.1038/s41598-019-54888-3.

[41] H. Tao *et al.*, "Global solar radiation prediction over North Dakota using air temperature: Development of novel hybrid intelligence model," *Energy Reports*, vol. 7, pp. 136–157, Nov. 2021, doi: 10.1016/j.egyr.2020.11.033.

[42] H. Tao *et al.*, "A newly developed integrative bio-inspired artificial intelligence model for wind speed prediction," *IEEE Access*, vol. 8, pp. 1–9, 2020, doi: 10.1109/ACCESS.2020.2990439.

[43] A. A. -Gizi, A. H. Miry, and M. A. Shehab, "Optimization of fuzzy photovoltaic maximum power point tracking controller using chimp algorithm," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 5, pp. 4549–4558, Oct. 2022, doi: 10.11591/ijece.v12i5.pp4549-4558.

[44] N. A. Windarko *et al.*, "Hybrid photovoltaic maximum power point tracking of Seagull optimizer and modified perturb and observe for complex partial shading," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 5, pp. 4571–4585, Oct. 2022, doi: 10.11591/ijece.v12i5.pp4571-4585.

[45] E. I. E. -Sayed, M. M. A. -Gazzar, M. S. Seif, and A. M. A. Soliman, "Energy management of renewable energy sources incorporating with energy storage device," *International Journal of Power Electronics and Drive Systems (IJPEDS)*, vol. 13, no. 2, pp. 883–899, Jun. 2022, doi: 10.11591/ijpeds.v13.i2.pp883-899.

[46] M. A. E. -Dabah, R. A. E. -Sehiemy, M. A. Ebrahim, Z. Alaas, and M. M. Ramadan, "Identification study of solar cell/module using recent optimization techniques," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 2, pp. 1189–1198, Apr. 2022, doi: 10.11591/ijece.v12i2.pp1189-1198.

[47] M. Traore, A. Ndiaye, and S. Mbodji, "A comparative study of meta-heuristic and conventional optimization techniques of grid connected photovoltaic system," *International Journal of Power Electronics and Drive Systems (IJPEDS)*, vol. 12, no. 4, pp. 2492–2500, Dec. 2021, doi: 10.11591/ijpeds.v12.i4.pp2492-2500.

[48] M. Zegrar, M. H. Zerhouni, M. T. Benmessaoud, and F. Z. Zerhouni, "Design and implementation of an I-V curvetracer dedicated to characterize PV panels," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 3, p. 2011, Jun. 2021, doi: 10.11591/ijece.v11i3.pp2011-2018.

[49] H. S. Alhadawi, S. Q. Salih, and Y. D. Salman, "Chaotic particle swarm optimization based on meeting room approach for designing bijective s-boxes," in *Proceedings of International Conference on Emerging Technologies and Intelligent Systems*, Cham: Springer, 2022, pp. 331–341.

[50] K. Z. Zamli, A. Kader, F. Din, and H. S. Alhadawi, "Selective chaotic maps Tiki-Taka algorithm for the S-box generation and optimization," *Neural Computing and Applications*, vol. 33, no. 23, pp. 16641–16658, Dec. 2021, doi: 10.1007/s00521-021-06260-8.

[51] A. H. Zahid *et al.*, "A novel construction of dynamic S-box with high nonlinearity using heuristic evolution," *IEEE Access*, vol. 9, pp. 67797–67812, 2021, doi: 10.1109/ACCESS.2021.3077194.

[52] S. Mirjalili, A. H. Gandomi, S. Z. Mirjalili, S. Saremi, H. Faris, and S. M. Mirjalili, "Salp swarm algorithm: A bio-inspired optimizer for engineering design problems," *Advances in Engineering Software*, vol. 114, pp. 163–191, Dec. 2017, doi: 10.1016/j.advengsoft.2017.07.002.

[53] D. Potnuru, T. S. L. V. Ayyarao, L. V. S. Kumar, Y. V. P. Kumar, D. J. Pradeep, and C. P. Reddy, "Salp swarm algorithm based optimal speed control for electric vehicles," *International Journal of Power Electronics and Drive Systems (IJPEDS)*, vol. 13, no. 2, pp. 755–763, Jun. 2022, doi: 10.11591/ijpeds.v13.i2.pp755-763.

[54] S. Salhi, D. Naimi, A. Salhi, S. Abujarad, and A. Necira, "A novel hybrid approach based artificial bee colony and salp swarm algorithms for solving ORPD problem," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 3, pp. 1825–1837, 2021, doi: 10.11591/ijeecs.v23.i3.pp1825-1837.

[55] M. A. -Shabi, C. Ghenai, M. Bettayeb, F. F. Ahmad, and M. E. H. Assad, "Estimating PV models using multi-group salp swarm algorithm," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 2, pp. 398–406, Jun. 2021, doi: 10.11591/ijai.v10.i2.pp398-406.

BIOGRAPHIES OF AUTHORS

Geehan Sabah Hassan received the B.Sc. degree in software engineering from Al-Rafidain University College in 2000, M.Sc. Degree in computer science from University Science Malaysia (USM) in 2014. Her interesting field of research has been concentrated in data mining, information retrieval and metaheuristics, algorithms. She is now working as an assistant teacher in Baghdad University. She can be contacted at email: Jihan.s@coeduw.uobaghdad.edu.iq.

Noora Jamal Ali received the B.Eng. degree in Medical Equipment Technology Engineering, Department of Medical Equipment Technology Engineering, College of Electrical and Electronic Technology, Middle Technical University, Iraq in 2009 and M.S. degree Medical Equipment Technology Engineering, Department of Medical Equipment Technology Engineering, College of Electrical and Electronic Technology, Middle Technical University, Iraq. She is currently Head of the Department of Electronic Technologies. Her research interests include power electronics, medical equipment, medical equipment electronics, artificial intelligence. She can be contacted at email: noora-jamal@mtu.edu.iq.

Asma Khazaal Abdulsahib holds a bachelor's degree in Software Engineering from Al-Rafidain College, Baghdad, Iraq and a master's degree from Utara University, Malaysia, in Artificial Intelligence. She has several papers published in Scopus journals. The area of research is data mining, machine learning, clustering algorithms, and IT applications. Currently. She works as an administrator for computers at the University of Baghdad, and she teach undergraduate students. She supervises a number of research studies for graduate students, and I have evaluated a number of research papers for journals located within ISI. She is currently a Ph.D. student in the field of artificial intelligence. She can be contacted at email: asma.khazaal@ircoedu.uobaghdad.edu.iq.

Farah Jasim Mohammed obtained a bachelor's degree in Computer Science from Baghdad University, Baghdad, Iraq, 2006. She works as a technician in the laboratories of the computer department since 2007. She participated in research published in the Journal of the College of Education for Girls in University of Baghdad. Her interests are in cyber security, computer network security, educational platforms, and cloud computing. She can be contacted at email: farah.hasan@coeduw.uobaghdad.edu.iq.

Hassan Muwafaq Gheni received his Bachelor (B.Sc) of Electrical and Electronic Engineering from Department of Electrical Engineering, Babylon University-Iraq-hilla in June 2016. In February 2018, he entered the master's program at the Faculty of Electrical and Electronic Engineer, Universiti Tun Hussein Malaysia. He is a lecturer at Al-Mustaqbal University College/Department of Computer Techniques Engineering. His research interest is optical communication, IoT, wireless sensor network, communications, V2V system, and artificial intelligent. He can be contacted at email: hasan.muwafaq@mustaqbal-college.edu.iq.