http://dx.doi.org/10.5755/j01.eie.25.6.24826

A New Heuristic Approach for Treating Missing Value: ABC$_{imp}$

Pinar Cihan[1, *], Zeynep Banu Ozger[2]
[1]*Department of Computer Engineering, Tekirdag Namik Kemal University,*
59860 Corlu, Tekirdag, Turkey
[2]*Department of Computer Engineering, Sutcu Imam University,*
46040 Kahramanmaras, Turkey
pkaya@nku.edu.tr

Abstract—**Missing values in datasets present an important problem for traditional and modern statistical methods. Many statistical methods have been developed to analyze the complete datasets. However, most of the real world datasets contain missing values. Therefore, in recent years, many methods have been developed to overcome the missing value problem. Heuristic methods have become popular in this field due to their superior performance in many other optimization problems. This paper introduces an Artificial Bee Colony algorithm based new approach for missing value imputation in the four real-world discrete datasets. At the proposed Artificial Bee Colony Imputation (ABC$_{imp}$) method, Bayesian Optimization is integrated into the Artificial Bee Colony algorithm. The performance of the proposed technique is compared with other well-known six methods, which are Mean, Median, k Nearest Neighbor (k-NN), Multivariate Equation by Chained Equation (MICE), Singular Value Decomposition (SVD), and MissForest (MF). The classification error and root mean square error are used as the evaluation criteria of the imputation methods performance and the Naive Bayes algorithm is used as the classifier. The empirical results show that state-of-the-art ABC$_{imp}$ performs better than the other most popular imputation methods at the variable missing rates ranging from 3 % to 15 %.**

Index Terms—**Data handling; Evolutionary computation; Heuristic algorithms; Bayes methods.**

I. INTRODUCTION

In several researches, it is quite possible to have missing values within the collected data. These missing values within datasets are known as lost values and this is a drawback encountered by many researchers. It is clear that these missing data influence research results and that they are one of the most common problems. Most of statistical data analysis software packages were developed under the assumption of complete data. Therefore, it is evident that the analyses performed with missing data are inaccurate and unreliable [1], [2].

Today, it is very difficult to find a complete dataset and the missing data issue is a disadvantage widely seen in the real world. In questionnaires, missing data exists because of some unanswered questions or some answers, which are partly incorrect [3], [4]. Missing values in control based applications, such as road traffic monitoring [5], industrial operations [6] or management of telecommunication and computer networks [7] arise due to the failure of monitoring equipment or data collectors, interrupted communication between data collectors and central management system, and failure at the archiving system. In the field of metabolomics, bio-samples, such as cell, tissue, biological fluids, and etc. commonly have missing values [8], [9]. In automatic speech recognition, the speech samples impaired with high background noise are assessed as missing data also [10], [11]. In DNA microarray and biological researches, genetic data may be missing due to various reasons, such as scratched slide or contaminated samples. In medical diagnosis, the physician may request a test, which either provides the exact result or a test not relevant to the diagnosis or, in cases, where measuring is difficult/harmful, the features may be missing [12]–[17].

Missing values are the disadvantage of almost all researches and there are a few alternative methods to overcome their drawbacks. The researchers may prevent potential problems using one of the methods, such as (i) extending the data with new observations, (ii) removing the observations with missing value from the dataset, and (iii) conducting predictions about missing value and substituting the missing value with obtained approximate values. This is generally not preferred, as new observations would generate time and labour costs. Removing the observations with missing value from the dataset might seriously reduce the number of observations and sufficient sampling might become an insufficient one. This might reduce the strength of the subsequent statistical analyses [18]. Besides, in some cases, where the missing values are associated with other variables included within the analysis, their deletion might result in a significant partiality [19]–[21]. When considered in this context, imputation methods by assigning the approximate values instead of missing values become the methods where the researchers shall be allowed to spare time and labour, while enabling them to preserve the collected data.

In literature, various imputation methods were used in order to successfully impute missing values in datasets and their performances were compared. Brock *et al.* [22] have assessed k-nearest neighbours (kNN), ordinary least squares (OLS), partial least squares (PLS), singular value decomposition (SVD), Bayesian principal component

Manuscript received 3 March, 2019; accepted 15 September, 2019.

analysis (bPCA), local least squares (LLS), and least squares adaptive (LSA) methods in order to determine, which imputation method is more successful in the imputation of missing values in the microarray dataset. Waljee *et al*. [23] imputed missing value in hepatocellular carcinoma patients' dataset through missForest, kNN, MICE, and mean imputation methods. They compared the influence of clinical anticipation models on accuracy. Celton *et al*. [24] imputed missing values in the dataset in order to interpret the microarray experiments and to improve clustering, and compared them. Schmitt *et al*. [25] tried to determine the most successful method by comparing performances of mean, kNN, SVD, fuzzy K-means (FKM), mice, and bPCA methods in imputing missing values of Iris, E. coli, Breast cancer 1, and Breast cancer 2 datasets.

Hron *et al*. [26] used two different versions of kNN for imputing missing values. Tutz and Ramzan [27], proposed a wNN to estimate the missing values. Betechuoh and Marwala [28] used Ant Colony Optimization (ACO) algorithm to estimate the missing values. Abdella and Marwala [29] used a combination of genetic algorithms and neural networks for approximate the missing values in dataset. Devi Priya *et al*. [30] implemented Dual repopulated Bayesian ant colony optimization (DPBACO) algorithm for imputing missing values in heterogeneous attributes of large datasets. As a result of the analysis, it was determined that the developed evolutionary method was more successful than other methods. Aydilek and Arslan [31] implemented a hybrid method for imputing missing values. They used optimized fuzzy c-means with support vector regression and a genetic algorithm. The proposed method yielded sufficient and sensible imputation performance results. Qui *et al*. [32] applied deep learning based method for the data imputation. They used a denoising autoencoder with partial loss (DAPL) method for imputing missing values in genomic data. Results showed that, the proposed method achieved comparable or better performance. McCoy *et al*. [33] used variational autoencoders (VAEs), which are deep learning techniques for the missing data imputation. In the study, VAEs are compared with traditional imputation methods (PCA and Mean) by using Root Mean Squared Error (RMSE). As a result of the analysis, VAE imputation method achieved lower error than the traditional methods. Cihan [34] used the ABC method to handle the missing values in the thesis study. However, in this study, the method was improved, hybridized, and the more successful method was obtained.

The aims of the study were to compare the proposed Artificial Bee Colony imputation (ABCImp) method to other existing methods, which are Mean, Median, kNN, MICE, SVD, and MissForest, under the MCAR pattern at 3 %, 5 %, 7 %, 10 %, 12 %, and 15 % missingness rates.

The rest of the study is organized as follows. Section II introduces the imputation methods, Bayes theorem, and artificial bee colony (ABC). The proposed method is explained at Section III. Datasets and evaluation criteria are introduced at Section IV. Section V is dedicated to the frequency of missingness in dataset, the imputation methods performance, the classification performance, and makes a comparison. The last section (Section VI) provides our conclusions.

II. BACKGROUND

A. Imputation Methods

ABC$_{\text{imp}}$ method is compared with seven imputation methods, which are, namely, mean imputation, median imputation, kNN, MICE, SVD, and missForest methods. We briefly introduce these methods below.

Mean imputation method is one of the simple and straightforward methods to impute the missing values. The average of existing values is taken with a missing value and the obtained result is assigned to the missing value. In this case, the average is left constant, while the variance shall be reduced. The negligence of variance makes the correlation structure of the dataset partial. Therefore, it might give quite bad results, when there is a correlation between variables [35].

Median imputation method is also one of the simple and straightforward methods as the mean imputation method. In this method, the median of existing values is taken for a variable with a missing value and the obtained result is assigned to the missing value.

Nearest neighbour algorithms were first proposed for the supervised pattern recognition. Then, Troyanskaya *et al*. [12] proposed the kNN and missing value imputation methods. The main idea is to measure the distance between each observation pair based on variables without missing values. Then, missing data are imputed through the weighted mean of k-nearest observations, which has non-missing data. To implement this method, the function "kNN" in R package Visualization and Imputation of Missing Values (VIM) was used [36], [37].

MICE was proposed by Van Buuren *et al*. [38]. This algorithm prompts the user for a conditional model of each variable. Other variables serve as predictors. Until a stopping criterion is satisfied, the algorithm imputes missing values iteratively based on conditional models fitted. As a general rule, continuous variables are analysed with a linear regression model and binary variables by a logistic regression model. We used R through the package "mice" [38].

SVD imputation algorithm was proposed by Troyanskaya *et al*. [12]. The idea behind the algorithm is to estimate the missing values as a linear combination of the k most significant eigenvalues. The missing values in the dataset are estimated using a low rank SVD approach estimated by the Expectation-Maximization (EM) algorithm. To implement this method, the function "impute.svd" in R package "bcv" was used.

MissForest imputation method was proposed by Stekhoven and Bühlmann [39]. A random forest model is created for each variable through remaining variables within the dataset. This model is used to estimate missing values of this variable. This process continues cyclically for all variables and it is iteratively repeated until the stop criterion is reached. This method was applied through the "missForest" package from the classifier R [39], [40].

B. Bayes Theorem

Bayes theorem shows the relationship between conditional probabilities and marginal probabilities in the probability distribution for a random variable. Namely, it shows how the probability of occurrence of an event will change if additional information is obtained. The theory is the most popular and widely used of all probability theories. It is given in (1)

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}, \tag{1}$$

where P(A) and P(B) are called as prior or marginal probabilities. They refer to the probabilities of occurrence of the events A and B, respectively. P(A|B) is the conditional probability of A, namely, the probability of event A when event B occurs. P(B|A) is the conditional probability of event B.

C. Artificial Bee Colony

Artificial bee colony (ABC) is a population-based stochastic algorithm proposed by Karaboga [41]. ABC models some foraging behaviours of honey bees. Each solution in search space is called as “food source”. The duty of the bees is finding an appropriate source. The quality of a source is problem dependent and calculated by the fitness function. Sources are represented by vector and the dimension of vector is the number of the parameter of the problem. There are three kinds of bees in population: employed bees, onlooker bees, and a scout bee. The number of employed bees is equal to the number of onlooker bees, and there is one scout bee in swarm. There is a trial value for each source. When sources are initialized, these values are equal to zero. The algorithm consists of four steps detailed below.

Initialization. Food sources are initialized randomly at the search space according to (2)

$$x_{ij} = x_j^{\min} + rand(0,1)(x_j^{\max} - x_i^{\min}), \tag{2}$$

where x is a food source, namely, a vector, i is the number of food source, and j is the number of parameter. Sources are initialized according to maximum (x_j^{max}) and minimum (x_j^{min}) value of the parameter.

Employed Bee Phase. Employed bees are responsible for exploitation. For each employed bee, a neighbour source is generated using the following formula and calculated its fitness value. It is given in (3)

$$v_{ij} = x_j^{\min} + rand(-1,1)(x_{ij} - x_{kj}). \tag{3}$$

If a new source is better than current one, the new source is memorized and the trial value of this source is set to zero. Otherwise, the trial value is incremented by one. v is the neighbour source, i is the current source, j is the current parameter of ith source, and x_k is a randomly selected source from swarm.

Onlooker Bee Phase. Employed bees share information about the sources with the onlooker bees. Onlooker bees select a source probabilistically using the Roulette-wheel scheme and try to optimize it. According to roulette-wheel scheme, the better source has a high probability of being selected. It is given in (4)

$$P_i = \frac{fitness_i}{\sum_{i=1}^{SN} fitness_i}, \tag{4}$$

where fitness$_i$ is the fitness value of ith source and SN is the size of a swarm. P_i is the selection probability of ith source. If it is higher than a random value that produced between zero and one, the onlooker bee selects this source and produces a new neighbour source using (3).

Scout Bee Phase. Scout Bee is responsible for the exploration in the swarm. After one employed and onlooker bee cycle is completed, the scout bee checks the trial values. If the trial value of any source exceeds the predefined limit value, this source is abandoned and a new source is generated using (2).

III. THE PROPOSED METHOD: ABC$_{IMP}$

Missing values are selected randomly from whole dataset. The samples that have no missing value generate the train data and the rest of the samples are test data. ABC$_{imp}$ is applied for each sample in the test set as shown in Fig. 1. If a sample includes two missing values, this means that the sources are two-dimensional vectors and there are two parameters that need to be optimized.

We determined a solution pool for each feature and initialized the sources according to this pool. The solution pool includes the minimum and maximum values for each feature in the train dataset. Namely, a missing value takes a value within these boundaries. For each parameter of a source, a discrete value between minimum and maximum values of the corresponding feature is randomly selected.

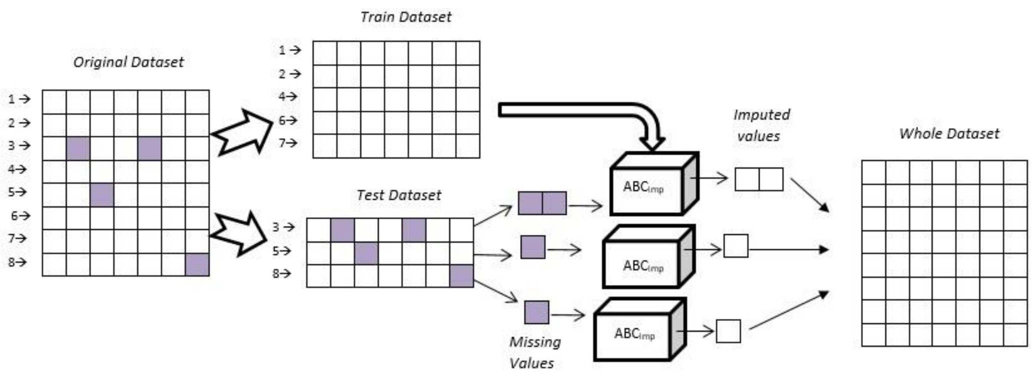

Fig. 1. Missing value imputation process.

ABC$_{imp}$ try to optimize one parameter at each iteration. If a source includes more than one parameter, the algorithm selects one of them randomly.

ABC algorithm is developed for continuous optimisation problems. Equation (3) used in the employed bee phase produces new sources by making vector addition and subtraction. It is not always suitable for the discrete space. Therefore, we changed new source generation strategy at the employed bee phase. As shown in Fig. 2, a parameter is selected randomly from the source. Three values for this parameter are generated by using mean imputation, distance imputation, and random imputation, and the average is taken. If this new value improved the source quality, it is changed by previous one.

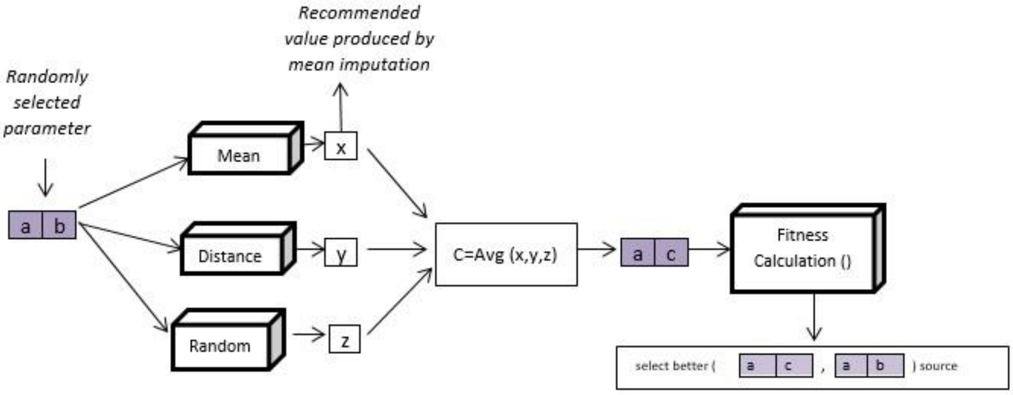
Recommended value produced by mean imputation
Randomly selected parameter
a b
Mean
Distance
Random
x
y
z
C=Avg (x,y,z)
a c
Fitness Calculation ()
select better (a c , a b) source

Fig. 2. Fitness evaluation process of ABC$_{imp}$.

The value produced by mean imputation is the average value of corresponding feature at the train dataset. The class information considered when taking the average of the feature. At distance imputation, the distance of the sample that includes missing data to all other samples in the train dataset is calculated by using the Euclidean Distance. The average of the three nearest samples is taken. According to random imputation, the new value is selected from the solution pool for the corresponding parameter. Initially, each value in the solution pool has the same probability of selection. The selection probability of a value in the solution pool increases if it improves the solution quality.

The Bayesian function is used as fitness function in the algorithm. Bayesian function uses posterior and prior probability values. When missing values are discrete, the Bayesian function is applied in (5) below

$$\frac{P(f1, f2, \ldots, f_{n-1} \mid MV_i)P(MV_i)}{P(f1, f2, \ldots, f_{n-1})}, \tag{5}$$

where MV$_i$ is ith missing value of the dataset, f$_1$, f$_2$, ... , f$_{n-1}$ are non-missing value attributes and n is the attribute number of the dataset. The Bayesian posterior probability is calculated as P(f$_1$, f$_2$, … , f$_{n-1}$|MV$_i$). P(MV$_i$) and P(f$_1$, f$_2$, ... , f$_{n-1}$) are prior probabilities for ith missing value and non-missing value attributes, respectively.

IV. DATASETS AND EVALUATION

A. Datasets

The datasets are selected from University of California Irvine (UCI) Machine Learning Repository. In the study, the balance scale, website phishing, nursery, and car datasets were used for to evaluate the imputation methods performance. The datasets have complete and discrete values. Characteristics of these datasets are summarized in Table I.

To evaluate the performance of the imputation methods, missing values simulated the MCAR pattern [42]. In this study, 3 %, 5 %, 7 %, 10 %, 12 %, and 15 % of data were randomly removed from all datasets. Then, these missing values were handled by ABC$_{imp}$ and other existing imputation methods in order to determine the method that has the closest estimated value to the actual value.

TABLE I. DATASETS USED IN EXPERIMENTS.

Dataset	#attributes	#instances	#classes
Balance Scale	4	625	3
Website Phishing	10	1353	3
Nursery	8	12690	5
Car	6	1728	4

B. Evaluation Criteria

The proposed ABC$_{imp}$ was compared with mean imputation, median imputation, kNN impute, MICE, SVD, and MissForest methods. For this, the results of the classification error and root mean square error of the methods were compared.

Classification error. This criterion measures the difference between the current subgroups and those, which were generated after the missing data imputation, and assesses if the discriminative or predictive capability is maintained. In the study, Naïve Bayes classification method is used. To increase reliability, the methods are repeated 30 times and 5-fold cross validation is being done. The classification error is defined in (6)

$$Error = \frac{\sum FalsePositive + \sum FalseNegative}{\sum TotalPopulation}, \tag{6}$$

where “False Positive” and “False Negative” are the number of incorrectly classified samples. The former refers to the number of positive samples that the system labelled as negative. The latter is the number of the negative samples that the system labelled as positive.

Root mean square error (RMSE). It measures the difference between the actual value and the estimated value. The smallest RMSE value is always desirable. Basically, the

RMSE is defined as follows

$$RMSE = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(t_{orig}^{m} + t_{reco}^{m})^{2}}, \tag{7}$$

where t_{orig} and t_{reco} are the m^{th} vectors, whose elements are the original values and the reconstructed values, respectively. M denotes the amount of missing value used.

V. Results and Discussion

Proposed ABC_{imp} algorithm is handled missing values at discrete space in four datasets that taken from UCI repository. The datasets and their information are given in Table I. In this study, MATLAB is used for developing ABC_{imp} and R programming is used for other imputation methods for to handle the missing values. In datasets, the missing fraction of 3 %, 5 %, 7 %, 10 %, 12 %, and 15 % of the MCAR values are simulated. Then, missing values are handled with ABC_{imp} and other existing imputation methods. For classification, the datasets are divided into two sets. Training set contains the complete records of datasets and testing set contains the incomplete records.

The performance of ABC_{imp} is compared with mean imputation, median imputation, kNN, MICE, SVD, and MissForest methods. For classification, the Naïve Bayes method is used. The classification process is repeated 30 times. The average classification error results are given in Tables II, III, IV, and V.

TABLE II. CLASSIFICATION ERROR IN BALANCE SCALE DATASET.

Missing rate	Mean	Median	kNN	Mice	SVD	MF	ABC_{imp}
3 %	5.27	4.82	5.94	2.62	1.59	0.55	**0.43**
5 %	5.09	5.51	6.66	3.1	1.72	0.59	**0.52**
7 %	7.15	6.66	9.52	3.87	1.99	0.62	**0.61**
10 %	9.59	8.61	11.21	3.07	2.68	1.05	**0.72**
12 %	11.43	9.7	12.99	4.19	2.89	1.24	**0.86**
15 %	11.4	10.68	13.88	5.53	2.95	1.39	**1.02**

TABLE III. CLASSIFICATION ERROR IN WEBSITE PHISHING DATASET.

Missing rate	Mean	Median	kNN	Mice	SVD	MF	ABC_{imp}
3 %	4.80	4.27	4.05	4.27	0.76	0.83	**0.75**
5 %	5.24	4.46	4.10	4.80	1.86	1.18	**0.94**
7 %	5.67	5.18	4.34	5.19	2.00	1.49	**1.08**
10 %	6.75	6.09	4.85	5.65	2.20	1.66	**1.61**
12 %	9.67	7.52	5.54	5.90	2.86	2.45	**2.22**
15 %	13.27	8.14	5.97	8.84	3.02	2.75	**2.37**

TABLE IV. CLASSIFICATION ERROR IN NURSERY DATASET.

Missing rate	Mean	Median	kNN	Mice	SVD	MF	ABC_{imp}
3 %	5.34	2.62	0.99	1.03	0.65	0.21	**0.01**
5 %	7.69	4.32	1.46	1.77	0.67	0.23	**0.01**
7 %	10.58	7.53	2.04	2.35	0.72	0.27	**0.04**
10 %	13.93	9.92	2.76	3.09	0.79	0.27	**0.05**
12 %	13.32	11.22	3.04	4.02	0.79	0.34	**0.09**
15 %	14.78	13.49	4.38	4.36	0.84	0.38	**0.09**

Comparing the performance of methods, the mean and median methods are found to be inferior in four datasets. The main reason that lies behind worst performance of mean and median is not preserved the relationships among the attributes. Also, the kNN and Mice datasets cannot obtain a good performance. Because of low availability of the valid records is increased, the classification error is increased also as the missing data rate in the datasets increases.

When the results are examined, ABC_{imp} evolutionary based algorithm shows better performance than other existing methods. When the missing values are around 3 %, percentage error for the balance scale is 0.43, website phishing - 0.075, nursery - 0.01, and car - 0.02. When the missing rate is increased to 15 %, the percentage of maximum classification error of ABC_{imp} is only 2.37. From the results of classification error, the MissForest shows only little underperformance compared to ABC_{imp} for all datasets. According to the results of classification error, it is seen that ABC_{imp} method outperforms the other existing methods.

TABLE V. CLASSIFICATION ERROR IN CAR DATASET.

Missing rate	Mean	Median	kNN	Mice	SVD	MF	ABC_{imp}
3 %	2.43	2.59	0.77	1.25	0.71	0.11	**0.02**
5 %	4.71	4.52	0.99	1.81	0.87	0.17	**0.02**
7 %	6.68	6.09	1.29	3.28	0.88	0.22	**0.02**
10 %	9.43	8.64	1.94	4.93	0.96	0.22	**0.03**
12 %	10.02	9.72	3.19	5.3	1.01	0.23	**0.03**
15 %	11.99	11.56	3.53	5.4	1.16	0.25	**0.04**

Tables VI, VII, VIII, and IX show the RMSE results obtained by applying ABC_{imp} and other methods in balance scale, website phishing, nursery, and car datasets. According to the results of RMSE values as the classification error results, the ABC_{imp} method is also successful compared to other existing methods. When the percentage of missing values is around 3 % in datasets, the RMSE values do not exceed 0.44. When the missing rate in datasets is 15 %, the RMSE values stands not more than 0.59. It is clear that ABC_{imp} can replace the missing data with optimal values by its strong search capability and multiple fitness processes.

TABLE VI. RMSE OF METHODS IN BALANCE SCALE DATASET.

Missing rate	Mean	Median	kNN	Mice	SVD	MF	ABC_{imp}
3 %	1.90	1.41	1.89	1.43	1.64	1.33	**0.44**
5 %	1.95	1.42	1.95	1.50	1.66	1.36	**0.45**
7 %	2.04	1.52	1.97	1.58	1.72	1.36	**0.49**
10 %	2.21	1.54	2.02	1.62	1.75	1.39	**0.51**
12 %	2.49	1.61	2.11	1.68	1.75	1.41	**0.49**
15 %	2.94	1.83	2.20	1.73	1.79	1.42	**0.59**

TABLE VII. RMSE OF METHODS IN WEBSITE PHISHING DATASET.

Missing rate	Mean	Median	kNN	Mice	SVD	MF	ABC_{imp}
3 %	1.01	0.93	0.81	0.90	0.77	0.63	**0.34**
5 %	1.02	0.93	0.81	0.93	0.76	0.65	**0.36**
7 %	1.11	0.94	0.81	0.91	0.76	0.64	**0.37**
10 %	1.23	0.95	0.82	0.94	0.77	0.67	**0.37**
12 %	1.48	0.96	0.83	0.94	0.76	0.67	**0.38**
15 %	1.88	0.99	0.84	0.97	0.78	0.68	**0.39**

RMSE focuses on the difference between real and estimated values. Therefore, if the difference is large, the RMSE value is also large. For balance scale dataset, the classification error and RMSE values are close to each other. The RMSE value is higher for the website phishing data, while the classification error for the other two datasets is lower.

TABLE VIII. RMSE OF METHODS IN NURSERY DATASET.

Missing rate	Mean	Median	kNN	Mice	SVD	MF	ABC$_{imp}$
3 %	1.27	1.09	1.13	1.22	1.02	0.96	**0.37**
5 %	1.29	1.16	1.13	1.23	1.05	0.96	**0.52**
7 %	1.43	1.22	1.14	1.24	1.05	0.96	**0.53**
10 %	1.54	1.25	1.15	1.24	1.06	0.97	**0.54**
12 %	1.63	1.37	1.15	1.25	1.07	0.97	**0.57**
15 %	1.91	1.41	1.16	1.27	1.09	0.98	**0.58**

TABLE IX. RMSE OF METHODS IN CAR DATASET.

Missing rate	Mean	Median	kNN	Mice	SVD	MF	ABC$_{imp}$
3 %	1.48	1.16	1.23	1.32	1.16	1.01	**0.44**
5 %	1.65	1.16	1.26	1.42	1.19	1.01	**0.47**
7 %	1.65	1.13	1.24	1.38	1.17	1.03	**0.47**
10 %	1.71	1.14	1.25	1.44	1.19	1.04	**0.55**
12 %	1.82	1.14	1.28	1.44	1.20	1.06	**0.56**
15 %	1.96	1.13	1.28	1.46	1.21	1.07	**0.59**

VI. CONCLUSIONS

A new heuristic approach is developed to handling missing value to alleviate the missing value imputation problems. By combining the artificial bee colony algorithm with Bayesian optimization, we proposed a strong method to estimate the missing values. Furthermore, multiple fitness functions are used to estimate the missing values. The competence of the ABC$_{imp}$ algorithm is measured by testing in four real-world discrete datasets. Missing values in datasets are generated using MCAR missingness mechanism in various size (from 3 % to 15 %). Proposed method is compared with six popular and success imputation methods: Mean, Median, kNN, Mice, SVM, and MissForest. The results clearly show that ABC$_{imp}$ outperforms the other six imputation methods in terms of classification error and RMSE for all datasets with different percentage of missing value rates.

According to experimental results, MissForest is the most successful imputation method of all compared methods. It obtained the closest results to ABC$_{imp}$ for both criteria. The average classification error results of ABC$_{imp}$ are: 0.69, 1.50, 0.05, and 0.03. However, the datasets imputed by MissForest classified them with 0.91, 1.73, 0.28, and 0.20 error rates, respectively.

With multiple fitness functions, the algorithm is provided to make a closer estimate to the real value Therefore, it obtains the lowest RMSE values. Additionally, with Bayesian optimization, the suggestions of ABC$_{imp}$ are more appropriate to the model of the dataset. That is why the proposed method classifies the imputed datasets with less classification error. As a further extension, ABC$_{imp}$ can be implemented for the handles heterogeneous attributes.

REFERENCES

[1] T. D. Pigott, “A review of methods for missing data”, *Educational research and evaluation*, vol. 7, no. 4, pp. 353–383, 2001. DOI: 10.1076/edre.7.4.353.8937.

[2] P. D. Allison, “Missing data techniques for structural equation modeling”, *Journal of abnormal psychology*, vol. 112, no. 4, pp. 545–557, 2003. DOI: 10.1037/0021-843X.112.4.545.

[3] D. B. Rubin, *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, vol. 81, 2004. DOI: 10.1002/9780470316696.

[4] L. Wang and X. Fan, “Missing data in disguise and implications for survey data analysis”, *Field Methods*, vol. 16, no. 3, pp. 332–351, 2004. DOI: 10.1177/1525822X03262276.

[5] L. N. Nguyen and W. T. Scherer, “Imputation techniques to account for missing data in support of intelligent transportation systems applications”, (No. UVACTS-13-0-78), Charlottesville, VA, Center for Transportation Studies, University of Virginia, 2003.

[6] K. Lakshminarayan, S. A. Harp, and T. Samad, “Imputation of missing data in industrial databases”, *Applied Intelligence*, vol. 11, no. 3, pp. 259–275, 1999. DOI: 10.1023/A:1008334909089.

[7] C. Ji and A. Elwalid, “Measurement-based network monitoring: Missing data formulation and scalability analysis”, in Proc. of *Information Theory, IEEE International Symposium on IEEE*, 2000, p. 78. DOI: 10.1109/ISIT.2000.866368.

[8] R. Wei, J. Wang, M. Su, E. Jia, S. Chen, T. Chen, and Y. Ni, “Missing value imputation approach for mass spectrometry-based metabolomics data”, *Scientific reports*, vol. 8, no. 1, p. 663, 2018. DOI: 10.1038/s41598-017-19120-0.

[9] R. Wei, J. Wang, M. Su, E. Jia, S. Chen, T. Chen, Y. Ni, and W. Jia, “GSimp: A gibbs sampler based left-censored missing value imputation approach for metabolomics studies”, *PLoS computational biology*, vol. 14, no. 1, 2018. DOI: 10.1371/journal.pcbi.1005973.

[10] M. Cooke, P. D. Green, and M. Crawford, “Handling missing data in speech recognition”, in *ICSLP*, 1994, pp. 1555–1558.

[11] S. Parveen and P. Green, “Speech enhancement with missing data techniques using recurrent neural networks”, in *Proc. of 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2004, I–733 p. DOI: 10.1109/ICASSP.2004.1326090.

[12] P. Cihan, E. Gökce, and O. Kalipsiz, “A review of machine learning applications in veterinary field”, *Kafkas Univ. Vet. Fak. Derg.*, vol. 23, no. 4, pp. 673–680, 2017. DOI: 10.9775/kvfd.2016.17281.

[13] P. Cihan, E. Gökce, and O. Kalipsiz, “Determination of computer aid diagnosis and/or risk factors using data mining methods in veterinary field: A review”, *Atatürk Üniversitesi Vet. Bil. Derg.*, vol. 14 no. 2, pp. 209–220, 2019. DOI: 10.17094/ataunivbd.462197.

[14] P. Liu, E. El-Darzi, L. Lei, C. Vasilakis, P. Chountas, and W. Huang, “An analysis of missing data treatment methods and their application to health care dataset”, in *Proc. of International Conference on Advanced Data Mining and Applications*, 2005, pp. 730–730. DOI: 10.1007/11527503_69.

[15] M. K. Markey and A. Patel, “Impact of missing data in training artificial neural networks for computer-aided diagnosis”, in *Proc. of IEEE International Conference on Machine Learning and Applications*, 2004, pp. 351–354. DOI: 10.1109/ICMLA.2004.1383534.

[16] M. A. Proschan, R. P. McMahon, J. H. Shih, S. A. Hunsberger, N. L. Geller, G. Knatterud, and J. Wittes, “Sensitivity analysis using an imputation method for missing binary data in clinical trials”, *Journal of Statistical Planning and Inference*, vol. 96, no. 1, pp. 155–165, 2001. DOI: 10.1016/S0378-3758(00)00332-3.

[17] J. M. Jerez, I. Molina, J. L. Subirats, and L. Franco, “Missing data imputation in breast cancer prognosis”, in *Proc. of. BioMed’06*, ACTA Press Anaheim, CA, pp. 323–328, 2006.

[18] P. L. Roth, “Missing data: A conceptual review for applied psychologists”, *Personnel psychology*, vol. 47, no. 3, pp. 537–560, 1994. DOI: 10.1111/j.1744-6570.1994.tb01736.x.

[19] B. G. Tabachnick, L. S. Fidell, and S. J. Osterlind, Using multivariate statistics, 2001.

[20] J. L. Schafer, “Multiple imputation: A primer”, *Statistical methods in medical research*, vol. 8, no. 1, pp. 3–15, 1999. DOI: 10.1177/096228029900800102.

[21] J. W. Osborne and A. Overbay, “Best practices in data cleaning”, *Best practices in quantitative methods*, SAGE, 2008. DOI: 10.4135/9781412995627.

[22] G. N. Brock, J. R. Shaffer, R. E. Blakesley, M. J. Lotz, and G. C. Tseng, “Which missing value imputation method to use in expression profiles: A comparative study and two selection schemes”, *BMC bioinformatics*, vol. 9, no. 1, p. 12, 2008. DOI: 10.1186/1471-2105-9-12.

[23] A. K. Waljee, A. Mukherjee, A. G. Singal, Y. Zhang, J. Warren, U. Balis, and P. D. Higgins, “Comparison of imputation methods for missing laboratory data in medicine”, *BMJ open*, vol. 3, no. 8, p. e002847, 2013. DOI: 10.1136/bmjopen-2013-002847.

[24] M. Celton, A. Malpertuy, G. Lelandais, and A. G. De Brevern, “Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments”, *BMC genomics*, vol. 11, no. 1, p. 15, 2010. DOI: 10.1186/1471-2164-11-15.

[25] P. Schmitt, J. Mandel, and M. Guedj, “A comparison of six methods for missing data imputation”, *Journal of Biometrics & Biostatistics*, vol. 6, no. 1, 2015. DOI: 10.4172/2155-6180.1000224.

[26] K. Hron, M. Templ, and P. Filzmoser, “Imputation of missing values for compositional data using classical and robust methods”, *Computational Statistics & Data Analysis*, vol. 54, no. 12, pp. 3095–3107, 2010. DOI: 10.1016/j.csda.2009.11.023.

[27] G. Tutz and S. Ramzan, “Improved methods for the imputation of missing data by nearest neighbor methods”, *Computational Statistics & Data Analysis*, vol. 90, pp. 84–99, 2015. DOI: 10.1016/j.csda.2015.04.009.

[28] B. L. Betechuoh and T. Marwala, “Ant colony optimization for missing data estimation”, in *Proc. of the Pattern recognition of South Africa*, pp.183–188, 2006.

[29] M. Abdella and T. Marwala, “The use of genetic algorithms and neural networks to approximate missing data in database”, *Comput. Inform*., vol. 24, pp. 577–589, 2005. DOI: 10.1109/ICCCYB.2005.1511574.

[30] R. D. Priya, R. Sivaraj, and N. S. Priyaa, “Heuristically repopulated Bayesian ant colony optimization for treating missing values in large databases”, *Knowledge-Based Systems*, vol. 133, pp. 107–121, 2017. DOI: 10.1016/j.knosys.2017.06.033.

[31] I. B. Aydilek and A. Arslan, “A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm”, *Information Sciences*, vol. 233, pp. 25–35, 2013. DOI: 10.1016/j.ins.2013.01.021.

[32] Y. L. Qiu, H. Zheng, and O. Gevaert, “A deep learning framework for imputing missing values in genomic data”, *bioRxiv*, 2018. DOI: 10.1101/406066.

[33] J. T. McCoy, S. Kroon, and L. Auret, “Variational autoencoders for missing data imputation with application to a simulated milling circuit”, *IFAC-PapersOnLine*, vol. 51, no. 21, pp. 141–146, 2018. DOI: 10.1016/j.ifacol.2018.09.406.

[34] P. Cihan, “Determination of diagnosis, prognosis and risk factors in animal diseases using by data mining methods”, Ph.D. dissertation, Dept. Comp. Eng., Yildiz Technical Univ., Istanbul, Turkey, 2018.

[35] P. Cihan, “A Comparison of five methods for missing value imputation in data sets”, *International Scientific And Vocational Journal (Isvos Journal)*, vol. 2, no. 2, pp. 80–85, 2018. DOI: 10.4172/2155-6180.1000224.

[36] M. Templ, A. Alfons, A. Kowarik, and B. Prantner, *VIM: Visualization and Imputation of Missing Values*. CRAN, 2015.

[37] J. L. Schafer, *Analysis of Incomplete Multivariate Data*. CRC press, 1997. DOI: 10.1201/9781439821862.

[38] S. van Buuren, H. C. Boshuizen, and D. L. Knook, “Multiple imputation of missing blood pressure covariates in survival analysis”, *Statistics in medicine*, vol. 18, no. 6, pp. 681–694, 1999. DOI: 10.1002/(SICI)1097-0258(19990330)18:6%3C681::AID-SIM71%3E3.0.CO;2-R.

[39] D. J. Stekhoven and P. Bühlmann, “MissForest – non-parametric missing value imputation for mixed-type data”, *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2011. DOI: 10.1093/bioinformatics/btr597.

[40] S. van Buuren and K. Groothuis-Oudshoorn, “MICE: Multivariate imputation by chained equations in R Journal of Statistical Software”, 2009. [Online]. Available: http://CRAN.R-project.org/package=mice/

[41] D. Karaboga and B. Akay, “A survey: Algorithms simulating bee swarm intelligence”, *Artificial Intelligence Review*, vol. 31, pp. 61–85, 2009. DOI: 10.1007/s10462-009-9127-4.

[42] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. John Wiley & Sons, 2014. DOI: 10.1002/9781119013563.