ExtraImpute: A Novel Machine Learning Method for Missing Data Imputation

Mustafa Alabadla, Fatimah Sidi, Iskandar Ishak, Hamidah Ibrahim, Lilly Suriani Affendey, and Hazlina Hamdan
Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor D.E., Malaysia
Email: gs59711@student.upm.edu.my, {fatimah, iskandar_i, hamidah.ibrahim, lilly, hazlina}@upm.edu.my

***Abstract*—Missing values are one of the common incidences that occurs in healthcare datasets. Its existence usually leads to undesirable results while conducting data analysis using machine learning methods. Recently, researchers have proposed several imputation approaches to deal with missing values in real-world datasets. Moreover, data imputation assists us to build a high-performance machine learning models to discover patterns in healthcare data that provides top-notch insights for a higher quality decision-making. In this paper, we propose a new imputation approach using Extremely Randomized Trees (Extra Trees) of machine learning ensemble learning methods named (ExtraImpute) to tackle numerical missing values in healthcare context. The proposed method has the ability to impute both continuous and discrete data features. This approach imputes each missing value that exists in features by predicting its value using other observed values in the dataset. To evaluate the efficiency of our algorithm, several experiments are conducted on five different benchmark healthcare datasets and compared to other commonly used imputation methods, viz. missForest, KNNImpute, Multivariate Imputation by Chained Equations (MICE), and SoftImpute. The results were validated using Root Mean Square Error (RMSE) and Coefficient of Determination (R^2) scores. From these results, it was observed that our proposed algorithm outperforms existing imputation techniques.**

***Index Terms*—imputation, missing values, extra trees, healthcare**

I. Introduction

The rapid development in technology in the last decade has increased the amount of data gathered in the daily basis. Industries in several domains analyzes the collected data in order to make decisions for the sake of their upgrowth. In this regard, data mining algorithms are utilized to extract significative information from hidden patterns in the data. Usually, these algorithms are accurate unless the used data is defective somehow. Thus, refining the dataset using data pre-processing is vital and considered as one of the most challenging part for most researchers [1]. Missing values is one of the most common problems in healthcare research. It can be caused by various reasons including manual entry error, equipment malfunction, and faulty measurement [2]. Hence, the existence of missing values leads to lack of efficiency, complications in analyzing the data, and a noticed bias which results from the significant variation between complete and missing data [3].

The mechanism of missing values is classified under three main categories including Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing not at Random (MNAR) [4], [5]. At MCAR settings the missing values have no dependency to other values. While in MAR, there is a relation between missing values and observed values. The last mechanism MNAR is only applicable if the previous mechanisms are not valid and in this case the missing values are usually related to unobserved factors or the missing values themselves [6].

Generally, there are two main approaches to address the missing values problem. The first approach is to remove them using listwise and casewise deletion methods. However, these methods can lead to a significant reduce in statistical performance especially if a large amount of data is omitted [7]. The second approach is to impute missing values using single and multiple imputation techniques. Mean substitution is one of the easiest and straightforward ways to replace missing values using the mean of observable values [8]. Nonetheless, the imputed value may be far from the average value of the selected variable which can produce bias in the data. More traditional single imputation approaches such as hot-deck and cold-deck methods are also used to replace missing values from an internal or external donor [9]. Unlike single imputation methods, multiple imputation replaces one or more values and provides better results in terms of handling uncertainty in the analyzed data [10].

In this paper, we propose a new data imputation algorithm based on Extremely Randomized Trees named (ExtraImpute) to handle numerical type of missing values in healthcare datasets. We used a complete training set to train our model first and a complete test set to evaluate its accuracy. Then we applied the trained model to features with non-missing values to predict the missing values in the dependent variable. The missing values in each feature are estimated starting from the feature with the lowest missingness ratio and ascending up until all the missing values in the dataset are imputed. Our approach imputes

Manuscript received February 29, 2022; revised June 30, 2022.

doi: 10.12720/jait.13.5.470-476

the missing values without creating a new dataset or parsing it to an array or changing any labels and indices.

II. Related Work

The effect of data imputation approaches on healthcare datasets is discussed in [11]. The authors indicated that there are some challenges and risks of imputing missing values in healthcare data. These issues usually arises when some medical data about a patient are missing or the truthfulness of data becomes questionable due to the vagueness in other values. They also discussed the usefulness of classification methods when replacing missing values using any imputation algorithm. Furthermore, identifying a disease while it is in early stages can only be done with a complete dataset since the existence of missing values can lead to inaccurate results [12]. Among various imputation methods, it was indicated that machine learning techniques are promising on the long-term for imputing missing values in healthcare domain [13].

In the literature, numerous imputation methods were proposed with the focus on machine learning approaches recently. A non-parametric and iterative imputation algorithm based on random forest named (missForest) was proposed to handle missing values for both categorical and continuous data types [14]. The missing values for the training set are first imputed with their mean value if the feature is numeric and with mode value if the feature is categorical. The greatest drawback of this method is that it iterates multiple times until the stopping criterion has met. Thus, the procedure is time-consuming especially when dealing with high-dimensional datasets. Another popular imputation method is called K-Nearest Neighbor Imputation referred to as (KNNI) imputes missing values by making an estimation of the k value or the number of similar records using a distance metric [15]. Similar to missForest, KNNI uses the mean and mode values to impute the training set and in case of large datasets it searches the whole dataset to find the k value which consumes more time based on the dataset size. Madhu *et al*. [16] showed that XGBoost could perform better than other existing methods in imputing missing values. The authors used different incomplete datasets and evaluated their proposed imputation method on the test set only. However, the original values of missing instances are still unknown and there is no way to measure the accuracy of imputed values against their original values if the collected dataset already contains missing values. Deep learning tech techniques were also utilized to estimate missing values. The study by Kim and Chung [17] adopted a multi-modal autoencoder method to impute missing values in healthcare big data. The main aim of this method is to save more time while handling large amount of data. However, there may be a loss in accuracy and execution time in case there are more relationships between input variables to be learned by the model. Missing values can be also imputed using unsupervised machine learning techniques. A novel K-means imputation method was proposed by Raja and Thangavel [18] to handle the uncertainty and discrepancy in the datasets by placing these objects in several clusters to improve the data imputation accuracy. The authors applied their method on four public datasets without mentioning the missing values data type and mechanism. Semi-supervised techniques were also used to deal with missing values. Fazakis *et al.* [19] proposed an iterative imputation algorithm based on semi-supervised ensemble learning methods that loops through all available features in the dataset, setting one of them as dependent variable and all the others as independent variables each time. Similar to missForest and KNNI, this method is computationally expensive and very sensitive to outliers.

III. Proposed Methods

The aim of this study is to propose an effective imputation approach to deal with missing values based on the extremely randomized trees of machine learning ensemble learning named (ExtraImpute). The extremely randomized trees method known as extra trees belongs to the decision trees family and was proposed by Geurts *et al*. [20]. It generates several decision trees to implement ensemble regression and classification tasks [21]. Both extra trees and random forest have achieved high performance in previous studies while dealing with high-dimensional data on regression tasks [21]. Random forests select a random subset of the training data set to generate an ensemble model of decision trees. For splitting the decision tree nodes, random subsets are chosen from the training set and the best features including its values are selected for the decision split using Gini criteria [20]. Alternatively, the extra trees method was proposed to reduce the computational resource consumption and provide more randomization than random forest. Unlike random forest, extra trees use the whole training set to train the model instead of selecting a random subset. Additionally, extra trees select the best attribute with its values to split the decision tree node. These differences make extra trees less prone to overfitting, thus achieve better performance [20]. In our problem, we train the extra trees regression model with 100 trees. Also, we have selected the square root of the total number of features to find the best split.

The idea of the proposed algorithm is to predict missing values using a trained Extra Trees on a complete dataset. Let D = $(\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{D}_3, \ldots, \boldsymbol{D}_j)$ be an i×j dataset than includes missing values. The algorithm starts by selecting the feature with the least missing values $\boldsymbol{y}_{orig}$; then, parse all variables to numeric data types. Make an initial guess for missing values in independent variables $\boldsymbol{x}_{orig}$ using Linear Interpolation method and drop all instances that have missing values from $\boldsymbol{y}_{orig}$. After that, standardization is applied to the complete independent and dependent variables denoted by $\boldsymbol{x}_{train}$, $\boldsymbol{y}_{train}$ respectively. The Extra Trees Regressor is fitted with the scaled predictors $\boldsymbol{x}_{train}$ and scaled target $\boldsymbol{y}_{train}$. After training the model, all the missing values from $\boldsymbol{y}_{\boldsymbol{orig}}$ are extracted and listed as $\boldsymbol{y}_{\boldsymbol{miss}}$ in order to be predicted using the correspondent predictors $\boldsymbol{x}_{\boldsymbol{miss}}$.

The predicted missing values $\boldsymbol{y}_{\boldsymbol{imp}}$ are imputed in the original dataset $\boldsymbol{y}_{orig}$ and the process is repeated for each

feature until all the missing values in the dataset are imputed. The complete representation of ExtraImpute method is shown in Algorithm 1.

Algorithm 1 Impute missing values with Extra Trees

D ← set of variables with missing values
NA ← missing values
while sum(NA) in **D** > 0 **do**
  $\boldsymbol{y_{orig}}$ ← dependent variable with the least NA
  $\boldsymbol{x_{orig}}$ ← independent variables
  for each variable v in **D do**
    parse v to numeric
    if v includes NA **and** v not $\boldsymbol{y_{orig}}$ **then**
      Make initial guess of NA;
    end if
  end for
  $\boldsymbol{x_{train}}$ ← $\boldsymbol{x_{orig}}$ with non-missing values
  $\boldsymbol{y_{train}}$ ← drop missing values from $\boldsymbol{y_{orig}}$
  reset index of **D**
  for each feature f in $\boldsymbol{x_{train}}$ **do**
    standardize $\boldsymbol{x_{train}}$[f]
  end for
  standardize $\boldsymbol{y_{train}}$
  fit Extra Trees Regressor: $\boldsymbol{y_{train}} \sim \boldsymbol{x_{train}}$
  $\boldsymbol{y_{miss}}$ ← list of NA in $\boldsymbol{y_{orig}}$
  $\boldsymbol{x_{miss}}$ ← list of $\boldsymbol{x_{train}}$ related to $\boldsymbol{y_{miss}}$
  Predict $\boldsymbol{y_{miss}}$ using $\boldsymbol{x_{miss}}$
  $\boldsymbol{y_{imp}}$ ← impute missing values in $\boldsymbol{y_{orig}}$
end while

IV. Experiments and Results

In order to evaluate our proposed imputation method, numerous experiments were designed and implemented using different settings. We have selected five medical datasets that are publicly available on Kaggle repository. The details of these datasets including the number of instances, features, and classes are provided in Table I.

TABLE I. Description of the Selected Healthcare Datasets

Dataset	# Instances	# Features	# Classes
Diabetes	768	9	2
Spine	310	14	1
Heart	303	14	4
Liver	583	11	1
Hepatitis	615	14	1

The datasets were selected to be diverse to ensure that the results are not biased to certain features. All of the chosen datasets are complete without any missing values. In order to test our algorithm, we have generated different proportions of missing values by deleting random values from the initial datasets. Missing values were created with 10% to 50% (with a step of 10% each time). Thus, five separate scenarios were represented. This mechanism of missing values in each scenario is MCAR according to the way the generation was performed.

To ensure getting precise results for our proposed algorithm, data was assigned as the following: 80% of total instances were randomly assigned to the training set, 10% for the validation set, and 10% for the test set. We implemented the experiments using Python on macOS Big Sur. The processor of the hardware is 6-Core Intel® Core(TM) i5 @3GHz and memory is 8GB. The Root Mean Square Error (RMSE) are obtained for each imputation method in all scenarios. The RMSE of imputation result is defined as the following.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}} \quad (1)$$

where N is the number of data points, $y(i)$ is the i-th observable, and $\hat{y}(i)$ is its estimated value from the prediction model.

For evaluation, Extra Trees is compared with several machine learning methods used for estimating the missing values including Random Forest, K-NN, XGBoost, and Stochastic Gradient Descent. The same training set are applied to all machine learning models and the results are assessed. Then the test set are implemented for each model and the predicted values are compared with the original values to measure the accuracy and mean square error. After training the selected machine learning models, we have assessed their prediction accuracy against the test set by calculating the Coefficient of Determination (R^2) for each feature [22]. The R^2 of imputation result is defined as the following.

$$R^2 = 1 - \frac{RSS}{TSS} \quad (2)$$

where RSS is the sum of squared residuals, and TSS is the total sum of squares which explains the degree of variation in the dependent variable.

To obtain a general result for the selected machine learning methods, the mean of R^2 for all predicted features values in each dataset is calculated. Table II shows the evaluation results for the machine learning methods used in predicting missing values.

TABLE II. Coefficient of Determination for Machine Learning Methods Used to Predict Missing Values

Dataset	Extra Trees	Random Forest	K-NN	XGB	SGD
Diabetes	0.245	0.199	0.208	0.001	0.228
Spine	0.237	0.224	0.141	0.064	0.222
Heart	0.140	0.083	0.080	0.133	0.123
Liver	0.612	0.556	0.561	0.492	0.580
Hepatitis	0.830	0.377	0.120	0.158	0.186

To test the performance of our proposed imputation algorithm (ExtraImpute), various imputation methods were employed including missForest, KNNImpute, Multivariate Imputation by Chained Equations (MICE), and SoftImpute for regression simulations [14], [23]-[25]. Each method was applied 10 times on every dataset under five different missing ratios with a total of 1,250 experiments. We have used the RMSE formula to measure the difference between the imputed values and the original values before generating missing values completely at

random. The mean for each scenario is calculated in order to avoid bias and present a neutral evaluation. The results obtained from experiments are shown in Table III, Table IV, Table V, Table VI, and Table VII.

TABLE III. AVERAGE RMSE OF THE PROPOSED MEHTOD AGAINST EXISTING IMPUTATION METHODS ON THE SELECTED DATASETS (10% MISSING VALUES)

Dataset	Extra-Impute	miss-Forest	KNNI	MICE	Soft-Impute
Diabetes	35.86	36.99	44.14	36.12	43.31
Spine	14.70	15.00	16.54	15.35	19.18
Heart	14.10	14.61	18.30	14.39	24.20
Liver	103.59	130.82	167.66	133.07	146.45
Hepatitis	24.98	26.65	29.12	29.23	28.91

TABLE IV. AVERAGE RMSE OF THE PROPOSED MEHTOD AGAINST EXISTING IMPUTATION METHODS ON THE SELECTED DATASETS (20% MISSING VALUES)

Dataset	Extra-Impute	miss-Forest	KNNI	MICE	Soft-Impute
Diabetes	34.49	37.22	43.53	34.95	42.60
Spine	11.97	12.71	15.05	12.55	17.99
Heart	14.79	15.53	17.99	15.08	26.22
Liver	110.97	120.61	152.09	136.04	151.15
Hepatitis	26.68	29.65	32.86	28.94	29.49

TABLE V. AVERAGE RMSE OF THE PROPOSED MEHTOD AGAINST EXISTING IMPUTATION METHODS ON THE SELECTED DATASETS (30% MISSING VALUES)

Dataset	Extra-Impute	miss-Forest	KNNI	MICE	Soft-Impute
Diabetes	37.80	40.11	46.36	41.17	46.86
Spine	12.65	13.19	16.53	13.90	18.74
Heart	16.45	17.40	19.36	19.31	33.67
Liver	121.05	124.58	171.48	156.44	159.32
Hepatitis	25.49	28.31	31.04	34.65	29.88

TABLE VI. AVERAGE RMSE OF THE PROPOSED MEHTOD AGAINST EXISTING IMPUTATION METHODS ON THE SELECTED DATASETS (40% MISSING VALUES)

Dataset	ExtraImpute	missForest	KNNI	MICE	Soft-Impute
Diabetes	37.02	40.06	51.23	42.62	45.33
Spine	12.12	12.9	16.72	13.655	20.23
Heart	17.93	18.32	18.56	17.43	39.91
Liver	116.7	126.48	152.4	163.43	157.07
Hepatitis	32.51	34.64	35.17	33.7	35.76

TABLE VII. AVERAGE RMSE OF THE PROPOSED MEHTOD AGAINST EXISTING IMPUTATION METHODS ON THE SELECTED DATASETS (50% MISSING VALUES)

Dataset	Extra-Impute	missForest	KNNI	MICE	Soft-Impute
Diabetes	37.48	45.78	45.63	44	48.59
Spine	10.87	12.75	17.66	14.23	24.67
Heart	13.78	15.58	16.41	17.62	45.2
Liver	105.5	121.62	128.46	115.68	137.82
Hepatitis	22.22	36.1	29.68	27.35	29.92

It is observed from Table III, Table IV, Table V, Table VI, and Table VII that ExtraImpute performs better than other imputation methods under different missing ratio. The RMSE of ExtraImpute in 10% missing ratio is the lowest among selected imputation methods in all datasets. The highest gap was in the Liver dataset with a difference of approximately 30 to 65 integers. While the lowest was in the Spine dataset with approximately 0.3 difference from missForest. The gap increased under 20% missing ratio and ExtraImpute continues to outperform other imputation methods in all datasets. In 30% missing ratio, ExtraImpute maintains the lead followed by missForest, MICE, KNNI, and SoftImpute respectively. Between 30% and 40% of missingness the increment rate of RMSE starts to lessen for ExtraImpute. Moreover, when dealing with over than 40% missing values it was noticed that four out of five imputation methods including our proposed method had a noticeable boost in performance despite the increase in missing ratio.

We have also calculated the average of RMSE obtained from the imputation methods for the five selected healthcare datasets in each missing ratio. Fig. 1 illustrates the overall performance of imputation methods under different missing ratio.

It is observed from Fig. 1 that the higher the percentage of missing values the more the error increases in all imputation methods. This happens due to the decrease of accurate observed variables which are used to predict missing values. From the results, it was found that ExtraImpute performs quite well in comparison to other imputation methods under 10% missingness and maintains the lead even when imputing 50% of missing values. Meanwhile, missForest was the most stable method amongst all imputation methods. It maintains its position in the first three missing ratios before it starts to lose some performance after the 30% missing rate. On the other hand, MICE started close with missForest then its performance decreased dramatically from 20% to 30% missing rates. The reduction continued until 40% before it regained it again at 50%. Both KNNI and SoftImpute performed similarly and had the highest RMSE under all missing ratios. However, KNNI had a remarkable improvement in performance after the 30% missing ratio. From these results, we can see that our proposed imputation method had the lowest error compared to existing imputation methods even when dealing with high missing ratios. ExtraImpute have the ability to handle a large portion of missing values by making an initial guess of all missing values from other observed values using Linear Interpolation method. Also, it takes advantage of the estimated values and use them for predicting missing values in other features starting from attributes with the lowest missing rate, making the increase in error rate less than other imputation methods.

After observing the time consumed for imputing the missing values using the selected methods, it was found that our proposed method achieved better timing than missForest. However, the processing time for KNNI, MICE, and SoftImpute was under one second in all datasets under different missing ratio scenarios which is

lower than our proposed method. This can be explained by the fact that ExtraImpute is an iterative method and it was developed to attain better results by imputing each feature carefully instead of using the whole sample to perform the prediction. Additionally, data imputation is a processing step where time is insignificant compared to prediction time using classification approaches. The average execution time for each imputation method under different settings is included in Table VIII. The mentioned execution time was estimated by calculating the average of 10 runs for each scenario.

TABLE VIII. AVERAGE EXECUTION TIME OF THE FIVE IMPUTATION METHODS ON THE SELECTED DATASETS (IN SECONDS)

Dataset	Missing ratio	ExtraImpute	missForest	KNNI	MICE	SoftImpute
Diabetes	10%	4.07	4.45	0.02	0.06	0.05
	20%	3.93	5.47	0.03	0.08	0.07
	30%	3.62	5.39	0.38	0.15	0.07
	40%	3.38	5.259	0.043	0.108	0.077
	50%	3.08	6.11	0.04	0.1	0.07
Spine	10%	3.93	6.83	0.015	0.074	0.043
	20%	3.76	9.61	0.015	0.086	0.047
	30%	3.465	8.158	0.017	0.154	0.067
	40%	3.44	11.8	0.0174	0.14	0.065
	50%	3.38	10.55	0.015	0.114	0.063
Heart	10%	4	7.65	0.016	0.1	0.0445
	20%	3.96	9.1	0.019	0.12	0.059
	30%	3.88	9.14	0.019	0.229	0.068
	40%	3.76	9.01	0.01	0.12	0.06
	50%	3.65	6.02	0.019	0.078	0.07
Liver	10%	3.61	5.27	0.02	0.05	0.06
	20%	3.25	5.14	0.02	0.11	0.06
	30%	3.14	7.13	0.02	0.12	0.06
	40%	2.99	4.07	0.02	0.116	0.069
	50%	2.74	5	0.02	0.1	0.07
Hepatitis	10%	4.78	7.76	0.02	0.05	0.04
	20%	4.43	8.76	0.02	0.13	0.06
	30%	4.09	8.48	0.025	0.116	0.06
	40%	3.8	5.95	0.02	0.115	0.07
	50%	3.4	6.89	0.031	0.05	0.076

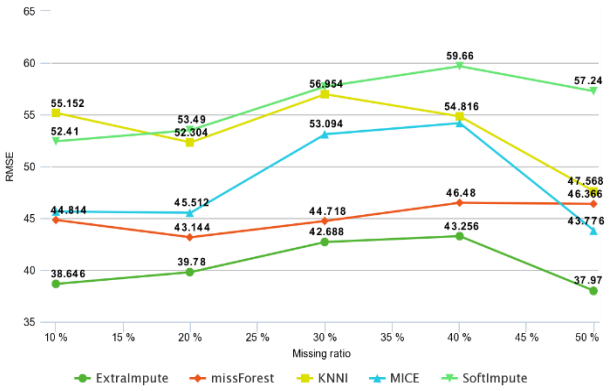

Figure 1. RMSE comparison among ExtraImpute and other imputation methods under different missing ratio.

V. CONCLUSION

In this study, we developed a novel imputation approach named ExtraImpute to solve the problem of missing values. This method is suitable for numerical features of healthcare datasets. ExtraImpute uses Extra Randomized Trees of ensemble machine learning methods to estimate continuous and discrete missing values. The proposed algorithm was tested on five benchmark healthcare datasets. Missing values were artificially generated under MCAR mechanism at 10%, 20%, 30%, 40%, and 50% to test the efficiency of ExtraImpute against other well-known imputation methods. Altogether, 1,250 experiments were conducted (5 methods × 5 datasets × 5 different scenarios 10 times). The performance of each imputation method was calculated using R^2 and RMSE on the selected datasets under different missing ratio. The experiment results shows that ExtraImpute performed better than other existing imputation methods under various missing rate. The number of performed experiments, in addition to the results verifies that the experimental procedure is reliable. Considering the fact that other imputation methods were selected to affirm the superiority of our proposed method. For future work, other missingness mechanisms including MAR and MNAR should be investigated since data related to these mechanisms are quite challenging. Thus, a robust imputation method able to handle such scenarios will be appealing to many imputation problems.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Mustafa Alabadla conducted and analyzed the research; Fatimah Sidi, Iskandar Ishak, Hamidah Ibrahim, Lilly Suriani Affendey and Hazlina Hamdan are supervised the research; all authors had approved the final version.

ACKNOWLEDGMENT

This work was supported by Universiti Putra Malaysia and Fundamental Research Grant Scheme (Grant No. FRGS/1/2020/ICT06/UPM/02/1) funded by Ministry of Higher Education, Malaysia.

REFERENCES

[1] M. H. Chowdhury, M. K. Islam, and S. I. Khan, "Imputation of missing healthcare data," in *Proc. 20th Int. Conf. Comput. Inf. Technol.*, 2018, pp. 1-6, 2018.
[2] J. Kaiser, "Dealing with missing values in data," *J. Syst. Integr.*, pp. 42-51, 2014.
[3] S. Phung, A. Kumar, and J. Kim, "A deep learning technique for imputing missing healthcare data," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, 2019, pp. 6513-6516.
[4] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581-592, 1976.
[5] N. J. Horton and K. P. Kleinman, "Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models," *Am. Stat.*, vol. 61, no. 1, pp. 79-90, 2007.
[6] B. L. Ford, "An overview of hot-deck procedures," *Incomplete Data in Sample Surveys*, vol. 2, part iv, 1983.
[7] J. W. Graham, "Missing data analysis: Making it work in the real world," *Annu. Rev. Psychol.*, vol. 60, pp. 549-576, 2009.
[8] W. Young, G. Weckman, and W. Holland, "A survey of methodologies for the treatment of missing values within datasets: Limitations and benefits," *Theor. Issues Ergon. Sci.*, vol. 12, no. 1, pp. 15-43, 2011.
[9] A. Farhangfar, L. A. Kurgan, and W. Pedrycz, "A novel framework for imputation of missing values in databases," *IEEE Trans. Syst. Man, Cybern. Part ASystems Humans*, vol. 37, no. 5, pp. 692-709, 2007.
[10] P. H. Rezvan, K. J. Lee, and J. A. Simpson, "The rise of multiple imputation: A review of the reporting and implementation of the method in medical research Data collection, quality, and reporting," *BMC Med. Res. Methodol.*, vol. 15, no. 1, pp. 1-14, 2015.
[11] S. I. Khan and A. S. M. L. Hoque, "An analysis of the problems for health data integration in Bangladesh," in *Proc. Int. Conf. Innov. Sci. Eng. Technol.*, 2016, pp. 10-13.
[12] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, pp. 8869-8879, 2017.
[13] T. Razzaghi, O. Roderick, I. Safro, and N. Marko, "Multilevel weighted support vector machine for classification on healthcare data with missing values," *PLoS One*, vol. 11, no. 5, pp. 1-18, 2016.
[14] D. J. Stekhoven and P. Bühlmann, "Missforest-Non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112-118, 2012.
[15] G. E. A. P. A. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," *Appl. Artif. Intell.*, vol. 17, no. 5-6, pp. 519-533, 2003.
[16] G. Madhu, B. L. Bharadwaj, G. Nagachandrika, and K. S. Vardhan, "A novel algorithm for missing data imputation on machine learning," in *Proc. 2nd Int. Conf. Smart Syst. Inven. Technol*., 2019, pp. 173-177.
[17] J. C. Kim and K. Chung, "Multi-modal stacked denoising autoencoder for handling missing data in healthcare big data," *IEEE Access*, vol. 8, pp. 104933-104943, 2020.
[18] P. S. Raja and K. Thangavel, "Missing value imputation using unsupervised machine learning techniques," *Soft Computing*, vol. 24, no. 6, 2020.
[19] N. Fazakis, G. Kostopoulos, S. Kotsiantis, and I. Mporas, "Iterative robust semi-supervised missing data imputation," *IEEE Access*, vol. 8, pp. 90555-90569, 2020.
[20] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3-42, 2006.
[21] J. Gall, A. Yao, N. Razavi, L. V. Gool, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2188-2202, 2011.
[22] N. S. Raju, R. Bilgic, J. E. Edwards, and P. F. Fleer, "Methodology review: Estimation of population validity and cross-validity, and the use of equal weights in prediction," *Appl. Psychol. Meas.*, vol. 21, no. 4, pp. 291-305, 1997.
[23] O. Troyanskaya, *et al.*, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520-525, 2001.
[24] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, "Multiple imputation by chained equations: What is it and how does it work?" *Int. J. Methods Psychiatr. Res.*, vol. 17, suppl 1, no. 1, pp. 40-49, 2011.
[25] T. Hastie, R. Mazumder, J. D. Lee, and R. Zadeh, "Matrix completion and low-rank SVD via fast alternating least squares," *J. Mach. Learn. Res.*, vol. 16, pp. 3367-3402, 2015.

Mustafa Alabadla is a Software Engineer with an extensive experience in developing cross-platform mobile apps and UI design. He received his BSc. in Information Technology from College of Science and Technology, Palestine, in 2011, And MSc. in Informatics from Universiti Sains Malasyia, in 2019. He is currently pursuing the Ph.D. degree with the Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. His research interests include machine learning, database systems, software engineering and information systems. He has worked in several IT fields including web development, database administration, computer networks, digital marketing and software engineering.

Fatimah Sidi received the Ph.D. degree in management information system from Universiti Putra Malaysia (UPM), Malaysia, in 2008. She is currently working as an Associate Professor in the discipline of computer science with the Department of Computer Science, Faculty of Computer Science and Information Technology, UPM. Her current research interests include knowledge and information management systems, data and knowledge engineering, database, data warehouse, big data and data analytics.

Iskandar Ishak received the Bach. of Information Technology from Universiti Tenaga Nasional, Malaysia. He received the Master of Technology (Information Technology) from the Royal Melbourne Institute of Technology Australia. He received the Ph.D. degree in Computer Science from the Universiti Teknologi Malaysia. His research interests are in the field of database systems, big data and data analytics.

Hamidah Ibrahim received the Ph.D. degree in computer science from the University of Wales, Cardiff, U.K., in 1998. She is currently a Full Professor with the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM). Her current research interests include databases (distributed, parallel, mobile, biomedical, and XML) focusing on issues related to integrity maintenance/checking, ontology/schema/data integration, ontology/schema/data.

Lilly Suriani Affendey is an Associate Professor in the Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. She received her Bachelor of Computer Science degree in 1991 from the Universiti Pertanian Malaysia and in 1994 received her MSc in Computing degree from the University of Bradford, UK. In 2007, she received her PhD degree from Universiti Putra Malaysia. Her current research interest is in Multimedia Databases, Video Content-based Retrieval, Data Science and Big Data Analytics. She teaches Database Application Development, Database Systems, Business Analytics, and Big Data Analytics. She has recently attended trainings on Data Science, RapidMiner, Talend, and Hadoop.

Hazlina Hamdan Hazlina Hamdan is a senior lecturer in the Department of Computer Science at Universiti Putra Malaysia. She graduated with a BSc (Hons) in Computer Science from Universiti Kebangsaan Malaysia. She holds a Master of Computer Science in Artificial Intelligence from Universiti Malaya and her PhD is from the University of Nottingham, UK. Her research areas are intelligent computing and application such as medical prognostic, pattern recognition, prediction system, optimization.