IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 July 2023, accepted 12 September 2023, date of publication 20 September 2023,
date of current version 26 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3317775

RESEARCH ARTICLE

Imputation of Missing Clinical Covariates for Downstream Classification Problems

BENJAMIN AGBO[1], **HUSSAIN AL-AQRABI**[2], **TARIQ ALSBOUI**[3], **MUHAMMAD HUSSAIN**[3], **AND RICHARD HILL**[3]

[1]School of Computer Science and Mathematics, Keele University, ST5 5BG Stoke on Trent, U.K.
[2]School of Computer Information Systems, Higher Colleges of Technology, Sharjah, United Arab Emirates
[3]School of Computing and Engineering, University of Huddersfield, HD1 3DH Huddersfield, U.K.

Corresponding author: Benjamin Agbo (b.agbo@keele.ac.uk)

ABSTRACT Noticeable growth in the use of intelligent devices has resulted in the generation of vast amounts of data from sensor devices. When dealing with large amounts of data, it is common to observe databases with large amounts of missing values. This is a challenge for data miners because various methods for data analysis only work well on complete databases. A traditional approach to handling missing data is to discard instances of missing values and only use complete cases for analysis. However, research has shown that this approach is not practical especially when large amounts of data are missing. This led to an increased need to develop strategies for replacing missing values with plausible values through imputation. This study presents an imputation strategy called *med.BFMVI* for recovering missing values before training downstream classification models. Experiments simulated missingness from 10% to 40% using MCAR and MAR mechanisms and the performance of the proposed technique was measured against state-of-the-art techniques. Overall, the proposed algorithm recorded the best imputation accuracy as opposed to benchmark techniques and showed significant improvements on downstream learning.

INDEX TERMS Electronic health records (EHR), imputation, Internet of Things, missing data.

I. INTRODUCTION

The application of machine learning techniques in healthcare can lead to the generation of actionable insights ranging from streamlining operations in hospitals, toxicity prediction to early detection of diseases [5], [38]. Statistical techniques that exploit the richness and variety of clinical data are relatively sparse, creating an avenue for further research in this area [1]. As vast amounts of information are available today, experts in the medical field have become reliant on machine learning techniques for performing various tasks [2], [8]. These information will be generated from numerous epidemiological and clinical sources such as claims records, data from longitudinal studies and clinical trials which over time have become invaluable assets for medical research. In most of these studies, data is gathered over time from individual subjects via repeated or continuous monitoring and assessment of both health outcomes and risk factors. For instance, longitudinal studies are used to find the correlation between levels of exposure and health effects such as chronic diseases. Originally, these studies collect prospective data, prior to the knowledge of future events, therefore mitigating bias from the responses of participants [3].

Another popular and valuable means of obtaining clinical data are Electronic Health Records (EHR). The widespread adoption of EHR over the years has led to the generation of massive amounts of data containing qualitative, quantitative and transactional data [4]. While primarily developed for storing patient records and enhancing administrative tasks, researchers explored secondary applications for EHRs in clinical informatics [5]. EHRs rely on data collected by various end devices, and data is collected instantly by physical objects that incorporate sensors, and network connectivity [7].

The issue of missing data is a prevalent challenge that poses significant challenges in the interpretation and analysis of longitudinal clinical data sets [9], potentially diminishing

The associate editor coordinating the review of this manuscript and approving it for publication was Alessandro Pozzebon.

their plausibility and producing biased conclusions. The presence of missing values may cause complications in the interpretation of important insights in a study or even invalidate the entire study [10]. As machine learning algorithms and statistical models rely on fully observed data sets, it is important to develop appropriate strategies to handle missing data effectively.

Classification algorithms such as Random Forests [11], Classification and Regression Trees (CART) [12], do not have built-in techniques for addressing missing values present in the training data. By ignoring instances with missing values and using only complete records in the classification algorithm, vital information may be lost in a given distribution [13]. The presence of missing data is a major challenge for experts aiming to solve classification problems in real-life studies [14].

The aim of classification problems is to develop a classifier from a sample of training data to ensure the correct classification of new test observations. In the training set, the class membership is assumed to be stated for each observation whereas missing values may be present in corresponding features/attributes. The test data on the other hand will consist of newly observed records with similar but unlabelled features. The goal of classification problems is to effectively assign class labels to the test data [15]. The problem formulation assumes a Missing Completely at Random (MCAR) and Missing at Random (MAR) mechanism in the training and test data set. This research identifies one approach to classification where instances with missing values are ignored before building a classifier. This approach can only yield effective results when the amount of missing data is relatively low. However, research has shown that adequate imputation techniques can improve classification accuracy even for a missing rate of 5% [16], [17].

A. LITERATURE REVIEW

The issue of missing data has gained attention from experts in various domains. Numerous research efforts have been directed towards tackling the problem of incomplete data by attempting to devise improved imputation methods that are more precise and dependable. In this section, we examine different research studies and recent undertakings that aim to address this issue.

Research conducted by [18] assessed the accuracy of random forest-based imputation for datasets with non-linearity, non-normality and interaction in biomedical research. To evaluate the effect of these factors, datasets were simulated based on the missing at random (MAR) mechanism and imputation was carried out using RF-based techniques (missForest and CALIBERrfimpute), and their performances were evaluated in comparison to predictive mean matching (PMM). Both RF-based imputation techniques showed high imputation accuracy. However, CALIBERrfimpute showed better performance when estimating regression coefficients as opposed to missForest which produced highly biased regression coefficients especially for skewed variables in non-linear models.

The KNN method is a machine learning algorithm which approaches imputation by classifying the closest neighbours to missing values and using these neighbours to impute missing values based on a distance measure between points. In [19], missing data was imputed base on the K-NN algorithm considering different mechanisms and missing data models. The authors further assessed the performance of imputation techniques using the Naive Bayes algorithm. Results from the study showed that the accuracy of the imputation technique closely matched the accuracy of the original complete data.

A study conducted by [34] assessed the performance of multiple imputation in clinical/epidemiological research. The results showed that the multiple imputation technique produced unbiased estimates for both Missing at Random and Missing Completely at Random mechanisms when compared with traditional techniques such as LOCF and mean substitution. [36] showed the merit of MI for imputing missing values, especially when the rate of missing data is above 10%. An advantage of multiple imputation over single imputation is that single imputation methods tend to underestimate the variance that may exist in a given distribution in some cases. Therefore, MI methods have been proposed to overcome this limitation [13].

The missForest imputation technique was introduced by [35], which builds upon the random forest approach. This method involves combining multiple unpruned regression and classification trees through averaging. The authors evaluated its effectiveness on numerous datasets from biological fields, introducing artificial missing values to evaluate the accuracy of the technique under various missing data rates. The results showed that the missForest method efficiently handles both continuous and categorical missing data. In comparison to alternative imputation methods like KNN, the study revealed that missForest outperformed other comparative techniques, particularly in scenarios where the dataset exhibited non-linear relationships and intricate interactions.

The authors in [37] {proposed a stochastic imputation technique based on the bayesian framework. An uncertainty aware attribute was introduced, which accounted for uncertainties in the prediction model. The technique showed a high performance as opposed to comparative techniques when evaluated using real-world EHR data, MIMIC-III.

A straightforward approach to handling missing data perhaps is by learning a different classifier on the different patterns of the observed values in a data set. A study conducted by [29] used this approach in conjunction with neural networks to investigate the diagnosis of thyroid diseases. This is referred to as the network reduction approach. Standard discriminative classifiers can be fitted to learn each model and [29] observed that each subspace of observed features learned on a neural network classifier produced better performance compared to regression imputation based on neural network (NN) combined with an NN classifier considering all the features as inputs. A major drawback of this approach however is that the number of missing patterns on features is exponential based on the number of features considered [30]. In the case of the data set considered by [29], four inputs with four missing patterns on features were considered which made the approach more feasible.

This study builds upon previous research conducted by [39] and presents a novel missing value imputation technique for classification problems. The algorithm exploits the class boundaries of a target variable and chooses the best fit objective function for imputation from a list of predefined sub-techniques.

B. CONTRIBUTIONS OF THE STUDY

As various imputation techniques exist in literature, when faced with real-life missing instances where there is no ground truth data, it is important to have imputation approaches that will embed the capability of selecting optimal algorithms for imputing missing values. This paper proposes an imputation algorithm called *med.BFMVI* that is capable of choosing appropriate techniques for filling-in missing instances based on established features and labels in a given distribution.

A validation technique was also developed using a reverse error score function RES(r) that is based on two error calculations between two final imputation estimates, which is used to obtain final imputation results for filling in missing instances. The performance of pre-defined sub-techniques were weighed against benchmark techniques and for each sub-technique, this study demonstrates that selecting the best plausible sub-technique improves the accuracy of downstream classification tasks in clinical data.

II. METHODOLOGY

This paper follows the recommendations of scientific research for simulating missing data, paying attention to the principles outlined by [31].

A. DATA COLLECTION

The data used for conducting experiments were reviewed and collected from secondary data sources through the University of California Irvine (UCI) machine learning repository. The datasets were carefully validated from the works of previous authors and selected to describe the problem statements considered in this study. For experimental purposes, we consider a dataset where all data entries are known. Instances with real missing values were disregarded from analyses because access to the true values of missing instances was required in order to measure the performance of the imputation algorithms. Therefore, artificial missing data was simulated at different rates ranging from 10% to 40% of the overall observations in a distribution. The choice of missing rates was inspired by the works of [32] which confirmed that missing data in IoT smart spaces reaches approximately 40% following the analysis of eight streams of IoT data. Missing data was also simulated based on Missing Completely at Random (MCAR) and Missing at Random (MAR) mechanisms.

TABLE 1. Original CVD dataset before the introduction of artificial missing data.

Observations	13,030
Attributes	75
Classes	2

B. PERFORMANCE EVALUATION

The performance of each imputation algorithm is evaluated based on the following metrics:

- **Root mean squared error (RMSE) and Mean Absolute Error (MAE)**: This measures the accuracy and precision of the imputation algorithms (how close the predicted values are to the ground truth) and is given by

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}(P_{ij} - \hat{P}_{ij})^2} \tag{1}$$

and

$$MAE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}(|P_{ij} - \hat{P}_{ij}|)} \tag{2}$$

where P_{ij} is the original (true) value and $\hat{P}_{ij}$ represents the imputed values. It is important to note that an error score closer to 0 indicates better performance.
- **Classification accuracy**: This is measured by the false negative and false positive error rates [33]. The false negative error is based on a probabilistic condition that a respondent is classed under a category that is lower than the respondent's true category. Similarly, the false positive error measures the probability of a respondent being classified under a higher category than the respondent's true category. This can be derived by the formula below;

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{3}$$

C. PROBLEM STATEMENT

Let $P = \{P_i\}_{i=1}^{n}$ be an $n \times p$- dimensional matrix of n distinct observations having p attributes/features and V is a response variable having class labels that are influenced by P. This research takes into account no dependence structure between the attributes in P. Let D be an $n \times p$ matrix showing the missingness of the corresponding features of P which was simulated at 10% - 40% missing rates. In practice, incomplete data is generated for a random size n of the population (P, V, D) set as the training data which was used to train the classifier

$$D = \{(P_i, V_i, D_i)\}_{i=1}^{n}, \tag{4}$$

where the class labels in $V_{i=1}^{n}$ are fully observed, $P_i = (P_{ij})_{j=1}^{m} = (P_{i1}, \ldots, P_{ip})$ denotes the m features of the

i-th observation measured along the indicator variable $D_i = (D_{ij})_{j=1}^m$ where

$$D_{ij} = \begin{cases} 0, & P_{ij} \text{ is missing} \\ 1, & \text{otherwise.} \end{cases} \tag{5}$$

without any loss of generality in the matrix, lets assume that for each i, the observation $P_i = (P_{ij})_{j=1}^m$ contains m_0 categorical attributes for $j \in \{1, 2, \ldots, m_0\}$ and m_1 continuous features for $j \in \{m_0+1, \ldots., m_0+m_1\}$ such that $m_0+m_1 = m$. Let the j-th categorical attribute contain k_j distinct values of the j-th continuous variable that represents the $(m_0 + j)$-th feature of P_i, indexed by $j \in \{1, \ldots, m_1\}$ and takes the values from a continuous set $C_j \subset \mathbb{R}$. We can map the k_j distinct values to the initial k_j natural values for each categorical features, such that $P_i \in \{1, \ldots, k_1\} \times \cdots \times \{1, \ldots, k_{po}\} \times C_1 \times \cdots \times C_{p1} \subset \mathbb{R}^m$.

Here, lets assume that the $\{(P_i, V_i)\}_{i=1}^n$ satisfies the model

$$V_i = g(P_i), \quad i = 1, 2, \ldots, n, \tag{6}$$

where $g(.)$ denotes an unknown function which maps a p-dimensional number (which belongs to a subspace of $\mathbb{R}^m$) to a discrete set R which represents the class labels and $V_i \in R$.We assume that R has m values and therefore, the classification problem is established from m classes before imputing missing value. In order to consider closely related covariates in the imputation model, the algorithm selects instances that fall within the same class and uses the representative covariates to estimate missing instances.

It is noteworthy that missing values can also be present in the test set V'.

III. CLASS-WEIGHTED med.BFMVI IMPUTATION

This study tailors the *med.BFMVI* for classification problems and encodes the target variable. Considering the data with m categories, the class variables are simply recreated based on m-1 assuming that the final class variable is already known or is dependent on other identified variables.

This dependence is mathematically expressed as:

$$\sum_{i=1}^{m} p_i, \quad p_i \in \{0, 1\} \tag{7}$$

where p_i represents the i-th binary variable and $p_i \in \{0, 1\}$ simply requires that the class variables must reflect the available boolean variables. The above equation can also be represented as:

$$p_j = 1 - \sum_{i \neq j} p_i \tag{8}$$

which clearly shows the linear dependence that exists between other variables and the j-th variable. The interaction between the binary variables in the classification problem will always be represented as

$$p_i.p_j = 0, \quad i \neq j \tag{9}$$

as both variables are mutually exclusive, meaning when the first value is 0, the other value is 1.

The goal of the classification task is to make use of a training set$\{(P_i, V_i)\}_{i=1}^n$ to produce estimates for $g(.)$. Considering a new set of L observations, $P' = \{P'_i\}_{i=1}^L$, the corresponding classes $V' = \{P'_i\}_{i=1}^L$ are predicted by the classifier using $\hat{V}'_i = \hat{g}(P'_i)$, where bootstrap samples were used to construct A number of trees from the training set at random and averaged to improve the accuracy of classification based on the equation:

$$\hat{f}_{bagg}(P) = \frac{1}{A} \sum_{a=1}^{a} \hat{f}_a(P) \tag{10}$$

The typical process involved in the imputation algorithm relies on three prescribed techniques. The first technique is based on the $k - NN$ algorithm where vectors $x^D_{(1)}, \ldots, x^D_n$ with $d(x_{i,}x_{(1)}) \leq, \ldots, \leq d(x_{i,}x_{(k)})$ represents the rows of the matrix X^D, and $d(x_{i,}x_{(k)})$ is the distance given by $Dist_{xy} = \sqrt{\sum_{k=1}^m (X_{ik} - Y_{jk})^2}$. For each point (y) in C_i, the distance (x,y) between the missing point and nearest imputed value is stored in a similarity array (S). The array (S) is sorted in descending order and the top K data for (y) in C_i is selected for imputation.

The process is repeated using a new imputation algorithm and missing values within clusters are imputed using a regression model where for each matrix C_i, the data was split and the regression model was trained on the response $y^{(s)}_{obs}$ and predictor variable $x^{(s)}_{obs}$. The trained regression model is then used to predict the missing values in X_s.

The final imputation technique (missForest) is initiated by pre-imputing the missing values in X with the mean of the distribution, after which the predictors $X_s = 1, \ldots, p$ are stacked in ascending order considering the amount of values that are missing. Each missing value in X_s is then imputed by first of all fitting the rf on the response $y^{(s)}_{obs}$ and predictor variable $x^{(s)}_{obs}$. Next, the trained rf model is applied to $x^{(s)}_{miss}$ to predict the missing values of $y^{(s)}_{miss}$. The imputation process is repeated until the set stopping criterion (γ) has been met. This is achieved when the difference between the most recent imputed data matrix and the old matrix has an increase for the first time, considering the variable types present. Lets take the $n \times p$ matrix to be a set of continuous variables in the proposed approach. Therefore, the difference in the new and previous imputed matrix $\boldsymbol{N}$ is defined as:

$$\Delta_N = \frac{\sum_{j \in N} (X^{imp}_{new} - X^{imp}_{old})^2}{\sum_{j \in N} (X^{imp}_{new})^2} \tag{11}$$

The class weight γ_j from the j-th attribute is used, which serves as a basis for weighing the *med.BFMVI* between observations as follows:

$$\text{med.}BFMVI(p_a, \ldots, p_n) = \sum_{j=1}^{m} \gamma_j \text{med.}BFMVI(p_{aj}, \ldots, p_{nj}). \tag{12}$$

Algorithm 1 Class-Weighted *med.BFMVI* Imputation

```
Input: (P,V,D) with P ⊂ $\mathbb{R}^{n\times p}$ having missing values, V contains $m$ class labels
Output: Imputed matrix $\widehat{P}$.
1: Pre-processing the Data: Lets first of all transform class variables of $V$ using (8).
2: Initialisation: Lets consider the class labels in $V$ and use them as a basis for splitting $P$ in to $\{P^v\}_{v=1}^m$ considering their weights $\gamma_j$. Take each class $v$ given $P^v$ and pre-impute missing continuous features $p_1$ using mean imputation.
3: Iterative Step: For each missing instance $i$ in class $v$, the imputed data matrix $\widehat{P}^{v,t}$ is derived using med.$BFMVI(p_a, \ldots, p_n) = \sum_{j=1}^m \gamma_j$ med.$BFMVI(p_{aj}, \ldots, p_{nj})$. This process is repeated for each $v$ using each sub-technique from III to obtain $\widehat{P}^t = \{\widehat{P}_{v,t}\}_{v=1}^m$.
4: Error Calculation: Lets use the error score function in 15 to determine the choice of imputation. The sub-technique with the lowest error is used to impute missing values for instances in $v$.
5: Stopping Criteria: Stop when the imputed values with the lowest RES(r) score is used to impute missing values in $\widehat{P}$
```

The result of the *rf* imputation is then set as the threshold, which is placed as the initial value of the first estimate generated in Algorithm 1. A reverse error score function RES(r) representing the error between γ and the final sequence β_{Ci} in each group C_i is expressed as:

$$M_{\gamma} = \frac{\partial(\gamma)}{\partial_n} = \sqrt{\frac{\sum_{i=1}^{N}(X_{\gamma} - \hat{y}_{\beta_{ci}})^2}{n}} \tag{13}$$

$$M_{\beta_{ci}} = \frac{\partial(\beta_{ci})}{\partial_n} = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_{\beta_{ci}} - X_{\gamma})^2}{n}} \tag{14}$$

$$RES(r) = M(\gamma, \beta_{ci}) = \left[\frac{\partial(\gamma)}{\partial_n}\frac{\partial(\beta_{ci})}{\partial_n}\right] \tag{15}$$

where γ is the imputation threshold for C_i and β_{Ci} is the best estimate from the previously chosen imputation techniques.

By using the class weights, imputation of the missing continuous variables is conducted based on the sub-techniques considered in section III and the score function in Equation 15.

IV. EXPERIMENTAL SETUP

For the purpose of reporting the comprehensive performance of classification techniques on real-world applications, this research evaluates the accuracy of imputation techniques on classification problems of real-world clinical data reporting Cardiovascular Diseases (CVD) among a range of participants obtained from the UCI Machine Learning Repository [23]. CVD are mostly identified as conditions that involve the blockage of blood vessels thereby causing ischemic heart disease (IHD) (angina, myocardial infraction) or stroke [24]. These conditions prevent the flow of blood to the brain or heart. To obtain the binary classification, various health and lifestyle factors of participants were considered in the dataset to identify plausible diagnosis of cardiovascular diseases.

TABLE 2. **Missing data mechanisms used for the generation of missing data *M* in the data set *P*. lets take *f* to be the density of the missing data pattern. P^{miss} and P^{obs} represent the missing and observed data respectively.**

Missing Data Mechanism	Statistical Assumption
Missing at Random (MAR)	$f(M\|P^{obs}, P^{miss}) = f(M)$
Missing Completely at Random (MCAR)	$f(M\|P^{obs}, P^{miss}) = f(M\|P^{obs})$
Not Missing at Random (NMAR)	$f(M\|P^{obs}, P^{miss})$ is a function of P^{miss}

For the purpose of experiments, missing data is generated at different rates ranging from 10% - 40% for different missing data mechanisms. The fully observed data generated from CCA was taken as the ground truth. The different imputation techniques were applied on the range of datasets generated and their performances were compared against all the techniques embedded in the optimised imputation algorithm.

Because different missing data mechanisms affect the quality of imputation, simulations were conducted considering two missing data mechanisms: missing at random (MAR) and missing completely at random (MCAR). These statistical assumptions are summarised in Table 2. In order to generate the MCAR mechanism, a subset of the records in P was sampled at random, assuming that each entry has equal probability of being chosen. The MAR mechanism was generated by sampling the entire dataset and modelling the missing data probability for the target variable. For instance, if $R_i = 1$ the corresponding value of P is deleted.

This study compares the proposed algorithm with six well established imputation methods as follows:

V. RESULTS

The imputation methods were tested on real-world clinical data obtained from the UCI Machine Learning Repository. This data contains $n = 10,000$ observations and $p = 11$ dimensions. Next, the results show that the quality of imputation produced from the *med.BFMVI* sub-methods is higher compared to other benchmark methods, which further leads to an improvement in the performance of downstream classification tasks. This research evaluates the performance of these methods considering different missing rates for MACR and MAR conditions.

A. IMPUTATION ACCURACY

The imputation accuracy for each technique was evaluated, assuming the MCAR condition. Among all the methods considered, Figure 1 and 2 shows that at least one of the *med.BFMVI* techniques record the lowest RMSE and MAE scores for 10% - 40% missing rate indicating strong performance, followed by the imp.mice technique which is closely

TABLE 3. Average computational complexity for benchmark and *med.BFMVI* imputation techniques at 30% Missing Rate.

			Time (s)				
Missing Pattern	imp.knn	imp.mice	imp.missforest	$med.BFMVI_{reg}$	$med.BFMVI_{knn}$	$med.BFMVI_{rf}$	imp.mean
MCAR	6.07	0.76	3.36	0.13	2.37	4.06	0.93
MAR	6.12	0.73	3.43	0.14	2.26	3.31	0.70

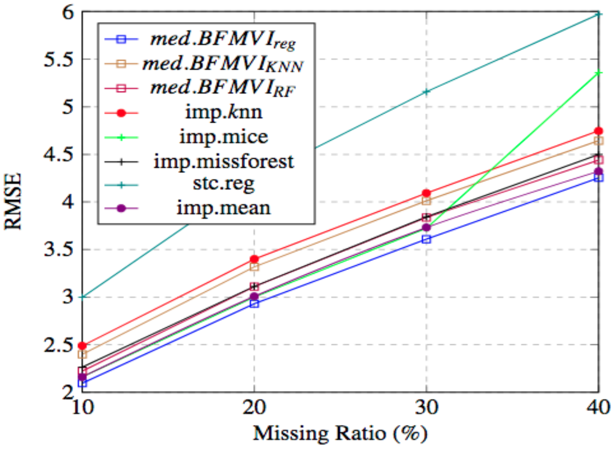


FIGURE 1. RMSE of imputation algorithms for MCAR data.

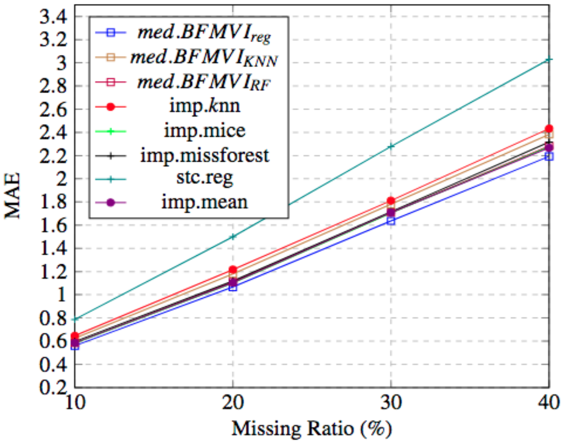


FIGURE 2. MAE of imputation algorithms for MCAR data.

followed by *missForest* outside the sub-methods embedded in the *med.BFMVI* technique. Comparatively, the *stc.reg* technique showed the weakest performance for the MCAR condition, followed by the benchmark *knn* technique and the *knn* based sub-technique in the *med.BFMVI* method.

The experiments were repeated for the MAR condition. Comparatively, for all the missing data ratios, the proposed method still shows the best performance in terms of imputation accuracy with the lowest RMSE and MAE for all missing data ratios as seen in Figures 3 and 4. Among the benchmark methods *stc.reg* still shows the weakest imputation

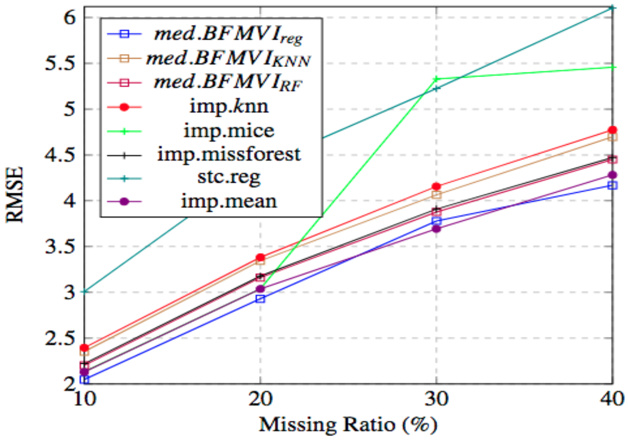


FIGURE 3. RMSE of imputation algorithms for MAR data.

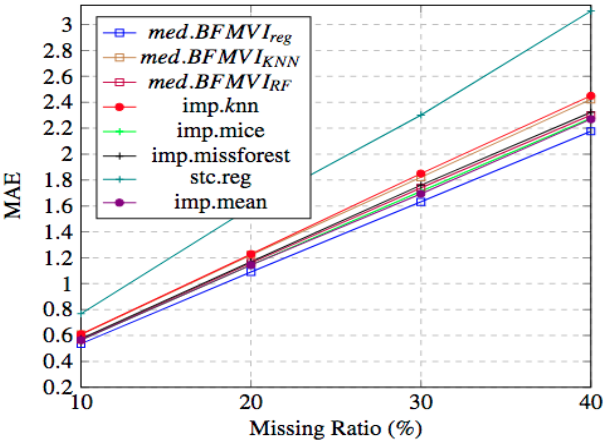


FIGURE 4. MAE of imputation algorithms for MAR data.

performance. It can be noted that the benchmark *imp.mice* approach performs well when 10% - 20% of the data is MAR. However, for higher missing rates 30% - 40%, the *regression* approach shows the weakest RMSE score.

B. COMPUTATIONAL COMPLEXITY

Next, the computational complexity of imputation methods were compared, showing the time required to complete a cycle of imputation for the dataset with $n = 10000$ observations across each identified missingness pattern. Simulations were still conducted on a machine having an Intel Core 2 Duo

TABLE 4. **Classification accuracy (%) of CVD data at 30% missing rate.**

Approach	MCAR			MCAR		
	CART	**RF**	**Bagging**	**CART**	**RF**	**Bagging**
No Imputation	85.2	88.1	90.9	84.9	88.3	90.6
imp.knn	85.4	88.3	91.0	85.3	88.4	90.8
imp.mice	85.3	88.2	91.3	80.5	88.3	90.9
imp.missforest	85.0	88.1	91.2	84.8	88.4	91.1
imp.mean	85.3	87.8	90.7	84.6	88.1	90.7
stc.reg	77.7	83.1	87.6	77.1	82.0	86.9
$med.BFMVI_{reg}$	**87.6**	**89.7**	**92.1**	**87.7**	**90.0**	**92.4**
$med.BFMVI_{knn}$	85.6	88.6	91.0	85.7	88.4	90.3
$med.BFMVI_{rf}$	84.8	88.0	90.5	85.0	88.0	90.2

(3.06 GHz) processor which is limited to 8 GB RAM. Results can be seen in Table 3 below.

Among the *med.BFMVI* methods, the regression based imputation scales very well considering the sample size *n* and dimension *p* for both MCAR and MAR mechanisms. Despite its imputation quality, the random forest based imputation performs relatively poor when compared to the other sub-techniques. Among the benchmark methods, *imp.knn* imputation performs poorly for both MCAR and MAR conditions.

C. IMPUTATION PERFORMANCE ON DOWNSTREAM CLASSIFICATION TASKS

In this section, the performance of machine learning classification algorithms trained on the range of imputed data is assessed. The challenge of classification tasks in completely observed data also differs widely across data sets.

In Table 4, the effect of the imputation methods on the accuracy of downstream classification tasks for MAR and MCAR scenarios are presented. Each benchmark methods and each individual *med*.BFMVI method were trained on classification problems to highlight performance gains across each imputation technique. Simulations were further conducted using 30% missing data rate.

Similarly, when trained on downstream classification algorithms, it can be observed that *med.BFMVI* shows comparative performance against benchmark techniques. As seen in Table 4, imputation done on MCAR scenario shows performance gains on all imputed data as compared to the unimputed data. Overall, from the sub-technique of the proposed approach, the $med.BFMVI_{reg}$ shows the best performance when trained on the classification algorithms with an accuracy of 89.7% when trained on the random forest classifier. An ensemble meta-estimator called bootstrap aggregation was used to train subsets of the imputed data and aggregate the individual predictions using voting technique. This led to a 1.32% increase in classification accuracy between the chosen sub-technique from *med.BFMVI* and unimputed data. Similar response can be observed in the MAR scenario with $med.BFMVI_{reg}$ showing a bagged classification accuracy of 92.4%, showing a 1.98% gain when compared to a scenario where the classifier is trained on the unimputed data set.

VI. CONCLUSION

This study presents an imputation approach that is capable of estimating missing values by exploiting the predefined class boundaries of a target variable and choosing the best fit objective function from defined sub-techniques. The algorithm first isolates the parameters of a given category and reproduces a distribution with closely tied membership covariates. The proposed technique empirically out performs benchmark techniques in terms of imputation accuracy for both MCAR and MAR conditions and results in performance gains when trained on classifiers. This technique is particularly useful for clinical researchers aiming to solve classification problems where low to high rates of missing values can be seen. The proposed approach can also be extended to other domains such as Photovolataic (PV) fault detection as this domain relies on historical data from live PV installations, which is not always available in full [25]. In addition to the above, the proposed mechanism can also be integrated into Machine vision application pipelines for pixel imputation due to discrepancies in images caused by scenarios such as occlusion, domians include automated pallet racking [26], industrial food inspection via Convolutional networks [27], and quality inspection within manufacturing facilities [28].

REFERENCES

[1] A. Callahan and N. Shah, "Machine learning in healthcare," in *Key Advances in Clinical Informatics*. New York, NY, USA: Academic, 2017, pp. 279–291.

[2] Z. Obermeyer and E. J. Emanuel, "Predicting the future—Big data, machine learning, and clinical medicine," *New England J. Med.*, vol. 375, no. 13, pp. 1216–1219, Sep. 2016.

[3] E. J. Caruana, M. Roman, J. Hernández-Sánchez, and P. Solli, "Longitudinal studies," *J. Thoracic Disease*, vol. 7, no. 11, pp. E537–E540, 2015. [Online]. Available: https://jtd.amegroups.com/article/view/5822

[4] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," *J. Amer. Med. Assoc.*, vol. 309, no. 13, p. 1351, Apr. 2013.

[5] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 5, pp. 1589–1604, Sep. 2018.

[6] H. Al-Aqrabi, A. P. Johnson, R. Hill, P. Lane, and L. Liu, "A multi-layer security model for 5G-enabled Industrial Internet of Things," in *Proc. Int. Conf. Smart City Informatization (iSCI)*, Guangzhou, China. Singapore: Springer, Nov. 2019, pp. 279–292.

[7] H. Al-Aqrabi, A. P. Johnson, R. Hill, P. Lane, and T. Alsboui, "Hardware-intrinsic multi-layer security: A new frontier for 5G enabled IIoT," *Sensors*, vol. 20, no. 7, p. 1963, Mar. 2020.

[8] H. Al-Aqrabi, R. Hill, P. Lane, and H. Aagela, "Securing manufacturing intelligence for the Industrial Internet of Things," in *Proc. 4th Int. Congr. Inf. Commun. Technol.*, 2020, pp. 267–282.

[9] A. M. Wood, I. R. White, and S. G. Thompson, "Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals," *Clin. Trials*, vol. 1, no. 4, pp. 368–376, Aug. 2004.

[10] J. H. Ware, D. Harrington, D. J. Hunter, and R. B. D'Agostino, "Missing data," *New England J. Med.*, vol. 367, pp. 1353–1354, Jan. 2012.

[11] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[12] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Monterey, CA, USA: Wadsworth, 1984.

[13] R. Little and D. Rubin, *Statistical Analysis With Missing Data*. Hoboken, NJ, USA: Wiley, 2019.

[14] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York, NY, USA: Wiley, 2012.

[15] E. Alpaydin, "*Introduction to Machine Learning*. Cambridge, MA, USA: MIT Press, 2020.

[16] A. Farhangfar, L. Kurgan, and J. Dy, "Impact of imputation of missing values on classification error for discrete data," *Pattern Recognit.*, vol. 41, no. 12, pp. 3692–3705, Dec. 2008.

[17] B. Agbo, H. Al-Aqrabi, R. Hill, and T. Alsboui, "Missing data imputation in the Internet of Things sensor networks," *Future Internet*, vol. 14, no. 5, p. 143, May 2022.

[18] S. Hong and H. S. Lynn, "Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction," *BMC Med. Res. Methodol.*, vol. 20, no. 1, p. 199, Dec. 2020.

[19] D. M. P. Murti, U. Pujianto, A. P. Wibawa, and M. I. Akbar, "K-nearest neighbor (K-NN) based missing data imputation," in *Proc. 5th Int. Conf. Sci. Inf. Technol. (ICSITech)*, Oct. 2019, pp. 83–88.

[20] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc., Ser. B Methodol.*, vol. 39, no. 1, pp. 1–22, Sep. 1977.

[21] B. Agbo, Y. Qin, and R. Hill, "Best fit missing value imputation (BFMVI) algorithm for incomplete data in the Internet of Things," in *Proc. 5th Int. Conf. Internet Things, Big Data Secur.*, 2020, pp. 130–137.

[22] J. Luengo, S. García, and F. Herrera, "On the choice of the best imputation methods for missing values considering three groups of classification methods," *Knowl. Inf. Syst.*, vol. 32, no. 1, pp. 77–108, Jul. 2012.

[23] A. Frank. (2010). *UCI Machine Learning Repository*. [Online]. Available: http://archive.Ics.Uci.Edu/ml

[24] A. Almas, *Depression and Cardiovascular Diseases*. Stockholm, Sweden: Karolinska Institutet Sweden, 2018.

[25] M. Hussain, H. Al-Aqrabi, and R. Hill, "Statistical analysis and development of an ensemble-based machine learning model for photovoltaic fault detection," *Energies*, vol. 15, no. 15, p. 5492, Jul. 2022.

[26] M. Hussain, T. Chen, and R. Hill, "Moving toward smart manufacturing with an autonomous pallet racking inspection system based on MobileNetV2," *J. Manuf. Mater. Process.*, vol. 6, no. 4, p. 75, Jul. 2022.

[27] M. Hussain, H. Al-Aqrabi, M. Munawar, and R. Hill, "Feature mapping for rice leaf defect detection based on a custom convolutional architecture," *Foods*, vol. 11, no. 23, p. 3914, Dec. 2022.

[28] M. Hussain, H. Al-Aqrabi, and R. Hill, "PV-CrackNet architecture for filter induced augmentation and micro-cracks detection within a photovoltaic manufacturing facility," *Energies*, vol. 15, no. 22, p. 8667, Nov. 2022.

[29] P. K. Sharpe and R. J. Solly, "Dealing with missing values in neural network-based diagnostic systems," *Neural Comput. Appl.*, vol. 3, no. 2, pp. 73–77, Jun. 1995.

[30] V. Tresp, S. Ahmad, and R. Neuneier, "Training neural networks with deficient data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 6, 1993, pp. 128–135.

[31] R. Little and D. Rubin, "*Statistical Analysis With Missing Data*. Hoboken, NJ, USA: Wiley, 1987.

[32] M. Lee, J. An, and Y. Lee, "Missing-value imputation of continuous missing based on deep imputation network using correlations among multiple IoT data streams in a smart space," *IEICE Trans. Inf. Syst.*, vol. E102.D, no. 2, pp. 289–298, Feb. 2019.

[33] W.-C. Lee, B. A. Hanson, and R. L. Brennan, "Estimating consistency and accuracy indices for multiple classifications," *Appl. Psychol. Meas.*, vol. 26, no. 4, pp. 412–432, Dec. 2002.

[34] M. C. M. de Goeij, M. van Diepen, K. J. Jager, G. Tripepi, C. Zoccali, and F. W. Dekker, "Multiple imputation: Dealing with missing data," *Nephrology Dialysis Transplantation*, vol. 28, no. 10, pp. 2415–2420, 2013.

[35] D. J. Stekhoven and M. D. J. Stekhoven, "Package 'missForest,'" *R Package Version*, vol. 1, p. 21, Dec. 2013.

[36] A. Marshall, D. G. Altman, and R. L. Holder, "Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: A resampling study," *BMC Med. Res. Methodol.*, vol. 10, no. 1, Dec. 2010.

[37] E. Jun, A. W. Mulyadi, and H.-I. Suk, "Stochastic imputation and uncertainty-aware attention to EHR for mortality prediction," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–7.

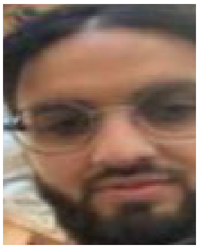
[38] C. N. Cavasotto and V. Scardino, "Machine learning toxicity prediction: Latest advances by toxicity end point," *ACS Omega*, vol. 7, no. 51, pp. 47536–47546, Dec. 2022.

[39] B. Agbo, "Downstream and privacy preserving missing data recovery for IoT systems," Ph.D. thesis, Dept. Comput. Sci. Inform., Univ. Huddersfield, Huddersfield, England, 2023.

BENJAMIN AGBO received the M.Sc. degree in information systems management and the Ph.D. degree in computer science and informatics from the University of Huddersfield, U.K., in 2017 and 2023, respectively, where he explored relevant techniques for data pre-processing. He is currently a Postdoctoral Research Associate with the School of Computer Science and Mathematics, Keele University.

HUSSAIN AL-AQRABI is an Assistant Professor with the Department of Computer Information Science at the Higher Colleges of Technology (HCT), UAE, since 2022. Prior to joining the HCT, he worked at the University of Huddersfield beginning in 2017. He also received a Postgraduate Certificate in Higher Education from the University of Huddersfield, U.K. He is a Fellow of the Higher Education Academy. In addition to his university education, he holds industry certifications, including EC-Council Certified Ethical Hacker, Microsoft Certified Educator, and Microsoft Certified IT Professional on Windows Server and he is also Cisco Certified in Routing and Switching. He has published over 50 publications in peer- reviewed journals, international conferences, and book series. He is a reviewer for many scientific journals, international conferences, and workshops. His research interests are cloud security, multiparty authentication, digital manufacturing, the Industrial Internet of Things, artificial intelligence, distributed ledger, network security, optimisation, secure protocol development, and evaluation.

TARIQ ALSBOUI received the B.Sc. degree in internet computing from Manchester Metropolitan University, U.K., in 2010, and the Ph.D. degree in computer science from the University of Huddersfield, U.K., in 2021. He is currently a Lecturer in computing with the School of Computing and Engineering, University of Huddersfield. He has authored several peer-reviewed international journals and conference papers. He is a fellow of the Higher Education Academy (FHEA). He is a Reviewer of high-impact-factor journals, such as IEEE Access and IEEE Internet of Things Journal.



MUHAMMAD HUSSAIN received the B.Eng. degree in electrical and electronic engineering and the M.S. degree in the Internet of Things from the University of Huddersfield, in 2019, and the Ph.D. degree in artificial intelligence for defect identification. His research is focused on the detection of various faults in particular micro-cracks forming on the surface of photovoltaic (PV) cells because of mechanical and thermal stress. His research interest includes machine vision, with a focus on the development of lightweight architectures that can be optimized for deployment on edge devices and ultimately on the production floor. He is also researching design-level architectural interpretability, with a focus on explainable AI for sensitive fields, such as medicine and healthcare.

RICHARD HILL is currently the Head of the Department of Computer Science and the Director of the Centre for Industrial Analytics, University of Huddersfield, U.K. He has published over 200 peer-reviewed articles. He was a recipient of several best paper awards, having been recognized by the IEEE for outstanding research leadership in the areas of big data, predictive analytics, the Internet of Things, cyber-physical systems security, and Industry 4.0, and has specific interests in digital manufacturing.