Missing Value Estimation using Clustering and Deep Learning within Multiple Imputation Framework

Manar D. Samad and Sakib Abrar
Department of Computer Science
Tennessee State University
Nashville, TN, USA
msamad@tnstate.edu

Norou Diawara
Department of Mathematics and Statistics
Old Dominion University
Norfolk, VA, USA

ABSTRACT

Missing values in tabular data restrict the use and performance of machine learning, requiring the imputation of missing values. The most popular imputation algorithm is arguably multiple imputations using chains of equations (MICE), which estimates missing values from linear conditioning on observed values. This paper proposes methods to improve both the imputation accuracy of MICE and the classification accuracy of imputed data by replacing MICE's linear conditioning with ensemble learning and deep neural networks (DNN). The imputation accuracy is further improved by characterizing individual samples with cluster labels (CISCL) obtained from the training data. Our extensive analyses involving six tabular data sets, up to 80% missingness, and three missingness types (missing completely at random, missing at random, missing not at random) reveal that ensemble or deep learning within MICE is superior to the baseline MICE (b-MICE), both of which are consistently outperformed by CISCL. Results show that CISCL + b-MICE outperforms b-MICE for all percentages and types of missingness. Our proposed DNN based MICE and gradient boosting MICE plus CISCL (GB-MICE-CISCL) outperform seven other baseline imputation algorithms in most experimental cases. The classification accuracy on the data imputed by GB-MICE is improved by proposed GB-MICE-CISCL imputed data across all missingness percentages. Results also reveal a shortcoming of the MICE framework at high missingness (>50%) and when the missing type is not random. This paper provides a generalized approach to identifying the best imputation model for a data set with a missingness percentage and type.

Keywords Missing value imputation, ensemble learning, deep learning, multiple imputations, MICE, clustering

1 INTRODUCTION

The inevitable presence of missing values is one major challenge in building machine learning-based predictive models using real-world data. Missing values in data are often imputed by an arbitrary imputation strategy to enable machine learning-based predictive modeling. However, imputation algorithms generate new synthetic values, which may alter the data distribution and mislead or inflate data-driven predictive models [1]. Therefore, the quality, statistics, and proportion of imputed values play a critical role in subsequent training and outcomes of predictive models. Many missing value imputation (MVI) studies limit their experiments on iteratively imputing a training data set without evaluating such models on imputing separate test data or classification tasks [2]. The type and pattern of data missingness are rarely analyzed to determine a suitable imputation model. On the other hand, the machine learning literature evaluates classification algorithms using benchmark tabular data sets that are clean, curated, complete, and free from the challenge of data missingness. Hence, there is a gap between the study of MVI and the machine learning literature innovating new data-driven predictive models. The problem of MVI primarily draws interests in statistics and domain research (bioinformatics, health science, sensors), aiming to bring machine learning-based solutions to respective do-

	Missing type	Max % of missing	Data domain
Nikfalazar et al. [4]	MCAR	10%	Multiple
Awan et al. [5]	MCAR	20%	Multiple
Camino et al. [6]	MCAR	80%	Multiple
Wang et al. [7]	Three types	30%	Multiple
Chen et al. [8]	MNAR, MAR	40%	Single
Ward et al. [9]	MNAR, MAR	27%	Single
Jones et al. [10]	Three types	50%	Single

Table 1: Recent articles on missing value imputation. Three types: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

main data with missing values. A substantial portion of the MVI literature proposes imputation models for a single domain data set considering a limited percentage of missingness.

In recent years, machine learning-based solutions to MVI problems have emerged, promising to outperform traditional statistical and matrix completion approaches. Existing data imputation strategies can be broadly categorized into single imputation (mean, median), non-MICE imputation (matrix factorization, k-nearest neighborhood), multiple imputations (using chained equations), imputation using ensemble learning (random forest), and deep learning (generative models, autoencoders). However, little or no study extensively compares all these model categories with different domain data, missingness types, and percentages, as we demonstrate in this paper. Instead, there is a consensus in the literature that multiple imputations using chained equations (MICE) are the most accurate models for data imputation because of its conditional modeling of missing variables on other variables [3]. Therefore, we take the conventional MICE model as the baseline to compare the performance of advanced machine learning models embedded in the MICE framework, including ensemble learning, deep learning, and our proposed model. We hypothesize that a hybrid framework that trains ensemble or deep learning models within the multiple imputation framework will improve the MVI accuracy. Additionally, we propose a novel approach to improve these hybrid model performance by data clustering and infusing the cluster information into individual samples for minimizing the variability in MVI estimates.

The remainder of the article is organized as follows. Section II provides an overview of MVI approaches and a literature review of the existing MVI algorithms. Section III discusses the proposed algorithm and experimental steps. Section IV provides the findings following the proposed experimentation. Section V discusses the findings with explanations and limitations of our work. The article concludes in Section VI with some future work.

2 Background review

This section reviews the theoretical underpinning of the MICE framework and the state-of-the-art literature in MVI studies.

2.1 Missing values in tabular data

Let's denote a data matrix, $X = \{X_{obs}, X_{miss}\}$ where $X_{obs} \subset X$ includes the observed or existing values, and $X_{mis} \subset X$ represents the missing values. A response indicator matrix ($R = [r_{ij}]$) identifies the location of missing values in X, taking $r_{ij} = 0$ for missing entries and $r_{ij} = 1$ for observed. A conditional probability distribution of $R = 0$ (missingness) given X can be defined as $Pr(R = 0|X_{obs}, X_{mis})$, which is used to model three types of missingness: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). First, MCAR occurs when the probability of missingness is completely independent of the observed (X_{obs}) or missing (X_{miss}) data as:

$$Pr(R = 0) = Pr(R = 0|X_{obs}, X_{mis}). \tag{1}$$

Second, when the probability of missingness depends on the observed data, it represents MAR as:

$$Pr(R = 0|X_{obs}) = Pr(R = 0|X_{obs}, X_{mis}). \tag{2}$$

Third, MNAR is identified when the missing probability depends on the unobserved or missing values themselves as follows.

$$Pr(R = 0|X_{mis}) = Pr(R = 0|X_{obs}, X_{mis}). \tag{3}$$

The literature has very limited studies that comprehensively analyze MVI algorithms for all missingness types, higher percentages of missingness, and data from multiple domains (Table 1). Most studies have considered only the MCAR

Algorithm 1 Multiple imputation with chained equations

Input: Data matrix $X = \{X_{obs}, X_{miss}\}$
Output: Imputed data matrix, $X = \{X_{obs}, X_{imputed}\}$
for i = 1 $\rightarrow$ 5 (multiple imputations) **do**
    $\{\bar{X}, \sigma^2\} \leftarrow X_{obs}$
    Initial imputation, $X_{miss}^0 \leftarrow \mathcal{N}(\bar{X}, \sigma^2)$
    for t = 1 $\rightarrow$ N **do**
        Estimate W^t, Equation 6
        Estimate X_{miss}^t, Equation 7
        $X_{miss}^t \sim P(X_{miss}^t | X_{obs}, X_{miss}^{-t})$
    end for
    $X_{imputed}^i \leftarrow X_{miss}^N$
end for
return $X_{imputed} \leftarrow$ Aggregate $(X_{imputed}^i)$

---

type because of the convenience of rendering completely random missing positions [4, 11, 12]. In a survey of the MVI literature, only 12 out of 111 articles (10.8%) have studied imputations of data with more than 50% missing values [13]. For example, Nikfalazar et al. have tested MCAR type missingness for up to 10% of missing data [4]. In practice, MAR and MNAR types are more common than the MCAR type. Table 1 highlights missingness types, percentages, and data domains of recent MVI studies. Intuitively, an MVI algorithm will be lesser accurate on data with a higher percentage of missing values and vice versa. Therefore, the robustness of MVI algorithms can be evaluated over a range of missing value percentages. For example, Jones et al. have imputed a large electronic health records (EHR) data set for up to 50% missingness [10]. Our previous study on a large EHR data set (sample size > 300,000) reveals that the most important predictor of patient mortality (out of 149 variables) can be missing for more than 50% of the samples [14]. Similar predictor variables with over 50% of missing values are often excluded before imputation and predictive modeling, limiting our ability to yield important insights from data. Therefore, the trade-off between data imputation and exclusion at a high missing percentage is critical for optimizing machine learning model performance.

2.2 Statistical imputation of missing values.

The field of MVI is predominantly researched by statisticians [3]. There are varying statistical patterns in missing values (e.g., monotone, k-monotone, arbitrary) [15] that may benefit from targeted imputation algorithms. One of the most successful MVI algorithms is built on the statistical concept of multiple imputations (MI) [3, 16], such as multiple imputations using chained equations (MICE) [17, 18]. As opposed to single value imputations (mean, median, or other point estimates such as k-nearest neighbors), multiple imputations generate multiple samples of imputed data sets. The variability across multiple imputed data sets captures the statistical uncertainty in data imputations, improving the imputation accuracy. The MICE algorithm estimates missing values of individual variables using observed values of other variables, which intuitively models the MAR type. Let a multivariate data matrix $X = [x_{ij}] = \{ x_1, x_2, ..., x_n \} \in \Re^{n \times d}$ with missing values has n samples and d variables. In MICE, the j-th variable $x_j \in \Re^n$ can be regressed from observed and imputed values of other $d-1$ variables x_k as below.

$$x_j = f(x_k, W), \; k \neq j, \tag{4}$$

where j-th (x_j) variable is regressed using a vector of parameters W as below.

$$W = [w_1, w_2, ..., w_{d-1}]^T. \tag{5}$$

The regression parameter vector $W \in \Re^{d-1}$ is drawn by a Gibbs sampler as shown in Equation 6. Given an estimation of $\hat{W}_j$ for variable x_j, missing values of this variable can be sampled from the conditional distribution in Equation 7.

$$\hat{W}_j^t \sim Pr(W_j^t | x_j^{obs}, x_{(k<j)}^t, x_{(k>j)}^{t-1}), \tag{6}$$

$$x_j^t \sim Pr(x_j^{miss} | x_j^{obs}, x_{(k<j)}^t, x_{(k>j)}^{t-1}, \hat{W}_j^t). \tag{7}$$

Here, t is the iteration index. Both $\hat{W}_j^t$ and x_j^t are iteratively updated using a chain of d equations (for d variables) until convergence following Algorithm 1. These sampling distributions generate multiple imputed data sets for multiple imputations. The conditional sampling distributions (Equations 6 and 7) of the MICE framework can be explained

by the Bayesian approach. In the Bayesian approach, the posterior distribution of imputation model parameter (W) given observed data (X_{obs}) and estimated missing values (y_{miss}) is likelihood $P(y_{miss}|X_{obs}, W)$ times the prior distribution, P(W).

$$P(W|y_{miss}, X_{obs}) = P(y_{miss}|X_{obs}, W)P(W) \tag{8}$$

We set the prior distribution of W as Gaussian.

$$P(W) = N(\mu_0, \sigma_0^2) \tag{9}$$

Considering zero mean μ_0=0 Gaussian distribution and ignoring the normalizing constant, the prior becomes:

$$P(W) = \frac{1}{\sqrt{2\sigma_0^2\pi}} \exp\left(-\frac{(W - \mu_0)^2}{2\sigma_0^2}\right)$$

$$\propto \exp\left(-\frac{W^2}{2\sigma_0^2}\right)$$

Therefore, the posterior (Equation 8) becomes as follows.

$$P(W|y_{miss}, X_{obs})$$
$$\propto \exp\left(-\frac{(y_{miss} - W^T X_{obs})^2}{2\sigma^2}\right) \exp\left(-\frac{W^2}{2\sigma_0^2}\right) \tag{10}$$

Taking log posterior we obtain:

$$\log P(W|y_{miss}, X_{obs})$$
$$\propto -\frac{1}{2\sigma^2}(y_{miss} - W^T X_{obs})^2 - \frac{1}{2\sigma_0^2}W^2$$
$$= -\frac{1}{2\sigma^2}\|y_{miss} - W^T X_{obs}\|_2^2 - \frac{1}{2\sigma_0^2}\|W\|_2^2 \tag{11}$$

We assume that σ^2=1 and $\lambda = \frac{1}{\sigma_0^2}$, then our log posterior becomes:

$$\log P(W|y_{miss}, X_{obs}) \propto -\frac{1}{2}\|y_{miss} - W^T X_{obs}\|_2^2 - \frac{\lambda}{2}\|W\|_2^2 \tag{12}$$

That is, the log posterior of the imputation model is equivalent to the objective function for ridge regression. However, this result yields a point estimate instead of a distribution. For the full Bayesian approach during testing, we marginalize the posterior predictive distribution. Given trained parameters (W) and a new data point with missing values (x'), the posterior distribution of imputed values given (x') and trained imputation model (W) is $P(y'|x', W)$. This distribution is marginalized by the posterior of weight distribution $P(W|y_{miss}, X_{obs})$ as below.

$$P(y'|y_{miss}, X_{obs}) = \int_W P(y'|x', W)P(W|y_{miss}, X_{obs})$$
$$= \mathbb{E}_W\left[P(y'|x', W)\right] \tag{13}$$

In practice, each component of W is a real number, which makes the problem of enumerating all possible values of W intractable. Therefore, approximation methods such as Monte Carlo Markov chains (MCMC) are used, which are computationally more expensive than point estimates but more accurate. The imputed missing values using linear regression models in multiple imputation framework (Algorithm 1) yield better classification accuracy than those imputed by single imputations such as K-nearest neighbors (KNN) (Fig. 1). The MICE algorithm is touted as one of the most successful imputation methods in state-of-the-art data-driven studies [19, 14, 3]. Therefore, the MICE model has been evaluated and compared in most recent MVI studies, including [20, 21, 5].

2.3 Machine learning in data imputation

The recent literature extends statistical MVI models with more advanced machine and deep learning (MnD) algorithms. Deep learning is proposed in the imputation of drug discovery data [22] and non-numerical tabular data [23]. Recently, deep generative models [24, 6], convolutional neural networks [25], recurrent neural networks [26] have been proposed to impute data with spatial and temporal dimensions. In another study, autoencoders are trained (encoded) using complete data set to estimate (decode) missing values from their high-dimensional latent spaces [27].

Data set	Sample size	# of variables	# of target classes	Type	Range
Breast cancer	569	30	2	Diagnostic	0 to 4254
Dermatology	358	35	6	Histopathological	0 to 9
SkillCraft1 master table	3395	19	NA	Game	0 to 1000000
Wine quality	4898	11	7	Food	0 to 440
Default of credit card clients	30000	21[a]	2	Financial	-339603 to 1684259
Mice protein expression	552	77	8	Proteomic	0 to 8.5

Table 2: A summary of the six data sets used to evaluate the performance of missing value imputation algorithms.

[a]Three categorical variables (sex, education, marriage) are excluded.
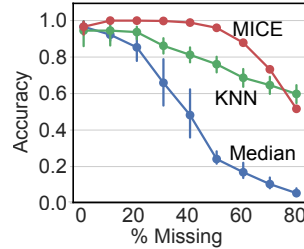

Figure 1: MICE imputed data have shown to yield superior and robust classification accuracy compared to single imputation methods, e.g., median and K-nearest neighbor.

Apart from these variants of deep models, fully connected deep neural networks with three or more hidden layers have not been studied for MVI. On the other hand, ensemble learning, such as extreme gradient boosting trees have been trained as regressors to estimate missing values for data imputation [28]. However, these MnD-based imputation methods do not provide the benefits of multiple imputations. In line with this notion, the multiple imputation performance of the MICE algorithm is improved by replacing the standard GLM with regression trees in categorical data imputation, [29]. A random forest classifier is used in the MICE framework to impute categorical and binary variables of a traffic crash report data set [30]. Decision tree-based regressors have been recently proposed within the MICE framework to impute missingness in brain function and anatomy data [31]. Hallam et al. have used gradient boosting (GB) tree regressors within MICE to impute geological data set with continuous variables [32]. The existing work reports findings on particular data types (binary or categorical), target domains, and within a limited range of missing values and types. These selective approaches may not generalize well across many other variants of data and missingness. A recent survey of the MVI literature has highlighted the importance of a) choice of data sets, b) missing rates, c) missing types, and d) evaluation metrics for missing value imputation methods [13]. We hypothesize that a hybrid framework of multiple imputations and ensemble learning can improve the performance and robustness of data imputation across various data sets, missing types, and percentages compared to the existing imputation models.

2.4 Data imputation algorithms

This section provides an overview of the state-of-the-art MVI algorithms.

2.4.1 Machine learning in MICE

Multiple imputations by chained equations (MICE) consist of variable-specific linear regression models, which are first seeded with initial imputation values. The chains of these equations are then iteratively updated to refine estimates of missing values (Algorithm 1). Machine learning-based non-linear regressors can replace these linear regression equations to improve the accuracy of missing value estimations [33]. The MI part of MICE models is known to capture the variability between models, whereas ensemble learning models the variability within a model [2]. An example of the MICE algorithm using an ensemble of decision trees for data imputation is MICEforest [34]. The iterative update in MICEforest is performed using predictive mean matching (PMM). In PMM, the model estimated missing value is compared with other observed values in the data set to determine the neighboring samples. The observed value of the nearest neighbor is then used to impute the missing value. Therefore, PMM performs single value imputation using the observed value of the nearest neighbor instead of performing multiple imputations. Our experiment with PMM shows no improvement in imputation accuracy compared to the MI-based imputation strategy.

2.5 Imputation for predictive modeling

One major objective of MVI is to retain maximum samples for predictive modeling (classification or regression) without excluding data due to missingness. The existing MVI approaches have two shortcomings in predictive modeling. First, Valdiviezo et al. [2] have argued that the MVI literature almost always evaluates MVI algorithms by comparing actual and imputed values. Unfortunately, many real-world data sets with missing values do not contain the ground truth actual values. Furthermore, it remains unclear how data quantity and quality, augmented by imputation algorithms, improve the performance of classification or regression models. Second, there is no established benchmark datasets for evaluating imputation algorithms. The imputation algorithms are iteratively executed and then evaluated on the same data set with missing values in a self-supervised manner. Ideally, a trained imputation model should be used to impute a standalone test data set with missing values to report the imputation accuracy. Third, a MICE model builds chains of regression equations using all heterogeneous data points at a time. We argue that the presence of heterogeneous and outlier samples may result in higher variance or errors in the final estimates of missing values. Therefore, data imputation modeling within individual clusters of homogeneous samples may reduce the variability and errors in the estimates of missing values.

2.6 Contributions

Based on our literature review, the contributions of this article is as follows. First, we provide algorithms to simulate three missing value types (MAR, MCAR, and MNAR) in tabular datasets. Second, we perform extensive analyses to demonstrate the effect of missing value percentages up to 80% across three missing value types on imputation algorithms. We hypothesize that no single imputation algorithm can be the best choice for imputing all types of missingness or data of any domain. Third, the imputation performance is reported on separate test data with missing values, and the quality of imputed data is subsequently evaluated in a supervised classification. We investigate if better imputation accuracy also translates to better classification performance. Fourth, we propose novel imputation algorithms by 1) deep learning in the MI framework and 2) ensemble learning within MI after infusing cluster information for homogeneous imputation. The information about data clusters is expected to improve errors in missing value estimation. Fifth, the performance of our proposed imputation algorithm is compared against non-MICE imputation algorithms, baseline MICE, and MICE with ensemble learning across MCAR, MAR, and MNAR types from 5% to 80% missing data.

3 Methodology

3.1 Data sets

We conduct all experiments on six tabular data sets obtained from the UCI machine learning data repository [35] as shown in Table 2. The datasets cover a good range of sample sizes (300-30,000), data dimensionality (11-77), and domains (diagnosis to game) to understand their effects on missing value estimation. The MVI models are trained using 70% of the data set, and the remaining 30% is used to evaluate the performance of trained MVI models. We standardize the data using the mean and standard deviation of individual variables in the training data set. The MVI model estimation of missing values is then transformed back to the original scale for comparing against the actual values.

3.2 Simulation of missingness types:

The three types of missingness (Section II-A) are simulated following the computational framework proposed in Algorithm 2. For MCAR, we randomly generate indices to remove values from the data matrix based on the missing rate. For MNAR, the random selection is made below the lower quantile and above the upper quantile of individual features. Here, the missingness is introduced to high or low values of the same features subjected to missingness. For MAR, the quantiles are determined on a separate set of observed feature values to remove corresponding values in other features. The total number of removed values is set by the number of samples times the number of features times the missing rate.

3.3 Multiple imputation using chained regressors

We exclude the target classification column from the data set before simulating the missingness and evaluating imputation algorithms. The procedure of multiple imputations (MI) is initialized by random sampling of multiple imputed datasets, which are slightly different from each other to capture uncertainties in data imputation (Section II-B). These imputed data sets are independently and iteratively updated using a chain of regressors, as shown in Algorithm 1.

Algorithm 2 Simulation of three types of missingness.

Input: $X_{complete}$, missing_rate, missing_type
Output: $X\_missing$
if missing_type == MCAR **then**
    [p, q] ← random_indices (missing_rate)
    $X_{missing}$ ← remove ($X_{complete}$, [p, q])
    return $X_{missing}$
end if
low = missing_rate / 2
high = 100 - missing_rate / 2
for i = 1 → n_features **do**
    feature ← $X_{complete}(:, i)$
    if missing_type == MNAR **then**
        low_val (i) ← Quantile (feature, low)
        high_val (i) ← Quantile (feature, high)
        missing←select (feature, <low_val ∪ >high_val)
        $X_{missing}$ ← remove (feature, missing)
    else if missing_type == MAR **then**
        obs_features$_{j \neq i}$ ⊂ {1, 2, ..., n_features}
        for j in obs_features **do**
            low_val (j) ← Quantile ($X_{complete}(:, j)$, low)
            high_val (j) ← Quantile ($X_{complete}(:, j)$, high)
            [p, q] ← indices (<low_val(j) ∪ >high_val(j))
        end for
        $X_{missing}$ ← remove(feature, [p, q])
    end if
end for
return $X_{missing}$

Buuren recommends 10 to 20 iterations to update the chains of regressors toward the final imputed values [3]. The final imputed data set is aggregated (averaged) from multiple imputed data sets. We implement and evaluate MI frameworks developed using the baseline linear regression (LR-MICE), ensemble learning regressors (random forest, gradient boosting), and chains of deep regressors (deep neural networks). It is worth noting that our LR-MICE is the popular MICE algorithm used across many domains, including in our prior work [14].

3.4 Regression using ensemble learning

Two of the most successful ensemble learning algorithms, random forest and gradient boosting trees, are used to build the regression chains within the MI framework. Many decision trees are trained using bootstrap samples in a random forest model. Estimates from from individual decision tree regressors are averaged to yield the final estimates. The split on feature values at each node of the decision tree is chosen optimally based on criteria derived from information theory. The gradient boosting trees algorithm uses a sequence of weak decision trees to learn the input-output mapping (regression or classification) and pass the error gradient to the following trees to boost the overall mapping accuracy. If e_i is the error of the first decision tree to map input X to output y (Equation 14), the next tree in the sequence will learn e_i from X and yield an error e_{i+1} (Equation 15). Recursively, the output y is learned from Equation 16.

$$y = a_i + (b_i * X) + e_i \tag{14}$$
$$e_i = a_{i+1} + (b_{i+1} * X) + e_{i+1} \tag{15}$$
$$y = \sum_{i=1}^{n} a_i + X * \sum_{i=1}^{n} b_i + e_n \tag{16}$$

3.5 Imputation aided by clustering

We hypothesize in Section II-E that the knowledge of sample clusters can aid more accurate estimates of missing values being treated by homogeneity of data samples. The data set with missing values is first completed with medians of individual variables to enable clustering. A clustering algorithm identifies each sample ($x_i \in \Re^d$) belonging to one

---

Algorithm 3 Ensemble learning and clustering within MI

Input: Data matrix $X = \{X_{obs}, X_{miss}\} \in \Re^{n \times d} = \{x_i\}$
Output: Imputed data matrix, $X' = \{X_{obs}, X_{imputed}\}$
X_{median} = median (X_{obs})
$X_{temp} \leftarrow \{X_{obs}, \text{impute } X_{miss} \text{ with } X_{median}\}$
Obtain k clusters $\{C_t\}$ in X_{temp}, t = 1, 2, ...,k
C $\leftarrow$ cluster labels for samples
if method == 'label' **then**
   X_G = concatenate (X, C)
else if method == 'one-hot' **then**
   $e_1, e_2, .., e_n$ = one-hot (C)
   X_G = concatenate (X, $e1, e2, .., e_n$)
else if method == 'MCMV' **then**
   for p = 1 $\rightarrow$ k **do**
     $X_p = X_{temp}$ $(C_t == \text{p})$
     cluster mean vector, $C_{mean} = \frac{1}{N_p} \sum_{i=1}^{N_p} X_p(i)$
     $G_p = C_{mean}^T . C_{mean}$
   end for
   X_G = concatenate (X, G_p)
end if
Under multiple imputations loop of Algorithm 1
for t = 1 $\rightarrow$ N **do**
   $X_{imputed}^t$ = Ensemble_MICE (X_G)
   $X^t \leftarrow \{X_{obs}, X_{imputed}^t\}$
   Update $X_G \leftarrow X^k$
   Continue until $|X_{imputed}^{t+1} - X_{imputed}^t| \leq \tau$
end for
return $X' = \{X_{obs}, X_{imputed}^N\}$

---

of k disjoint groups of samples as $c_i \in \{c_1, c_2, ..., c_k\}$ with a cluster centroid, $\bar{x}_i \in \{\bar{x}_1, \bar{x}_2, ..., \bar{x}_k\}$. A distance-based clustering algorithm updates the cluster centroids to minimize the sum of squared distance (SSE) between the samples within a cluster and the cluster centroid as below.

$$SSE = \sum_{i=1}^{n} \sum_{p=1}^{k} w_{ip} ||x_i - \bar{x}_p||_2^2, \tag{17}$$

where $\bar{x}_p$ is the centroid of the p-th cluster. The association between the i-th sample with the p-th cluster is

$$w_{ip} = \begin{cases} 1, \text{if } x_i \text{ belongs to the p-th cluster} \\ 0, \text{otherwise.} \end{cases} \tag{18}$$

Thus, the data set is clustered into k groups labeling each sample under one cluster. The cluster label is infused into the training data samples in three ways, as shown in Algorithm 3. First, the cluster labels are appended as a new feature column to the samples as $X = \{x_i, c_i\}$. Second, one-hot encoded labels are appended as a vector of features, $X = \{x_i, e_{1i}, e_{2i}, ..., e_{ni}\}$, where n = ceil (log_2^k). Third, the magnitude of cluster mean vector (MCMV) is proposed as a feature column, $X = \{x_i, \bar{x}_i^T \bar{x}_i\}$. Thus, each variable (x_j) is also regressed from sample cluster information to introduce more homogeneity in MVI. The revised regression model for variable x_j, following Equation 4, becomes as below.

$$x_j = f(x_k, c_i, W,), \; k \neq j. \tag{19}$$

In other words, the regression model for variable x_j is now a function of other variables x_k and the cluster variable $c_i \in \{c_1, c_2, ..., c_k\}$ corresponding to the data sample i, weighted by the model parameter vector W.

3.6 Experimental design

The imputation accuracies of the baseline and proposed algorithms are obtained and compared through extensive experimentation. First, the robustness of the algorithms is evaluated for data sets with 5% to 80% missing values. It
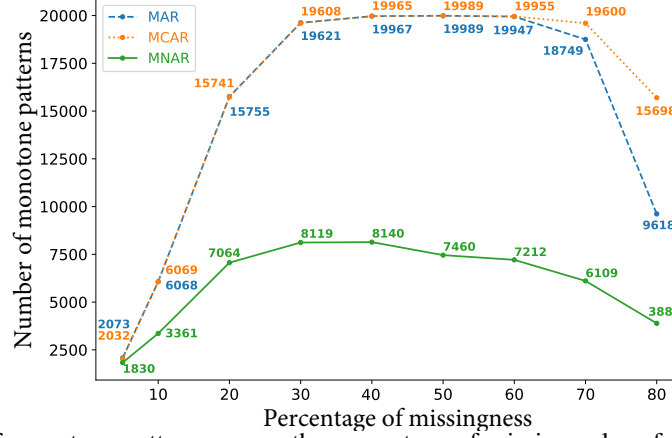
Figure 2: The number of monotone patterns versus the percentage of missing values for MAR, MCAR and MNAR type of missingness for default of credit card clients data set.
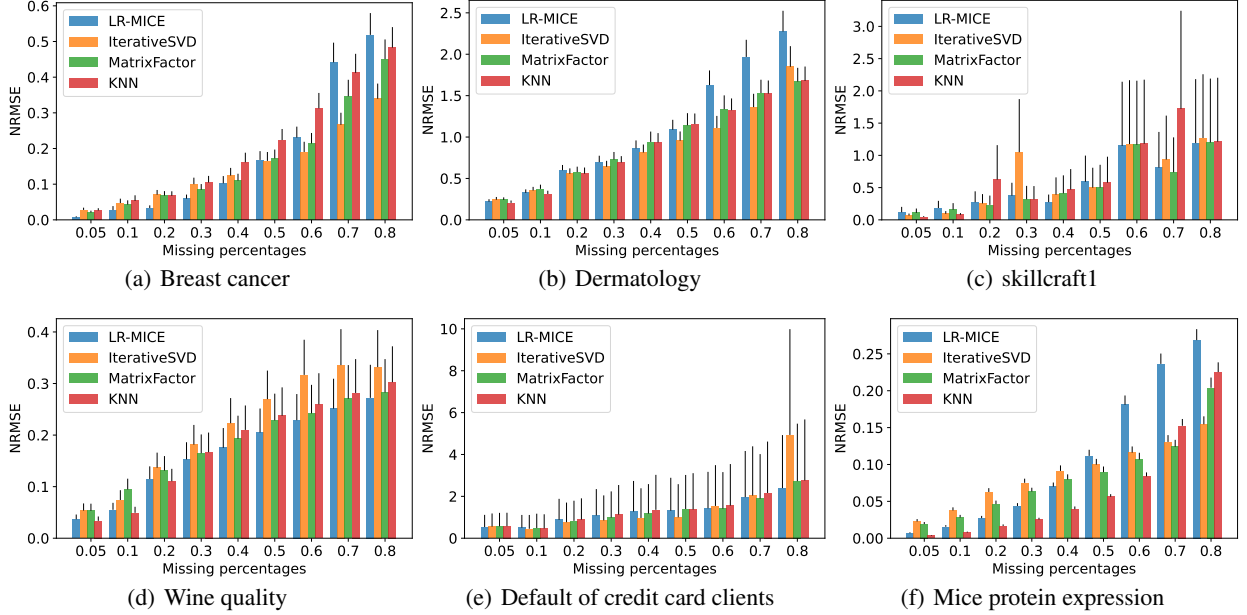

Figure 3: Normalized root mean squared error (NRMSE) comparison between MICE (linear regression-MICE or LR-MICE) and non-MICE algorithms (iterative SVD, matrix factorization, K-nearest neighbor (KNN)). The test data are missing at random (MAR).

should be noted that only 10.8% of the literature have performed imputation for more than 50% missing values [13]. Second, we compare five categories of imputation algorithms: 1) non-MICE imputations (KNN, iterative SVD, matrix factorization) using the fancyImpute package [36], 2) the baseline MICE algorithm (MI with chains of linear regression), 3) MI with chains of ensemble learning (gradient boosting and random forest), 4) MI with chains of deep regressors, and 5) MI with ensemble learning plus the cluster information. The first two categories will serve as the baseline models. Third, the experimentation will be conducted for imputing three missingness types (MAR, MCAR, and MNAR) across data sets of different domains.

3.7 Model evaluation

The imputation of missing values is performed on standardized data set with zero mean and unit standard deviation. Following the imputation, the data set is transformed back to its original scale to compare the imputed values against their actual counterparts. The performance of the imputation models is evaluated in two ways. First, the trained

Datasets	Best Model		
	MAR	MCAR	MNAR
Breast cancer	LR-MICE + Label	LR-MICE + One-hot	LR-MICE + One-hot
Dermatology	RF-MICE + One-hot	RF-MICE + One-hot	RF-MICE + One-hot
SkillCraft1 master	DR-MICE	DR-MICE	DR-MICE
Wine quality	GB-MICE + Label	GB-MICE + MCMV	LR-MICE + One-hot
Default of credit card	RF-MICE	GB-MICE + One-hot	GB-MICE + One-hot
Mice protein expression	LR-MICE + Label	RF-MICE + MCMV	RF-MICE + MCMV

Table 3: Best MICE models for each data set after voting across all missingness percentages.

	MAR				MCAR				MNAR			
Missing	LR-MICE	LR-MICE + Label	LR-MICE + One-hot	LR-MICE + MCMV	LR-MICE	LR-MICE + Label	LR-MICE + One-hot	LR-MICE + MCMV	LR-MICE	LR-MICE + Label	LR-MICE + One-hot	LR-MICE + MCMV
5%	0.0377	**0.0367**	0.0369	0.0368	**0.0397**	0.0398	0.0402	0.0399	0.146	0.1452	**0.1445**	0.1449
10%	0.0551	0.0555	**0.0548**	0.055	0.0595	0.0587	**0.0577**	0.0589	0.1754	0.1746	**0.174**	0.1746
20%	0.1147	**0.1136**	0.1136	0.1147	0.1125	0.1119	**0.111**	0.1119	0.2164	0.2153	**0.215**	0.2157
30%	0.1529	0.1515	**0.1515**	0.1519	0.1486	0.1529	0.1479	**0.1468**	0.2417	0.2411	**0.2409**	0.2412
40%	0.1754	0.1756	**0.1748**	0.1748	0.1823	0.1814	**0.1809**	0.182	0.2655	0.2652	**0.2652**	0.2656
50%	0.2051	0.2043	**0.2032**	0.2043	0.2032	0.2031	**0.2021**	0.2028	0.2772	0.2766	**0.2766**	0.2767
60%	0.2289	0.2292	**0.2264**	0.2274	0.2217	0.2244	0.2227	**0.2214**	0.2883	0.2874	0.2874	**0.2871**
70%	0.2513	0.2567	0.2569	**0.2488**	**0.2496**	0.256	0.2578	0.2504	0.2958	0.2957	**0.2956**	0.2957
80%	0.2722	0.2724	0.2724	**0.2716**	0.272	0.2802	0.2795	**0.2708**	0.2993	0.2992	**0.2992**	0.2993

Table 4: Comparison of NRMSE scores of proposed inclusion of cluster information (label, one-hot, MCMV) with the baseline MICE (LR-MICE) on the wine data set.

imputation model imputes a separate test data set with missing values. The error between imputed and actual values on the test data set is measured by the root mean squared error (RMSE) (Equation 20) divided by the mean of the actual values as the normalized root mean square error (NRMSE) (Equation 21).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \hat{X}_i)^2}{N}} \tag{20}$$

$$NRMSE = \frac{RMSE}{\frac{1}{N}\sum_{i=1}^{N}X_i} \tag{21}$$

Second, the quality of imputed data is evaluated in supervised classification tasks using the random forest classifier. A 10x2 fold nested cross-validation is performed first to search the best model hyperparameters (two inner folds) and then report the mean classification accuracy across the test folds (ten outer folds). The outcomes of the two evaluation approaches will reveal if better imputation accuracy translates to better classification accuracy.

4 Results

The proposed experiments are executed on a Dell Precision 5820 workstation running Ubuntu 20.04 with 64GB RAM and an Nvidia Quadro 5000 GPU with 16GB memory. All algorithms and evaluation steps are implemented in Python. The findings of the proposed simulation and experimentation are discussed below.

4.1 Simulating missing values

We simulate three missingness types following Algorithm 2. The simulation of missingness is evaluated by counting monotone patterns in missing values as a function of the percentage of missingness [15]. More monotone patterns indicate that the missingness in data has more more uniquely and randomly distributed patterns with minimum matching. A low number of monotone patterns indicates more matching or similar patterns in missingness. Figure 2 shows that MNAR has the least number of monotone patterns suggesting minimal randomness in the missing value pattern. On the other hand, MCAR has more monotone patterns than MAR, as observed in data with 60% or more missing values. The number of monotone patterns (or the number of unique patterns) in missing values peaks between 30% and 50% data missingness.

	MAR				MCAR	MNAR
Missing	GB-MICE (baseline)	GB-MICE + Label	GB-MICE + One-hot	GB-MICE + MCMV	GB-MICE + MCMV	GB-MICE + MCMV
5%	0.964 (0.03)	0.969 (0.03)	0.967 (0.03)	**0.969 (0.02)**	0.975 (0.03)	0.967(0.03)
10%	0.969 (0.02)	0.969 (0.02)	0.972 (0.02)	**0.975 (0.02)**	0.978 (0.02)	0.955 (0.02)
20%	0.972 (0.02)	**0.978 (0.02)**	0.972 (0.02)	0.969 (0.02)	0.975 (0.04)	0.927 (0.03)
30%	0.955 (0.02)	0.953 (0.02)	0.958 (0.02)	**0.964 (0.02)**	0.958 (0.02)	0.899 (0.03)
40%	0.930 (0.04)	0.922 (0.04)	0.936 (0.05)	**0.939 (0.04)**	0.922 (0.02)	0.804 (0.02)
50%	0.924 (0.07)	**0.927 (0.05)**	0.927 (0.05)	0.927 (0.06)	0.916 (0.02)	0.793 (0.04)
60%	0.846 (0.06)	0.857 (0.04)	0.852 (0.06)	**0.863 (0.07)**	0.874 (0.01)	0.752 (0.03)
70%	0.757 (0.06)	0.779 (0.06)	**0.782 (0.06)**	0.751 (0.03)	0.805 (0.06)	0.470 (0.08)
80%	0.662 (0.09)	0.676 (0.10)	0.657 (0.08)	**0.696 (0.08)**	0.670 (0.05)	0.327 (0.07)

Table 5: The accuracy of a random forest classifier in classifying the dermatology data set imputed by the gradient boosting (GB)-MICE algorithms. The mean classification accuracy is obtained across ten cross-validation folds. The accuracy on the actual dermatology data set (without missing values) is 98.04%.

Data sets	Best model			Lowest sum of NRMSE		
	MAR	MCAR	MNAR	MAR	MCAR	MNAR
Breast cancer	Iterative SVD	RF-MICE + MCMV	Iterative SVD	1.33	1.24	2.77
Dermatology	RF-MICE + One-hot	GB-MICE + One-hot	Iterative SVD	7.47	7.36	9.70
SkillCraft1	DR-MICE	DR-MICE	DR-MICE	3.32	3.38	9.57
Wine quality	GB-MICE + MCMV	GB-MICE + One-hot	LR-MICE + One-hot	1.47	1.47	2.20
Default of credit	RF-MICE + One-hot	GB-MICE + One-hot	LR-MICE	9.36	16.47	19.12
Mice protein	DR-MICE	KNN	DR-MICE	0.59	0.56	1.29

Table 6: The best missing value imputation (MVI) algorithm for each data set and missingness type pair. The values represent the lowest sum of NRMSEs (therefore the best model) across all missing percentages. The best model is chosen from six baseline MVI models (Iterative SVD, matrix factorization, KNN, LR-MICE, RF-MICE, GB-MICE) and three proposed models (RF-MICE-Cluster, GB-MICE-Cluster, and DR-MICE). Here, the 'Cluster' represents cluster labels (label), one-hot encoded labels (one-hot), and the magnitude of cluster mean vector (MCMV).

4.2 Linear MICE versus non-MICE models

We compare the popular MICE model (LR-MICE) with other non-MICE imputation methods, including iterative singular value decomposition (Iterative SVD), matrix factorization, and K-nearest neighbor (KNN). Figure 3 shows the effect of the percentage of missingness on NRMSE scores for the MAR type data. Intuitively, all data sets show a linear relationship between the NRMSE scores and the percentage of missing values. A general observation is that the performance of the MVI algorithms depends on the data set. Therefore, selecting one MVI model as the best for all datasets is likely to yield suboptimal results. The LR-MICE model generally performs well at a lower percentage of missing values ($< 30\%$). LR-MICE appears to be the best for the the wine quality data set at any percentages of missingness. The KNN algorithm outperforms LR-MICE up to 60% missingness in the mice protein expression data set. The iterative SVD algorithm yields the best NRMSE scores at higher missing percentages ($>50\%$), especially with the breast cancer and dermatology datasets. These observations contradict the general notion that the baseline MICE is the best model for imputing missing values in tabular data.

4.3 Machine learning in MICE

We replace the linear equations of the baseline MICE model with non-linear machine learning regression models, including decision tree (DT), random forest (RF), gradient boosting (GB) trees, and deep regression (DR) with three hidden layers. The number of neurons in the first hidden layer is set based on the number of input variables. Each successive layer has half the number of neurons compared to the previous layer. Understandably, the DT-MICE model performs the worst among all non-linear MICE models because of not being an ensemble learning with a single decision tree. The MVI methods with ensemble learning in the MI framework (GB-MICE and RF-MICE) appear most frequently in the best model chart compared to the baseline LR-MICE model (Table 3).

The MCAR type mostly benefits from ensemble learning within the MI framework (RF-MICE, GB-MICE). This benefit appears to be the least in imputing data with the MNAR type. Interestingly, the skillcraft1 data set yields the best imputation results with the DR-MICE imputation model regardless of the missing value types, as shown in Table 3. However, the DR-MICE model takes a much higher computing time compared to other models. For example, our workstation takes six hours on average to complete GB-MICE imputation for all nine missing percentages. In contrast, DR-MICE takes approximately 48 hours to impute the same data, which is eight times more. In almost all cases, the inclusion of cluster information (one-hot encoding or labels or MCMV) in the imputation process augments the performance of ensemble learning within the MI framework. The only exception is the default of credit data set with MAR values where RF-MICE outperforms those with the cluster information.

Furthermore, we investigate the effect of including the proposed cluster information in the baseline LR-MICE model alone. The number of clusters is chosen to be three based on the silhouette scores. The comparison of NRMSE scores in Table 4 reveals that data with the MNAR type are almost always best imputed by LR-MICE plus one-hot encoded cluster information. The inclusion of cluster information to LR-MICE consistently outperforms the baseline LR-MICE for any missingness types or percentage of missing values. The only exceptions are found at 5% and 70% of the MCAR type, where the baseline LR-MICE outperforms the models with cluster information. However, these percentage values are less common, and so is the MCAR type in real-world data sets.

4.4 Comparison of classification accuracy

We investigate if improved data imputation also yields improved classification accuracy. That is, the imputation model yielding the lowest NRMSE score is expected to contribute to the best classification accuracy. Table 4 shows that MNAR is the most challenging type of missingness, resulting in the highest NRMSE values, which is followed by MAR and MCAR types. However, the differences between the missingness types diminish in classification accuracy. Table 5 shows the quality of GB-MICE imputed data in a six-class classification task using the dermatology data set. As expected, the classification accuracy obtained using imputed data sets never exceeds the one with actual or complete data set (98.04%). However, the classification accuracy is substantially affected (drops below 90%) when the data set is imputed for more than 50% missing values (MCAR, MAR types) and more than 30% missing values (MNAR type). The inclusion of data cluster information yields superior classification accuracy for all missing percentages than the baseline GB-MICE without the cluster information. Therefore, our proposed inclusion of data cluster information improves the MVI accuracy, which contributes to superior classification accuracy using the imputed data set. A data set, imputed with the aid of the cluster information represented by MCMV, yields the most robust improvement in the classification accuracy, as observed in Table 5. It is noteworthy that the cluster information is only used during the imputation step and excluded at the time of classification.

4.5 Overall comparison of imputation models

We finally compare the performance of ten MVI models: 1)the baseline LR-MICE, 2) LR-MICE with the cluster information, 3-5) non-MICE imputation algorithms (iterative SVD, matrix factorization, KNN), 6-7) ensemble learning within the MI framework (RF-MICE, GB-MICE), 8) deep learning within the MI framework (DR-MICE), and 9-10) MI with ensemble learning plus the cluster information (MI-EL-CI). Here, MI-EL-CI includes the cluster information in GB-MICE and RF-MICE models. Table 6 sums the NRMSE scores across all missing percentages for each imputation model. DR-MICE yields the best performance when imputing the skillcraft1 data set regardless of missingness types.

Three out of six data sets with MAR values show superior performance with our proposed MI-EL-CI imputation algorithms. The proposed DR-MICE performs the best for two other data sets with MAR. Five out of six data sets with MCAR values find our proposed MI-EL-CI or DR-MICE imputation superior to all other algorithms. Conversely, the imputation of data MNAR type yields mixed findings. In general, ensemble learning within MI does not perform the best in imputing any of the six datasets with MNAR values. Iterative SVD, LR-MICE, and our proposed DR-MICE individually have ranked the best in imputing two datasets with MNAR. The NRMSE values in Table 6 once again confirm that imputing MNAR is the most challenging type with the highest errors among the three missingness types.

5 Discussion

This paper aims to improve the performance of the baseline MICE model by incorporating deep learning and cluster information. Extensive evaluations of the proposed models across six data sets, three missingness types, and up to 80% missingness have revealed several important findings as follows. First, ensemble learning, replacing the linear regression equations in the MICE framework, improves the MVI accuracy for MAR and MCAR types (Tables 3, 6)

but not for the MNAR type. Second, the inclusion of cluster labels almost always improves the imputation accuracy of ensemble-based MICE (Table 3) and baseline MICE (LR-MICE) models (Table 4). This observation is across all percentages of missingness. Third, the inclusion of cluster labels in MVI yields datasets that achieve a better classification accuracy compared to the data set imputed without the cluster information (Table 5). The improvement in classification accuracy is consistently observed in one of the best MICE model with ensemble learning (GB-MICE). Fourth, missing not at random (MNAR) is the most challenging among the three types of missingness, resulting in the highest NRMSE and lowest classification accuracy (Tables 4-6). The best MVI and classification accuracies are observed in the MCAR type, followed by the MAR type. Fifth, both MAR and MCAR types are best imputed by our proposed ensemble plus cluster-based MICE and DR-MICE models (Table 6). In contrast, the MNAR type is better imputed with iterative SVD and DR-MICE than ensemble-based MICE models. Sixth, MICE-based MVI models are superior to non-MICE models with up to 30%-50% missing data. Non-MICE MVI models, such as iterative SVD and KNN, have been found to outperform the baseline MICE model at higher missingness percentages ($>50\%$). (Figure 3).

5.1 Effects of missingness types

The missingness type in data manifests several notable effects on the performance of MVI algorithms. Figure 2 shows the varying number of monotone patterns across different percentages of missingness for different missingness types. This variability in monotone patterns may have impacted the MVI performance. According to [13], only 15 out of their 111 surveyed papers (13.5%) have studied all three missingness types, including MNAR. Therefore, MVI models in the literature are often limited to only MCAR type, which may not perform equally well on more realistic types of missingness such as MAR and MNAR as observed in our study. It is noteworthy that the MAR type is defined under the same conditional dependence on observed data that is used to formulate MICE-based imputation models. Therefore, it is no wonder that MICE will yield superior performance with MAR or MCAR types. However, the superiority of the MICE framework in previous studies may not be generalized for the MNAR type. In MNAR, missingness depends on the missing values rather than the observed part of the data. Therefore, MICE-based models can yield suboptimal performance because missing values do not depend on the observed values in the MNAR type. Our results consistently show that the imputation and classification accuracies for the MNAR type are substantially more vulnerable to increasing percentages of missingness than MAR or MCAR types.

5.2 Effects of clustering on imputation

The addition of cluster information to individual data samples has consistently improved the MVI accuracy over the baseline MICE and ensemble learning in MICE models. That is, a missing value is regressed from its sample cluster information in addition to other observed variables. In standard MICE, a missing value is regressed from heterogeneous samples that can be widely different from the sample under imputation. The chains of equations in MICE remain indifferent about the contribution of heterogeneous data samples, which may introduce high variability in estimating the missing values. Including the cluster information in data treats samples of the same cluster more equally than those from different clusters, improving the imputation accuracy. The improved imputation accuracy has also resulted in better quality data for training and testing classification models.

5.3 Effects of data sets

The performance of the baseline and proposed MVI algorithms has revealed some dataset-specific trends. It is obvious that not all data sets are imputed equally well by one MVI algorithm. Therefore, papers reporting imputation accuracy on a single data set or domain data may not generalize well on other data domains. The wine quality data set shows the most robust performance of the baseline MICE model compared to non-MICE models in Figure 3(d), even for a high percentage of missingness in data. In contrast, the MICE model yields the worst performance at high missingness percentages ($> 50\%$) when imputing the mice protein expression data set. It is noteworthy that the wine quality and mice datasets have the lowest (11) and highest (77) number of variables. Intuitively, a large number of variables may complicate the performance of the MICE model because of involving one regression model per variable. The most conspicuous discrepancy is noted with the skillcraft1 data set in Figure 3(c) at varying percentages of missingness. The skillcraft1 data set is invariably best imputed by DR-MICE regardless of the missingness types (Tables 3,6). It appears that deep neural networks in the MI framework are robust for a particular data set (skillcraft1) and also in imputing datasets with MNAR type. Unlike other data sets, the skillcraft1 data set contains all positive values within a large range between zero and one million. Eight out of 19 variables of this data set have mean values less than 0.01, starting at the third decimal point. Data sets with similar distribution and statistics may be better imputed by the DR-MICE model. The DR-MICE model also performs the best in imputing the mice protein expression data set with the highest number of variables (Table 6). The deep neural network sequentially down samples the data dimensionality by half at each of the three hidden layers, which may benefit the imputation of high dimensional data.

5.4 Selection of imputation models

The baseline MICE (LR-MICE) is often compared against any newly proposed MVI algorithms in the literature [20, 5]. We improve the MICE model by incorporating more accurate ensemble and deep regressors, which can be taken for future model comparisons. The concept of ensemble learning in the MI framework appears in some sporadic papers on specific data domains, which are yet to receive attention similar to the baseline MICE model. Our extensive analyses on both MICE and non-MICE MVI algorithms provide several recommendations considering missingness types, missingness percentages, and data domains.

The baseline MICE model appears to be effective in most MVI studies with missingness within 30%. Our results reveal that the baseline MICE model performance substantially deteriorates for data with more than 50% missing values compared to non-MICE models. In two extreme cases: 1) high missingness percentages ($>$50%) and 2) MNAR type, we identify that non-MICE algorithms, such as iterative SVD and KNN, outperform MICE-based models. This observation is explainable because the number of observed values also drops with increasing missing data, whereas the MNAR type does not depend on observed values. In contrast, the chains of equations in the MICE model are built by conditioning on observed values. Furthermore, with increasing missingness, the data matrix is proportionally filled with more default values (e.g., median) to initiate an MVI algorithm. The presence of default values results in low-rank data matrices. Therefore, low-rank matrix decomposition techniques, such as SVD, may more effectively estimate missing values in such data.

5.5 Limitations

Despite promising results, this work has several limitations, like any other study. Our proposed models are susceptible to a higher percentage of missingness and the MNAR type, which is an inherent limitation of the MICE framework. Despite the promising performance, the proposed DR-MICE model takes excessive time to train, losing its practical value. Our models are built on regressors, which are suitable for real-valued data, not categorical variables. The performance of MVI models does not explain how these models correlate with different data distributions, which would be imperative for selecting the appropriate MVI model for a given data set. The clustering is performed on median imputed data where the separation between data clusters diminishes with increasing missingness imputed by default median values. Therefore, the performance of our cluster-based MVI approach is limited by the percentage of missingness.

6 Conclusions

This paper aims to improve one of the most successful MVI algorithms, MICE, by modifying its linear mapping between observed and missing values using ensemble learning, deep learning, and cluster information. Our extensive analyses across a wide range of missingness, six data sets, and three missingness types conclude that ensemble learning and cluster information together consistently improve the imputation accuracy of the MICE framework. However, the MICE-based models are vulnerable to high missingness and the MNAR type compared to non-MICE models. This shortcoming of MICE is unknown because MVI studies are often limited up to 50% missingness and MAR/MCAR types. The deep learning-based MICE framework may be superior with specific data sets and missingness types, including MNAR, but at substantially higher computational costs than other models. In the future, the issues listed in the limitation section will be addressed to improve the robustness and explainability of the proposed imputation models.

Acknowledgements

References

[1] Collins Leke, Tshilidzi Marwala, and Satyakama Paul. Proposition of a theoretical model for missing data imputation using deep learning and evolutionary algorithms. *arXiv preprint arXiv:1512.01362*, 2015.

[2] H. Cevallos Valdiviezo and S. Van Aelst. Tree-based prediction on incomplete data using imputation or surrogate decisions. *Information Sciences*, 311:163–181, 2015.

[3] Stef van Buuren. *Flexible imputation of missing data*. Chapman and Hall/CRC, 2018.

[4] Sanaz Nikfalazar, Chung Hsing Yeh, Susan Bedingfield, and Hadi A. Khorshidi. Missing data imputation using decision trees and fuzzy clustering with iterative learning. *Knowledge and Information Systems*, 62(6):2419–2437, 2020.

[5] Saqib Ejaz Awan, Mohammed Bennamoun, Ferdous Sohel, Frank Sanfilippo, and Girish Dwivedi. Imputation of missing data with class imbalance using conditional generative adversarial networks. *Neurocomputing*, 453:164–171, 2021.

[6] Ramiro D. Camino, Christian A. Hammerschmidt, and Radu State. Improving Missing Data Imputation with Deep Generative Models. 2019.

[7] Zhenhua Wang, Olanrewaju Akande, Jason Poulos, and Fan Li. Are deep learning models superior for missing data imputation in large surveys? Evidence from an empirical comparison. *arXiv preprint arXiv:2103.09316*, 2021.

[8] Nan Chen, Meijuan Li, and Hongyun Liu. Comparison of maximum likelihood approach, Diggle–Kenward selection model, pattern mixture model with MAR and MNAR dropout data. *Communications in Statistics-Simulation and Computation*, 49(7):1746–1767, 2020.

[9] Ralph C Ward, Robert Neal Axon, and Mulugeta Gebregziabher. Approaches for missing covariate data in logistic regression with MNAR sensitivity analyses. *Biometrical Journal*, 62(4):1025–1037, 2020.

[10] Brett K Beaulieu-Jones, Daniel R Lavage, John W Snyder, Jason H Moore, Sarah A Pendergrass, and Christopher R Bauer. Characterizing and managing missing structured data in electronic health records: Data analysis. *JMIR medical informatics*, 6(1):e8960, 2018.

[11] Aurora Gonzalez-Vidal, Punit Rathore, Aravinda S. Rao, Jose Mendoza-Bernal, Marimuthu Palaniswami, and Antonio F. Skarmeta-Gomez. Missing Data Imputation with Bayesian Maximum Entropy for Internet of Things Applications. *IEEE Internet of Things Journal*, 4662(i):1–1, 2020.

[12] Harshad Hegde, Neel Shimpi, Aloksagar Panny, Ingrid Glurich, Pamela Christie, and Amit Acharya. MICE vs PPCA: Missing data imputation in healthcare. *Informatics in Medicine Unlocked*, 17:100275, 2019.

[13] Wei Chao Lin and Chih Fong Tsai. Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53(2):1487–1509, 2020.

[14] Manar D. Samad, Alvaro Ulloa, Gregory J. Wehner, Linyuan Jing, Dustin Hartzel, Christopher W. Good, Brent A. Williams, Christopher M. Haggerty, and Brandon K. Fornwalt. Predicting Survival From Large Echocardiography and Electronic Health Record Datasets. *JACC: Cardiovascular Imaging*, 12(4):681–689, jun 2018.

[15] Bas van Stein, Wojtek Kowalczyk, and Thomas Bäck. Analysis and visualization of missing value patterns. In *Communications in Computer and Information Science*, volume 611, pages 187–198. Springer Verlag, 2016.

[16] Vahid Nassiri, Geert Molenberghs, Geert Verbeke, and João Barbosa-Breda. Iterative Multiple Imputation: A Framework to Determine the Number of Imputed Datasets. *American Statistician*, 74(2):125–136, 2020.

[17] Matthieu Resche-Rigon and Ian R. White. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research*, 27(6):1634–1649, 2018.

[18] Yuan Luo, Peter Szolovits, Anand S. Dighe, and Jason M. Baron. 3D-MICE: Integration of cross-sectional and longitudinal imputation for multi-analyte longitudinal clinical data. *Journal of the American Medical Informatics Association*, 25(6):645–653, 2018.

[19] Timur Köse, Su Özgür, Erdal Coşgun, Ahmet Keskinoğlu, Pembe Keskinoğlu, and Dariusz Mrozek. Effect of Missing Data Imputation on Deep Learning Prediction Performance for Vesicoureteral Reflux and Recurrent Urinary Tract Infection Clinical Study. *BioMed Research International*, 2020, 2020.

[20] Marek Smieja, Łukasz Struski, Jacek Tabor, Bartosz Zieliński, and Przemysław Spurek. Processing of missing data by neural networks. *arXiv preprint arXiv:1805.07405*, 2018.

[21] Thu Nguyen, Duy HM Nguyen, Huy Nguyen, Binh T Nguyen, and Bruce A Wade. EPEM: Efficient Parameter Estimation for Multiple Class Monotone Missing Data. *Information Sciences*, 567:1–22, 2021.

[22] Benedict W.J. Irwin, Julian R. Levell, Thomas M. Whitehead, Matthew D. Segall, and Gareth J. Conduit. Practical Applications of Deep Learning to Impute Heterogeneous Drug Discovery Data. *Journal of Chemical Information and Modeling*, 60(6):2848–2857, 2020.

[23] Felix Biessmann, David Salinas, Sebastian Schelter, Philipp Schmidt, and Dustin Lange. Deep learning for missing value imputation in tables with non-numerical data. *International Conference on Information and Knowledge Management, Proceedings*, pages 2017–2026, 2018.

[24] Hongbao Zhang, Pengtao Xie, and Eric Xing. Missing value imputation based on deep generative models. *arXiv preprint arXiv:1808.01684*, pages 1–9, 2018.

[25] Yifan Zhuang, Ruimin Ke, and Yinhai Wang. Innovative method for traffic data imputation based on convolutional neural network. *IET Intelligent Transport Systems*, 13(4):605–613, apr 2019.

[26] M. Sangeetha and M. Senthil Kumaran. Deep learning-based data imputation on time-variant data using recurrent neural network. *arXiv preprint arXiv:1808.01684*, (17):13369–13380.

[27] Suvra Jyoti Choudhury and Nikhil R. Pal. Imputation of missing data with neural networks for classification. *Knowledge-Based Systems*, 182:104838, 2019.

[28] G. Madhu, B. Lalith Bharadwaj, G. Nagachandrika, and K. Sai Vardhan. A Novel Algorithm for Missing Data Imputation on Machine Learning. *Proceedings of the 2nd International Conference on Smart Systems and Inventive Technology, ICSSIT 2019*, (Icssit):173–177, 2019.

[29] Olanrewaju Akande, Fan Li, and Jerome Reiter. An Empirical Comparison of Multiple Imputation Methods for Categorical Data. *The American Statistician*, 71(2):162–170, apr 2017.

[30] Linchao Li, Carlo G Prato, and Yonggang Wang. Ranking contributors to traffic crashes on mountainous freeways from an incomplete dataset: A sequential approach of multivariate imputation by chained equations and random forest classifier. *Accident Analysis & Prevention*, 146:105744, 2020.

[31] Emily Slade and Melissa G. Naylor. A fair comparison of tree-based and parametric methods in multiple imputation by chained equations. *Statistics in Medicine*, 39(8):1156–1166, 2020.

[32] Antony Hallam, Debajoy Mukherjee, and Romain Chassagne. Multiple imputation via chained equations for elastic welllog imputation and prediction. 2021.

[33] Manar D Samad and Linlin Yin. Non-linear regression models for imputing longitudinal missing data. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–3. IEEE, 2019.

[34] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

[35] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. Accessed on July 18, 2021.

[36] Alex Rubinsteyn and Sergey Feldman. Fancyimpute. an imputation library for python. *URL https://github.com/iskandr/fancyimpute*, 2016.