

Imputation using training labels and classification via label imputation

Thu Nguyen^{1*}, Tuan L. Vo², Pål Halvorsen¹, Michael A. Riegler¹

¹*Department of Holistic Systems, SimulaMet, Oslo, Norway.

²Department of Mathematics and Computer Science, University Of
Science, Ho Chi Minh City, Vietnam.

*Corresponding author(s). E-mail(s): thu@simula.no;
Contributing authors: tuanlvo293@gmail.com; paalh@simula.no;
michael@simula.no;

Missing data is a common problem in practical data science settings. Various imputation methods have been developed to deal with missing data. However, even though in many situations, the labels are available in the training data, the common practice of imputation usually only relies on the input and ignores the label. We propose the *imputation using labels (IUL)* algorithm, a strategy that stacks the label into the input and illustrate how it can significantly improve the imputation quality of the input. In addition, we propose *Classification Based on MissForest Imputation (CBMI)*, a classification strategy that initializes the predicted test label with missing values and stacks the label with the input for imputation, and the imputation is based on the well-known MissForest algorithm [Stekhoven and Bühlmann \(2012\)](#). This allows imputing the label and the input at the same time. Also, the technique is capable of handling data training with missing labels without any prior imputation and is applicable to continuous, categorical, or mixed-type data. Experiments show promising improvement in terms of mean squared error/classification accuracy, especially for imbalanced data, categorical data, and small sample data. Importantly, the experiment results suggest that when a dataset has too little sample to build a good classification model upon, merging the training and testing data and imputing the label as CBMI can be a significantly better option than building a classification model on the training set and use it to classify the test samples.

1 Introduction

Data can come in various forms, ranging from continuous numerical values to discrete values and mixed between categorical and continuous. For example, a dataset on housing may contain information about the price and areas, which are continuous, while the number of rooms and number of bathrooms are discrete features. In addition, it is also common to encounter missing data. This can happen for various reasons, such as malfunctioning devices, broken sensors, or people declining to fill in survey information, etc. This can undermine the integrity and robustness of analyses, potentially leading to biased results and incomplete insights. Consequently, the need for methodologies to address missing data becomes paramount, prompting the widespread adoption of imputation techniques.

Imputation, as a remedial practice, involves the estimation or prediction of missing values based on observed information. This process becomes particularly intricate when dealing with datasets that encompass diverse data types, including continuous, categorical, and mixed data. In response to this complexity, a spectrum of imputation techniques has been developed, each tailored to handle specific data types and patterns [Stekhoven and Bühlmann \(2012\)](#); [Vu et al. \(2023\)](#). In addition, in recent years, classification or regression models that can handle missing data directly such as Decision Tree Classifier [Breiman \(2017\)](#) and LSTM [Ghazi et al. \(2018\)](#) have also captured a lot of attention. However, imputation remained a popular choice as it renders the data complete not only for the classification or regression task but also data visualization or other inferences [Pham et al. \(2023\)](#).

In this work, we propose stacking training labels to training input to improve the imputation of the training input and illustrate how this can significantly improve the imputation of training input. Moreover, we propose Classification Based on MissForest Imputation (CBMI) to predict the label on testing data without building a classification model on the training set. The method starts by initializing the predicted testing labels with missing values. Then, it stacks the input and the label, the training and testing data together, and uses missForest algorithm [Stekhoven and Bühlmann \(2012\)](#) to impute all the missing values in the matrix. The imputation returns a matrix that contains imputed training and testing input, imputed training labels if the training labels contain missing data, and predicted testing labels.

Our contributions are as following: (i) We propose using training labels to aid the imputation of the training input and illustrate how this can significantly improve the imputation of training input; (ii) We propose CBMI algorithm for predicting the output of a classification task via imputation, instead of building a model on training data and then predicting on a test set; (ii) we conduct various experiments to illustrate that most of the time, the proposed approaches provide better accuracy for classification.

The rest of this paper is organized as follows: In section 2, we detail some related works in this area. Next, in section 3, we review the fundamentals of the MissForest algorithm [Stekhoven and Bühlmann \(2012\)](#). Then, in section 4, we detail our proposed methods. Finally, we conduct the experiments to illustrate the power of the proposed techniques in section 5, and we summarize the main ideas of the paper in section 6.

2 Related works

A diverse array of techniques, ranging from traditional methods to sophisticated algorithms, has been developed to handle missing data. Each method brings unique strengths, making them suitable for different types of data and analysis scenarios. Traditional approaches, for example, complete case analysis, have been widely used due to their simplicity but are prone to bias due to the exclusion of observations with missing values. More sophisticated methods, such as multiple imputation [Buuren and Groothuis-Oudshoorn \(2010\)](#) or multiple imputation using Deep Denoising Autoencoders [Gondara and Wang \(2017\)](#), have emerged to mitigate these challenges. The methods acknowledge the uncertainty associated with the imputation process by generating multiple plausible values for each missing data point. Another notable work is the Conditional Distribution-based Imputation of Missing Values (DIMV) [Vu et al. \(2023\)](#) algorithm. The algorithm finds the conditional distribution of features with missing values based on fully observed features, and its imputation step provides coefficients with direct explainability similar to regression coefficients.

Regression and clustering-based methods also play a significant role in addressing missing data. The CBRL and CBRC algorithms [M Mostafa, S Eladimy, Hamad, and Amano \(2020\)](#), utilizing Bayesian Ridge Regression, showcase the efficacy of regression-based imputation methods. Additionally, the cluster-based local least square method [Keerin, Kurutach, and Boongoen \(2013\)](#) offers an alternative approach to imputation based on clustering techniques. For big datasets, deep learning imputation techniques are also of great use due to their powerful performance [Choudhury and Pal \(2019\)](#); [Garg, Naryani, Aggarwal, and Aggarwal \(\(2018\)\)](#); [Mohan and Pearl \(2021\)](#).

For continuous data, techniques based on matrix decomposition or matrix completion, such as Polynomial Matrix Completion [Fan, Zhang, and Udell \(\(2020\)\)](#), Alternating Least Squares (ALS) [Hastie, Mazumder, Lee, and Zadeh \(2015\)](#), and Nuclear Norm Minimization [Candès and Recht \(2009\)](#), have been employed to make continuous data complete, enabling subsequent analysis using regular data analysis procedures. For categorical data, various techniques have been explored. The use of k-Nearest Neighbors imputation involves identifying missing values, selecting neighbors based on a similarity metric (e.g., Hamming distance), and imputing missing values by a majority vote from the neighbors' known values. Moreover, the missForest imputation method [Stekhoven and Bühlmann \(2012\)](#) has shown efficacy in handling datasets that comprise a mix of continuous and categorical features. In addition, some other methods that can handle mixed data include FEMI [Rahman and Islam \(2016\)](#), SICE [Khan and Hoque \(2020\)](#), and HCMM-LD [Murray and Reiter \(2016\)](#).

In recent years, some other aspects of imputation such as scalability also are being considered. For example, in [T. Nguyen, Ly, Riegler, Halvorsen, and Hammer \(2023\)](#); [T.H. Nguyen et al. \(2023\)](#), the authors use Principle Component Analysis (PCA) to conduct dimension reduction on fully observed features and merged this part with the not fully observed features to impute. This help resulting in significant improvement in speed. A similar examination is conducted in [P. Nguyen et al. \(2023\)](#) where Singular Value Decomposition is used instead of PCA. Next, in [Do et al. \(2023\)](#), the authors propose a Blockwise principal component analysis Imputation for monotone missing data, where Principal Component Analysis (PCA) is conducted on the observed part

of each monotone block of the data and it use a chosen imputation technique to impute on merging the obtained principal components. Experiments also show a significant improvement in imputation speed, at a minor cost of imputation quality.

In addition, there has been a considerable development of classification methods and models capable of directly handling missing data, circumventing the need for imputation. These approaches recognize that imputation introduces a layer of uncertainty and potential bias and instead integrates mechanisms to utilize the available information effectively. Typically tree-based methods that have such capabilities are Gradient-boosted trees [Friedman \(2001\)](#), Decision Tree Classifier [Breiman \(2017\)](#), and their implementation are readily available in the sklearn package [Pedregosa et al. \(2011\)](#).

The Random Forest classifier has been a pivotal contribution in the field of machine learning, renowned for its robustness and versatility. Originating from ensemble learning principles, Random Forests combine the predictions of multiple decision trees to enhance overall predictive accuracy and mitigate overfitting. This approach was introduced by Breiman in 2001 [Breiman \(2001\)](#), and since then, it has gained widespread adoption across diverse domains. The classifier’s ability to handle high-dimensional data, capture complex relationships, and provide estimates of feature importance has made it particularly valuable. Various extensions and modifications to the original Random Forest algorithm have been proposed to address specific challenges. For instance, methods like Extremely Randomized Trees [Geurts, Ernst, and Wehenkel \(2006\)](#) introduce additional randomness during the tree-building process, contributing to increased diversity and potentially improved generalization. Additionally, research efforts have explored adapting Random Forests for specialized tasks, such as imbalanced classification, time-series analysis, and feature selection.

3 Preliminary: missForest algorithm

MissForest [Stekhoven and Bühlmann \(2012\)](#) is an imputation algorithm used for handling missing data in datasets. It is suitable not only for continuous or categorical but also for mixed-type data. The algorithm works by iteratively constructing a random forest based on the observed data, and then using this forest to predict missing values in the dataset. The process is repeated until convergence, refining the imputations in each iteration. MissForest is advantageous for its ability to handle non-linear relationships and interactions in the data, making it suitable for a wide range of applications. In addition, one key feature of MissForest is its adaptability to various types of missing data patterns, such as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). This flexibility makes it a versatile tool for imputing missing values in datasets with complex structures.

For the setting of the proposed CBMI method, we use the default missForest setting for γ . That means that, a set of continuous features F , γ is governed by

$$\Delta = \frac{\sum_{j \in F} (\mathbf{X}_{new}^{imp} - \mathbf{X}_{old}^{imp})^2}{\sum_{j \in F} (\mathbf{X}_{new}^{imp})^2}, \quad (1)$$

Algorithm 1 MissForest algorithm

Input:

1. a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$,
2. stopping criterion γ

Procedure:

- 1: initialize missing values in $\mathbf{X}$ by some imputation method
- 2: $\mathbf{k} \leftarrow$ sorted indices of features in $\mathbf{X}$ in an increasing order of missing values
- 3: **while not** γ **do**:
- 4: $\mathbf{X}_{old}^{imp} \leftarrow$ previously imputed version of $\mathbf{X}$,
- 5: **for** s in $\mathbf{k}$ **do**:
- 6: fit a random forest where the observed values in $\mathbf{X}_s$ is the response, and the corresponding values of features in $\mathbf{X}$ other than $\mathbf{X}_s$ as the input
- 7: predict $\mathbf{y}_{miss}^{(s)}$ (the missing values in $\mathbf{X}_s$) based on the features other $\mathbf{X}_s$.
- 8: $\mathbf{X}_{new}^{imp} \leftarrow$ updated version of $\mathbf{X}$ using $\mathbf{y}_{miss}^{(s)}$
- 9: **end for**
- 10: Update γ .
- 11: **end while**

Return: the imputed version of $\mathbf{X}$.

while for a set of categorical features G , γ is governed by

$$\Delta_G = \frac{\sum_{j \in G} \sum_{i=1}^n \mathbf{I}_{\mathbf{X}_{new}^{imp} \neq \mathbf{X}_{old}^{imp}}}{\#NA}. \quad (2)$$

Here, $\#NA$ is the number of missing entries in the categorical features.

4 Methodology

4.1 Using labels for imputation

In this section, we detail the idea of *imputation using labels (IUL)*. For clarity, from now, we refer to the common imputation procedure that only relies on the input itself, without the labels, as *direct imputation (DI)*.

The IUL process is straightforward and presented in algorithm 2. Suppose that we have input data $\mathbf{X}_{train}$ and labels $\mathbf{y}_{train}$. Then, the procedure starts by stacking the input training data $\mathbf{X}_{train}$ and labels $\mathbf{y}_{train}$ in a column-wise manner in step 1. Then, in step 2, the algorithm uses the imputation method I to impute $\mathcal{D}_{train}$. This gives $\mathcal{D}_{train}^{imp}$. Then in step 3, the algorithm assigns $\mathbf{X}_{train}^{imp}$, the imputed version of $\mathbf{X}_{train}$, to be the submatrix of $\mathcal{D}_{train}^{imp}$ that consists of all columns except the last column of $\mathcal{D}_{train}^{imp}$. In addition in step 4, the algorithm assigns $\mathbf{y}_{train}^{imp}$, the imputed version of $\mathbf{y}_{train}$ (if $\mathbf{y}_{train}$ contains missing values), to be the last column of $\mathcal{D}_{train}^{imp}$. The algorithm ends by returning $\mathbf{X}_{train}^{imp}, \mathbf{y}_{train}^{imp}$.

Algorithm 2 IUL algorithm

Input:

1. input training data $\mathbf{X}_{train}$, labels $\mathbf{y}_{train}$
2. imputation algorithm I .

Procedure:

- 1: $\mathcal{D}_{train} = [\mathbf{X}_{train} | \mathbf{y}_{train}]$
- 2: $\mathcal{D}_{train}^{imp} \leftarrow$ the imputed version of $\mathcal{D}_{train}$ using the imputation algorithm I .
- 3: $\mathbf{X}_{train}^{imp} \leftarrow$ imputed version of $\mathbf{X}_{train}$ that consists of all columns except the last column of $\mathcal{D}_{train}$,
- 4: $\mathbf{y}_{train}^{imp} \leftarrow$ imputed version of $\mathbf{y}_{train}$ if $\mathbf{y}_{train}$ contains missing values. This corresponds to the last column of $\mathcal{D}_{train}^{imp}$.

Return: $\mathbf{X}_{train}^{imp}, \mathbf{y}_{train}^{imp}$

4.1.1 Remarks

It is important to note that *the IUL algorithm can handle not only $\mathbf{X}_{train}$ with missing data, but also $\mathbf{y}_{train}$ with missing data.* In case the labels contain missing data, the returned $\mathbf{y}_{train}^{imp}$ is the imputed version of the training labels. So, this also has a great potential to be applied to semi-supervised learning, as in such a setting, the samples without labels can be considered as samples with missing labels.

Some intuition about why IUL and CBMI work in practice can be seen from the feature selection aspect: when we use labels for imputing a feature f in the input, the amount of information increases. Therefore, one can expect the imputation quality to improve unless the feature is not useful or the information from the label that is useful for imputing f can already be retrieved by other feature(s) in the input. For an Ordinary Least Square problem, for example, the MSE is known not to decrease as the number of features increases. However, this may not hold true for other regression or classification models.

It is also important to notice that *IUL can be used with other imputation techniques than MissForest.* Also, *the difference in the number of choices for building a random forest when stacking the label to the input as in IUL can be significantly higher than imputing based only on $\mathbf{X}_{train}$ itself.* To be more specific, assuming we have p features in $\mathbf{X}_{train}$ and one label $\mathbf{y}$. With IUL, the label can be utilized to aid the imputation of $\mathbf{X}_{train}$, which means that we have $p + 1$ features. In particular, if we are to build m model RandomForest and choose k features for each model, then we can have $\binom{p+1}{k}$ combinations. Meanwhile, DI only work on $\mathbf{X}_{train}$ so there are just $\binom{p}{k}$ options. The difference is

$$\binom{p+1}{k} - \binom{p}{k} = \binom{p}{k-1}, \quad (3)$$

which can be extremely large.

Last but not least, while it will be shown in the experiments that IUL has a competitive performance compared to DI, IUL is not theoretically guaranteed to outperform DI all the time. The following results illustrate that

Theorem 1. Assume that we have the data $\mathcal{D} = (\mathbf{I}, \mathbf{y})$ of n samples. Here, $\mathbf{I} = (\mathbf{x}, \mathbf{z})$, $\mathbf{x} \in \mathbf{R}$ contains missing values, $\mathbf{z} \in \mathbf{R}$ is fully observed and $\mathbf{y}$ is the label. For MICE imputation, with the IUL strategy, we construct a model

$$\hat{x} = \hat{\gamma}_o + \hat{\gamma}_1 z + \hat{\gamma}_2 y. \quad (4)$$

Meanwhile, DI ignores label $\mathbf{y}$ and use the following model

$$\hat{x}' = \hat{\gamma}_o + \hat{\gamma}_1 z. \quad (5)$$

Without loss of generality, assume that $y_i \geq 0, i = 1, 2, \dots, n$ for label encoding. Least square method yields the following estimates for the coefficients,

$$\begin{aligned} \hat{\gamma}_1 &= \frac{(\sum y^2)(\sum xz) - (\sum zy)(\sum yx)}{(\sum z^2)(\sum y^2) - (\sum zy)^2}, \\ \hat{\gamma}_2 &= \frac{(\sum z^2)(\sum yz) - (\sum zy)(\sum zx)}{(\sum z^2)(\sum y^2) - (\sum zy)^2}, \\ \hat{\gamma}_o &= \mu_{\mathbf{x}} - \hat{\gamma}_1 \mu_{\mathbf{z}} - \hat{\gamma}_2 \mu_{\mathbf{y}}. \end{aligned}$$

Let denote

$$\mathcal{E}_i = \hat{\gamma}_2^2 (y_i - 2\mu_{\mathbf{y}}) + 2\hat{\gamma}_1 \hat{\gamma}_2 (z_i - \mu_{\mathbf{z}}), \quad i = 1, 2, \dots, n, \quad (6)$$

$$\mathcal{I}^+ = \{i : \mathcal{E}_i \geq 0\}, \quad \mathcal{V}^+ = \sum_{i \in \mathcal{I}^+} y_i \mathcal{E}_i, \quad (7)$$

$$\mathcal{I}^- = \{i : \mathcal{E}_i < 0\}, \quad \mathcal{V}^- = \sum_{i \in \mathcal{I}^-} y_i \mathcal{E}_i. \quad (8)$$

Then, $SSE_{IUL} \leq SSE_{DI}$ if and only if $\mathcal{V}^+ + \mathcal{V}^- \geq 0$, where SSE_{IUL}, SSE_{DI} are the Sum of Square Error for the model based on IUL and DI, respectively.

The proof of this theorem is provided in Appendix A.

4.2 Classification via imputation

In this section, we detail the methodology of the proposed **Classification Based on MissForest Imputation (CBMI)** method. The algorithm is presented in algorithm 3.

Suppose that we have a dataset that consists of the training set $\{\mathbf{X}_{train}, \mathbf{y}_{train}\}$ that has n_{train} samples and the testing set $\{\mathbf{X}_{test}, \mathbf{y}_{test}\}$ that has n_{test} samples. Then, the procedure is as follows:

At the first step of the algorithm, we stack the training input $\mathbf{X}_{train}$ and the training label $\mathbf{y}_{train}$ in a column-wise manner. Next, in step 2, we initiate the predicted

label of the test set $\hat{\mathbf{y}}$ with missing values, denoted by *. After that, in step 3, we stack the input of the test set $\mathbf{X}_{test}$ and $\hat{\mathbf{y}}$ in a column-wise manner, and we call this $\mathcal{D}_{test}^*$.

In the fourth step, we stack $\mathcal{D}_{train}$ and $\mathcal{D}_{test}^*$ in a row-wise manner. This gives us $\mathcal{D}^*$. Then, at step 5, we impute $\mathcal{D}^*$ using the MissForest algorithm. This gives us an imputed matrix $\mathcal{D}$ that contains the imputed training, testing input, and the imputed label of the training set if the training data has missing labels and the imputed label $\hat{\mathbf{y}}$ of the test set.

Algorithm 3 CBMI algorithm

Input:

1. a dataset consists of the training set $\{\mathbf{X}_{train}, \mathbf{y}_{train}\}$ that has n_{train} samples and the testing set $\{\mathbf{X}_{test}, \mathbf{y}_{test}\}$ that has n_{test} samples,

Procedure:

- 1: $\mathcal{D}_{train} = [\mathbf{X}_{train} | \mathbf{y}_{train}]$
- 2: $\hat{\mathbf{y}} = \begin{pmatrix} * \\ \vdots \\ * \end{pmatrix}$: an NA-initiated vector of size n_{test} where n_{test} is the number of sample in the test set
- 3: $\mathcal{D}_{test}^* = [\mathbf{X}_{test} | \hat{\mathbf{y}}]$
- 4: $\mathcal{D}^* = \begin{pmatrix} \mathcal{D}_{train} \\ \mathcal{D}_{test}^* \end{pmatrix}$
- 5: $\mathcal{D} \leftarrow$ the imputed version of $\mathcal{D}^*$ using the missForest algorithm

Return: $\mathcal{D}$, the imputed version of $\hat{\mathbf{y}}$ obtained from $\mathcal{D}$.

So, instead of following the conventional way of imputing the data and then build a model on a training set, CBMI conducts no prior imputation and build no model on training data to predict the label of the test set. Instead, CBMI combine the training and testing data, including the input and the available and NA-initialized labels together to impute the label of the test set.

5 Experiments

5.1 Experiment set up

We compare the imputation using labels (IUL) with the direct imputation on training data approach (DI) by splitting a dataset into training and testing sets with a ratio of 6:4 and simulating missing rates 20% – 80% on the training input. Here, the missing rate is defined as the ratio between the number of missing entries and the total number of entries. After missing data generation, the data is scaled to $[-1, 1]$. Each experiment is repeated 10 times, and we report the mean and standard deviation of accuracy and running time. We use MissForest as the imputation method with default configuration and Random Forest as the classifier.

Table 1 Datasets for experiments

Datasets	# samples	# features
Iris	150	4
liver	345	6
soybean	47	35
Parkinson	195	22
heart	267	44
glass	214	9
car	1728	6
California housing	20640	8
diabetes	442	10

In addition, we compare CBMI with the two-step approach of *Imputation* using missForest and then *Classification* using Random Forest Classifier Breiman (2001), abbreviated as IClf. The reason for choosing Random Forest as the Classifier is to facilitate fair comparison with CBMI, which relies on Random Forest to impute missing values. The stopping criterion for missForest is the default in the missingpy package¹.

The datasets for the experiments are from the UCI Machine Learning repository Dua and Graff (2017), and are described in table 1. For each dataset, we simulate missing data with missing rates r ranging between 0% - 80%. Each experiment is repeated 10 times, and the training-to-testing ratio is 6:4, and we report the mean $\pm$ standard deviation from the runs for both the accuracy and the running time. We also conducted the experiments in two settings: the test set is fully observed and the test set has missing values. In case the test set is fully observed, the missing rate r is the missing rate for simulating missing values on training data only. In the case where there are missing values in the dataset, the missing values are generated with the same missing rates as in the training input.

For a fair comparison, for IClf, if the test set contains missing data then it is merged with the train set for imputation. Note that this is only for the imputation of the test set only. The imputation of the training data is independent of the testing data. The codes for this paper will be made available at <https://github.com/thunguyen177/iul-cbmi>.

5.2 Results and Analysis for IUL

The results of the experiments on the performance of IUL are shown in figures 2, 3, 4, and 5. Note that the soybean dataset and the car dataset contain solely categorical features. Therefore, the MSE plot for that dataset is not available. From the figures, we can see that while in some cases, IUL and DI have similar performance, a clear improvement of IUL compared to DI can be seen in the glass dataset (figure 5) and the Iris dataset (figure 1).

5.3 Results and Analysis for CBMI

The results for the experiments on the performance of IUL are shown in tables 3, 4, 5, 6, 10, 11, 12, and 13.

¹<https://pypi.org/project/missingpy/>

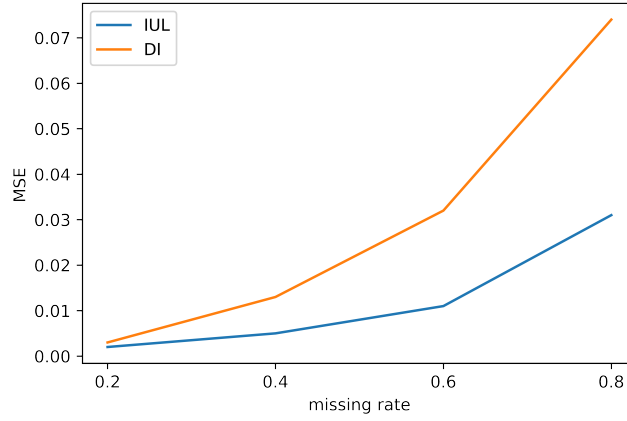


Fig. 1 MSE of IUL and DI on the Iris dataset.

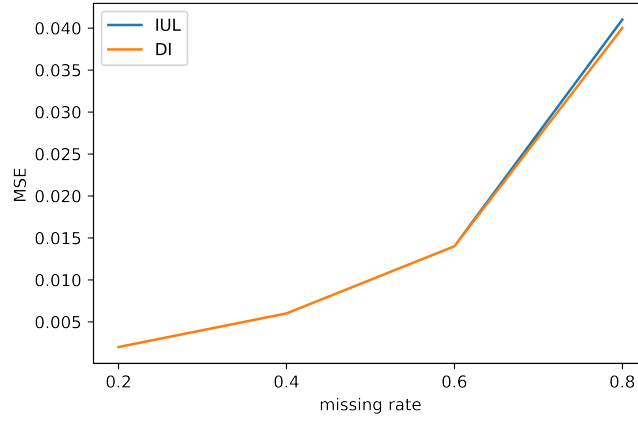


Fig. 2 MSE of IUL and DI on the heart dataset.

From these tables, we see that most of the time, CBMI achieves better results than ICIf. For example, at missing rate $r = 80\%$, for the soybean dataset when missing data present in the test set (table 4), the accuracy of CBMI is 0.547, while it is 0.5 for ICIf. This implies a significant improvement of 4.7% when using CBMI compared to ICIf. In addition, most of the time, the standard deviation of the accuracy obtained from CBMI is also smaller than ICIf.

Interestingly, CBMI achieves better results than ICIf for many cases where the missing rate r is zero, i.e., there is no missing data in the training set. For example, when the training dataset is fully observed ($r = 0$) on the heart dataset with missing data present in the test set (table 6), CBMI gives an accuracy of 0.821, while ICIf gives 0.817; or on the glass dataset where the test set is fully observed (table (table

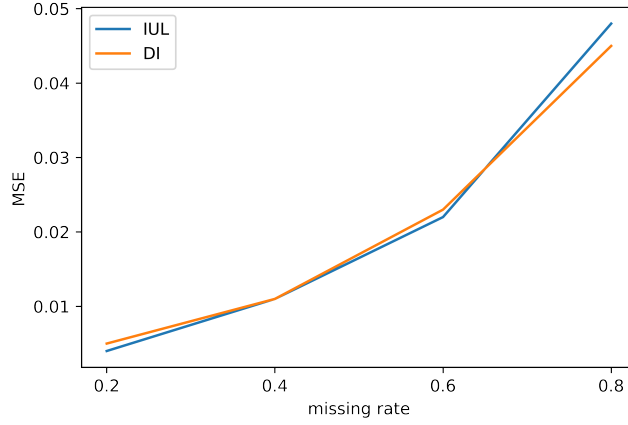


Fig. 3 MSE of IUL and DI on the liver dataset.

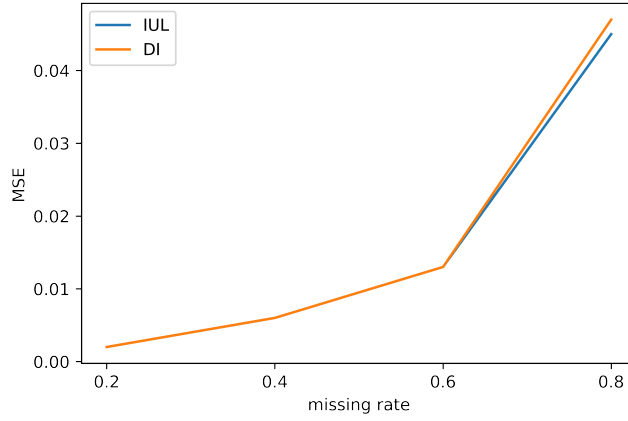


Fig. 4 MSE of IUL and DI on the Parkinson dataset.

14), CBMI's accuracy is 0.762, while ICIf's accuracy is 0.751. Note that such an improvement is small but in many applications such as healthcare or banking, such a small improvement is still significant. In addition, these comparisons show that even for a dataset without missing data, CBMI may provide better test accuracy than building a Random Forest classification model on training data and predict the labels of the test set.

Regarding the running time, when missing data present in the test set, from tables 3, 4, 5 and 6, we can see that when missing data presents in the test set, CBMI is clearly faster. However, from tables 10, 11, 12, and 13, when the testing data is fully observed, ICIf is slightly faster than CBMI most of the time. This is possibly because when the testing data is fully observed, the CBMI's imputation process has to impute

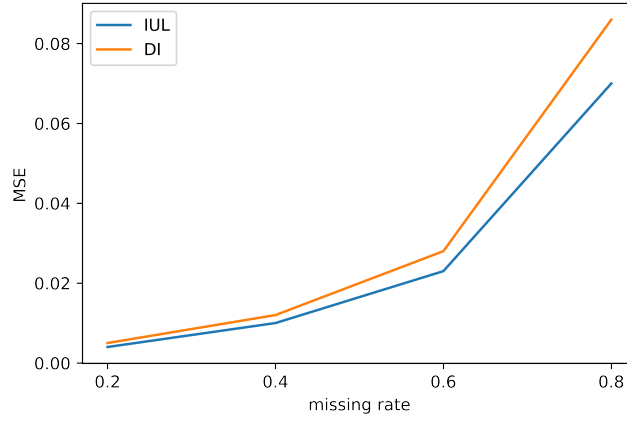


Fig. 5 MSE of IUL and DI on the glass dataset.

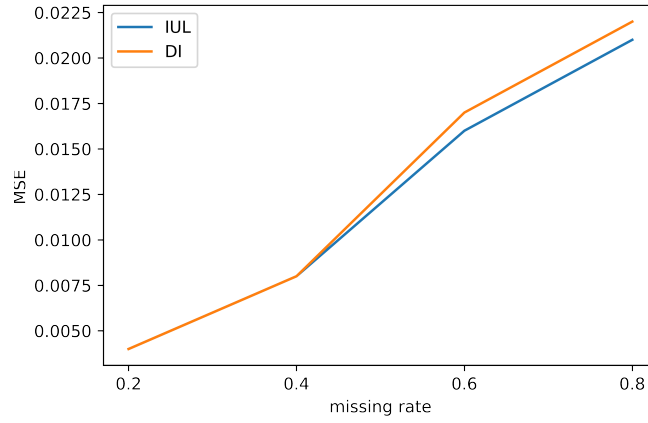


Fig. 6 MSE of IUL and DI under MICE imputation on the California housing dataset.

a larger matrix, which consists of the training and testing input and labels, while ICIf imputes only the training input.

5.4 Results on imbalanced data

Among the dataset for experiments, note that the number of samples for each class in the heart dataset is (55, 212), while for the Parkinson dataset is (48,147), for the car dataset is (384, 69, 1210, 65), and for the glass dataset is (70, 76, 17, 13, 9, 29). Regardless of whether the test set has missing values or not, the CBMI presents better results than ICIf in most cases. For example, when the test input is fully observed, take the result in the glass dataset (table 14), there is a significant difference in the

Table 2 Classification results on the Iris dataset when missing data presents in the test set

r	accuracy		running time	
	CBMI (our)	IClf	CBMI (our)	IClf
20 %	0.922±0.045	0.935±0.026	3.165±0.87	4.346±0.79
40 %	0.873±0.032	0.873±0.047	3.57±0.61	5.312±1.079
60 %	0.752±0.065	0.75±0.049	3.234±0.769	4.875±0.497
80 %	0.583±0.064	0.585±0.035	3.389±1.158	4.172±0.955

Table 3 Classification results on the liver dataset when missing data presents in the test set

r	accuracy		running time	
	CBMI (our)	IClf	CBMI (our)	IClf
20 %	0.649±0.022	0.642±0.019	5.379±1.778	7.094±1.237
40 %	0.587±0.046	0.583±0.028	5.658±1.421	7.662±1.659
60 %	0.541±0.041	0.562±0.042	5.715±1.594	7.502±1.428
80 %	0.554±0.025	0.547±0.048	5.704±1.685	7.774±1.775

Table 4 Classification results on the soybean dataset when missing data presents in the test set

r	accuracy		running time	
	CBMI (our)	IClf	CBMI (our)	IClf
20 %	0.989±0.021	0.984±0.034	24.579±5.383	45.279±6.573
40 %	0.916±0.116	0.9±0.106	27.251±6.12	44.467±5.554
60 %	0.821±0.092	0.8±0.091	29.487±5.299	44.25±9.486
80 %	0.547±0.142	0.5±0.127	26.82±6.247	48.91±8.085

Table 5 Classification results on the Parkinson dataset when missing data presents in the test set

r	accuracy		running time	
	CBMI (our)	IClf	CBMI (our)	IClf
20 %	0.877±0.024	0.869±0.038	15.456±4.69	27.058±5.667
40 %	0.862±0.045	0.858±0.043	18.927±3.824	26.745±2.528
60 %	0.795±0.035	0.813±0.039	18.949±4.077	22.418±5.805
80 %	0.778±0.037	0.769±0.06	18.711±4.516	27.507±5.776

Table 6 Classification results on the heart dataset when missing data presents in the test set

r	accuracy		running time	
	CBMI (our)	IClf	CBMI (our)	IClf
20 %	0.83±0.038	0.83±0.036	42.814±8.138	74.018±12.008
40 %	0.809±0.048	0.813±0.037	50.46±5.902	87.255±11.24
60 %	0.806±0.025	0.801±0.029	47.827±5.954	86.801±9.228
80 %	0.779±0.036	0.783±0.05	48.81±4.614	90.821±7.51

Table 7 Classification results on the glass dataset when missing data presents in the test set

r	accuracy		running time	
	CBMI (our)	IClf	CBMI (our)	IClf
20 %	0.705±0.036	0.692±0.052	8.161±1.42	11.92±1.64
40 %	0.633±0.048	0.598±0.074	9.097±2.013	12.305±1.235
60 %	0.527±0.039	0.535±0.031	6.633±1.839	12.21±1.989
80 %	0.422±0.049	0.35±0.055	9.734±2.092	11.475±2.754

Table 8 Classification results on the car dataset when missing data presents in the test set

r	accuracy		running time	
	CBMI (our)	IClf	CBMI (our)	IClf
20 %	0.768±0.012	0.758±0.012	6.278±1.198	11.405±1.771
40 %	0.707±0.019	0.694±0.020	6.564±1.147	11.673±1.593
60 %	0.678±0.020	0.670±0.040	6.450±1.160	11.846±2.382
80 %	0.682±0.015	0.678±0.015	5.835±1.524	9.369±1.681

Table 9 Classification results on the Iris dataset when the test input is fully observed

r	accuracy		running time	
	CBMI (our)	IClf	CBMI (our)	IClf
0 %	0.955±0.024	0.953±0.023	0.350±0.013	0.083±0.001
20 %	0.963±0.024	0.952±0.028	3.911±1.039	2.755±0.632
40 %	0.945±0.020	0.943±0.015	3.576±0.913	2.483±0.510
60 %	0.943±0.020	0.917±0.041	3.376±0.872	2.423±0.653
80 %	0.948±0.020	0.798±0.108	2.810±0.707	1.964±0.792

Table 10 Classification results on the liver dataset when the test input is fully observed

r	accuracy		running time	
	CBMI (our)	IClf	CBMI (our)	IClf
0 %	0.707±0.029	0.712±0.035	0.395±0.002	0.095±0.002
20 %	0.687±0.032	0.67±0.015	5.898±1.426	3.664±0.886
40 %	0.664±0.037	0.624±0.037	5.209±2.157	4.223±0.565
60 %	0.646±0.056	0.632±0.049	6.262±1.813	4.115±1.322
80 %	0.568±0.033	0.565±0.046	6.162±1.465	2.558±1.065

Table 11 Classification results on the soybean dataset when the test input is fully observed. Missing rate 80% gives some samples that are completely empty. So we discard this case.

r	accuracy		running time	
	CBMI (our)	IClf	CBMI (our)	IClf
0 %	0.995±0.016	1.0±0.0	0.709±0.052	0.116±0.014
20 %	0.995±0.016	0.995±0.016	57.483±14.496	63.033±8.178
40 %	0.989±0.021	0.963±0.062	50.181±15.803	43.733±9.249
60 %	0.995±0.016	0.937±0.061	27.561±11.688	25.348±9.366

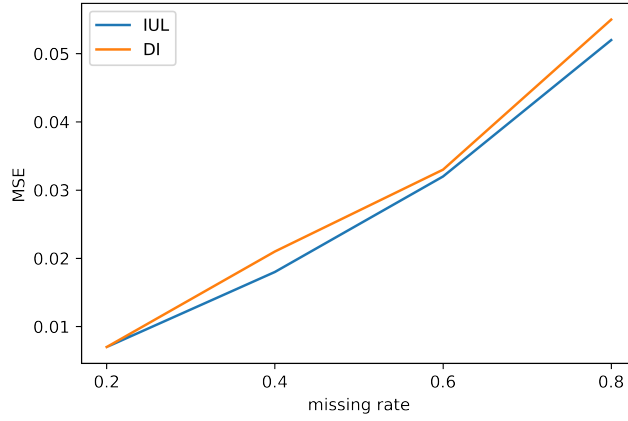


Fig. 7 MSE of IUL and DI under MICE imputation on the diabetes dataset.

Table 12 Classification results on the parkinson dataset when the test input is fully observed

r	accuracy		running time	
	CBMI (our)	IClf	CBMI (our)	IClf
0 %	0.904±0.037	0.895±0.045	0.368±0.006	0.092±0.003
20 %	0.863±0.041	0.871±0.05	16.72±3.905	15.741±4.267
40 %	0.865±0.026	0.862±0.033	18.888±4.622	13.52±1.636
60 %	0.873±0.032	0.847±0.045	16.582±5.786	10.715±3.629
80 %	0.79±0.071	0.799±0.05	19.093±5.147	13.792±5.38

Table 13 Classification results on the heart dataset when the test input is fully observed

r	accuracy		running time	
	CBMI (our)	IClf	CBMI (our)	IClf
0 %	0.82±0.017	0.818±0.023	0.375±0.004	0.101±0.001
20 %	0.813±0.04	0.808±0.039	38.631±8.816	35.641±5.884
40 %	0.819±0.031	0.815±0.039	46.706±9.855	41.273±6.889
60 %	0.825±0.032	0.811±0.032	44.298±8.777	42.332±8.138
80 %	0.777±0.04	0.78±0.038	45.646±7.224	37.522±7.628

accuracy between CBMI and IClf that is 1.1% – 1.8% – 3.7% – 4% – 8.5% when the missing rate is 0% – 20% – 40% – 60% – 80%, respectively.

Regarding the running time, CBMI has shorter running times than IClf when missing data is present in the test set, and sometimes may be up to twice as much. For example, in the glass dataset, at a missing rate of 60% (table 7), the average running time of CBMI and IClf is 6.633 and 12.21 respectively. As another example, in table 15, the differences in running times between CBMI and IClf are 0.329 – 1.193 – 1.806 – 0.798 – 1.445 with the corresponding missing rates 0% – 20% – 40% – 60% – 80%.

Table 14 Classification results on the glass dataset when the test input is fully observed

r	accuracy		running time	
	CBMI (our)	IClf	CBMI (our)	IClf
0 %	0.762±0.055	0.751±0.063	0.377±0.003	0.092±0.001
20 %	0.763±0.049	0.745±0.048	7.41±1.166	5.791±1.228
40 %	0.738±0.022	0.701±0.035	8.132±2.086	5.787±1.558
60 %	0.669±0.068	0.629±0.039	7.655±1.919	5.266±1.384
80 %	0.588±0.075	0.503±0.027	8.056±2.292	5.662±1.476

Table 15 Classification results on the car dataset when the test input is fully observed

r	accuracy		running time	
	CBMI (our)	IClf	CBMI (our)	IClf
0 %	0.962±0.009	0.959±0.012	0.431±0.012	0.112±0.001
20 %	0.892±0.01	0.867±0.018	7.202±0.988	6.009±1.498
40 %	0.828±0.023	0.793±0.02	7.282±1.258	5.476±0.675
60 %	0.79±0.018	0.73±0.01	6.557±1.273	5.759±1.34
80 %	0.72±0.035	0.716±0.012	6.326±1.813	4.881±0.911

In summary, CBMI illustrates significantly more promising results in terms of accuracy for imbalanced data. For the running time, CBMI is faster than IClf when missing data is present in the test set, while this may not be the case when the test data is fully observed.

5.5 Results on categorical data

The soybean dataset and the car dataset are categorical. From the results of these two datasets, we can see that CBMI gives higher accuracy in many cases, although the running time may be higher. For instance, in table 11 when the test set is fully observed and the missing rate in the training set is increasing, CBMI shows an extremely stable and high performance $0.995 - 0.995 - 0.989 - 0.995$, while the accuracy of IClf is decreasing $1.0 - 0.995 - 0.963 - 0.937$. In addition, when missing data is present in the test set (tables 4 and 8), CBMI shows higher accuracy and lower running time. Specifically, focus on running time in table 4, the discrepancy of CBMI compared to IClf is significant, CBMI gives the runtime in the range $[24.579; 29.487]$, while the interval of IClf's runtime is $[44.25; 48.91]$. Thus, CBMI can provide significantly better performance compared to IClf for categorical data.

5.6 Results on small sample data

CBMI shows a more promising performance compared to IClf when the ratio between the number of samples and the number of features is small. For example, for the heart dataset, the ratio between the number of samples and the number of features is $267 : 4 \approx 6.07$, CBMI gives an accuracy of 0.633 at missing rate 40% when missing data presents in the test set (table 7), while it is 0.598 for IClf. As another example,

for the soybean dataset, the ratio between the number of samples and the number of features is 47 : 35, CBMI also illustrates a prominent performance. For example, at missing rate 60%, when missing data is present in the test set (table 4), CBMI gives an accuracy of 0.821, while it is 0.8 for IClf. From these examples, we can see that CBMI illustrates significantly more promising results in terms of accuracy for small sample data.

6 Conclusion and Future Works

In conclusion, this study has presented two novel strategies, *IUL* and *CBMI*, to address the ubiquitous issue of missing data in supervised learning. These algorithms leverage the often overlooked labels in the training data, leading to a significant improvement in imputation/classification quality. Specifically, CBMI, which integrates the MissForest algorithm, enables simultaneous imputation of the training and testing label and input and is capable of handling data training with missing labels without prior imputation. Both IUL and CBMI have demonstrated their versatility by being applicable to continuous, categorical, or mixed-type data. Experimental results have shown promising improvements in mean squared error or classification accuracy, especially in dealing with imbalanced data, categorical data, and small sample data, indicating their potential to enhance data analysis and model performance.

In the future, we want to further explore the capability of CBMI for handling input with missing labels. In addition, I also want to study the application of CBMI in semi-supervised learning. Specifically, in semi-supervised learning, there are samples without labels. So, we can regard these samples as samples with missing labels. Moreover, we will investigate if similar strategies work for regression. Moreover, since CBMI is based on MissForest, an algorithm that is built upon Random Forests, it is also likely that CBMI also inherits the properties of Random Forest, such as robustness to noise, the ability to handle outliers, and multicollinearity. Therefore, in the future, we will conduct experiments to confirm whether CBMI also inherits these qualities.

Appendix A Proof of theorem 1

Theorem 1. Assume that we have the data $\mathcal{D} = (\mathbf{I}, \mathbf{y})$ of n samples. Here, $\mathbf{I} = (\mathbf{x}, \mathbf{z})$, $\mathbf{x} \in \mathbf{R}$ contains missing values, $\mathbf{z} \in \mathbf{R}$ is fully observed and $\mathbf{y}$ is the label. For MICE imputation, with the IUL strategy, we construct a model

$$\hat{x} = \hat{\gamma}_o + \hat{\gamma}_1 z + \hat{\gamma}_2 y. \quad (\text{A1})$$

Meanwhile, DI ignores label $\mathbf{y}$ and use the following model

$$\hat{x}' = \hat{\gamma}_o + \hat{\gamma}_1 z. \quad (\text{A2})$$

Without loss of generality, assume that $y_i \geq 0, i = 1, 2, \dots, n$ for label encoding. Least square method yields the following estimates for the coefficients,

$$\begin{aligned}\hat{\gamma}_1 &= \frac{(\sum y^2)(\sum xz) - (\sum zy)(\sum yx)}{(\sum z^2)(\sum y^2) - (\sum zy)^2}, \\ \hat{\gamma}_2 &= \frac{(\sum z^2)(\sum yz) - (\sum zy)(\sum zx)}{(\sum z^2)(\sum y^2) - (\sum zy)^2}, \\ \hat{\gamma}_o &= \mu_{\mathbf{x}} - \hat{\gamma}_1\mu_{\mathbf{z}} - \hat{\gamma}_2\mu_{\mathbf{y}}.\end{aligned}$$

Let denote

$$\mathcal{E}_i = \hat{\gamma}_2^2(y_i - 2\mu_{\mathbf{y}}) + 2\hat{\gamma}_1\hat{\gamma}_2(z_i - \mu_{\mathbf{z}}), \quad i = 1, 2, \dots, n, \quad (\text{A3})$$

$$\mathcal{I}^+ = \{i : \mathcal{E}_i \geq 0\}, \quad \mathcal{V}^+ = \sum_{i \in \mathcal{I}^+} y_i \mathcal{E}_i, \quad (\text{A4})$$

$$\mathcal{I}^- = \{i : \mathcal{E}_i < 0\}, \quad \mathcal{V}^- = \sum_{i \in \mathcal{I}^-} y_i \mathcal{E}_i. \quad (\text{A5})$$

Then, $SSE_{IUL} \leq SSE_{DI}$ if and only if $\mathcal{V}^+ + \mathcal{V}^- \geq 0$, where SSE_{IUL}, SSE_{DI} are the Sum of Square Error for the model based on IUL and DI, respectively.

Proof. We first recall in linear regression, $SSE = SST - SSR$ where SST is the Sum of Squares Total and SSR is the Sum of Squares due to Regression. For each model relies on IUL and DI we have

$$SSE_{IUL} = \sum_{i=1}^n (\hat{x}_i - x_i)^2 = \sum_{i=1}^n (x_i - \mu_{\mathbf{x}})^2 - \sum_{i=1}^n (\hat{x}_i - \mu_{\mathbf{x}})^2, \quad (\text{A6})$$

$$SSE_{DI} = \sum_{i=1}^n (\hat{x}'_i - x_i)^2 = \sum_{i=1}^n (x_i - \mu_{\mathbf{x}})^2 - \sum_{i=1}^n (\hat{x}'_i - \mu_{\mathbf{x}})^2. \quad (\text{A7})$$

Then,

$$\begin{aligned}SSE_{DI} - SSE_{IUL} &= \sum_{i=1}^n (\hat{x}_i - \mu_{\mathbf{x}})^2 - \sum_{i=1}^n (\hat{x}'_i - \mu_{\mathbf{x}})^2 \\ &= \sum_{i=1}^n (\hat{x}_i - \hat{x}'_i + \hat{x}'_i - \mu_{\mathbf{x}})^2 - \sum_{i=1}^n (\hat{x}'_i - \mu_{\mathbf{x}})^2 \\ &= \sum_{i=1}^n (\hat{x}_i - \hat{x}'_i)^2 + 2 \sum_{i=1}^n (\hat{x}_i - \hat{x}'_i)(\hat{x}'_i - \mu_{\mathbf{x}}) \\ &= \sum_{i=1}^n (\hat{\gamma}_2 y_i)^2 + 2 \sum_{i=1}^n \hat{\gamma}_2 y_i (\hat{\gamma}_1 z_i + \hat{\gamma}_o - \mu_{\mathbf{x}}) \\ &= \sum_{i=1}^n \hat{\gamma}_2 y_i (\hat{\gamma}_2 y_i + 2(\hat{\gamma}_1 z_i + \hat{\gamma}_o - \mu_{\mathbf{x}}))\end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^n \hat{\gamma}_2 y_i (\hat{\gamma}_2 y_i + 2(\hat{\gamma}_1 z_i - \hat{\gamma}_1 \mu_{\mathbf{z}} - \hat{\gamma}_2 \mu_{\mathbf{y}})) \\
&= \sum_{i=1}^n y_i (\hat{\gamma}_2^2 (y_i - 2\mu_{\mathbf{y}}) + 2\hat{\gamma}_1 \hat{\gamma}_2 (z_i - \mu_{\mathbf{z}})) \\
&= \sum_{i=1}^n y_i \mathcal{E}_i \\
&= \sum_{i \in \mathcal{I}^+} y_i \mathcal{E}_i + \sum_{i \in \mathcal{I}^-} y_i \mathcal{E}_i \\
&= \mathcal{V}^+ + \mathcal{V}^-
\end{aligned}$$

Therefore, $SSE_{IUL} \leq SSE_{DI}$ if and only if $\mathcal{V}^+ + \mathcal{V}^- \geq 0$.

References

- Breiman, L. (2001). Random forests. *Machine learning*, 45, 5–32,
- Breiman, L. (2017). *Classification and regression trees*. Routledge.
- Buuren, S.v., & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 1–68,
- Candès, E.J., & Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6), 717,
- Choudhury, S.J., & Pal, N.R. (2019). Imputation of missing data with neural networks for classification. *Knowledge-Based Systems*, 182, 104838,
- Do, T.T., Vu, M.A., Ly, H.T., Nguyen, T., Hicks, S.A., Riegler, M.A., ... Nguyen, B.T. (2023). Blockwise principal component analysis for monotone missing data imputation and dimensionality reduction. *arXiv preprint arXiv:2305.06042*, ,
- Dua, D., & Graff, C. (2017). *UCI machine learning repository*. Retrieved from <http://archive.ics.uci.edu/ml>
- Fan, J., Zhang, Y., Udell, M. ((2020)). Polynomial matrix completion for missing data imputation and transductive learning. *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 3842–3849).

- Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232,
- Garg, A., Naryani, D., Aggarwal, G., Aggarwal, S. ((2018)). Dl-gsa: a deep learning metaheuristic approach to missing data imputation. *International conference on sensing and imaging* (pp. 513–521).
- Geurts, P., Ernst, D., Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63, 3–42,
- Ghazi, M.M., Nielsen, M., Pai, A., Cardoso, M.J., Modat, M., Ourselin, S., Sørensen, L. (2018). Robust training of recurrent neural networks to handle missing data for disease progression modeling. *arXiv preprint arXiv:1808.05500*, ,
- Gondara, L., & Wang, K. (2017). Multiple imputation using deep denoising autoencoders. *arXiv preprint arXiv:1705.02737*, ,
- Hastie, T., Mazumder, R., Lee, J.D., Zadeh, R. (2015). Matrix completion and low-rank svd via fast alternating least squares. *The Journal of Machine Learning Research*, 16(1), 3367–3402,
- Keerin, P., Kurutach, W., Boongoen, T. (2013). An improvement of missing value imputation in dna microarray data using cluster-based lls method. *2013 13th international symposium on communications and information technologies (iscit)* (pp. 559–564).
- Khan, S.I., & Hoque, A.S.M.L. (2020). Sice: an improved missing data imputation technique. *Journal of big data*, 7(1), 1–21,
- M Mostafa, S., S Eladimy, A., Hamad, S., Amano, H. (2020). Cbri and cbrc: Novel algorithms for improving missing value imputation accuracy based on bayesian ridge regression. *Symmetry*, 12(10), 1594,
- Mohan, K., & Pearl, J. (2021). Graphical models for processing missing data. *Journal of the American Statistical Association*, 1–42,
- Murray, J.S., & Reiter, J.P. (2016). Multiple imputation of missing categorical and continuous values via bayesian mixture models with local dependence. *Journal of the American Statistical Association*, 111(516), 1466–1479,

- Nguyen, P., Tran, L.G., Le, B.H., Nguyen, T.H., Nguyen, T., Nguyen, H.D., Nguyen, B.T. (2023). Faster imputation using singular value decomposition for sparse data. *Asian conference on intelligent information and database systems* (pp. 135–146).
- Nguyen, T., Ly, H.T., Riegler, M.A., Halvorsen, P., Hammer, H.L. (2023). Principal components analysis based frameworks for efficient missing data imputation algorithms. *Asian conference on intelligent information and database systems* (pp. 254–266).
- Nguyen, T.H., Le, B., Nguyen, P., Tran, L.G., Nguyen, T., Nguyen, B.T. (2023). Principal components analysis based imputation for logistic regression. *International conference on industrial, engineering and other applications of applied intelligent systems* (pp. 28–36).
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830,
- Pham, N.-H., Vo, K.-L., Vu, M.A., Nguyen, T., Riegler, M.A., Halvorsen, P., Nguyen, B.T. (2023). Correlation visualization under missing values: a comparison between imputation and direct parameter estimation methods. *arXiv preprint arXiv:2305.06044*, ,
- Rahman, M.G., & Islam, M.Z. (2016). Missing value imputation using a fuzzy clustering-based em approach. *Knowledge and Information Systems*, 46(2), 389–422,
- Stekhoven, D.J., & Bühlmann, P. (2012). Missforest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112–118,
- Vu, M.A., Nguyen, T., Do, T.T., Phan, N., Halvorsen, P., Riegler, M.A., Nguyen, B.T. (2023). Conditional expectation for missing data imputation. *arXiv preprint arXiv:2302.00911*, ,