BIMICE: Bayesian Influenced Multiple Imputations by Chained Equations

Yoav Zelinger
Ben-Gurion University of the Negev
Beer-Sheva, Israel
yoavzel@post.bgu.ac.il

Yehonatan Kidushim
Ben-Gurion University of the Negev
Beer-Sheva, Israel
kidushim@post.bgu.ac.il

Tiltan Doron Gilat
Ben-Gurion University of the Negev
Beer-Sheva, Israel
tiltang@post.bgu.ac.il

Gil Ari Agmon
Ben-Gurion University of the Negev
Beer-Sheva, Israel
gilagm@post.bgu.ac.il

Shir Rozenfeld
Ben-Gurion University of the Negev
Beer-Sheva, Israel
shirmord@post.bgu.ac.il

Abstract

Handling missing data is a fundamental challenge in data preprocessing, directly impacting the accuracy and reliability of machine learning models. Traditional imputation methods, such as mean substitution and the Multiple Imputation by Chained Equations (MICE), often fail to capture intricate variable dependencies, leading to inefficiencies and potential biases. This paper introduces Bayesian Influenced Multiple Imputations by Chained Equations (BIMICE), an enhanced framework that leverages Bayesian networks to optimize variable ordering and predictor selection during the imputation process. By prioritizing relevant predictors and respecting probabilistic dependencies, BIMICE reduces computational overhead while maintaining high imputation accuracy. Empirical evaluations on benchmark datasets demonstrate the method's efficiency and robustness, achieving superior Mean Absolute per Predictor Error (MAPE) under varying levels of missingness compared to baseline methods. This work highlights the potential of integrating probabilistic graphical models into traditional imputation techniques, providing a pathway for future innovations in data preprocessing.

Keywords

Missing data handling, Data imputation, Data preprocessing, Feature selection, Bayesian networks, Multiple imputations, Chained equations

ACM Reference Format:
Yoav Zelinger, Yehonatan Kidushim, Tiltan Doron Gilat, Gil Ari Agmon, and Shir Rozenfeld. 2025. BIMICE: Bayesian Influenced Multiple Imputations by Chained Equations. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 10 pages.

1 Introduction

In the era of data-driven decision-making, the integrity of datasets is crucial for deriving accurate and reliable insights [4]. However, missing values are a ubiquitous challenge across domains, ranging from healthcare and finance to social sciences. Incomplete data not only compromises statistical analyses but also impairs the performance of machine learning models by introducing bias and reducing predictive accuracy [11]. Addressing this issue requires robust and efficient data imputation techniques.

While traditional imputation methods such as mean or median substitution are computationally inexpensive, they often fail to capture the underlying relationships between variables, leading to biased results [1]. Advanced techniques, including Multiple Imputation by Chained Equations (MICE), offer a more sophisticated approach by iteratively modeling missing values based on other observed variables. MICE operates by creating multiple imputed datasets through a series of conditional models, where each variable with missing data is predicted in turn using other variables as predictors. The results are then pooled to produce estimates that account for uncertainty in the imputations. This iterative and probabilistic framework allows MICE to outperform simpler methods, particularly in datasets with complex relationships.

Despite its advantages, MICE presents challenges, such as computational inefficiency when dealing with large-scale or high-dimensional data, as well as potential difficulties in modeling complex variable dependencies. To address these limitations, this study proposes Bayesian Influenced Multiple Imputations by Chained Equations (BIMICE), which enhances MICE by integrating Bayesian networks to optimize variable ordering and predictor selection. This approach not only improves computational efficiency but also maintains high accuracy by leveraging probabilistic dependencies between variables, making it particularly suitable for large and complex datasets.

2 Related Work

2.1 Regression Models for Data Imputation

Regression models provide a structured approach to estimating missing values by utilizing observed variable relationships.

An example is the Iterative Stepwise Regression Imputation [13], which proposes a single imputation and prioritizes robustness to outliers and model deviations to enhance stability in extreme datasets. This approach attempts to include only the most relevant predictors through automatic variable selection based on statistical criteria, such as AIC or p-values. Although it improves computational complexity, it ignores the variability inherent in imputation, which can lead to biased conclusions. In contrast, MICE employs

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), A. El Fallah Seghrouchni, Y. Vorobeychik, S. Das, A. Nowe (eds.), May 19 – 23, 2025, Detroit, Michigan, USA.

multiple imputation, offering a richer representation of missing data variability and producing statistically robust results.

Some other imputation techniques, such as Imputation VIA Clusterwise Linear Regression (IVIACLR) [8], utilize Clusterwise Linear Regression (CLR), which uses similar data points for regression of incomplete objects. Each cluster is associated with a regression function that minimizes its regression error. Data points are assigned to the cluster that minimize their error. The process is repeated until the missing values changes are minimal. While IVIACLR outperforms MICE on datasets with a small amount of missing values, MICE demonstrates superior performance when dealing with datasets containing a large proportion of missing values.

2.2 Bayesian networks for Imputation

BN-K2Iχ^2 and 1BN-K2Iχ^2 [6] are methods for imputing missing values using Bayesian networks, designed to preserve causal relationships between variables. These methods rely on conditional probabilities learned from complete data: the former building separate networks for each missing variable and the latter constructing a single network for all missing variables. Compared to traditional approaches, like EM and Decision Trees, they offer advantages in maintaining relationships, reducing bias, and improving imputation. However, the inclusion of indirect relationships may reduce accuracy compared to MICE, which focuses on direct relationships.

2.3 Reuse of Imputed Data

Some researchers have explored usages of imputed values for subsequent imputations. One approach, Sequential K-Nearest Neighbor (SKNN) [10], enhances the K-Nearest Neighbor (KNN) algorithm. Unlike traditional KNN, which predicts a data point's value based on its nearest neighbors from complete data, SKNN sequentially imputs missing values and incorporates previously imputed values. While SKNN outperforms traditional KNN in accuracy, it carries the risk of error propagation, where early errors can be carried forward and reduce accuracy, whereas MICE imputes missing values iteratively for each variable, ensuring that errors do not propagate.

2.4 Different Applications of MICE

Multivariate Imputation by Chained Equations (MICE) has been adapted and extended to improve performance and address challenges in imputing missing data in complex datasets.

Costantini et al. [3] tackled the "many variables" problem in MICE by integrating Principal Component Regression (PCR) to reduce predictor dimensionality. Their study, which included simulations and a real-world case, demonstrated that PCR-based MICE (MI-PCR) enhances both computational efficiency and imputation accuracy for high-dimensional datasets. The findings underscore the trade-offs between imputation performance and computational cost, providing practical guidance for applications in disciplines such as social science and psychology. Our approach tackles the "many variables" problem by excluding variables that are not correlated with the target variable.

Similarly, Shah et al. [12] evaluated a Random Forest-based adaptation of MICE and compared its performance to traditional parametric MICE. Their simulation studies on datasets with nonlinear relationships revealed that the Random Forest-based approach yielded more accurate imputations, particularly for complex data, by reducing bias and enhancing efficiency. However, this method leads to higher computational costs, especially for large-scale datasets. Our proposed method aims to reduce computational cost by retaining traditional parametric MICE models while decreasing the number of times those models are run.

Another enhancement of MICE includes Single Center Imputation from Multiple Chained Equation (SICE) [9], which extends the MICE algorithm by combining single and multiple imputation methods to handle categorical and numerical data. It iteratively refines imputed values, selecting the most frequent for categorical data or the mean for numerical data, achieving a 20% F-measure improvement for binary data and 11% error reduction for numerical data. However, SICE does not utilize the relationships between variables, which could significantly reduce computational time. Incorporating feature relationships into our proposed method by using smaller subsets of features for each imputation will enhance efficiency.

3 Background

3.1 The Importance of Data Preprocessing

Raw data often contains inconsistencies, redundancies, and missing values, which can hinder the learning process and lead to suboptimal results. Data preprocessing [5] is therefore an essential step in the machine learning pipeline, aimed at transforming raw data into a clean, structured, and meaningful format. Preprocessing enhances model performance, accuracy, and generalization, ensuring reliable and accurate analysis. Typical data preprocessing tasks include data cleaning, normalization, encoding categorical variables, handling missing values, and feature scaling, all aimed at making the data suitable for machine learning algorithms.

3.2 Handling Missing Values with Data Imputation

Missing values are a common challenge in data preprocessing, as they can disort analyses and compromise the accuracy of predictive models. While straightforward methods, such as deleting incomplete rows or columns may suffice in some cases, they often lead to loss of valuable information, making more sophisticated approaches necessary. This is where data imputation comes into play, as it provides a more robust solution. Data imputation [7] involves filling in missing values using statistical or machine learning methods, ensuring the dataset remains complete and usable for analysis or modeling. Approaches to imputation range from basic techniques, such as replacing missing values with the mean, median, or mode, to advanced methods like k-nearest neighbors, random forests, deep learning models, and generative approaches.

3.3 Improving MICE: A Focused Approach to Efficiency

The MICE [2] method addresses missing data by generating multiple complete datasets, accounting for statistical uncertainty, and improving the reliability of analyses compared to single imputation. MICE is highly flexible and suitable for various types of variables and complexities such as bounded ranges or skip patterns in surveys. The method assumes that missing data are Missing At Random

(MAR) and uses regression-based iterations where each missing variable is imputed conditionally based on other variables. In our research, we propose an approach focused on improving process efficiency by using only variables correlated with the target variable, enabling higher accuracy and reducing model complexity. Compared to the traditional MICE, in *BIMICE* using causal relationships between variables learned from the Bayesian network, in each iteration the features with missing values are imputed using only the features correlated to it. In addition, learning the imputation order from the Bayesian network, multiple imputations are redundant and only a single dataset imputation is required.

4 Methodology

This study presents a methodology for improving data imputation by integrating Bayesian networks into the Multiple Imputation by Chained Equations (MICE) framework. Traditional imputation methods often face challenges such as inefficient variable selection, high computational demands, and potential inaccuracies. Using the inherent structure of Bayesian networks, which provide a structured representation of variable dependencies, the methodology addresses these challenges and enables a more systematic, accurate, and efficient approach to imputation. Furthermore, this integration improves the interpretability of the imputation process by focusing on relationships and dependencies within the data.
To enhance the accuracy and efficiency of the traditional MICE algorithm, we incorporate a Bayesian network approach.

Definition 4.1 *(Bayesian Networks, BNs).* A Bayesian network (BN) is a directed acyclic graph (DAG) where the set of nodes represents variables, and directed edges represent conditional dependencies between these variables.

Utilizing a Bayesian network provides three key advantages in the imputation process:

(1) **Guiding Variable Order** is critical to achieve accurate imputation.

Definition 4.2 *(Variable Ordering).* Variable Ordering is the prioritized sequence in which variables are processed during regression-based imputation. This order is determined based on the Bayesian network, which identifies the probabilistic dependencies and relationships among variables.

In traditional MICE frameworks, variables are processed sequentially without explicit prioritization, often leading to suboptimal results. In contrast, Bayesian networks inherently capture the conditional dependencies between variables, enabling the prioritization of variables based on their predictive significance. Our approach ensures that variables with greater predictive power are imputed first, thereby enhancing the stability and accuracy of the overall process. The structure of the network determines the imputation sequence, prioritizing variables with stronger predictive dependencies. For example, in Figure 1, missing values are introduced in the totChol and sysBP columns. The Bayesian network (BN) structure derived from this dataset may reveal that Age influences both sysBP and totChol, BMI affects sysBP, and cigsPerDay impacts totChol. This structure guides the variables ordering for imputation, such as Age, BMI, cigsPerDay, totChol, and sysBP.

(2) **Focusing on Relevant Predictors** ensures accuracy by focusing on directly related variables.

Definition 4.3 *(Conditional Independence).* Conditional Independence is a concept in Bayesian networks where two variables are independent given the knowledge of a third variable.

A common limitation of traditional imputation methods is their reliance on all available variables as predictors, which often leads to increased noise, overfitting, and computational inefficiency. The Bayesian network structure addresses this issue by modeling conditional independencies and selecting only the most relevant predictors for each target variable. This streamlined approach reduces variable redundancy and improves both the interpretability and the performance of the imputation model. In Figure 1, the Bayesian network highlights that sysBP is influenced by Age and BMI, while totChol is influenced by Age and cigsPerDay. By focusing solely on the variables directly related to each missing value, the network reduces the number of required predictors, enhancing the regression process efficiency by excluding irrelevant features.

(3) **Improving Computational Efficiency**: Enhancing the selection and ordering of predictors in regression significantly reduces computational overhead and improves time efficiency compared to the traditional MICE method. For the given example in figure 1, the regression equations would be:

$$\text{sysBP} = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{BMI}$$
$$\text{totChol} = \alpha_0 + \alpha_1 \cdot \text{Age} + \alpha_2 \cdot \text{cigsPerDay}$$

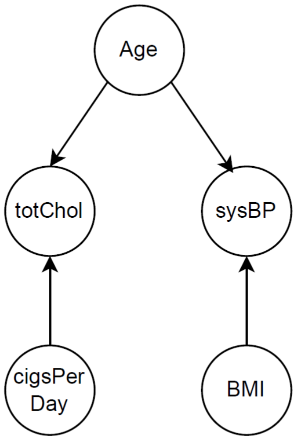

Figure 1: A Bayesian network constructed from a subset of the Framingham Heart Study dataset. The features include Age, Total Cholesterol (totChol), Systolic Blood Pressure (sysBP), Body Mass Index (BMI), the number of cigarettes smoked per day (cigsPerDay) and it illustrates the dependencies between these variables.

The BIMICE algorithm with Bayesian networks for Missing Data Imputation is detailed in **Algorithm 1**. It begins by defining the input requirements: the dataset D with features F, the number of iterations E and the convergence criterion ϵ (line 1). The algorithm constructs a Bayesian network BN to capture probabilistic dependencies among features (line 3). Features with missing values are identified and stored in F_{miss}, ensuring that only these features are targeted for imputation (line 4). An optimal variable ordering $O = \langle v_1, \ldots, v_n \rangle$ is then determined using BN (using topological sorting), ensuring imputation respects dependencies between variables (line 5). A duplicated dataset D^* is created to hold to resulted dataset (line 6) and initializes missing values with basic estimates to serve as a starting point (line 7). Within each iteration, the algorithm refines the imputed values over E iterations (line 8). For each variable v_j in O, if v_j has missing values (line 9-10), the algorithm extracts predictors $P(v_j)$ using BN (line 11), fits a regression model R_j based on these predictors (line 12). Using R_j, the imputed missing values of v_j are updated in D^*(line 13). Convergence is checked by comparing the imputed values in the current and previous iterations.

Definition 4.4 *(Imputation Convergence).* The stabilization of imputed values during iterations, where further changes become negligible.

Variables meeting the convergence criterion are removed from F_{miss}, indicating stability (lines 14-15). After completing all iterations, the imputed dataset is returned (line 16). By integrating Bayesian networks with MICE, this algorithm effectively and accurately handles missing data.

Algorithm 1: BIMICE

```
1  Dataset $D$ with features $F = \{f_1, \ldots, f_n\}$, number of
   iterations $E$, convergence criterion $\epsilon$
2  begin
3      Construct Bayesian network $BN$ based on $D$;
4      Store features with missing values in $F_{\text{miss}}$;
5      Determine variable ordering $O = \langle v_1, \ldots, v_n \rangle$ using $BN$;
6      Create $D^*$ a copy of dataset $D$;
7      Initialize missing values in $D^*$ with basic estimates;
8      for $e \leftarrow 1$ to $E$ do
9          foreach $v_j \in O$ do
10             if $v_j \in F_{miss}$ then
11                 Extract predictors $P(v_j)$ for $v_j$ from $BN$;
12                 Fit a regression model $R_j$ using $P(v_j)$;
13                 Update missing values of $v_j$ in $D^*$ using $R_j$;
14             if $|D_e^*(v_j) - D_{e-1}^*(v_j)| < \epsilon$ then
15                 Remove $v_j$ from $F_{\text{miss}}$;
16     Return $D^*$;
```

5 Empirical Evaluation

Data imputation is a critical challenge in data preprocessing, as incomplete datasets can compromise both analysis and model accuracy. To address this issue, we propose BIMICE, a new method based on the MICE framework that utilizes variable dependencies for predictor selection and imputation sequencing. BIMICE aims to reduce redundancy, improve computational efficiency, and enhance imputation accuracy while addressing limitations inherent in traditional methods.

To evaluate the proposed approach, we investigate the following research questions:

RQ1 How does manipulating variable ordering in MICE improve the data imputation performance?

RQ2 How does leveraging variable dependencies improve MICE performance for data imputation?

RQ3 Can BIMICE outperform traditional techniques in terms of accuracy and efficiency?

RQ4 How effectively does BIMICE address varying levels of missing data?

The performance of BIMICE is evaluated using the mean absolute error (MAE) and mean squared error (MSE) across varying levels of missing data injection and initial imputation techniques. Those metrics measure predictive accuracy. For all metrics, a lower value indicates better model performance.

We aim to measure the efficiency of the model using only the relevant predictors for data imputation. To better understand the error distribution relative to the complexity of the model, we introduce the *Mean Absolute per Predictor Error (MAPE)* metric, which normalizes the Mean Absolute Error (MAE) based on the number of **unused** predictors in the imputation of each feature (Equation 1). Since a feature cannot be used as a predictor for its own imputation, for each feature $f \in F$, the unused predictors defined such that $Predicators^f \cup (Predicators^f)^C = F \setminus \{f\}$:

$$(Predicators^f)^C = \{g | g \in F \wedge g \notin Predicators^f \wedge g \neq f\} \quad (1)$$

Then for each feature $f \in F$ we define it's MAPE score as the division between it's MAE and the number of unused predictors:

$$MAPE^f = \frac{MAE^f}{max(|(Predicators^f)^C|, 1)} \quad (2)$$

The overall efficiency of the model is calculated with the average $MAPE$ among all the features (F):

$$MAPE = \frac{1}{|F|} \sum_{f \in F} MAPE_f \quad (3)$$

Similarly to the previous metrics, a lower value indicates greater efficiency.

5.1 Experimental Setup

The overall design of our imputation and evaluation system is shown in Fig. 2.

The imputation and evaluation process follows these steps:

(1) **Dataset preprocessing:** The process begins with a base dataset. Preprocessing involves removing records with missing values to ensure that imputation performance can be compared against known true values for evaluation. Additionally, we remove from the datasets features that contain non-numeric data because they are not supported in the original MICE algorithm. In addition, we removed from the

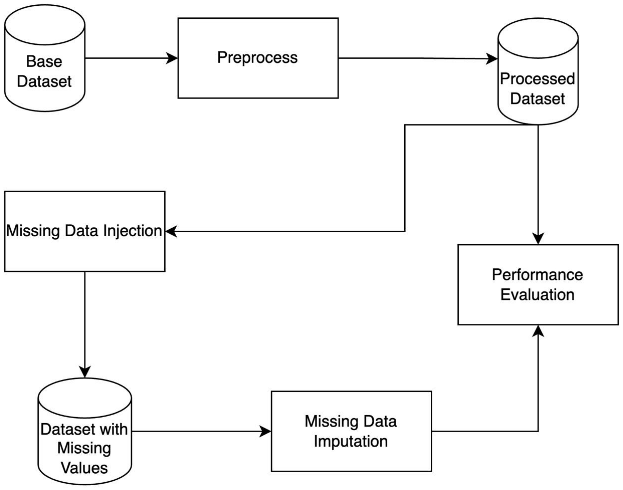

Figure 2: Block diagram of the experimental design.

datasets the target column since our goal is not to predict the label, but to impute missing values in the dataset's features. The resulting dataset, referred to as the *Processed Dataset*, is used to evaluate the performance of imputation algorithms. Splitting the dataset to train and test is not required here since all the compared algorithms are lazy learning algorithms.

(2) **Missing Data Injection:** Artificially injecting missing values into the dataset. To ensure the proposed algorithm is not biased, missing data is systematically simulated by removing a specified proportion (ranging from 10% to 40%) of random values from different number of features (ranging from one feature with missing values to all features). Each resulting dataset (with different amount of missing values) is referred to as the *Dataset with Missing Values*. This enables a comprehensive evaluation of our method's performance under varying conditions.

(3) **Missing Data Imputation:** Missing values are imputed using the different variations of BIMICE on the injected missing values. In addition, For comparison and evaluation, the missing values were imputed using 3 missing values imputation algorithms such as the original MICE, SICE and KNN. Each imputed dataset is referred to as the *Dataset With Imputations*.

(4) **Performance evaluation:** The performance of the imputation algorithms is evaluated using standard metrics for numerical imputations such as MAE, MSE and our proposed metric Mean Absolute per Predictor Error (MAPE). The metrics were computed between the *Processed Dataset* and each *Dataset With Imputations*.

In addition, we experimented the different MICE methods (MICE, SICE and our different variations of *BIMICE*) with different hyperparameters to get the best algorithm. We checked the different algorithms with different number of iterations, using 5, 10, and 15 iterations of the run. For better comparison, we explored MICE and SICE with different number of imputations, and took the ones that provided the best results.

5.1.1 Performance Comparisons:

The performance of BIMICE is evaluated using MAE, MSE and MAPE.

Paired t-test for normally distributed data is applied to determine whether observed performance differences are statistically significant at a significance level of $\alpha = 0.05$. The test evaluates whether the differences between our approach (BIMICE) and the baseline algorithms (MICE, KNN, and SICE) are meaningful and not attributable to random variation.

5.1.2 Ablation Study:

An ablation study is conducted to isolate and evaluate the individual contributions of two key enhancements introduced in the approach. Each variant is compared directly to the different imputation algorithms to assess its impact on imputation performance. The study is structured as follows:

- **Ablation 1: Variable ordering.** This experiment examines the effect of introducing variable ordering based on the Bayesian network structure. The regression process remains unchanged, using all available variables as in standard MICE. In the results, this ablation will be marked as $BIMICE^1$.
- **Ablation 2: Subset selection for regression.** This experiment examines the effect of restricting the regression process to use only the subset of correlated variables identified by the Bayesian network, while maintaining the default variable ordering of MICE. In the results, this ablation will be marked as $BIMICE^2$.

The ablation experiments aim to objectively evaluate the individual contributions of variable ordering and subset selection for regression. Isolating these enhancements enables a clearer understanding of their specific roles in influencing imputation performance. Compared to our full algorithm, the ablation tests do not provide the assurance that all predictors of a feature have been imputed prior to imputing its values. Therefore, similarly to the original MICE algorithm, in each ablation test we fill the missing values with imputation place holders with the average of the feature.

5.1.3 Datasets:

The evaluation of the proposed method is performed using two publicly available datasets from the Kaggle platform:

(1) **Framingham Heart Study Dataset:** This dataset is used to predict whether a patient has a 10-year risk of developing coronary heart disease (CHD). It is publicly available on Kaggle[1]. The dataset originally contains 4,240 records with 15 feature columns and one binary target column. The features include a mix of ordinal, numeric, and binary data types. After removing irrelevant data during preprocessing (as described earlier), the dataset consists of 3,671 records and 9 features.

(2) **Financial Risk Dataset:** This dataset, available on Kaggle[2], contains data on individual financial profiles, including demographic, financial and behavioral attributes. It initially consists of 15,000 records with 19 features, comprising both

[1] https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset/data

[2] https://www.kaggle.com/datasets/preethamgouda/financial-risk/data

numeric and categorical data types, along with a categorical target column representing the risk rating. After preprocessing, the dataset comprises 7,839 records and 7 numeric features.

5.1.4 Implementation:

All tests were conducted on a computer with an Intel Core i7-1370P CPU and 64GB RAM.

For our benchmarking of baseline missing value imputation methods, we employed the open-source Python package **Reparo**[3]. Reparo offers a robust suite of state-of-the-art algorithms designed specifically for handling missing data. Its availability not only facilitated the implementation of various imputation methods but also ensured the standardization and reproducibility of our experimental evaluations. This consistency was critical for drawing meaningful comparisons between the baseline methods and our proposed approach. For the initialization of the Bayesian network, we relied on **causalnex**[4], a comprehensive Python library that supports learning of causal relationships between variables and inference within Bayesian networks. This tool allowed us to efficiently design and implement the probabilistic graphical model used in our approach. Additionally, we used the **graphlib**[5] package for performing topological sorting, which was essential for managing the dependencies and ensuring that the features were imputed in the required order.

We employed the widely known **scikit-learn** library, using its linear regression model. Additionally, we utilized its well-recognized evaluation metrics, including Mean Absolute Error (MAE) and Mean Squared Error (MSE), to assess the performance of the imputation methods.

To promote transparency and reproducibility in research, the complete code for our algorithms and experiments has been made publicly available on GitHub[6]. This repository provides access to all implementation details, allowing other researchers to replicate our results and further explore the effectiveness of our proposed methodologies.

5.2 Results

Tables 1 and 2 provide a detailed summary of the results obtained for the Framingham and financial-risk datasets, respectively. In this study, we have chosen to include results for 5 and 15 iterations, as the outcomes for 10 iterations exhibit the same trends and do not offer additional insights. This selection ensures a concise yet comprehensive presentation of the findings. The first three rows in each table represent the baseline results, while the last three rows correspond to the variations of our proposed algorithm. These include two ablation tests and the full variation of the algorithm. The ablation tests, denoted as *BIMICE*^{1} and *BIMICE*^{2}, are specifically designed to address research questions **RQ1** and **RQ2**, respectively. The MAPE columns directly address research question **RQ3**, providing an evaluation of the efficiency of our approach relative to the baseline methods. For clarity, the best-performing values in each category are highlighted in bold.

	5 iterations			15 iterations		
	MSE	MAE	MAPE	MSE	MAE	MAPE
KNN	116.49	3.05	3.05	113.25	3.05	3.05
SICE	2765.73	18.70	18.70	2762.17	18.68	18.68
MICE	**85.51**	**2.51**	2.51	**85.18**	**2.51**	2.51
BIMICE^1	98.25	2.79	2.79	97.78	2.79	2.79
BIMICE^2	87.05	2.58	**0.98**	86.66	2.58	**0.98**
BIMICE	87.05	2.58	**0.98**	86.66	2.58	**0.98**

Table 1: Summary of Framingham dataset results, with 5 and 15 iterations of the algorithm, aggregated over the different missing values severities.

Based on the results summarized in Table 1, which focuses on the Framingham dataset, it can be observed that the MICE method consistently achieved the best results in terms of accuracy (MSE and MAE), while *BIMICE* and *BIMICE*^{2} demonstrated better performance in terms of efficiency (MAPE).

The results of Framingham dataset reveal differing trends between accuracy (MSE and MAE) and efficiency (MAPE). In terms of accuracy, MICE demonstrated significantly better performance compared to all other algorithms, including both baseline methods and the *BIMICE* variations. Conversely, in terms of efficiency (MAPE), *BIMICE* and *BIMICE*^{2} showed a statistically significant improvement over all other baseline algorithms.

All the algorithms demonstrated significantly larger errors when applied to the financial-risk dataset, particularly in terms of MSE. To better illustrate the results, the MSE values in this table are presented on a scale of 10^6.

	5 iterations			15 iterations		
	MSE·*e*6	MAE	MAPE	MSE·*e*6	MAE	MAPE
KNN	331	4086	4086	364	4214	4214
SICE	1366	9144	9144	1370	9215	9215
MICE	267	3770	3770	269	3802	3802
BIMICE^1	**266**	**3765**	3765	**268**	**3799**	3799
BIMICE^2	**266**	3767	**2887**	269	**3799**	**2928**
BIMICE	**266**	3767	**2887**	269	**3799**	**2928**

Table 2: Summary of financial-risk dataset results, with 5 and 15 iterations of the algorithm, aggregated over the different missing values severities.

Examining the results of the financial-risk dataset reveals different trends compared to the Framingham dataset (a relatively smaller dataset). For both accuracy and efficiency, all *BIMICE* variations demonstrated statistically significant improvements over all other baseline algorithms. In terms of accuracy, there is no significant evidence of a single *BIMICE* variant outperforming the others. However, when measuring efficiency, the results remain consistent with the previous findings, with *BIMICE* and *BIMICE*^{2} achieving the best performance, supported by statistical significance.

[3] https://github.com/SigmoidAI/reparo

[4] https://causalnex.readthedocs.io/en/latest/01_introduction/01_introduction.html

[5] https://docs.python.org/3/library/graphlib.html

[6] https://github.com/yoavzelinger/BIMICE

To address **RQ4** and better observe the results, we examined the performance of the different algorithms for increasing severity of missing values, either by the total size of missing data or the number of features requiring imputation.

The results in Figures 3, 4 emphasize the effect of total missing values on the Framingham dataset. Since the trend is consistent across all iterations, we present only the results for 15 iterations.

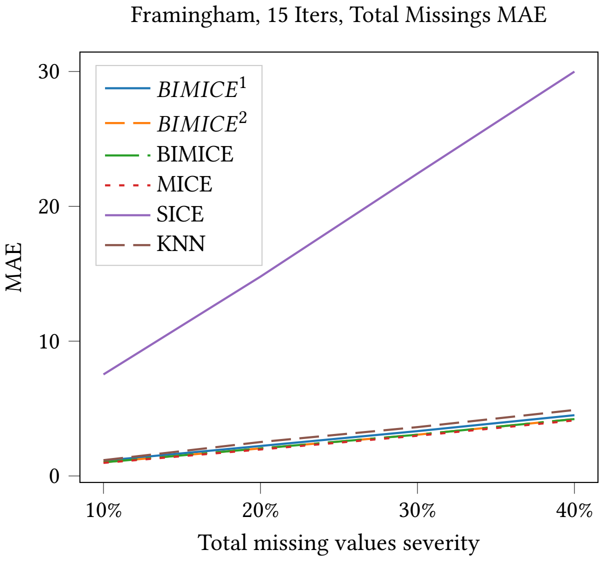

Figure 3: MAE results over Framingham dataset, running with 15 iterations of imputations, over different severities of total missing values.

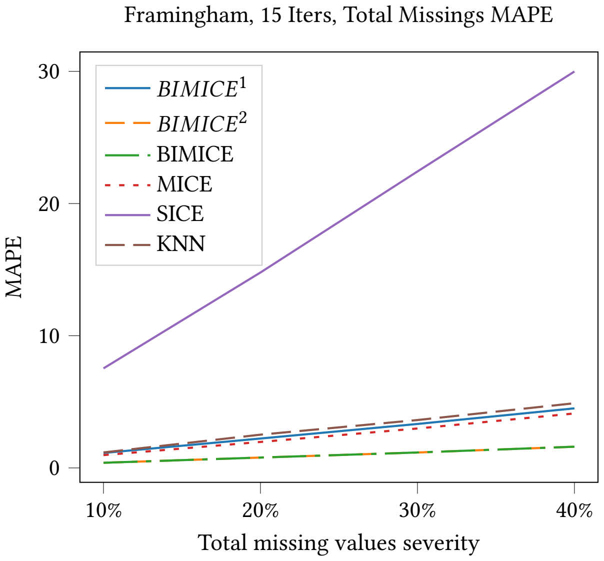

Figure 4: MAPE results over Framingham dataset, running with 15 iterations of imputations, over different severities of total missing values.

The results shown in Figures 5 and 6 highlight the significant impact of total missing values on the financial-risk dataset (medium-sized dataset). Similarly, as the trend remains consistent across all number of iterations, we have chosen to present only the results for 15 iterations of imputations. This approach allows for a clearer interpretation of the results without redundant data, ensuring that the key insights are emphasized.

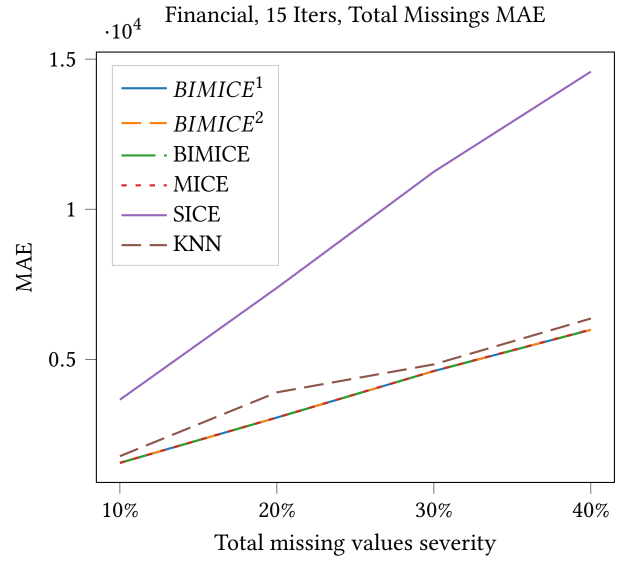

Figure 5: MAE results over Financial-Risk dataset dataset, running with 15 iterations of imputations, over different severities of total missing values.

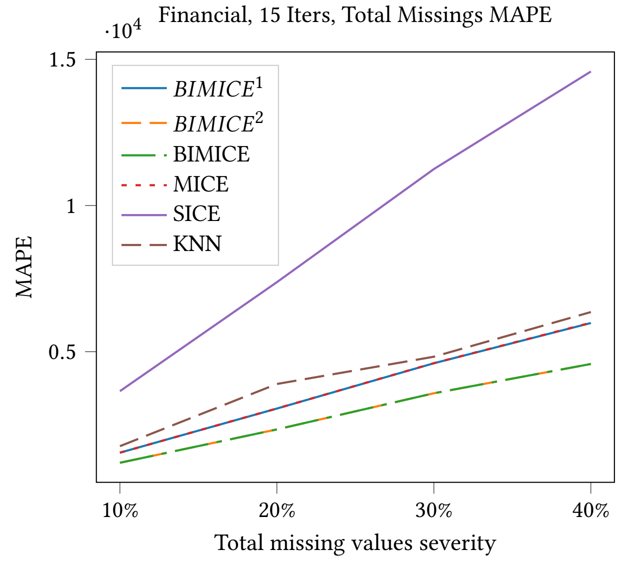

Figure 6: MAPE results over Financial-Risk dataset dataset, running with 15 iterations of imputations, over different severities of total missing values.

Together with this, the results in Figures 7, 8, 9, 10 represent the effect of the number of features with missing values on the accuracy of the imputations. Consistent with the previous, since the trend remains consistent across all iteration numbers, we present only the results for 15 iterations of imputations.

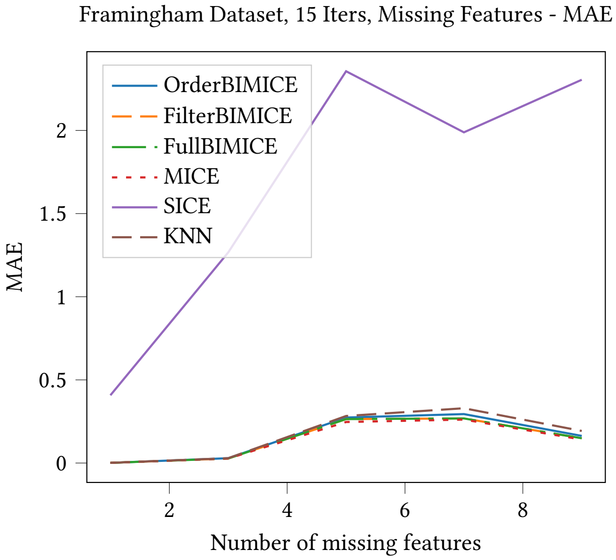

Figure 7: MAE results over Framingham dataset, running with 15 iterations of imputations, over different amount of features with missing values.

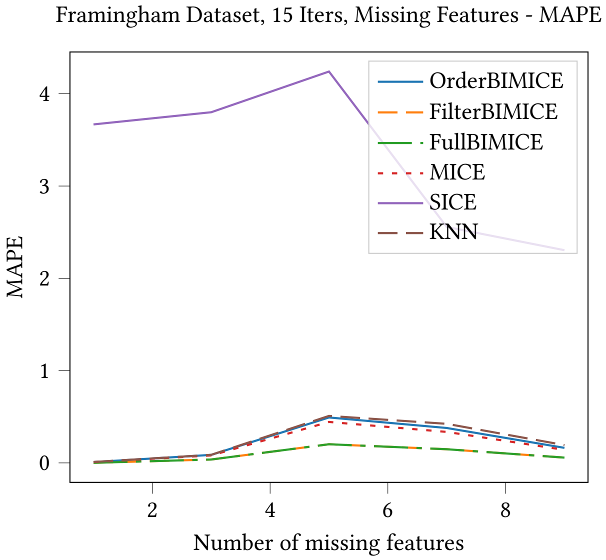

Figure 8: MAPE results over Framingham dataset, running with 15 iterations of imputations, over different amount of features with missing values.

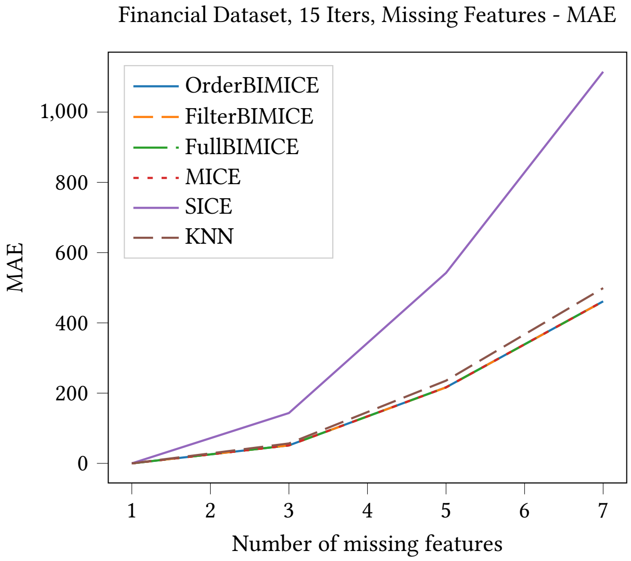

Figure 9: MAE results over Financial-Risk dataset, running with 15 iterations of imputations, over different amount of features with missing values.

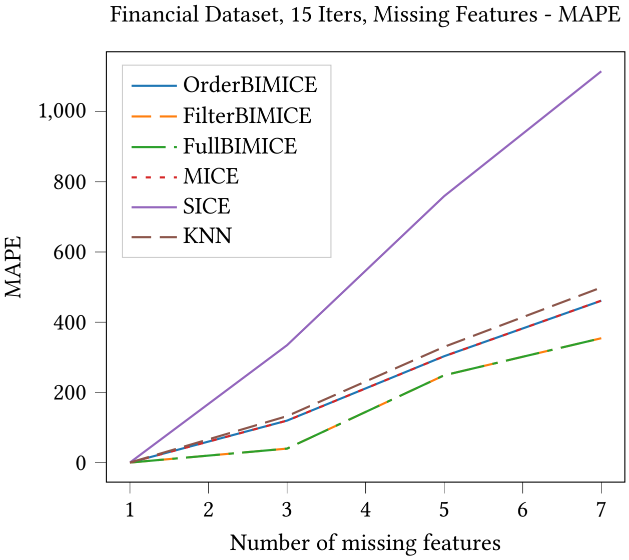

Figure 10: MAPE results over Financial-Risk dataset, running with 15 iterations of imputations, over different amount of features with missing values.

By looking at the figures we can infer that similarly to the baseline algorithms, the total portion of missing values in the data has a clear effect on the imputation accuracy, while the number of features with missing values does not represent a consistent trend.

5.3 Discussion

In this section, we provide a comprehensive discussion of the strengths, weaknesses, and limitations of our proposed algorithm, BIMICE.

5.3.1 Strengths: Several advantages of our approach can be observed:

- *BIMICE* consistently achieved the lowest MAPE across all tests, highlighting its efficiency in utilizing relevant predictors. This efficiency was particularly evident in scenarios with higher proportions of missing values, where other methods like *KNN* and *SICE* struggled, as illustrated in the figures 4 and 6. While *BIMICE*'s accuracy performance (the MAE and MSE results) were comparable to or slightly better than those of traditional MICE, its superior MAPE highlights its ability to optimize imputation without compromising accuracy.
- *BIMICE* demonstrated robust performance across datasets of varying sizes and complexities. In tests exploring different missingness scenarios, it consistently maintained strong results on both the small (Framingham) and large (Financial Risk) datasets. Even with 40% missing values or missing data across all columns, *BIMICE* exhibited minimal degradation in both MAE and MAPE. This highlights its resilience and reliability across diverse conditions, as well as its ability to handle varying data complexity and scale effectively.
- A third observation arises from the ablation tests. By analyzing the results in the figures (varying proportions of missing data) and in the tables (average scores), a significant observation is highlighted. Both enhancements, guided variable ordering (*BIMICE*^{1}) and selective predictor (*BIMICE*^{2}), led to improved performance compared to the traditional MICE algorithm.

5.3.2 Limitations: Despite its strengths, our work has several limitations:

- From the results we can observe that the full *BIMICE* algorithm contains redundant parts. Specifically, we can observe that the impact of *BIMICE*^{1} was relatively modest. Looking on the results, best results achieved by both the full *BIMICE* algorithm and *BIMICE*^{2}. This finding suggests that, in practice, the contribution of *BIMICE*^{1} may be redundant. Thus, we can conclude that a topological sorting is not necessary, and the imputation order can be determined randomly.
- When working with small-sized datasets, the accuracy performance of *BIMICE* tends to be inferior when compared to the traditional MICE algorithm. However, in the case of small datasets, the efficiency advantage of *BIMICE* becomes less significant, as the computational cost and time savings are less pronounced. Therefore, when dealing with smaller datasets, the traditional MICE algorithm, with its simpler and more direct approach, might be preferable due to its superior accuracy and the lower impact of efficiency concerns.
- By looking at table 2, a medium-sized dataset, we can observe interesting trends. All the variations of *BIMICE* suffered from larger errors as the number of iterations increased. Thus, we can infer that the additional complexity introduced by the iterative process may lead to diminishing returns in accuracy, especially as the algorithm becomes more susceptible to overfitting or instability with increasing iterations. This suggests that a balance between the number of iterations and the algorithm's performance needs to be carefully considered to avoid exacerbating errors.
- By looking at the results related to the varying number of features with missing values, we can observe that imputing a small number of missing features resulted in a relatively small error. However, as the number of features with missing values increased, the error also grew significantly. This suggests that the imputation process becomes more challenging and less accurate as the number of missing features rises, highlighting the need for more sophisticated techniques or adjustments to handle datasets with a higher proportion of missing values.

6 Conclusions and Future Work

Handling missing data is a critical challenge in data analytics, impacting the accuracy and reliability of predictive models. Traditional imputation methods, such as mean substitution and standard MICE, often struggle with capturing complex variable relationships, leading to potential biases and inefficiencies. In this paper, we introduced Bayesian Influenced Multiple Imputations by Chained Equations (*BIMICE*), which enhances the traditional MICE framework by leveraging Bayesian networks to optimize variable ordering and predictor selection. Our empirical evaluation demonstrated that *BIMICE* significantly enhances computational efficiency compared to baseline methods like KNN and SICE, consistently achieving lower Mean Absolute per Predictor Error (MAPE). The efficiency gains are particularly notable in scenarios with higher proportions of missing values. While maintaining comparable accuracy to traditional methods, *BIMICE* offers a more efficient use of relevant predictors without compromising performance. Ablation tests highlight that Selective Predictor Inclusion plays a key role in improving imputation efficiency. Despite its advantages, *BIMICE* is currently restricted to continuous data and relies heavily on the quality of the Bayesian network, which can be computationally intensive and sensitive to data quality and modeling decisions. In addition, from the empirical tests we can infer that the selective predictor enhancement to the MICE algorithm provides the most impact, compared to the optimized imputation ordering. Moreover, generally *BIMICE* becomes more effective with larger datasets with missing values.

In future work, the proposed method can be extended to support categorical data in addition to continuous data, thereby enhancing its flexibility and ability to handle diverse datasets. Another potential enhancement involves developing advanced heuristics for selecting relevant variables, which would optimize the process and reduce computational overhead. Additionally, incorporating domain-specific knowledge or prior insights into the imputation process could significantly enhance accuracy. These advancements have the potential to broaden the method's applicability and improve its suitability for various use cases across different fields.

References

[1] Majed Alwateer, El-Sayed Atlam, Mahmoud Mohammed Abd El-Raouf, Osama A Ghoneim, and Ibrahim Gad. 2024. Missing Data Imputation: A Comprehensive Review. *Journal of Computer and Communications* 12, 11 (2024), 53–75.

[2] Melissa J. Azur, Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. 2011. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research* 20, 1 (March 2011), 40–49. https://doi.org/10.1002/mpr.329

[3] Edoardo Costantini, Kyle M Lang, Klaas Sijtsma, and Tim Reeskens. 2024. Solving the many-variables problem in MICE with principal component regression. *Behavior Research Methods* 56, 3 (2024), 1715–1737.

[4] Widad Elouataoui, Saida El Mendili, and Youssef Gahi. 2023. An Automated Big Data Quality Anomaly Correction Framework Using Predictive Analysis. *Data* 8, 12 (2023), 182.

[5] Salvador García, Julián Luengo, and Francisco Herrera. 2015. *Data Preprocessing in Data Mining*. Intelligent Systems Reference Library, Vol. 72. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-10247-4

[6] Estevam R. Hruschka, Eduardo R. Hruschka, and Nelson F. F. Ebecken. 2007. Bayesian networks for imputation in classification problems. *Journal of Intelligent Information Systems* 29, 3 (2007), 231–252. https://doi.org/10.1007/s10844-006-0016-x

[7] Sebastian Jäger, Arndt Allhorn, and Felix Bießmann. 2021. A Benchmark for Data Imputation Methods. *Frontiers in Big Data* 4 (July 2021), 693674. https://doi.org/10.3389/fdata.2021.693674

[8] Napsu Karmitsa, Sona Taheri, Adil Bagirov, and Pauliina Mäkinen. 2022. Missing Value Imputation via Clusterwise Linear Regression. *IEEE Transactions on Knowledge and Data Engineering* 34, 4 (2022), 1889–1901. https://doi.org/10.1109/TKDE.2020.3001694

[9] Shahidul Khan and Abu Latiful Haque. 2020. SICE: an improved missing data imputation technique. *Journal of Big Data* 7 (06 2020). https://doi.org/10.1186/s40537-020-00313-w

[10] Ki-Yeol Kim, Byoung-Jin Kim, and Gwan su Yi. 2004. Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC Bioinformatics* 5 (2004), 160 – 160. https://api.semanticscholar.org/CorpusID:5127318

[11] Tolou Shadbahr, Michael Roberts, Jan Stanczuk, Julian Gilbey, Philip Teare, Sören Dittmer, Matthew Thorpe, Ramon Viñas Torné, Evis Sala, Pietro Lió, et al. 2023. The impact of imputation quality on machine learning classifiers for datasets with missing values. *Communications Medicine* 3, 1 (2023), 139.

[12] Anoop D Shah, Jonathan W Bartlett, James Carpenter, Owen Nicholas, and Harry Hemingway. 2014. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American journal of epidemiology* 179, 6 (2014), 764–774.

[13] Matthias Templ, Alexander Kowarik, and Peter Filzmoser. 2011. Iterative stepwise regression imputation using standard and robust methods. *Computational Statistics & Data Analysis* 55, 10 (2011), 2793–2806.