FairImputeAE: A Fairness-Aware Autoencoder for Missing Data Imputation

Amal Koodoruth, Dasen Ye, Hassen Kammoun

December 2023

1 Abstract

Addressing missing data challenges is paramount for ensuring the robustness and reliability of research outcomes across diverse disciplines. In this work, we explored the application of autoencoders for missing data imputation, with a focus on fairness considerations. It addresses the fundamental types of missing data mechanisms, emphasizing Missing Completely at Random (MCAR) scenarios. The study introduces a tailored binary cross-entropy cost function that incorporates a "missingness" vector to enhance imputation. Additionally, a novel Imputation Fairness Risk (IFR) metric, initially proposed by [13], is introduced for a comprehensive evaluation. Empirical investigations on the Adult dataset involve four distinct models, including a "vanilla" model with a modified Binary Cross Entropy (BCE) loss function, and three FairImputeAE models one of which being the proposed model of this work. As a result, the proposed model guided by fairness considerations, achieves a commendable balance between accuracy and fairness. More specifically, we have demonstrated that under large missing probabilities, the proposed model significantly improves imputation fairness while maintaining high accuracy.

2 Introduction

2.1 Missingness

Missing data is a pervasive challenge in research across various disciplines, ranging from social sciences to healthcare and beyond. The presence of missing values can compromise the integrity and validity of analyses, leading to biased results and potentially erroneous conclusions. As researchers strive to draw meaningful inferences from incomplete datasets, it becomes imperative to understand the nature of missingness, as different mechanisms govern the occurrence of missing values.

The classification of missing data mechanisms is crucial for selecting appropriate imputation methods and understanding the potential biases introduced during the analysis. Three fundamental types of missing data mechanisms are commonly recognized: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) [7].

Data are said to be MCAR if the missingness is independent of both observed and missing values. In other words, the missingness is entirely random and does not depend on any measured or unmeasured variables. In this case, the missing values are essentially a random subset of the observed data. Missing at random (MAR) occurs when the probability of missingness depends on observed variables but not on the missing values themselves. In MAR scenarios, the missingness can be explained by the values of other variables in the dataset. For example, let *sex* be a binary feature observed in all samples. In a MAR scenario, the probability of missingness of the other features is different for $sex = 0$ and for $sex = 1$. In a Missing not at random (MNAR) setting, the probability of a data point being missing is related to its own value. In MNAR situations, the missingness is influenced by unobserved factors, and the missing values are systematically different from the observed values. For example, in this case, let *sex* be missing in some samples. The probability of missingness of other features is different for $sex = 0$ and for $sex = 1$. This scenario is the most challenging because the probability of missingness of features, in our case, *sex*, depends on a missing value itself.

In this paper, especially because of time constraints, we will consider only the MCAR case.

2.2 Related Work

Missing values are a pervasive issue across real-world datasets, often manifesting as NaNs or blanks. Addressing this challenge through imputation is a fundamental step in the data processing pipeline. A multitude of imputation methods have been developed to tackle missing data, each exhibiting unique advantages and limitations.

Multiple Imputation

In the domain of imputation techniques, the method of multiple imputation stands out as a significant approach. This technique involves the generation of multiple datasets with imputed values, and subsequent statistical analyses conducted on each dataset provide unbiased estimates and standard errors [9].

While multiple imputation proves to be a robust strategy for handling missing data, it is essential to underscore certain considerations. The application of this method demands substantial computational resources, making it imperative to assess the computational capacity available for its implementation. Additionally, meticulous implementation is crucial to ensure the reliability and validity of the imputed values, emphasizing the need for careful attention to detail in the execution of the methodology.

Deep Learning Models

In the realm of imputation methodologies, the integration of neural networks, particularly autoencoders and Generative Adversarial Networks (GANs), stands out in related works. These models excel in capturing non-linear relationships and intricate patterns, showing promise for addressing missing data challenges.

However, their efficacy depends on substantial datasets, requiring diverse and extensive data for effective generalization. Researchers must ensure the dataset size aligns with the demands of these sophisticated neural network architectures.

Additionally, the vulnerability of neural networks to overfitting is a crucial consideration. Despite their ability to capture complex relationships, careful model tuning and regularization techniques are essential to mitigate overfitting risks and enhance robustness.

As the field progresses, ongoing research plays a key role in understanding and optimizing the application of neural networks in imputation tasks, refining their effectiveness for missing value imputation within the broader landscape of related works.

Data Augmentation Techniques

In addressing class imbalances and imputing missing values, data augmentation techniques have emerged as pivotal tools in the realm of related works. These methods aim to enhance the representation of underrepresented classes and augment available data, thereby contributing to more robust model training.

Despite their widespread application, lingering questions persist regarding the efficacy of data augmentation techniques in capturing true values. The nuances of their performance are intricately tied to the specific algorithm employed, emphasizing the importance of careful algorithmic selection to align with the characteristics of the dataset and the nature of missing values.

Moreover, the equality of samples resulting from data augmentation is contingent upon the algorithmic choices made during the process. Concerns about potential biases or artifacts introduced during augmentation underscore the need for researchers and practitioners to exercise diligence in selecting and configuring these methods, ensuring the fairness and representativeness of imputed data.

As the field of data augmentation evolves, ongoing research and experimentation are crucial for refining and validating these techniques. By addressing persistent questions surrounding their efficacy and sample equality, researchers contribute to the advancement of reliable and trustworthy data augmentation approaches, enhancing their utility in handling class imbalances and missing values across diverse datasets.

Common Baseline Models

Common baseline imputation models often resort to mean or median imputation for continuous data, as discussed in [6]. While straightforward and efficient, these methods come with limitations, notably neglecting covariance between features and the potential introduction of bias. To address categorical values in time series data, Forward/Backward fill is a popular choice. However, it too disregards covariance and may not perform optimally with irregular time intervals.

In addition to these baseline methods, more advanced techniques like IterativeImputer have gained prominence. IterativeImputer, a multivariate algorithm, supports a range of models including linear regression, Bayesian Ridge regression, k-nearest neighbors regression, decision trees, and random forests. This flexibility makes it a valuable asset in scenarios where diverse feature dimensions need to be considered for comprehensive imputation.

MICE (Multiple Imputation by Chained Equations) stands out as another widely adopted technique for imputation. Particularly advantageous when dealing with complex datasets exhibiting non-random missing data patterns, MICE creates multiple imputed datasets, each accounting for the uncertainty in imputed values. This approach provides a nuanced and robust imputation strategy for intricate data scenarios.

MissForest

Operationalizing the Random Forest algorithm, MissForest consistently outperforms alternative imputation methods, demonstrating superior efficacy across diverse metrics. In a pivotal 2011 study by Stekhoven and Buhlmann, it surpassed competing approaches, including KNN-Impute, by over 50

MissForest's strength lies in its ensemble learning approach, utilizing multiple decision trees to capture complex relationships within the data. This adaptability proves advantageous across various data structures and enhances resilience to outliers, making it a versatile and reliable choice for imputation tasks.

Stekhoven and Buhlmann's findings underscore MissForest's reliability, solidifying its status as a leading imputation method. Its consistent performance across different domains positions it as a valuable tool for researchers and practitioners seeking accurate and comprehensive data analysis.

K-Nearest Neighbours

The K-nearest neighbors (KNN) algorithm, renowned for its efficiency, excels in imputing missing values by leveraging information from non-missing neighbors within uniformly distributed data [1]. This approach offers a straightforward and intuitive imputation strategy, particularly suited for scenarios where proximity in feature space indicates similarity.

Nevertheless, it is imperative to underscore a critical consideration associated with KNN imputation. While effective in random missing data scenarios, the algorithm's reliance on proximity can lead to bias when the missing data mechanism is non-random. This potential bias arises from the assumption that the distribution of non-missing neighbors adequately represents the true characteristics of the missing values. Therefore, caution is warranted, and researchers should be mindful of the underlying data patterns and the implications of non-random missingness.

2.2.1 Autoencoders

Autoencoders, a specialized category within the domain of artificial neural networks, are designed to acquire a distributed representation of input data [12]. These networks undergo a learning process to determine optimal parameters, facilitating the transformation of data into a concealed or "hidden" layer, and subsequently reconstructing the original input. A pivotal aspect of autoencoders lies in the deliberate imposition of a hidden layer with fewer dimensions than the input features – colloquially known as the "bottleneck" layer. This strategic architectural choice compels the autoencoder to discern and internalize the most salient patterns inherent in the data [10].

To circumvent the potential pitfalls of overreliance on specific features, autoencoders employ well-established techniques. In denoising autoencoders, a deliberate introduction of noise corrupts a subset of input features, compelling the network to discern robust patterns amidst this added complexity [11]. An alternative approach involves the incorporation of dropout, a technique wherein random units and connections are temporarily removed during training, compelling the network to generalize more effectively [8].

2.2.2 Fair Imputation

Fair imputation is the process of handling missing data in a manner that aligns with fairness principles in Machine Learning. It is paramount to account for the existence of protected/sensitive attributes such as sex, age, or race. Its main goal is to impute missing values without any discrimination or bias.

Different imputation techniques affect the fairness differently. It depends on the assumptions of the imputation method, the nature of the missing data, and whether the imputation process aligns with fairness principles. Therefore, careful consideration and evaluation of different imputation techniques are essential to ensure that fairness is preserved in machine learning models dealing with missing data. [5]

3 Preliminaries

3.1 Modified BCE

Addressing missing data using Autoencoders involved the creation of an autoencoder architecture incorporating a tailored binary cross-entropy cost function. This modification, inspired by the work of Beaulieu-Jones and Greene (2016)[2], aimed to enhance the handling of missing data. The adjusted cost function takes into consideration the presence or absence of data through a "missingness" vector denoted as '$\boldsymbol{m}$'. Here, '$\boldsymbol{m}$' takes on a value of 1 where the data is available and 0 when the data is missing. Modifying the Binary Cross Entropy (BCE) function by multiplying by $\boldsymbol{m}$ translates into calculating the error in prediction based only on observed values and not missing values. Finally, dividing by *count(*$\boldsymbol{m}$*)* averages the loss across observed values.

$$\mathcal{L}_{obs} = \sum_{k=1}^{K} [x_k \log(z_k) m_k + (1 - x_k) \log(1 - z_k) m_k] \,/\, count(\boldsymbol{m}), \tag{1}$$

where K is the total number of samples in the training set, x_k is the k^{th} sample and z_k is its reconstruction. While this loss function, used in [3], provides promising results, it however does not take into consideration the fairness aspect when imputing values.

3.2 Imputation Fairness Risk

Similar to [13], to assess the fairness of an imputation model, we define the modified BCE loss within each sensitive group:

$$\mathcal{L}_{obs,\ A=a} = \sum_{k=1}^{K^a} [x_k^a \log(z_k^a) m_k^a + (1 - x_k^a) \log(1 - z_k^a) m_k^a] \ / \ count(m^a), \tag{2}$$

where K^a is the number of samples with the sensitive attribute a, x_k^a is a sample with the sensitive attribute a and z_k^a is the reconstruction of x_k^a.

With this and without loss of generality, we build the imputation fairness risk [13] function:

$$IFR = |BCE_{obs}^{A=0} - BCE_{obs}^{A=1}| \tag{3}$$

which is the difference between the loss function of groups defined by the sensitive attribute. Intuitively, the Imputation Fairness Risk (IFR) serves as an indicator of the "fairness" embedded in our model's imputed values. It translates into the absolute difference between the misclassification error across the sensitive groups. A smaller IFR corresponds to a higher degree of fairness in our model, suggesting that the imputation process aligns more closely with fair outcomes.

3.3 Loss Function

In order to improve training, Zhang et al. (2022) [13] also penalize wrong imputations by a factor λ_{acc}.

$$\mathcal{L}_{miss} = \sum_{k=1}^{K} [x_k \log(z_k) m'_k + (1 - x_k) \log(1 - z_k) m'_k] \ / \ count(\boldsymbol{m'}), \tag{4}$$

where $\boldsymbol{m'} = \mathbb{1} - \boldsymbol{m}$.

Equipped with the above, we define our loss function as:

$$\mathcal{L} = \mathcal{L}_{obs} + \lambda_{acc} \mathcal{L}_{miss} + \lambda_{fair} IFR \tag{5}$$

4 Methodology

4.1 Data Preparation

To study the performance of our model, we conduct experiments on the Adult dataset [4]. We first drop all the rows with NAs, and split the dataset into 60% for training, 20% for validation and 20% for testing. The 8 categorical columns were selected and one-hot encoded for each category. We selected the *sex* feature to be the sensitive attribute. For each sample $x^{(i)}$, we generate a random binary mask, $m^{(i)}$, with 1 indicating that the column at that index is "missing". We investigate the performance of the model on different missingness levels, P_{miss}.

4.2 Models

We performed a parameter sweep to determine the best model at each missingness level. Each layer in the autoencoder is followed by a "LeakyReLU" activation function with a dropout

probability of 0.2, except the last layer of the decoder, where the activation function was sigmoid. The choice of sigmoid here was natural, forcing the values to be between 0 and 1. During the evaluation, we decoded the one-hot encoded dataset by splitting the resulting matrix and by selecting the "argmax" for each original feature.

In total, we investigate 4 types of models at different missingness levels: a "vanilla" model, similar to the one proposed by [3], FairImputeAE with $\lambda_{acc} = 0$, FairImputeAE with $\lambda_{fair} = 0$ and FairImputeAE with $\lambda_{acc} \neq 0$ and $\lambda_{fair} \neq 0$. We used the AdamW optimizer in each case.

During the training of each model, we meticulously recorded the values of the loss function for both the training and validation sets after each epoch. This practice enabled us to closely monitor the training process, facilitating an assessment of the appropriateness of hyper-parameters, such as the learning rate. Figure 1 shows a particular example of the training and validation loss of the vanilla model during the training process, plotted as a function of the number of epochs. The losses exhibited a notable improvement after 20 epochs; however, this enhancement gradually diminished, with minimal progress observed between epochs 60 and 100. This suggests that the number of epochs should not be larger than 100 when training the final models.

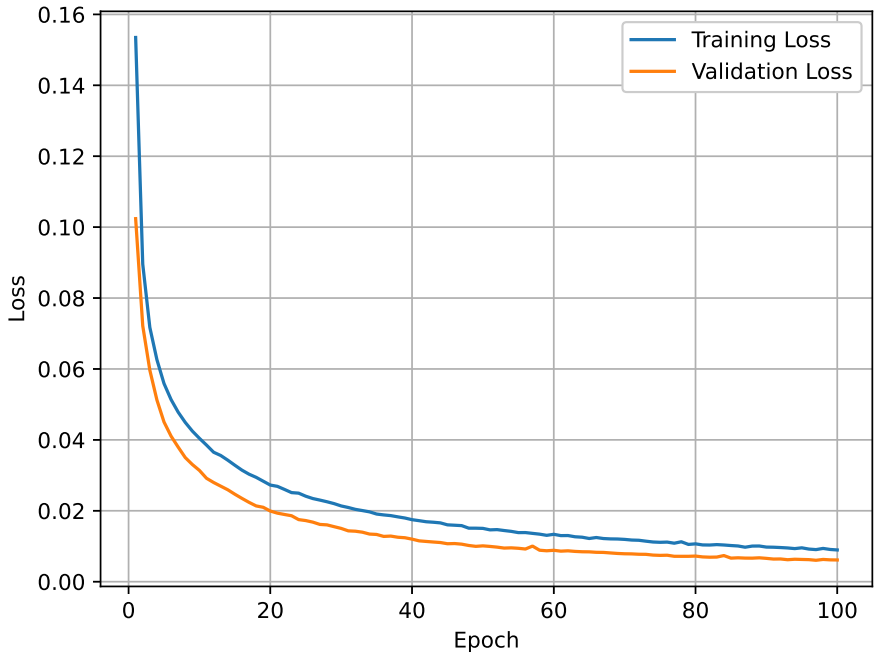

Figure 1: Training and validation loss of the vanilla model for $P_{miss} = 0.1$ with respect to training epochs.

4.3 Hyper-parameter Selection

As part of the training process, it is also necessary to perform hyper-parameter tuning of the models. For each missing probability $P_{miss} \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$, the hyper-parameters of the model take the following possible values:

$$\begin{aligned} \lambda_{fair} &= \{0.1, 0.2, 0.5\} \\ \lambda_{acc} &= \{0.1, 0.2, 0.5\} \\ Encoder\ layers &= \{[99, 128, 256, 256], [99, 64, 128, 256]\} \\ Decoder\ layers &= \{[256, 256, 128, 99], [256, 128, 64, 99]\} \\ bottleneck\ size &= \{5, 32\} \\ learning\ rate = &= \{0.001, 0.005\} \end{aligned}$$

while $dropout = 0.2$, $input\_size = 99$, and $epoch = 100$ are fixed to a single value to control the computational cost. Note that there are only 2 possible combined choices for the encoder and decoder layer since they need to be symmetric to each other.

5 Results

The final results were generated on the test set using the "best-tuned" hyper-parameters determined in the hyper-parameter selection step. We are mainly interested in comparing the imputation accuracy and the IFR value of our complete model (with loss function $\mathcal{L}$) with other models. Table 1 shows the imputation accuracy for each model with different missing probabilities, and Table 2 shows the IFR values.

P_{miss}	Vanilla Model	Model with $\lambda_{fair} = 0$	Model with $\lambda_{acc} = 0$)	Full Model
0.1	97.13%	97.33%	**97.50**%	97.16%
0.25	97.23%	97.48%	**97.49**%	97.23%
0.5	96.67%	**97.34**%	96.91%	97.19%
0.75	94.74%	**96.95**%	95.18%	96.70%
0.9	91.74%	95.27%	92.21%	**95.48**%

Table 1: Imputation Accuracy of all models on the test dataset for each missing probability, using the selected hyper-parameters.

P_{miss}	Vanilla Model	Model with $\lambda_{fair} = 0$	Model with $\lambda_{acc} = 0$	Full Model
0.1	0.00044	0.00042	0.00044	**0.00039**
0.25	0.00094	0.00086	0.00094	**0.00078**
0.5	0.00226	**0.00135**	0.00203	0.00159
0.75	0.00576	0.00344	0.00398	**0.00305**
0.9	0.01522	0.0065	0.00975	**0.00471**

Table 2: Imputation fairness risk (IFR) of all models on the test dataset for each missing probability, using the selected hyper-parameters.

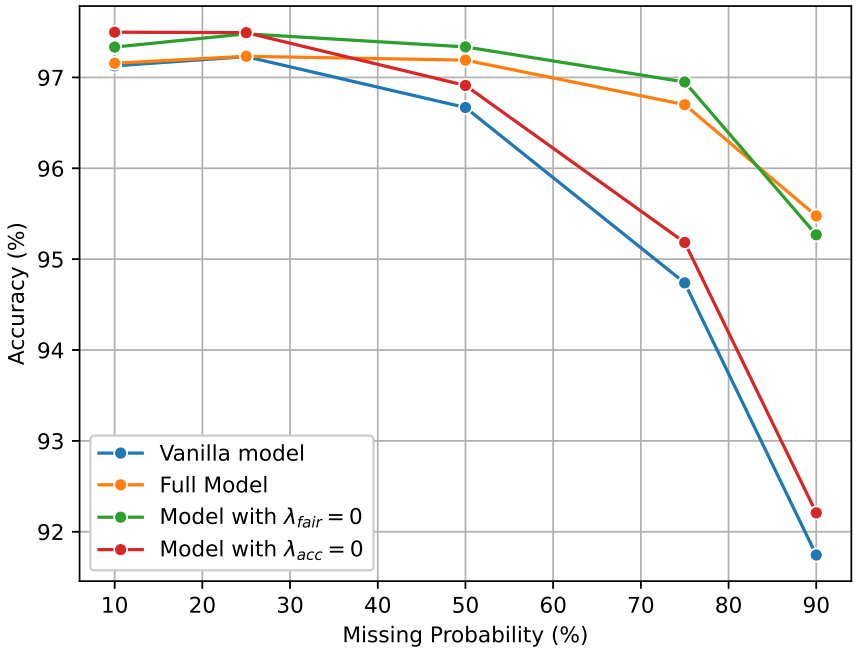

Figure 2: Imputation accuracy of all models with respect to the missing probability.

The above tables are also presented graphically in Figure 2 and 3.

6 Discussion

In Figure 2, one can see that the accuracy of all models decreases as the missing probability increases. This is expected since a larger missing probability implies more missing data, and therefore less available information for the models, leading to a drop in imputation accuracy. The model that we are mainly interested in, the one using the full loss function $\mathcal{L}$, has the highest accuracy when the missing probability $P_{miss} = 90\%$. However, we also see that the model with $\lambda_{fair} = 0$ has very similar performance. This suggests that the IFR regularization has a limited effect on improving the imputation accuracy. Furthermore, one can also notice that the model with $\lambda_{acc} = 0$ has a relatively high performance (around 92 to 97%), this implies that we can in fact train an autoencoder using a dataset with missing values since the loss for missing values is not backpropagated.

Importantly, in Figure 3 we see that the vanilla model has the worst IFR values across all missingness levels. This is because the vanilla model is not regularized to take into account the imputation fairness. The regularized model with full $\mathcal{L}$, the main result of this work, has the lowest IFR among all models for large missing probabilities, with a significant improvement compared to the other models. This result suggests that the model we developed in this work significantly improves the imputation fairness on categorical data while maintaining a high imputation accuracy.

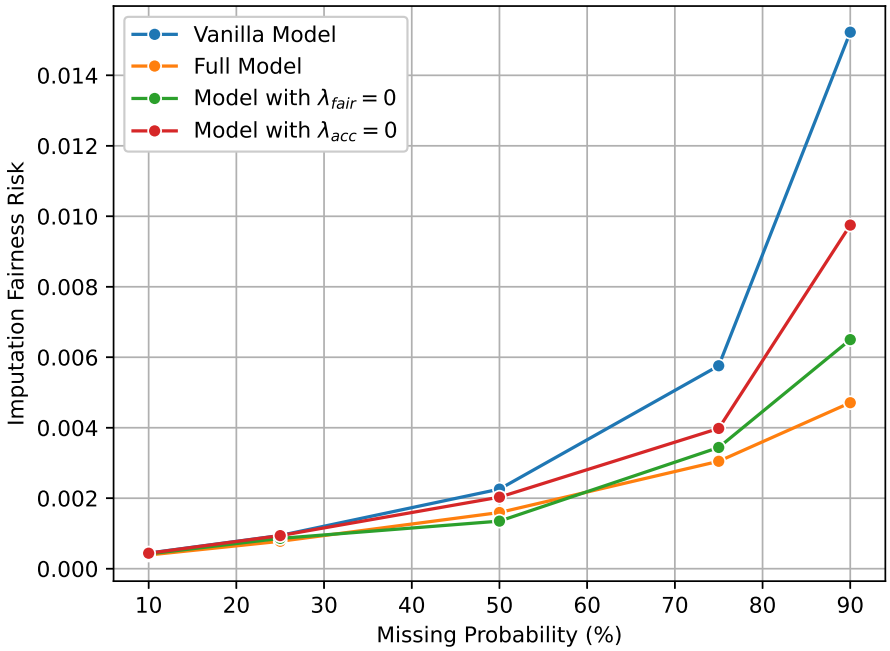

Figure 3: Imputation fairness risk of all models with respect to the missing probability.

7 Limitations and Future Work

Although the results we obtained are inspiring, there are a few limitations and shortcomings that can be further studied in the future. The first limitation is that the missing data that were generated in this work are only of type MCAR, which means missing completely at random. In the future, other types of missing data such as MAR and MNAR may be explored with the model developed in this work.

The second limitation is that the number of candidates is small for all hyper-parameters during the hyper-parameter selection. This is due to the limited computational power that we have. This can be improved by using high-performance GPU and/or parallel computing.

Moreover, the model we described in this paper is suitable only for categorical data, while many datasets in practice contain mixed datatypes (continuous and categorical). This, however, can be mitigated, although not really addressing the problem, by having a similar architecture for continuous data and concatenating outputs.

Lastly, the model should be evaluated on more datasets to get a more comprehensive understanding of its performance. If all the above limitations can be investigated and improved, then we might be able to increase the model's performance and robustness further.

8 Conclusion

This study explores missing data imputation methodologies with a specific emphasis on fairness, focusing on Missing Completely at Random (MCAR) scenarios. Autoencoders,

a subset of artificial neural networks, are employed with a modified binary cross-entropy cost function to enhance imputation. Also, the introduction of the Imputation Fairness Risk (IFR) metric, along with a comprehensive loss function, demonstrates the potential to improve accuracy and fairness in categorical data scenarios.

Empirical experiments on the Adult dataset have revealed that the proposed model maintains a balance between accuracy and fairness, proving robust in the MCAR scenario. In addition, the vanilla model serves as an accuracy benchmark, while the proposed model with the IFR metric exhibits notable fairness improvements without significant accuracy compromise, especially with large missing probabilities.

In conclusion, this work contributes to the discourse on missing data imputation by integrating fairness considerations in the MCAR scenario. Also, we acknowledge limitations and propose future directions, encouraging continued exploration of diverse missing data mechanisms and enhanced model functionalities.

References

[1] Ahmed Abulkhair. Data imputation demystified — time series data, 2023.

[2] Beaulieu-Jones B. and Greene C.S. Semi-supervised learning of the electronic health record for phenotype stratification. *Journal of Biomedical Informatics*, 2016.

[3] BRETT BEAULIEU-JONES and Jason Moore. Missing data imputation in the electronic health record using deeply learned autoencoders. volume 22, pages 207–218, 02 2017.

[4] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[5] Simon Caton, Saiteja Malisetty, and Christian Haas. Impact of imputation strategies on fairness in machine learning. *Journal of Artificial Intelligence Research (JAIR)*, 2022.

[6] Satyam Kumar. 7 ways to handle missing values in machine learning, 2020.

[7] R. Little, D. Rubin, and an O'Reilly Media Company Safari. *Statistical Analysis with Missing Data., 3rd Edition*. Wiley, 2019.

[8] Srivastava N., Hinton G., Krizhevsky A., Sutskever I., and Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 2014.

[9] Neri Van Otten. Imputation of missing values comprehensive practical guide, 2023.

[10] Vincent P., Larochelle H., Lajoie I., and Manzagol P-A. Bengio Y. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, 2010.

[11] Vincent P., Larochelle H., Bengio Y., and Manzagol P-A. Extracting and composing robust features with denoising autoencoders. 2008.

[12] Bengio Y. *Learning Deep Architectures for AI*. Now Publishers Inc, 2009.

[13] Yiliang Zhang and Qi Long. Fairness-aware missing data imputation. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*, 2022.