The Impact of Missing Data Imputation on Model Performance and Explainability

Işıl Baysal Erez[1 [0000-0002-0636-5372]], Jan Flokstra[1], Mannes Poel[1 [0000-0002-3813-9732]], and Maurice van Keulen[1 [0000-0003-2436-1372]]

EEMCS, University of Twente, 7500 AE Enschede, The Netherlands

Abstract. Missing data is a common problem in data analysis processes. Data may be missing due to data collection or recording, which can occur in any field, such as healthcare, transportation, telecommunications, biology, and more. Several machine learning models require complete data. Therefore, imputation — the process of filling in missing data — is typically the preferred method. Missing data imputation methods are useful for predictive model analysis by filling in the missing values in the data. However, the impact of imputation techniques on model explainability is not yet known. In this study, we investigate the impact of white-box and black-box imputation methods on the explainability of the black-box Random Forest classification algorithm. Moreover, we applied the glass box classification model that allows us to see the model's decisions in a transparent and understandable way, such as the Explainable Boosting Machine, to compare performance differences. We conducted an experiment with seven imputation methods, using three missing levels (10%, 50%, and 70%) and three datasets. We investigated the effect of the imputers on the feature importance of using SHAP values and applying surrogate models per imputed dataset. Summarize findings show that striving for the best predictive model performance, best explainability, or imputation performance leads to a different best choice for the imputer. It is, therefore, important, contrary to common practice, to regard the selection of imputer and classifier as a combined selection problem and to explicitly choose what to optimize for.

Keywords: Missing data · Imputation · Explanability.

1 Introduction

Missing data is a critical issue across many fields, including healthcare, transportation, and business, and it can significantly impact the performance of predictive models [2, 28]. Missing data can arise from problems with data collection and human mistakes, such as lack of participation in research and surveys [19], to technical and systemic issues, device malfunctions, and sensor failures [17]. Additionally, data may be missing simply because it does not exist, such as in recommendation model systems [13]. Missing data can be divided into three categories: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) [26].

Most machine learning algorithms cannot effectively handle datasets with missing values. When the proportion of missing data is very small, deleting data (items) with missing values can be a practical approach to enable smooth analysis [27]. Another common technique for handling missing values is imputation, where plausible values are assigned to missing data before training predictive models. A variety of imputation methods have been proposed, ranging from statistical-based imputation methods such as mean and Multivariate Imputation by Chained Equations (MICE) to machine learning and deep learning-based imputation methods such as k-nearest Neighbors (k-NN), Multilayer Perceptron (MLP), Generative Adversarial Imputation Networks (GAIN). Some tree-based algorithms, such as Classification and Regression Trees (CART) [4], eXtreme Gradient Boosting (XGBoost) [5], and Light Gradient-Boosting Machine (LightGBM) [15] algorithms, can natively handle missing data.

Many machine learning algorithms are used for data analysis and continue to be developed to meet the demands of increasingly complex data. In the pursuit of predictive performance, achieving high model performance remains a fundamental goal within the machine learning community [22]. However, for the decision-making mechanism to operate optimally, the quality of model explainability is as important as the quality of model performance. These two areas are considered to have a trade-off between each other [10].

In particular, Machine Learning (ML) algorithms tend to lack transparency, making it challenging to understand how they operate internally. To solve this opacity, several Explainable AI (XAI) methods and tools have been proposed [22]. They can be categorized as global (explaining the model) versus local (explaining an individual prediction of the model) or as model-agnostic vs. model-specific [20]. An alternative approach to overcome the explainability problem of black-box models is to use a glass-box (e.g., Generalized Additive Model) or a white-box (e.g., linear regression, decision tree) machine learning method.

Although imputation methods have been developed to handle missing data, a research gap still exists regarding their impact on both the performance and explainability of black-box models, as well as interpretable glass-box models. To illustrate a real-world scenario, we can consider a pneumonia mortality risk dataset [9] [6], where a heart rate in the range of 38-125 bpm is considered normal. However, if we assume that some heart rate values are missing and we impute the dataset using various imputation methods, it is crucial that the normal range of 38-125 bpm remains intact. This is essential to ensure that decision-support systems for healthcare professionals do not result in potentially life-threatening consequences.

In this study, we study the effect of different families of imputation methods on predictive model explainability. We focus on the performance of model-agnostic methods on global and local explainability. In our experiments, we use as representatives for black-box, glass-box and white-box, Random Forest (RF), Explainable Boosting Machine (EBM), and RF in combination with a decision tree as a surrogate model for the explanations. To summarize, we seek answers to the following questions:

RQ1: How do different imputation methods affect the explainability of a machine-learning model?

RQ2: Can high predictive performance be achieved using glass-box models in the presence of missing data?

2 Related Work on Explainability and Imputation

In this section, related work is presented about the use of XAI methods in the context of dealing with missing data.

Hans et al. [12] focus on human-readable explanations for imputation methods. They introduced data imputation techniques by considering diverse data types (categorical, numeric, text, and date) along with their associated constraints.

Chen et al. [6] use EBM to provide users with insights into the missing data problem and demonstrate potential hazards arising from different imputation methods (mean, k-NN, MissForest) for several medical datasets. They illustrate that their recommended interpretability approach can aid in understanding missing data mechanisms, such as MCAR or not, to understand the relationship between features and missingness. EBM is also used to predict missing values like an imputer.

Cinquini et al. [8] proposed Local Interpretable Model Agnostic Explanations (LIMEMV) with Missing Values approaches that can handle missing data using the LIME method. They produced synthetic data by extracting statistical features such as mean and standard deviation obtained from missing data, and they used a surrogate tree model instead of a Lasso model. They tested the explainability performance of the LIMEMV method using k-NN and MICE imputation methods and obtained very competitive results.

Ahmad et al. [1] emphasized that imputation methods should be applied carefully, especially in the presence of missing data that may occur in Criminal Justice and patient safety systems. Their focus is on the fidelity of explanations.

Recent work by Vo et al. [34] focusing on regression discussed the XGBoost model explainability with imputation and without. They used two datasets and applied five different imputation methods (Mean, MICE, Conditional Distribution-based Imputation of Missing Values with Regularization (DIMV), MissForest, and SoftImpute) to data with missingness levels of 20%, 40%, 60%, and 80%. They calculated mean square error (MSE) in Shapley values between the original and imputed datasets. According to their results, model's interpretability can be strongly impacted by the imputation method selection. It has been observed that the Shapley values are not robust and show variability in methods with high imputation accuracy. The authors suggest that, when choosing an imputer, a choice should be made by taking into account the model explainability criterion and dataset meta-features.

In the literature on explainability methods, it has often been observed that either a single explainability method or an interpretable model like a glass-box model is used in the presence of missing data. However, our approach goes a

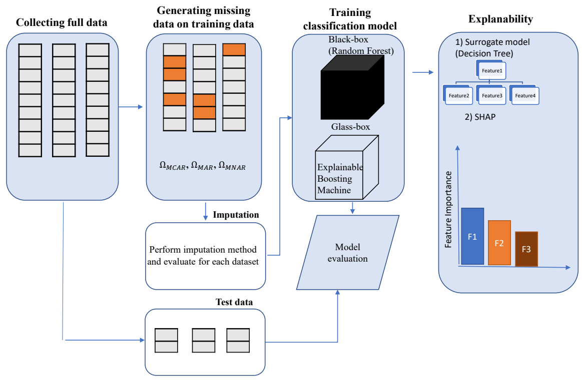

Fig. 1. Methodology

step further by integrating multiple methods—such as a surrogate model, SHAP method, and glass-box model—thus providing a more comprehensive understanding of model behaviour with imputed datasets.

3 Methodology

We selected complete datasets to measure the performances of imputation methods and regression models. We use 80%/20% train/test split. Then, we induce different missingness ratios in the training set.

An overview of the approach and pipeline for the experiments is given in Figure 1.

Datasets: We use datasets from the healthcare and biology domain from [33] [16] (see Table 1). Data selection was made by taking into account the number of features.

The "Lung discrete" datasets were high-dimensional with few instances, while the "Iris" and "Appendicitis" were lower-dimensional with more instances.

Dataset	Features	Instances	Classes
Lung discrete [16] [24]	325	73	7
Iris [11]	4	150	3
Appendicitis [35]	7	106	2

Table 1. Summary of Datasets

Missing Data Mechanisms: We experiment with the missing completely at random (MCAR) data mechanism, where the probability of missing data is the same for all features [31], and the missing not at random (MNAR) data mechanism, where the probability of missing data may be caused by the missing data itself [31]. Different missing rates of 10%, 50% and 70% were synthetically generated. We simulate MCAR and MNAR settings as employed in [21]. MCAR data mechanism, the probability of missing data is independent of observed and non-observed values. To generate MCAR settings, we used a binary mask $\forall \omega_{ij} \in \Omega_{MCAR}$, $\omega_{ij} \in \{0,1\}^{n \times d}$, a Bernoulli distribution, $\mathcal{B}(p)$, was used to sample the mask ω_{ij}, so that $\omega_{ij} \sim \mathcal{B}(p)$, where p represents the likelihood that a data point will be present.

The MAR data mechanism depends on the values of the observed variables. Missing data is generated by keeping a certain proportion of observed data points fixed (p_obs). A subset of the fully observed variables was selected, and missing values were then introduced in the remaining variables using a logistic model. $\omega_{ij} \sim \frac{1}{1+e^{-\langle \mathbf{w}, \mathbf{X} \rangle}}$. We set (p_obs)=0.5. For the MNAR data mechanism, the missingness is conditional on the unobserved data, and to set MNAR, a logistic model was employed similar to MAR, but instead of separating a certain percentage of the observed data, it used the entire dataset as inputs. The variables were split into two groups: one set was used to determine missing probabilities based on a logistic model, and the other was used as inputs for a logistic model, following similar settings as employed in [21]. We applied the missing data mechanisms for the MCAR and MNAR in our experiments. The MAR mechanism was not included in this study, since we considered the MAR assumption to be less generalizable in real-world scenarios. Additionally, a similar reasoning for excluding MAR was observed in a related study [2].

Imputation methods: We selected imputation methods based on different levels of complexity, namely, the statistical methods mean imputation, matrix completion (Soft Impute) [18] and ICE [32], machine learning methods k-NN [30] and MissForest [29], and deep learning methods MLP [14] and GAIN [36]. For mean imputation, missing values are replaced using the mean of each column. For the ICE, MissForest, and MLP imputation methods, mean imputation was used as an initial strategy.

Classification methods:

- **Black-box representative**: Random Forest.
  As an example of a black-box model, we selected the Random Forest method [3], Random forest is a type of ensemble model formed by combining multiple decision trees. The bagging method in RF takes random samples from different data subsets and trains models (e.g., decision trees) independently on these samples. The final prediction is made by taking the majority vote or an average of these models.
- **Glass-box representative**: Explainable Boosting Machine.
  To compare the performance of Random Forest on imputed data with a

more interpretable model, the EBM method [23] was chosen. It belongs to the Generalized Additive Model family and uses a tree-based and cyclic approach [23].
- **Black-box + surrogate representative**: RF + Decision tree surrogate. A surrogate model can be used to obtain the explainability of the black-box model. We selected decision trees [4] for the surrogate. First, we trained a Random Forest black-box model and obtained its predictions for full and imputed datasets. We trained a surrogate decision tree model using these black-box model predictions Then we measured how successfully the black-box model's predictions were reproduced by the surrogate model [7].

We applied the imputation methods after splitting to prevent data leakage. All experiments were performed independently 5 times using different random seeds.

Evaluation Metrics: We used root mean squared error (RMSE) metrics to evaluate imputers and model explainability and accuracy for comparison of the performance of classifiers. The RMSE for imputers is calculated only between the imputed (completed) values and the original data points. We calculated the RMSE of the SHAP values to quantify the difference in explanations between the original data and the imputed data across all samples (n).

4 Results

Variation in black box model explainability We show the RMSE of SHAP values, denoted as $\text{SHAP}_{\text{rmse}}$, in Figure 2 for the Iris, Appendicitis, and Lung discrete datasets. As the missing level increases, the $\text{SHAP}_{\text{rmse}}$ values increase in all datasets. For the MCAR and MNAR missing data mechanisms, only in the Iris data set, the $\text{SHAP}_{\text{rmse}}$ values for the data imputed with the GAIN imputer are smaller at a 70% missingness level compared to 50%. We observe that for the MNAR missing data mechanism, the mean imputer leads to the biggest change in model explanation. The GAIN imputer is the second most influential in terms of changing the explanation of the model, except for the high-dimensional lung discrete dataset. For the MNAR missing data mechanism, most $\text{SHAP}_{\text{rmse}}$ values are higher in the small-sized Iris dataset compared to MCAR, indicating that the type of missing data mechanism is important for this data set. The GAIN imputer performed better for the lung discrete dataset, especially for the MCAR data mechanism. It appears that the low $\text{SHAP}_{\text{rmse}}$ values were achieved with the larger dataset.

Comparison of black-box and glass-box model performance First, we conducted a performance comparison for imputation methods and accuracies of imputed datasets for the predictive models Random Forest and Explainable Boosting Machine to analyze their efficacy under a low (e.g., 10%), medium (e.g., 50%), and high (e.g., 70%) missingness level. The results are in Table 2 and 3. The two tables compare the results of low-dimensional vs. high-dimensional data sets. The imputer demonstrating the best performance varies according

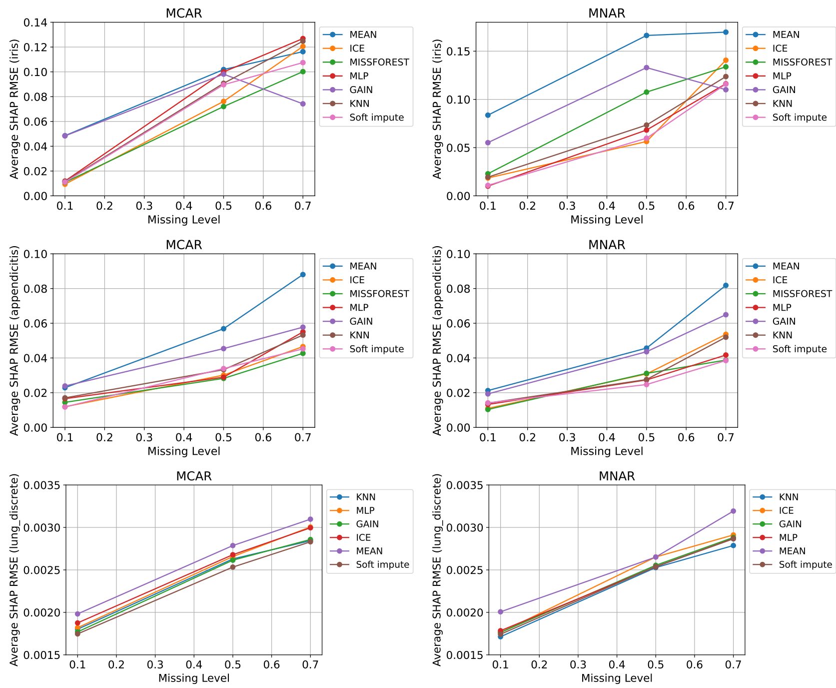

Fig. 2. Average RMSE of SHAP values for the Iris (top), Appendicitis (middle) and Lung discrete (bottom) datasets and MCAR (left) and MNAR (right) missing mechanisms. Experiments run 5 times. Low $\text{SHAP}_{\text{rmse}}$ is better.

to the missing levels. For the low missing level, the mean imputer achieved the best performance. At the medium and high missing levels, the Soft Impute method showed the best results. It can be observed that the best imputation methods based on RMSE values do not always improve the classification models' performances for all missing levels. Datasets imputed with k-NN showed the best performance in the classification model for low missing levels. Datasets imputed with MLP imputer showed the highest accuracy presence of a high missing level for black and glass box models. For a high-dimensional dataset, k-NN was the best imputer at a low level of missing data, while Soft Impute showed the best imputation performance at higher missing rates, such as 50% and 70%. Compared to small-dimensional data, high-dimensional data showed lower performance in both black box and glass box classification models. The performance of the black box model exceeds the glass box model in all missing levels.

Surrogate Model Performance In Figure 3, We compare the performance of the surrogate model trained on both imputed and original datasets, as well as across low-dimensional and high-dimensional datasets. While the performance of the surrogate model is above 90% in small datasets, the model's performance

Table 2. Imputation results (RMSE) and classification model performances (ACC) across imputed *low-dimensional* datasets (Iris+Appendicitis) over three missing levels using MCAR data mechanisms for black-box RF ($\mathbf{ACC_{BB}}$) and glass-box EBM ($\mathbf{ACC_{GB}}$). Lower RMSE and higher accuracy (ACC) are better. RMSE and ACC results are presented with mean and standard deviation. Bold numbers represent the best performer per missing level. Low RMSE and high ACCs are better. Experiments were repeated 5 times.

Data	10			50		
	RMSE ↓	$\mathbf{ACC_{BB}}$ ↑	$\mathbf{ACC_{GB}}$ ↑	**RMSE** ↓	$\mathbf{ACC_{BB}}$ ↑	$\mathbf{ACC_{GB}}$ ↑
MEAN	**2.28±1.88**	0.92±0.08	**0.92±0.05**	2.51±1.55	0.89±0.05	0.88±0.07
ICE	2.45±1.65	0.92±0.06	0.90±0.06	2.36±1.68	0.89±0.04	**0.89±0.06**
k-NN	2.36±1.73	**0.93±0.05**	0.91±0.06	2.28±1.74	0.89±0.06	0.88±0.05
MISSFOREST	2.43±1.68	0.92±0.05	0.90±0.06	2.44±1.65	0.89±0.05	**0.89±0.07**
MLP	2.43±1.66	0.91±0.08	0.91±0.06	2.51±1.55	**0.90±0.07**	0.88±0.07
GAIN	2.38±1.67	0.92±0.06	0.91±0.05	2.36±1.65	0.88±0.08	**0.89±0.07**
SoftImpute	2.37±1.72	0.92±0.08	0.90±0.07	**2.25±1.76**	0.89±0.05	0.88±0.05

Data	70		
	RMSE ↓	$\mathbf{ACC_{BB}}$↑	$\mathbf{ACC_{GB}}$ ↑
MEAN	2.23±1.81	0.86±0.08	0.89±0.04
ICE	2.34±1.68	0.84±0.10	0.86±0.09
k-NN	2.24±1.75	0.86±0.08	0.87±0.06
MISSFOREST	2.42±1.60	0.87±0.07	0.86±0.05
MLP	2.72±1.26	**0.90±0.04**	**0.91±0.05**
GAIN	2.35±1.63	0.84±0.07	0.84±0.06
SoftImpute	**2.22±1.81**	0.82±0.05	0.85±0.07

is lower in large data sets per data mechanism. It can be said that the surrogate model matches the explanation for the original data with better performance for small-sized data sets. ICE and MLP imputer show similar results for MCAR and MNAR missing data mechanisms. The surrogate model performance of data completed with the mean imputer shows lower results for both MCAR and MNAR data mechanisms. The performance of models trained on datasets imputed with GAIN imputer increased for the MCAR data mechanism. In Figure 3, the bottom section shows the surrogate model performance for the lung discrete dataset. The black-box model appears not to mimic the behaviour very well, even without missing data, so it seems that the surrogate model is not sufficient in interpreting the black-box model. When lung discrete datasets completed with k-NN imputation are trained with a surrogate model, they show better results for the MCAR data mechanism, while datasets filled with ICE imputation performed the best

Table 3. Imputation results (RMSE) and classification model performances (ACC) on imputed *high-dimensional* dataset (Lung discrete) over three missing levels using MCAR data mechanisms for black-box Random forest ($\mathbf{ACC_{BB}}$) and glass-box model EBM ($\mathbf{ACC_{GB}}$). Lower RMSE and higher accuracy (ACC) are better. RMSE and ACC results are presented with mean and standard deviation. Bold numbers represent the best performer per missing level. Experiments run 5 times.

Data	10			50		
	RMSE ↓	$\mathbf{ACC_{BB}}$ ↑	$\mathbf{ACC_{GB}}$ ↑	**RMSE** ↓	$\mathbf{ACC_{BB}}$ ↑	$\mathbf{ACC_{GB}}$ ↑
MEAN	1.60±0.01	0.83±0.08	0.72±0.12	1.60±0.01	0.76±0.10	0.67±0.08
ICE	1.33±0.07	0.85±0.13	0.77±0.11	1.56±0.02	0.71±0.09	0.68±0.09
k-NN	**1.30±0.01**	**0.85±0.06**	0.76±0.10	1.36±0.02	**0.79±0.09**	**0.73±0.11**
MLP	1.50±0.04	0.83±0.04	0.77±0.11	1.66±0.04	0.77±0.13	0.71±0.10
GAIN	1.31±0.03	0.81±0.06	0.77±0.10	1.46±0.02	0.73±0.08	0.69±0.11
Soft Impute	1.25±0.02	0.84± 0.08	**0.79±0.11**	**1.34±0.01**	**0.79±0.09**	0.73±0.11

Data	70		
	RMSE ↓	$\mathbf{ACC_{BB}}$↑	$\mathbf{ACC_{GB}}$ ↑
MEAN	1.61±0.01	0.61±0.12	0.61±0.12
ICE	1.60±0.03	0.65±0.11	0.63±0.12
k-NN	**1.43±0.01**	**0.72±0.06**	0.65±0.09
MLP	1.68±0.05	0.61±0.09	0.59±0.10
GAIN	1.56±0.02	0.63±0.08	0.61±0.12
Soft Impute	**1.43±0.02**	0.69±0.06	**0.68±0.09**

performance compared to other imputation techniques under the MNAR data mechanism.

5 Limitations and Future Work

This study investigates how imputation affects a black-box model's explainability when employing datasets that have missing values. Furthermore, we chose the EBM as the interpretable model and compared its results with those of the black-box model for imputed datasets. We conducted our experiments based on two small and one high-dimensional dataset, using seven different imputers (Mean, k-NN, ICE, MissForest, MLP, GAIN, and Soft impute) and three missing data levels (10%, 50%, 70%). The missing data were simulated according to MCAR and MNAR missing data mechanisms. Since the model explainability and missing data issues have broad and multidisciplinary research areas, this study has limitations, particularly regarding dataset selection and the choice of imputation methods and black-box models. Future work will focus on expanding to further strengthen and generalize our findings. This includes different datasets (characteristics such as statistical and information-theoretic), dataset types (image

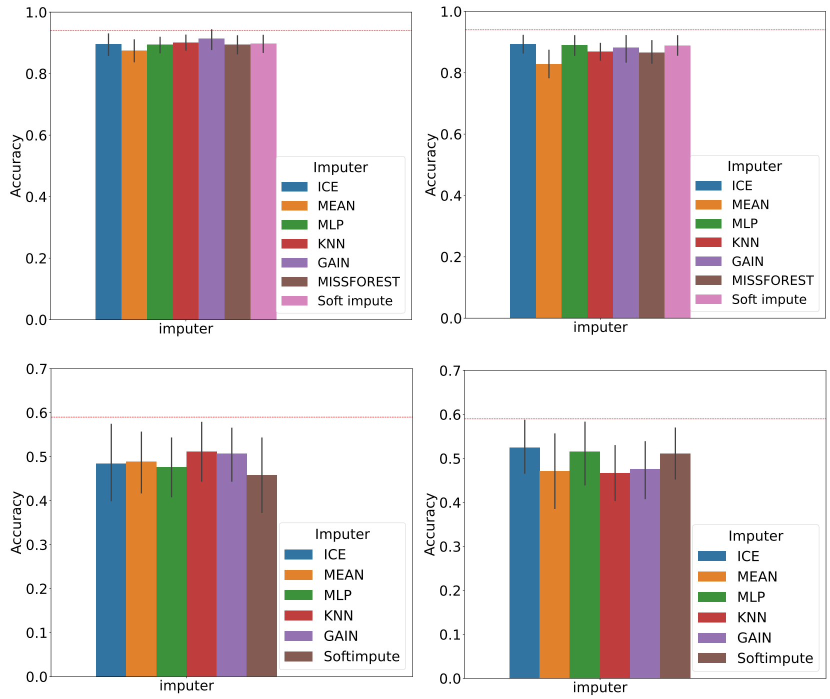

Fig. 3. Average accuracy of surrogate models per imputer for low-dimensional datasets Iris and Appendicitis (top) and high-dimensional dataset Lung discrete (bottom). Experiments run 5 times for MCAR (left) and MNAR (right) missing data mechanisms. 80-20 train-test split. The red line represents the original test dataset surrogate model.

and text-domain datasets), expansion of the data preprocessing pipeline (more imputers based on deep learning) and other explainability values (LIME [25] method). Moreover, the feature ranking based on the SHAP method can be analyzed before and after imputation to observe whether there is any change in the ranking. Since there is a time constraint of 60 minutes for imputation, the lung discrete dataset contains a large number of features; therefore, the MissForest imputer was very costly in terms of time, so the MissForest imputation was removed from the analysis for this dataset.

6 Conclusion

In this study, two research questions were formulated. The first research question (RQ1) was: How do different imputation methods affect the explainability of a machine learning model? Model explainability after imputation was worse in low-dimensional data (See Figure 2). There is no single best imputer for this analysis, but it can be said that the mean imputer changed the black box model explainability in most cases. It has been observed that simple imputation methods such as Mean imputer harm the explainability of the model, especially in small datasets.

The surrogate model was trained using the decision tree algorithm with the predictions of the black-box model. The GAIN and ICE imputers were the closest to the performance of the original surrogate model under the MCAR and MNAR missing data mechanisms for the low-dimensional dataset. It was observed that the surrogate model trained on high-dimensional data could not simulate very well the black box model in both the original data set and the imputed datasets.

The second research question (RQ2) was: Can high predictive performance be achieved using glass-box models in the presence of missing data? The Random Forest black box model performed better, especially for high-dimensional data (see Table 3). However, it has been noted that the glass-box model outperforms the black-box model at a 70% missing data level for low-dimensional datasets (see Table 2). This indicates that the number of features and the mechanisms of missing data influence model performance.

To sum up our findings, imputation affects model explainability and accuracy based on data characteristics, such as missing level, number of features, and so on, so the imputer and classifier selection problem should be considered as a combined selection problem, determining what to optimize for and then making an optimal choice.

References

1. Ahmad, M.A., Eckert, C., Teredesai, A.: The challenge of imputation in explainable artificial intelligence models. arXiv preprint arXiv:1907.12669 (2019)
2. Bertsimas, D., Pawlowski, C., Zhuo, Y.D.: From predictive methods to missing data imputation: an optimization approach. Journal of Machine Learning Research **18**(196), 1–39 (2018)
3. Breiman, L.: Random forests. Machine learning **45**, 5–32 (2001)
4. Breiman, L.: Classification and regression trees. Routledge (2017)
5. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)
6. Chen, Z., Tan, S., Chajewska, U., Rudin, C., Caruna, R.: Missing values and imputation in healthcare data: Can interpretable machine learning help? In: Conference on Health, Inference, and Learning. pp. 86–99. PMLR (2023)
7. Christoph, M.: Interpretable machine learning: A guide for making black box models explainable. Leanpub (2020)
8. Cinquini, M., Giannotti, F., Guidotti, R., Mattei, A.: Handling missing values in local post-hoc explainability. In: World Conference on Explainable Artificial Intelligence. pp. 256–278. Springer (2023)
9. Cooper, G.F., Abraham, V., Aliferis, C.F., Aronis, J.M., Buchanan, B.G., Caruana, R., Fine, M.J., Janosky, J.E., Livingston, G., Mitchell, T., et al.: Predicting dire outcomes of patients with community acquired pneumonia. Journal of biomedical informatics **38**(5), 347–366 (2005)
10. Crook, B., Schlüter, M., Speith, T.: Revisiting the performance-explainability trade-off in explainable artificial intelligence (xai). In: 2023 IEEE 31st Interna-

tional Requirements Engineering Conference Workshops (REW). pp. 316–324. IEEE (2023)
11. Fisher, R.A.: Iris. UCI Machine Learning Repository (1988), DOI: https://doi.org/10.24432/C56C76
12. Hans, S., Saha, D., Aggarwal, A.: Explainable data imputation using constraints. In: Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD). pp. 128–132 (2023)
13. Heinrich, B., Hopf, M., Lohninger, D., Schiller, A., Szubartowicz, M.: Something's missing? a procedure for extending item content data sets in the context of recommender systems. Information Systems Frontiers pp. 1–20 (2022)
14. Hinton, G.E.: Connectionist learning procedures. In: Machine learning, pp. 555–610. Elsevier (1990)
15. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems **30** (2017)
16. Li, J.: Scikit-feature feature selection repository - data. https://github.com/jundongl/scikit-feature/tree/master/skfeature/data (2024), accessed: 2024-08-22
17. Li, Y., Bao, T., Chen, Z., Gao, Z., Shu, X., Zhang, K.: A missing sensor measurement data reconstruction framework powered by multi-task gaussian process regression for dam structural health monitoring systems. Measurement **186**, 110085 (2021)
18. Mazumder, R., Hastie, T., Tibshirani, R.: Spectral regularization algorithms for learning large incomplete matrices. The Journal of Machine Learning Research **11**, 2287–2322 (2010)
19. Mirzaei, A., Carter, S.R., Patanwala, A.E., Schneider, C.R.: Missing data in surveys: Key concepts, approaches, and applications. Research in Social and Administrative Pharmacy **18**(2), 2308–2316 (2022)
20. Molnar, C.: Interpretable machine learning. Lulu. com (2020)
21. Muzellec, B., Josse, J., Boyer, C., Cuturi, M.: Missing data imputation using optimal transport. In: International Conference on Machine Learning. pp. 7130–7140. PMLR (2020)
22. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., Seifert, C.: From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. ACM Computing Surveys **55**(13s), 1–42 (2023)
23. Nori, H., Jenkins, S., Koch, P., Caruana, R.: Interpretml: A unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223 (2019)
24. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence **27**(8), 1226–1238 (2005)
25. Ribeiro, M.T., Singh, S., Guestrin, C.: " Why Should I Trust You?" Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
26. Rubin, D.B.: Inference and missing data. Biometrika **63**(3), 581–592 (1976)
27. Schafer, J.L., Graham, J.W.: Missing data: our view of the state of the art. Psychological methods **7**(2), 147 (2002)
28. Shadbahr, T., Roberts, M., Stanczuk, J., Gilbey, J., Teare, P., Dittmer, S., Thorpe, M., Torne, R.V., Sala, E., Lio, P., et al.: Classification of datasets with imputed

missing values: does imputation quality matter? arXiv preprint arXiv:2206.08478 (2022)
29. Stekhoven, D.J., Bühlmann, P.: Missforest—non-parametric missing value imputation for mixed-type data. Bioinformatics **28**(1), 112–118 (2012)
30. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B.: Missing value estimation methods for dna microarrays. Bioinformatics **17**(6), 520–525 (2001)
31. Van Buuren, S.: Flexible imputation of missing data. CRC press (2018)
32. Van Buuren, S., Groothuis-Oudshoorn, K.: mice: Multivariate imputation by chained equations in r. Journal of statistical software **45**, 1–67 (2011)
33. Vanschoren, J., van Rijn, J.N., Bischl, B., Torgo, L.: Openml: Networked science in machine learning. SIGKDD Explorations **15**(2), 49–60 (2013). https://doi.org/10.1145/2641190.2641198, http://doi.acm.org/10.1145/2641190.2641198
34. Vo, T.L., Nguyen, T., Hammer, H.L., Riegler, M.A., Halvorsen, P.: Explainability of machine learning models under missing data. arXiv preprint arXiv:2407.00411 (2024)
35. Weiss, S.M., Kulikowski, C.A.: Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems. Morgan Kaufmann Publishers Inc. (1991)
36. Yoon, J., Jordon, J., Schaar, M.: GAIN: Missing Data Imputation using Generative Adversarial Nets. In: International conference on machine learning. pp. 5689–5698. PMLR (2018)