2025 IEEE International Conference on Artificial Intelligence Testing (AITest) | 979-8-3315-8913-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/AITest66680.2025.00032

The Impact of Imputation Timing on Model Performance Estimation

Ben Mitchell
Computer Science Department
Swarthmore College
Swarthmore, USA
bmitche1@swarthmore.edu

Shikha Shrestha
Computer Science Department
Swarthmore College
Swarthmore, USA
dshrest2@swarthmore.edu

***Abstract*—Handling missing data is a critical challenge in applying machine learning, as most algorithms assume complete data. Imputation, the process of replacing missing values with estimates from available data, is a common solution. This study investigates the impact of imputation timing (before vs. after train-test split) on machine learning classifier performance estimates, particularly focusing on the biases introduced by different imputation strategies. We evaluate the effects of imputation before train-test split (IBS) and imputation after train-test split (IAS) across multiple datasets and imputation methods, including Random Forest (RF), KNN, and Mean Imputation. Our findings reveal that IBS consistently *overestimates* generalization performance, with severity worsening as the proportion of missing data increases, while IAS *underestimates* performance, again worsening as missing data fractions grow. These discrepancies highlight the potential for bias and instability in performance estimates, emphasizing the need for careful handling of imputation techniques to avoid misleading conclusions about model robustness. Our results further underscore the influence of missing data rates and dataset characteristics on classifier performance, suggesting that no single imputation method is universally appropriate.**

***Index Terms*—Missing Data, Imputation, Performance Estimation, Data Preprocessing, Data Quality**

I. Introduction

There are many ways in which data may require modification before being practical or appropriate for use with machine learning (ML) algorithms. In general, the math underlying ML algorithms makes a fairly stringent set of assumptions about the data (e.g. that it is sampled IID from a distribution which is homoskedastic and stationary, that features are conditionally independent given the target, that there is additive Gaussian noise, etc.) [19]. Some assumptions are specific to particular types of model or algorithm (e.g. conditional independence), but others are common across nearly all ML techniques (e.g. stationarity). These assumptions will typically not hold perfectly for real-world data, but mostly they are 'soft' assumptions, meaning that mild violations will have only mild impacts on the generalization performance of the learned model [19].

Some assumptions, on the other hand, are 'hard,' meaning that the algorithm cannot be applied at all if there is even a slight violation. For most algorithms, the assumption that each sample will be 'complete' (i.e. containing no missing feature values) is one of these [15]. The violation of this assumption is referred to as the 'missing value' problem, and in most cases it must be resolved before any traditional ML approach can be used [37]. Unfortunately, while it is a common problem in many domains, any solution runs the risk of skewing the data distribution in ways that can lead to biased predictions and decisions [13]. This problem extends beyond traditional machine learning to a range of other data-related fields, and leads to issues including performance degradation, data analysis challenges, and biased outcomes due to disparities between missing and complete data [4], [17], [25], [26]. There are many real-world causes for missing values, such as data collection errors, patient dropouts in medical studies, skipped survey questions, or technical failures in sensor-based systems. These occurrences follow distinct patterns, commonly categorized as Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR) [35]. In MCAR, the likelihood of a value being missing is unrelated to both observed and unobserved data [35]; for example, a sensor might randomly fail and miss recording a temperature reading due to a power outage. In MAR, the missingness depends only on observed data [35]; for instance, younger patients may be less likely to report income in a survey, so missing income values can be predicted based on age. In MNAR, the probability of missingness depends on the unobserved, missing data itself [35]; for example, patients with severe depression may be less likely to answer mental health questionnaires, making the missingness related to the actual severity of their depression.

Since most methods require complete data, missingness must typically be resolved as a pre-processing step. The simplest approach for handling missing data is *list-wise deletion*, which discards any samples with missing values. While this method is straightforward, it can introduce significant biases into the study, particularly when the missing data is not randomly distributed [34], [36], or when a large fraction of the available samples have at least one missing value. An alternative is *imputation*, where missing values are replaced with fabricated ones, often using the available data to make these fabrications [27]. The simplest method is single imputation by mean, in which the missing value is filled in with the mean value for that feature over the data set [3], [27].

Researchers have explored more sophisticated imputation

979-8-3315-8913-4/25/$31.00 ©2025 IEEE
DOI 10.1109/AITest66680.2025.00032

Authorized licensed use limited to: Swarthmore College Library. Downloaded on November 04,2026 at 18:20:12 UTC from IEEE Xplore. Restrictions apply.

methods, such as multiple imputation (MI) [34], singular value decomposition (SVD), K-nearest neighbor (KNN) [42], random forests [40], and deep learning techniques [12], [23]. As always, the correct choice will depend on the specific scenario; no-free-lunch applies here as it does everywhere [43].

Several studies have been conducted to help practitioners and data scientists choose the right imputation methods based on the characteristics of their datasets and the patterns of missing data [6], [21]. These studies often introduce missing values artificially to simulate specific data patterns, allowing researchers to compare how well different imputation techniques replicate the original data and how they affect the performance of machine learning models. The approaches vary, with some studies introducing missing values and performing imputation only in the training set, while others do so in both the training and test sets. For example, Poulos and Valle [31] focused on missing values in the training set alone. They found that using imputation methods like K-nearest neighbors (KNN) and random forest improved the accuracy of neural network classifiers compared to one-hot encoding, as imputation acted as a regularizer to prevent overfitting. On the other hand, studies by Jager et al. [21] and Bertsimas et al. [6] applied imputation to both training and test sets. They observed that this approach could lead to a significant drop in accuracy: up to 20% when using mean imputation on a dataset with 40% missing data, compared to performance on the original (complete) dataset. In some cases, imputing values in both sets resulted in worse performance compared to simple listwise deletion.

In a real-world scenario, we cannot guarantee a test set entirely free of missing values, so we will need to deal with imputation of the test set values at some point. To mimic this, studies like this one make use of 'synthetic' missingness, meaning we start with a complete data set and then remove values. This methodology can be problematic, in that real-world data may display patterns of missingness that are difficult to capture in synthetic examples. Unfortunately, using data with real-world missingness does not allow us to calculate 'imputation-free' baseline scores, so using synthetic missingness is the only practical option for this type of study.

It is obvious that any attempt at imputation risks introducing bias to the data distribution, and thus to any model trained on that data. What is less obvious is the way that the *timing* of imputation can impact that bias. In general, the two options are to apply imputation to the entire available dataset, *before* splitting it up into train/test/validation sets, or to apply imputation to each subset individually *after* splitting for validation (using whatever validation method is desired). Both approaches are represented in the literature, with some researchers opting to impute before splitting [2], [16] and others after splitting [31], [41]. Our study seeks to systematically compare the impact of imputing before versus after splitting the data for validation, and how performance estimates vary depending on whether imputed values are present in just the train set or in both the train and test sets.

However, the concept of a "performance estimate" becomes more nuanced in the presence of missing data, as imputed values in the test set can further distort the assessment of model performance. While constructing a test set exclusively from non-imputed, non-missing data is possible, this approach is not advisable under MAR or MNAR missingness conditions. Unlike MCAR, where missingness is truly random, MAR and MNAR involve systematic patterns tied to observed or unobserved variables. This makes it difficult to construct unbiased test sets, as imputing or removing values without accounting for these dependencies can distort model evaluation. Therefore, the second key objective is to evaluate discrepancies between *estimated* and *actual* generalization performance when imputed values are present in the test set. In this case, *actual* performance is calculated using examples with complete and correct values, where there is no missingness or imputation.

II. METHODS

A. Experimental Design

We used three complete benchmark datasets for our experiments (details of the data sets are given in the following sub-section). To simulate real-world scenarios with missing data, we artificially introduced missing values at rates of 5%, 10%, 25%, 35%, and 50% under MCAR, MAR, and MNAR patterns, resulting in 15 distinct dataset-variants per original dataset based on the combination of missing data patterns and rates. Missing values were introduced in 80% of randomly selected columns using `pyampute` [38]. `pyampute` employs Multivariate Amputation, a transformation method that generates missing values in complete datasets.

To evaluate the impact of imputation timing on classifier performance, we compared imputation before train/test split (IBS) and imputation after train/test split (IAS). In the IBS approach, missing values were first introduced to the complete dataset, and then imputation was performed before dividing the data into train and test sets. We then trained classifiers and performed testing and analysis. Conversely, in the IAS approach, missing values were introduced into the complete dataset, the data was split into train and test sets, and imputation was then performed separately on each set before classifier training and evaluation.

Each dataset was evaluated using both approaches (IBS and IAS). For handling missing values, we compared four imputation methods from `scikit-learn`: Mean Imputation [3], [27] (`SimpleImputer`), Random Forest Imputation [40] (`RandomForestRegressor`), KNN Imputation [5] (`KNNImputer`), and Multiple Imputation [3], [34] (`IterativeImputer`). In addition to these imputation methods, we also applied listwise deletion, where rows containing missing values were removed instead of imputing the values. Since no imputation occurs in this case, comparisons between IBS and IAS were not possible. Instead, performance estimates obtained via listwise deletion were analyzed solely across missing data patterns to assess the impact of data removal. Each dataset was split into 70% training and 30% testing data for each experiment.

Authorized licensed use limited to: Swarthmore College Library. Downloaded on November 04,2026 at 18:20:12 UTC from IEEE Xplore. Restrictions apply.

To evaluate the impact of different imputation methods on downstream machine learning tasks, we tested four classifiers: Logistic Regression [20], Random Forest [7], Support Vector Machine (SVM) [9], and Multi-Layer Perceptron (MLP) [18]. Hyperparameter tuning was performed via grid search, optimizing for accuracy, which served as the primary evaluation metric. To ensure robustness, each experiment was repeated five times, allowing for the reporting of mean and variance in performance estimates.

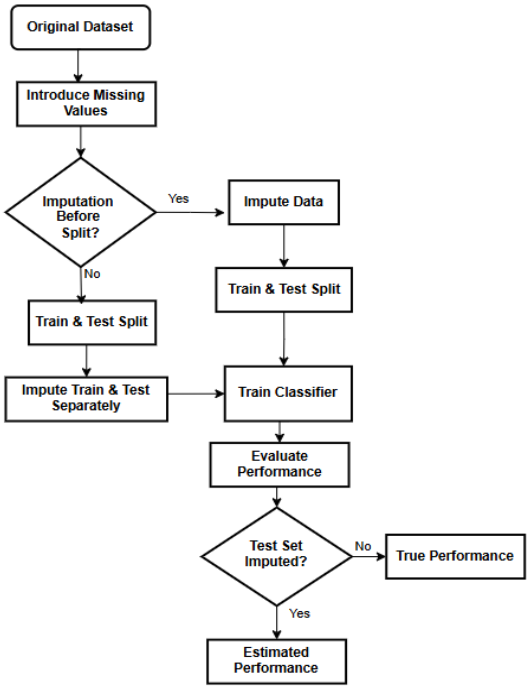


Fig. 1. Flowchart illustrating the experimental design

Baseline performance for each classifier was determined using the original dataset without any missing values. To examine discrepancies between *estimated* and *actual* generalization performance when imputed values were present in the dataset, we assessed estimated performance using the imputed test set and true performance by replacing missing values with their original values in the test set. We recognize that the latter approach is not feasible in real-world scenarios due to the unavailability of complete data, but it remains the only way to get a proper baseline. This distinction between estimated and true performance allows us to evaluate the impact of imputation on the correctness and robustness of model performance estimates.

B. Datasets

We selected three benchmark datasets, each containing both numerical and categorical features, with binary and multiclass target variables. The datasets were chosen to balance complexity and efficiency; they were large enough to provide meaningful insights, but not so large that training times became impractical given the number of experiments we planned to run. Additionally, to accurately compare estimated performance with true performance, we required datasets with complete ground truth, free of missing values. All three datasets were sourced from the UCI Machine Learning Repository [11].

To prepare the datasets for analysis, we applied several preprocessing steps consistently across all three datasets. Irrelevant or redundant features (e.g. unique IDs) were removed based on domain knowledge or exploratory analysis. Categorical features were converted into numerical format using label encoding, where each unique category value is assigned an integer. For example, a binary feature would be label-encoded by mapping values to integers (e.g., 'yes' to 1, 'no' to 0); a feature with three values would map to the integers $[0, 1, 2]$, etc. We chose label encoding over one-hot encoding to keep the feature dimensionality low, especially since many categorical variables had only two or three levels and we were primarily focused on evaluating imputation strategies rather than modeling categorical relationships.

The first dataset, the Student Performance dataset [10], consists of 649 instances and 30 features, with the target variable `Dalc` representing daytime alcohol consumption. We preprocessed this dataset by removing irrelevant features, encoding categorical variables into numeric codes, and converting `Dalc` into a binary class (values ≤ 1 as 0, and values > 1 as 1), resulting in 451 instances labeled as 0 and 198 labeled as 1.

The second dataset, the Indian Liver Patient Dataset [32], includes 583 patient records with 11 features used to determine the presence of liver disease. We preprocessed this dataset by removing irrelevant features, resulting in 416 records of participants with liver disease and 167 records of patients without liver disease. Additionally, we encoded all categorical variables into numeric codes.

The third dataset, the Maternal Health Risk Dataset [1], contains 1,013 instances and 6 features, with the target variable `RiskLevel` predicting the risk intensity during pregnancy. The `RiskLevel` variable has 406 instances labeled as 1 (`low risk`), 336 as 2 (`mid risk`), and 272 as 0 (`other levels`).

III. RESULTS

Our results showed distinct trends in performance across the two experimental settings: imputation before train/test split (IBS) and imputation after train/test split (IAS). The baseline performance, representing the classifier's accuracy on the original dataset without missing values, varied depending on the dataset and classifier, but the pattern of behavior between IBS and IAS in relation to the baseline performance remained consistent.

A. Comparison For IBS and IAS In Relation To Baseline

Overall, the *estimated* performance for IBS improved as the proportion of missing data increased and remained consistently above the baseline at all levels (see Figure 2). In contrast, for IAS, where imputation was applied separately to the training

Authorized licensed use limited to: Swarthmore College Library. Downloaded on November 04,2026 at 18:20:12 UTC from IEEE Xplore. Restrictions apply.

and test sets after the split, the estimated performance fell, and declined further as the missing data fraction increased. While the degree of estimated performance improvement or decline varied by dataset, within any given dataset this pattern held across all classifiers, imputation methods, and missing data patterns. Notably, for Random Forest, Iterative, and KNN imputation methods, there was no substantial difference in performance trends across the three missing data patterns (MCAR, MAR, MNAR), which was surprising. However, differences were observed for mean imputation and row-wise deletion, which are discussed in the later sections.

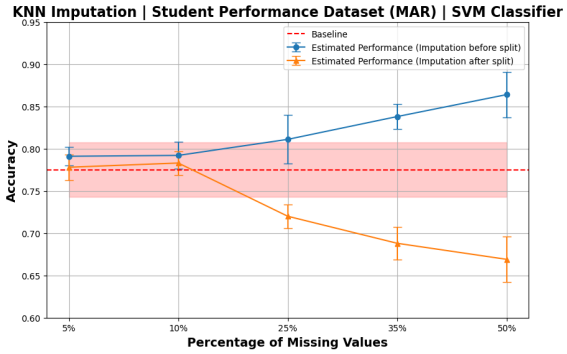


Fig. 2. Accuracy of a Random Forest classifier on the Student Performance Dataset (MAR) using KNN imputation. The graph compares true and estimated performance when imputation is applied before the train-test split (IBS) and after the split (IAS) across different missingness levels. IBS overestimates accuracy (blue), while IAS underestimates it (orange), with the gap widening as missingness increases. The baseline (dashed red line) represents the accuracy without missing data, and the red shaded region indicates the variance in baseline performance. The error bars represent the variance in performance estimates at each level of missingness.

B. Comparison For IBS and IAS In Relation To True Performance

True performance, measured on the testing set where missing values were replaced with their original data values instead of imputed ones, generally declined as the proportion of missing data increased (see Figure 3). Both IBS and IAS followed similar patterns, with only a slight difference in accuracy between them across all levels of missing values. In most cases, true performance remained within one standard deviation of the baseline, except at 50 percent missing values, where it occasionally fell within two standard deviations. These trends were consistent across all classifiers, imputation methods, and missing data patterns. As the fraction of missing values increased, the variance in performance also generally increased. This rise in variance was observed in both IBS and IAS settings, indicating reduced stability of performance estimates under higher missingness. Keeping the classifier, imputation method, and missing data pattern constant between IBS and IAS, there was no meaningful difference in performance variance. The size of the dataset also did not appear to influence variance. For example, the average variance for the Liver dataset (583 instances), using the SVM classifier and KNN imputation under IBS across all levels of missingness, was only 0.002. In contrast, under the same settings, the Maternal Health dataset (1,013 instances) had an average variance of 0.008, suggesting that factors beyond dataset size, such as feature distribution or class balance, may have a stronger impact on performance variability.

Comparing true performance with estimated performance, IBS consistently overestimated performance, with the gap widening as the fraction of missing values increased. For example, in the student performance dataset with 35% missing data, using the Random Forest Classifier and KNN imputation, the estimated performance in the IBS setting was, on average, 8% higher than the true performance, indicating a clear trend of overestimation. In contrast, IAS showed the opposite pattern. The estimated performance in IAS was consistently lower than the true performance and continued to decline as missing values increased. For instance, with 35% missing data using RF imputation and Random Forest Classifier on the same dataset, the estimated performance in the IAS setting was, on average, 6% lower than the true performance, reflecting a trend of underestimation.

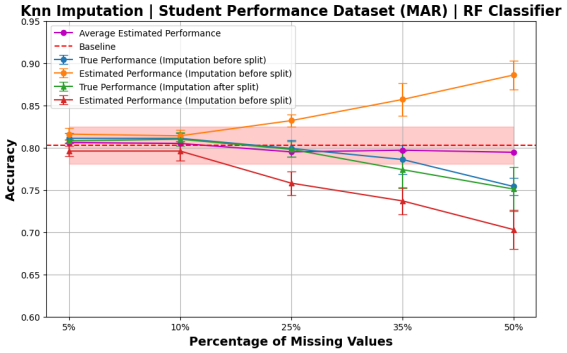


Fig. 3. Performance estimates of the RF classifier for the Student Performance dataset under the MAR missing data pattern with KNN Imputation. The red shaded region represents the variance in baseline performance. The error bars represent variance in performance estimates at each level of missingness. While IBS and IAS exhibit opposite trends in estimated performance, their deviations are symmetric around the true performance. The purple line represents the averaged estimated performance from both IBS and IAS, which may offer a more balanced estimate, reducing the effects of overestimation and underestimation.

In both IBS and IAS, the difference between estimated and true performance increased as the fraction of missing values grew. Similarly, higher levels of missing data were associated with greater variance, and the extent of performance overestimation or underestimation varied by dataset. Consistent trends were observed across all classifiers and the three main imputation methods (RF, KNN, and multiple imputation). Notable differences appeared only when applying mean imputation and listwise deletion.

C. Results For Listwise Deletion

Performance under listwise deletion showed only slight variations as the fraction of missing values increased. Generally, accuracy remained within one standard deviation of the baseline for up to 35% missing data and within two

Authorized licensed use limited to: Swarthmore College Library. Downloaded on November 04,2026 at 18:20:12 UTC from IEEE Xplore. Restrictions apply.

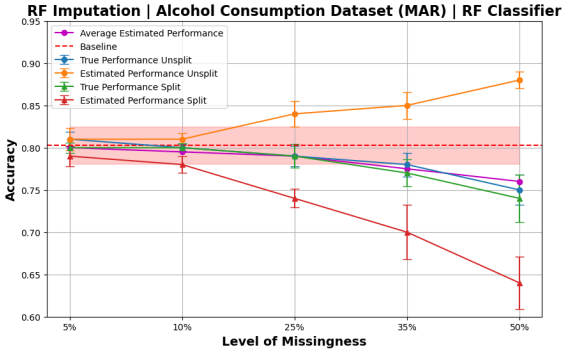


Fig. 4. Performance estimates of RF classifier for Student Performance dataset under the MAR missing data pattern with RF Imputation. The red shaded region represents the variance in baseline performance.

standard deviations for 35%-50% missing data. However, under the MAR pattern, variance was higher and less stable compared to MCAR and MNAR. Despite this, trends largely stayed within two standard deviations of baseline performance, except in rare cases with 50% missing data, where deviations extended to three standard deviations. Overall, performance declined as missing values increased, aligning with trends in true performance observed for both IBS and IAS, with most declines remaining within one standard deviation of the baseline accuracy.

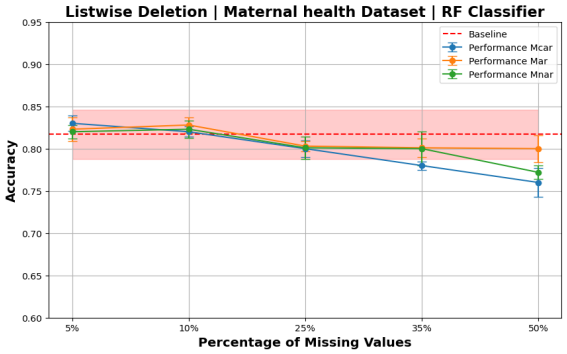


Fig. 5. Performance estimates of the RF classifier for the Maternal Health dataset under different missing data patterns with listwise Deletion. Performance remained largely consistent across the three missing data patterns and stayed within one standard deviation of the baseline for missing rates up to 35

D. Results For Mean Imputation

Performance trends with mean imputation were inconsistent, varying by dataset and classifier. In most cases, performance either showed a slight downward trend as missing values increased or remained largely unchanged, even at higher levels of missing data. For example, in the liver disease dataset with the RF classifier, estimated and true performances for both IBS and IAS remained within one standard deviation of the baseline, showing no significant change. In contrast, in the student performance dataset with the RF classifier, estimated performance degradation for IBS and IAS exceeded two standard deviations from the baseline at 35% missing values and three standard deviations at 50%. True performance also declined but less drastically, staying within two standard deviations. Unlike other imputation methods, which led to overestimation in IBS, mean imputation did not exhibit this pattern. Instead, estimated performances for IBS and IAS followed a similar downward trajectory, closely aligning with each other.

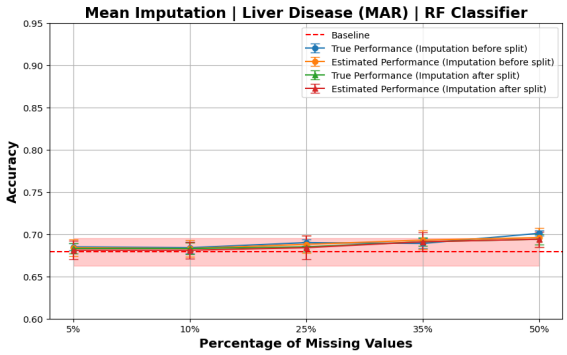


Fig. 6. Performance estimates of the RF classifier for the Liver Disease dataset under the MAR missing data pattern with Mean Imputation. The red shaded region denotes the variance in baseline performance. At all levels of missingness, the estimated and true performance for both IBS and IAS remained within one standard deviation of the baseline.

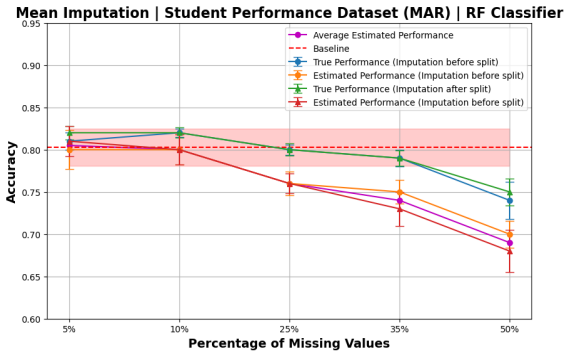


Fig. 7. Performance estimates of the RF classifier for the Student Performance dataset under the MAR missing data pattern with Mean Imputation. The red shaded region indicates the variance in baseline performance. Unlike in the case of the Liver Disease dataset, trends show that true performance for both IBS and IAS remains mostly within one standard deviation of the baseline, while estimated performance deviates beyond two standard deviations at higher missing value rates. Unlike other imputation methods, the average estimated performance (purple line) from IBS and IAS did not provide a good estimate of the true performance.

IV. DISCUSSION

The observed differences in performance trends when imputing before or after the train-test split highlight the potential for bias in performance estimates. When imputation is performed before splitting the data, the model can learn patterns from the imputed values that appear in both the training and

Authorized licensed use limited to: Swarthmore College Library. Downloaded on November 04,2026 at 18:20:12 UTC from IEEE Xplore. Restrictions apply.

test sets. This overlap results in a form of data leakage, artificially inflating performance metrics and introducing an optimistic bias, the severity of which increases with the proportion of missing values. This result helps to quantify an effect that has been noted elsewhere; Kapoor and Narayanan [22] similarly emphasize that pre-processing the entire dataset, including imputation, can lead to performance overestimation due to data leakage, which disrupts the intended separation between the training and test sets. Rosenblatt et al. (2024) [44] note that data leakage, such as conducting feature selection before splitting the data or including the same subjects in both the training and test sets, can significantly inflate prediction accuracy in connectome-based models.

When imputation is performed separately on the training and test sets, the imputed values in the test set do not affect the model during training. This approach is intended to provide a more realistic evaluation of model performance by avoiding data leakage, as the test set remains an independent measure of the model's generalization ability. However, our experiments show that as the fraction of missing values increases, the quality of the imputation on the test set decreases, leading to a decline in estimated performance relative to actual performance. Thus, while separate imputation in the train and test set can mitigate the optimism bias, it can introduce a countervailing bias that can be just as misleading about the model's true quality.

If we view the process of generating imputed values as an ML modeling problem, this effect can be seen as a version of the classic bias-variance dilemma, in this case applied to the *imputation* model, not the *classification* model. In a case where the imputation model has a strong inductive bias, such as mean imputation, the imputation model will tend to *underfit* the distribution of missing values. For MCAR data, we would expect an underfitting imputation model to produce largely similar results regardless of variation in which examples are used (i.e. train set or validation set). At the same time, an underfitting model is likely to skew the distribution of imputed values significantly away from the 'true' distribution of values for the missing feature(s). This should result in lowered *actual* performance of the classification model, but should tend to minimize differences between IBS and IAS performance estimates. Our results using mean imputation conform well to this theory.

On the other hand, if the imputation model has a weak inductive bias, it is likely to *underfit* the true distribution of missing values. In this case, for MCAR data we would expect most of the error in estimated values to be variance-type error, and thus the distributions estimated by learning the model on different samples (i.e. training vs validation) are likely to be significantly different. In the IAS case, the overfitting imputation model will skew the training and validation sets in different directions, with the magnitude of the shift based on the amount of missing values. As the training and validation sets become more and more dissimilar, the performance of a classification model on the validation set will grow worse and worse. In the IBS case, since the *same* overfitting imputation model is applied to both sets, the distributions will be shifted closer and closer together (a manifestation of the data leakage problem), resulting in classifier performance appearing to get better and better. In both IBS and IAS, the training set will be shifted away from the 'true' distribution by a similar amount, so the 'actual' generalization performance of the model will tend to decrease, but by a much less extreme amount. Our results using the more complex imputation models conform well to this theory.

It is worth note that this theory based explanation of the behavior we observe applies most naturally to the case of MCAR data, in which values are missing in a completely random fashion. If there are patterns in the missingness, then we would expect to see larger amounts of systematic bias-type error from *any* type of imputation model, simply because whatever errors it makes are no longer distributed at random across the set of examples. In general, this should not impact the overall trends if we rely on accuracy as our metric, since the same phenomena of over- and under-fitting will apply regardless of whether the missing values are distributed uniformly or not. However, it is important to recognize that 'overall trends' tend to be less robustly representative in situations like this. While it was not the focus of this study, we view the closer inspection of distributional shifts in non-MCAR missingness as an important direction for future work.

Our analysis shows that improperly managed imputation can destabilize performance estimates, similar to how minor distribution shifts do [28] [24]. This underscores the importance of careful handling of imputation to prevent amplifying the effects of such shifts, ensuring models remain robust and reliable under realistic evaluation scenarios. Moreover, this finding ties into the broader issue of the reproducibility crisis in machine learning [22], as it demonstrates how different imputation strategies and the timing of imputation can yield conflicting results, complicating efforts to reproduce and compare findings across studies. Similarly, Rosenblatt et al. (2024) [44] underscore that leakage effects can vary in severity depending on the data size and modeling choices, and that small datasets are particularly vulnerable to inflated performance due to leakage. This observation aligns closely with our findings on imputation bias under high missingness, though data set size did not have as significant an impact in our experiments.

Since IBS tends to overestimate performance and IAS underestimates it, averaging performance estimates from both approaches appears promising. This averaged estimate better aligns with true performance, suggesting it could help mitigate the biases introduced by different imputation timings. However, its effectiveness must be validated across diverse datasets, classifiers, and imputation methods to confirm its general applicability.

In comparing various imputation methods, we found no substantial difference in performance trends, except for mean imputation, which while being more robust to leakage, consistently underperformed in terms of true model performance. This finding aligns with previous studies, such as Cihan et al.

Authorized licensed use limited to: Swarthmore College Library. Downloaded on November 04,2026 at 18:20:12 UTC from IEEE Xplore. Restrictions apply.

[8] and Farhangfar et al. [14], who reported that mean imputation often leads to higher error rates and lower classification accuracy compared to more sophisticated methods like KNN, random forest, or multiple imputation. These results suggest that mean imputation may introduce bias that negatively impacts classifier performance, especially in datasets with higher missing rates. Furthermore, Alam et al. (2023) [46] found that mean imputation not only reduced accuracy but also resulted in the highest variance across performance metrics, indicating its tendency to destabilize the data structure and introduce inconsistency across missingness levels.

Moreover, our analysis supports the notion that the missing rate and dataset characteristics have a more significant influence on performance estimates than the choice of imputation method. This observation aligns with the findings of Platias & Petasis [30] and Shadbahr et al. [39], who documented a consistent decline in classifier performance as the missing rate increased, regardless of the imputation method used.

Finally, while methods like listwise deletion were not the focus of our study, the consistent performance within a standard deviation of the baseline observed in our experiments mirrors findings from Pan and Chen [29]. They found that certain imputation methods, including regression-based and hot deck, did not outperform listwise deletion, suggesting that the benefits of imputation may be dataset-specific. Ren et al. [33] also noted that machine learning-based imputation methods sometimes yielded slightly better classification accuracy than listwise deletion, further emphasizing the context-dependent nature of imputation effectiveness.

V. LIMITATIONS

Introducing missing data synthetically is a common approach in studies focused on evaluating imputation techniques. However, synthetic missingness can differ significantly from missing data encountered in real-world settings. While this difference may be minimal for MCAR, or Missing Completely At Random, scenarios where randomly selecting data points to remove closely mirrors the underlying assumption, it becomes more problematic in the case of MAR, or Missing At Random, and especially MNAR, or Missing Not At Random, where the missingness is dependent on observed or unobserved values. Synthetic missing data is typically introduced in a systematic and uniform way, such as by removing a fixed percentage of values from each variable or record. In contrast, real-world missingness often results from more complex and heterogeneous causes. These can include structural factors, such as certain medical tests not being performed for specific patient subgroups, contextual influences, or unpredictable occurrences. As a result, the patterns of missing data found in natural settings are often irregular and more challenging to model.

While our findings are consistent across a range of different scenarios, their generalizability to datasets, imputation techniques, or machine learning models beyond those tested in this study remains uncertain. Of particular note, the data sets used were relatively small by the standards of some data domains, and more work needs to be done to evaluate how these phenomena scale in big-data settings. Additionally, the study primarily focuses on accuracy as the evaluation metric; while this is common in work studying imputation techniques, other metrics (e.g. F1, confusion matrices, etc.) may be more appropriate in some domains, particularly in the presence of imbalanced classes or varying levels of missing data. The study excludes more computationally expensive imputation and modeling techniques, such as deep learning-based methods, which could potentially yield different results. Finally, the proposed averaging strategy for performance estimates from IBS and IAS, while promising, has not been fully validated across different scenarios, and there is a risk that it may not perform as effectively in datasets with different structures or missing data patterns.

VI. CONCLUSION

Handling missing values is a critical step in machine learning pipelines, yet the timing of imputation remains an overlooked factor that can significantly influence performance estimation. Our study systematically examined the effects of imputing missing values before versus after the train-test split (IBS vs. IAS) across multiple datasets, classifiers, and imputation methods. By artificially introducing missing values under MCAR, MAR, and MNAR patterns at varying rates, we assessed how different imputation strategies impact both estimated and true performance.

Our results demonstrated that IBS consistently led to overestimation, with performance estimates exceeding true performance as missingness increased. In contrast, IAS tended to underestimate performance, with estimates declining below true performance. Given the countervailing trends shown in these two settings, we propose that performing both and averaging the results may result in a better estimate of the true model generalization performance.

While these broad trends were consistent across datasets and classifiers, the *degree* of overestimation or underestimation varied depending on the imputation method. RF, KNN, and multiple imputation followed predictable patterns, whereas mean imputation exhibited more erratic behavior, and listwise deletion aligned closely with true performance but resulted in greater data loss.

Importantly, while IBS and IAS produced divergent performance estimates, at lower levels of missingness true performance remained relatively stable, generally staying within one standard deviation of the baseline. However, at extreme missingness levels, performance degradation became more pronounced. These findings underscore the importance of carefully considering imputation timing when evaluating model performance, as overlooking this factor can lead to misleading conclusions about a classifier's generalization ability. Future work should explore additional imputation techniques, alternative performance metrics, and real-world applications to better understand the broader implications of imputation strategies in machine learning.

Authorized licensed use limited to: Swarthmore College Library. Downloaded on November 04,2026 at 18:20:12 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] M. Ahmed, "Maternal Health Risk [Dataset]," UCI Machine Learning Repository, 2020.

[2] S. Alam, G. Dobbie, and X. Sun, "Improving imputation accuracy in ordinal data using classification," in *Intelligent systems design and applications. ISDA 2016. Advances in intelligent systems and computing*, vol. 557, pp. 57–67, Springer, Cham, 2017.

[3] P. D. Allison, *Missing Data*, Sage Publications, Inc., 2001.

[4] O. F. Ayilara, L. Zhang, T. T. Sajobi, R. Sawatzky, E. Bohm, and L. M. Lix, "Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry," *Health Qual. Life Outcomes*, vol. 17, no. 1, p. 106, 2019.

[5] G. E. A. P. A. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," *Applied Artificial Intelligence*, vol. 17, no. 5, pp. 519–533, 2003.

[6] D. Bertsimas, C. Pawlowski, and Y. D. Zhuo, "From predictive methods to missing data imputation: An optimization approach," *J. Mach. Learn. Res.*, vol. 18, no. 196, pp. 1–39, 2018.

[7] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[8] P. Cihan, O. Kalıpsız, and E. Gökçe, "Effect of imputation methods in the classifier performance," *Sakarya Univ. J. Sci.*, vol. 23, no. 6, pp. 1225–1236, 2019.

[9] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[10] P. Cortez, "Student Performance [Dataset]," UCI Machine Learning Repository, 2008.

[11] D. Dua and C. Graff, "UCI machine learning repository," University of California, Irvine, 2017.

[12] Y. Duan, Y. Lv, W. Kang, and Y. Zhao, "A deep learning based approach for traffic data imputation," in *Proc. 17th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, 2014, pp. 912–917.

[13] Y. Dong and C.-Y. J. Peng, "Principled missing data methods for researchers," *SpringerPlus*, vol. 2, p. 222, 2013.

[14] A. Farhangfar, L. Kurgan, and J. Dy, "Impact of imputation of missing values on classification error for discrete data," *Pattern Recognit.*, vol. 41, no. 12, pp. 3692–3705, 2008.

[15] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: A review," *Neural Comput. Appl.*, vol. 19, no. 2, pp. 263–282, Mar. 2010. [Online]. Available: https://doi.org/10.1007/s00521-009-0295-6.

[16] S. Ghorbani and M. C. Desmarais, "Performance comparison of recent imputation methods for classification tasks over binary data," *Appl. Artif. Intell.*, vol. 31, no. 1, pp. 1–22, 2017.

[17] S. Greenland and W. D. Finkle, "A critical look at methods for handling missing covariates in epidemiologic regression analyses," *American Journal of Epidemiology*, vol. 142, no. 12, pp. 1255–1264, 1995.

[18] S. Haykin, *Neural networks: A comprehensive foundation*, 2nd ed. Prentice Hall, 1999.

[19] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, pp. 7776–7797, 2017.

[20] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*, 3rd ed. Wiley, 2013.

[21] S. Jäger, A. Allhorn, and F. Bießmann, "A benchmark for data imputation methods," *Front. Big Data*, vol. 4, p. 693674, 2021.

[22] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, Article 100804, 2023.

[23] J.-S. Kim, X. Gao, and A. Rzhetsky, "RIDDLE: Race and ethnicity imputation from disease history with deep learning," *PLOS Comput. Biol.*, vol. 14, no. 4, e1006106, 2018.

[24] T. Kyono, Y. Zhang, A. Bellot, and M. van der Schaar, "MIRACLE: Causally-aware imputation via learning missing data mechanisms," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[25] R. A. Little, "Regression with missing X's: A review," *Journal of the American Statistical Association*, vol. 87, no. 420, pp. 1227–1237, 1992.

[26] R. J. Little, R. D'Agostino, M. L. Cohen, K. Dickersin, S. S. Emerson, J. T. Farrar, C. Frangakis, and H. Stern, "The prevention and treatment of missing data in clinical trials," *The New England Journal of Medicine*, vol. 367, no. 14, pp. 1355–1360, 2012.

[27] R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*, John Wiley & Sons, 1987.

[28] E. S. Nordholt, "Imputation: Methods, simulation experiments, and practical examples," *International Statistical Review*, vol. 75, no. 1, pp. 103–114, 2007.

[29] S. Pan and S. Chen, "Empirical comparison of imputation methods for multivariate missing data in public health," *Int. J. Environ. Res. Public Health*, vol. 20, no. 2, p. 1524, 2023.

[30] C. Platias and G. Petasis, "A comparison of machine learning methods for data imputation," in *Proc. SETN 2020: 11th Hellenic Conf. Artif. Intell.*, 2020, pp. 150–159.

[31] J. Poulos and R. Valle, "Missing data imputation for supervised learning," *Appl. Artif. Intell.*, vol. 32, no. 2, pp. 186–196, 2018.

[32] B. Ramana and N. Venkateswarlu, "ILPD (Indian Liver Patient Dataset) [Dataset]," UCI Machine Learning Repository, 2022.

[33] L. Ren, T. Wang, A. Sekhari Seklouli, H. Zhang, and A. Bouras, "Missing values for classification of machine learning in medical data," in *Proc. 2022 IEEE 5th Int. Conf. Artif. Intell. Big Data (AIBD)*, 2022.

[34] D. B. Rubin, *Multiple imputation for nonresponse in surveys*, John Wiley & Sons, 1987.

[35] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[36] J. L. Schafer, *Analysis of incomplete multivariate data*, Chapman & Hall/CRC, 1997.

[37] Scikit-learn developers, "Imputation of missing values," *Scikit-learn*, accessed Mar. 27, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/impute.html.

[38] R. M. Schouten, D. Zamanzadeh, and P. Singh, "pyampute: A Python library for data amputation," 2022.

[39] T. Shadbahr, M. Roberts, J. Stanczuk, J. Gilbey, P. Teare, S. Dittmer, M. Thorpe, R. Viñas Torné, E. Sala, P. Liò, M. Patel, J. Preller, AIX-COVNET Collaboration, J. H. F. Rudd, T. Mirtti, A. S. Rannikko, J. A. D. Aston, J. Tang, and C.-B. Schönlieb, "The impact of imputation quality on machine learning classifiers for datasets with missing values," *Commun. Med.*, vol. 3, Article 139, 2023.

[40] D. J. Stekhoven and P. Bühlmann, "MissForest—Non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.

[41] X. Su, T. M. Khoshgoftaar, and R. Greiner, "Using imputation techniques to help learn accurate classifiers," in *Proc. 2008 20th IEEE Int. Conf. Tools Artif. Intell.*, 2008, pp. 156–161.

[42] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[43] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997.

[44] M. Rosenblatt, L. Tejavibulya, R. Jiang *et al.*, "Data leakage inflates prediction performance in connectome-based machine learning models," *Nature Communications*, vol. 15, no. 1829, 2024. doi:10.1038/s41467-024-46150-w.

[45] K. John, D. D. Saurette, and B. Heung, "The problematic case of data leakage: A case for leave-profile-out cross-validation in 3-dimensional digital soil mapping," *Geoderma*, vol. 455, p. 117223, 2025. doi:10.1016/j.geoderma.2025.117223.

[46] S. Alam, M. S. Ayub, S. Arora, and M. A. Khan, "An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity," *Decision Analytics Journal*, vol. 9, p. 100341, 2023. doi:10.1016/j.dajour.2023.100341.

Authorized licensed use limited to: Swarthmore College Library. Downloaded on November 04,2026 at 18:20:12 UTC from IEEE Xplore. Restrictions apply.