https://doi.org/10.31891/csit-2026-2-21

Volodymyr SLIPCHENKO
Doctor of Engineering Sciences, Professor, Professor at the Department of Digital Technologies in Energy,
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
e-mail: ddpolytechnic2016@gmail.com
https://orcid.org/0000-0002-3405-0781

Liubov POLIAHUSHKO
Candidate of Engineering Sciences, Associate Professor, Associate Professor at the Department of Digital Technologies in Energy,
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
e-mail: liubovpoliagushko@gmail.com
https://orcid.org/0000-0003-3287-8523

Oleksandr VOLKOV
PhD Student of the Department of Digital Technologies in Energy,
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
e-mail: volkov1aleksander@gmail.com
https://orcid.org/0009-0003-6834-8118

Received: 12/04/2026
Accepted: 02/05/2026
Published: 31/05/2026

UDC 004.8:519.24:616-071

COMPARATIVE ANALYSIS OF MISSING DATA IMPUTATION METHODS IN BIOMEDICAL RESEARCH: IMPACT ON BIOLOGICAL AGE PREDICTION

Missing data remain a major challenge in biomedical research because they can bias statistical estimates, reduce predictive accuracy, and compromise the robustness of scientific conclusions. The present study provides a comparative evaluation of five imputation approaches: IterativeImputer with RandomForest, ExtraTrees, and BayesianRidge estimators, together with KNNImputer and median-based SimpleImputer. The methods were assessed on two biomedical datasets, Bones (3,285 records, 11 biomarkers, n/p = 299) and NHANES (11,016 records after reduction from 55,081, 85 biomarkers, n/p = 130), with an n/p gradient ranging from 19 to 299. The experimental design incorporated three missingness mechanisms, MCAR, MAR, and MNAR, and three missingness levels: 10%, 40%, and 80%. Imputation quality was quantified using RMSE, while downstream effects were examined through biological age prediction based on ElasticNet and PCA models. IterativeImputer with ExtraTrees achieved the lowest average RMSE (9.275), whereas BayesianRidge and RandomForest demonstrated the strongest average rank (2.19-2.20), indicating more stable overall performance across heterogeneous scenarios. Under MNAR conditions, RandomForest produced the best results (RMSE 10.896), while ExtraTrees was most effective for MAR (RMSE 8.704). Downstream analysis showed that PCA yielded lower prediction RMSE than ElasticNet (2.14 versus 5.86), although 34% of cases exhibited negative correlations. A paradoxical improvement in imputation quality with increasing missingness was observed in 55-75% of scenarios. Median imputation was the fastest method (0.0075 s), whereas RandomForest was the slowest (261 s). The findings support practical recommendations for selecting imputation strategies according to dataset structure, missingness mechanism, and computational constraints in biomedical applications.

Keywords: data imputation, missing data, biological age, machine learning, MCAR, MAR, MNAR.

Introduction

Missing data are a common issue in biomedical datasets due to participant non-response, technical failures, sample loss, and budget limitations. Improper handling of missing values may lead to biased estimates and reduced statistical power; therefore, the choice of an imputation strategy is a crucial stage of data preprocessing [1, 2, 3, 4].

This study considers the applied task of biological age prediction based on biomarkers, where missing values can substantially affect both the accuracy and the interpretability of downstream models. Particular attention is given to the three missingness mechanisms, namely MCAR, MAR, and MNAR, since they determine the risk of systematic bias [14].

Literature Review and Problem Statement

Missing data is a common problem in biomedical research, and their improper handling can introduce systematic bias and reduce the statistical power of the analysis [1, 2]. The fundamental classification of missingness mechanisms, namely MCAR, MAR, and MNAR, determines which imputation strategies are statistically valid [4, 14]. Under MCAR, complete-case analysis remains unbiased, although less efficient; under MAR, properly specified imputation models yield unbiased results; under MNAR, eliminating bias is fundamentally impossible without explicitly modeling the missingness process. Since it is impossible to distinguish MAR from MNAR based solely on the observed data, it is recommended to assume MAR and to perform sensitivity analysis [1].

Classical approaches, such as the deletion of incomplete observations or median substitution, remain the most used in practice despite their well-documented limitations, including underestimated variance, weakened correlations between variables, and overly narrow confidence intervals [4]. Multiple imputation, particularly the MICE algorithm [7], is generally regarded as the methodological gold standard. This approach iteratively constructs regression models for each variable containing missing values. Azur et al. [8] provided a detailed description of the practical conditions required for the proper application of MICE, emphasizing the importance of model specification and sensitivity analysis. In parallel, Jakobsen et al. [5] developed step-by-step guidelines for choosing among different imputation strategies in clinical research.

Among machine learning based methods, the work of Stekhoven and Bühlmann [15], who proposed the MissForest algorithm, is of particular importance. MissForest is an iterative imputation method based on ensembles of decision trees and has been shown to outperform MICE and kNN on mixed-type data even for relatively small sample sizes. Beretta and Santaniello [9] provided a critical evaluation of kNN imputation, highlighting its sensitivity to the choice of distance metric in high-dimensional spaces. Using cohort medical data, Li et al. [3] demonstrated that there is no universally best method, as the performance of a particular approach depends on the dependency structure among variables and on the missingness mechanism. A fundamentally important conclusion was drawn by Madley-Dowd et al. [17]: the proportion of missing data alone cannot serve as a criterion for method selection, since the determining factor is the mechanism by which the missing data arise.

A separate line of research concerns the downstream effect of imputation, that is, its impact on the quality of subsequent modelling, which is not evaluated in most comparative studies. This issue is particularly relevant to biological age prediction. Jylhävä et al. [12] and Belsky et al. [13] showed that clinical biomarkers are sensitive predictors of morbidity, while Salgado et al. [6] emphasized that imputation quality affects not only point estimates but also the ranking of predictors within models. Regarding neural network-based imputation approaches, Liang et al. [16] and Antonenko et al. [11] demonstrated the potential of hybrid methods, Casella et al. [10] showed the applicability of transformer-based models for the imputation of psychometric scales, and the authors' previous study [28] investigated hybrid imputation of biomedical data based on transformers and autoencoders for biological age estimation. However, the performance of such approaches under limited sample sizes and different missingness mechanisms, as well as their comparison with classical methods, remains insufficiently studied.

Thus, the literature lacks comparative studies that systematically examine the behavior of iterative imputers with different base regressors across a wide n/p gradient (from 19 to 299) under three missingness mechanisms, with a quantitative evaluation of the downstream effect on biological age prediction. Addressing this gap is the aim of the present study.

Aim and Objectives of the Study

The aim of this study was to perform a comprehensive comparison of five missing data imputation methods in biomedical research and to evaluate their impact on biological age prediction under different missingness mechanisms and levels of missing data.

To achieve this aim, the following objectives were defined:

1. To implement and test five imputation methods, namely IterativeImputer with RandomForest, ExtraTrees, and BayesianRidge estimators, as well as KNNImputer and SimpleImputer with median imputation, on two datasets of different dimensionality.
2. To evaluate the primary imputation quality using RMSE, MAE, and R^2 under three missingness mechanisms, namely MCAR, MAR, and MNAR, and at three levels of missing data, namely 10%, 40%, and 80%.
3. To investigate the downstream effect of the imputation methods on biological age prediction using two approaches, namely ElasticNet and PCA.
4. To analyze the trade-off between imputation accuracy and computational efficiency in terms of execution time.
5. To develop practical recommendations for selecting an imputation method depending on data characteristics and study constraints.

Materials and Methods

Two datasets of different dimensionality were used in this study. Together, they form a gradient of the observation-to-feature ratio (n/p) and make it possible to analyze the behavior of imputation methods under different levels of statistical complexity.

The Bones dataset [20] contains 3,285 records and 11 biomarkers, including 10 variables used for imputation and age as the target variable. The n/p ratio in the full dataset is 299, while a subsample of 493 records (15% of the data) has an n/p ratio of 45, representing a regime of moderate statistical reliability.

The second data source was the NHANES dataset [21], which contains 55,081 records and 85 biomarkers, including age. To optimize the analysis, the dataset size was reduced to 11,016 records by random sampling, corresponding to 20% of the original data. In this reduced version, 84 biomarkers were used for imputation, while age served as the target variable, yielding an n/p ratio of 130. In addition, a subsample of 1,652 records (15%) was created, in which the n/p ratio decreased to 19, reflecting a near-limit analytical regime.

This combination of datasets and sampling variants makes it possible to systematically evaluate the stability, accuracy, and computational efficiency of the imputation methods under conditions that approximate real biomedical and epidemiological studies.

Five imputation methods from the scikit-learn library [18] were analyzed in this study. IterativeImputer with RandomForest (iter_rf) and IterativeImputer with ExtraTrees (iter_et) model missing values as functions of the other features using ensembles of decision trees; for both methods, the following parameters were used: max_iter=20, n_estimators=100, max_depth=10, random_state=42. IterativeImputer with BayesianRidge (iter_bayes) uses a regularized linear model (max_iter=20, random_state=42) and serves as a substantially faster alternative to ensemble based methods. KNNImputer reconstructs missing values as the mean of the 5 nearest neighbors based on Euclidean distance. SimpleImputer with median imputation was included as a baseline method for comparison.

To simulate missing values, three missingness mechanisms were considered: MCAR, in which the probability of missingness does not depend on the data; MAR, in which it depends on the observed variables; and MNAR, in which it depends on the missing values themselves [14]. For each mechanism, missing data were generated at three levels: 10%, 40%, and 80%.

The experiment was based on a factorial design comprising twelve scenarios: two datasets (Bones and NHANES) × two sample sizes (full dataset and subsample) × three missingness mechanisms × three missingness levels. For each combination, five imputation methods were applied, resulting in a total of 180 runs.

Imputation quality was evaluated using RMSE, MAE, and R^2, while execution time was additionally recorded as an indicator of computational efficiency. To assess the downstream effect on biological age prediction, Pearson correlation, RMSE, and MAE were used with two models: ElasticNet regression (α=0.1, ρ=0.5, random_state=42) and PCA with automatic selection of the number of components. The statistical analysis included mean values and standard deviations, method ranking within each scenario, 95% confidence intervals based on 1,000 bootstrap resamples, Pareto analysis of the quality-time trade-off, and linear regression to estimate the effect of missingness level on the RMSE of each method.

Implementation and Testing of Five Imputation Methods on Datasets of Different Dimensionality

At this stage of the study, the correctness of implementation and the baseline performance of five imputation methods were evaluated on two datasets with contrasting dimensionality and structure, namely Bones and NHANES. The aim was to assess the overall quality of missing value reconstruction, the stability of the methods across different scenarios, and their computational efficiency.

Table 1.

Overall performance of the imputation methods (averaged across all scenarios)

Method	Mean RMSE	Mean Rank	Number of Wins	Near-best (5%)	Near-best (10%)	Mean Execution Time (s)
iter_et	9.275	2.528	4	47.2%	83.3%	83.12
iter_rf	9.368	2.200	15	51.4%	77.1%	260.99
median	9.910	3.972	2	–	–	0.0075
iter_bayes	10.241	2.194	15	61.1%	80.6%	2.415
knn	10.469	4.028	0	–	–	7.48

An aggregated analysis of all 180 experiments made it possible to obtain a generalized view of the performance of the imputation methods, which is presented in Table 1.

The obtained results confirm that all five methods were correctly implemented and successfully tested on both datasets. The lowest mean RMSE was achieved by iter_et (9.275), whereas iter_rf and iter_bayes showed the best mean rank (approximately 2.2) and the highest number of wins, with 15 out of 36 possible scenarios each. The median imputation method was the fastest, but its accuracy was lower than that of the more advanced approaches. KNN showed the worst performance across all metrics, with no wins in any scenario. The distribution of wins across methods is presented in Figure 1.

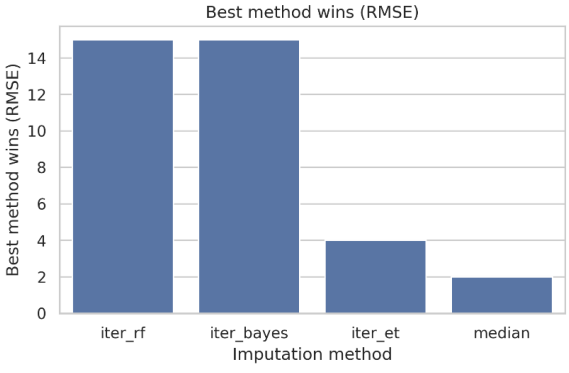

Fig. 1. Number of wins of the imputation methods across all 36 scenarios

A separate analysis for each dataset revealed substantial differences associated with the n/p ratio. The summary results for the Bones and NHANES datasets are presented in Tables 2 and 3, respectively.

Table 2.

Best-performing methods for the Bones dataset (mean RMSE across missingness levels)

Scenario	Mechanism	Best-performing method	RMSE
Bones full (n/p = 299)	MCAR	iter_bayes	1.713
Bones full (n/p = 299)	MAR	iter_bayes	2.261
Bones full (n/p = 299)	MNAR	iter_et	2.693
Bones subsampled (n/p = 45)	MCAR	iter_et	2.001
Bones subsampled (n/p = 45)	MAR	iter_bayes	2.101
Bones subsampled (n/p = 45)	MNAR	iter_rf	2.865

Table 3.

Best-performing methods for the NHANES dataset (mean RMSE across missingness levels)

Scenario	Mechanism	Best-performing method	RMSE
NHANES full (n/p = 130)	MCAR	iter_bayes	12.544
NHANES full (n/p = 130)	MAR	iter_bayes	13.748
NHANES full (n/p = 130)	MNAR	iter_rf	18.677
NHANES subsampled (n/p = 19)	MCAR	median	13.662
NHANES subsampled (n/p = 19)	MAR	iter_rf	15.239
NHANES subsampled (n/p = 19)	MNAR	iter_rf	19.340

The comparison of the two datasets reveals fundamental differences driven by the dimensionality of the feature space. For the Bones dataset, iter_bayes was the dominant method under MCAR and MAR, whereas under MNAR the ensemble based methods iter_et and iter_rf performed best, which can be explained by their ability to capture nonlinear relationships even in the presence of systematic missingness. The overall error level for NHANES was 6-7 times higher than that for Bones due to the substantially larger number of features (85 versus 11) and the lower n/p ratio. A particularly notable result was observed in the subsampled NHANES scenario under MCAR with a critically low n/p ratio of 19, where the median imputation method performed best, likely because of the regularization effect of a simple replacement strategy under conditions of excessive dimensionality. The corresponding heatmaps of method ranks for both datasets are presented in Figure 2.

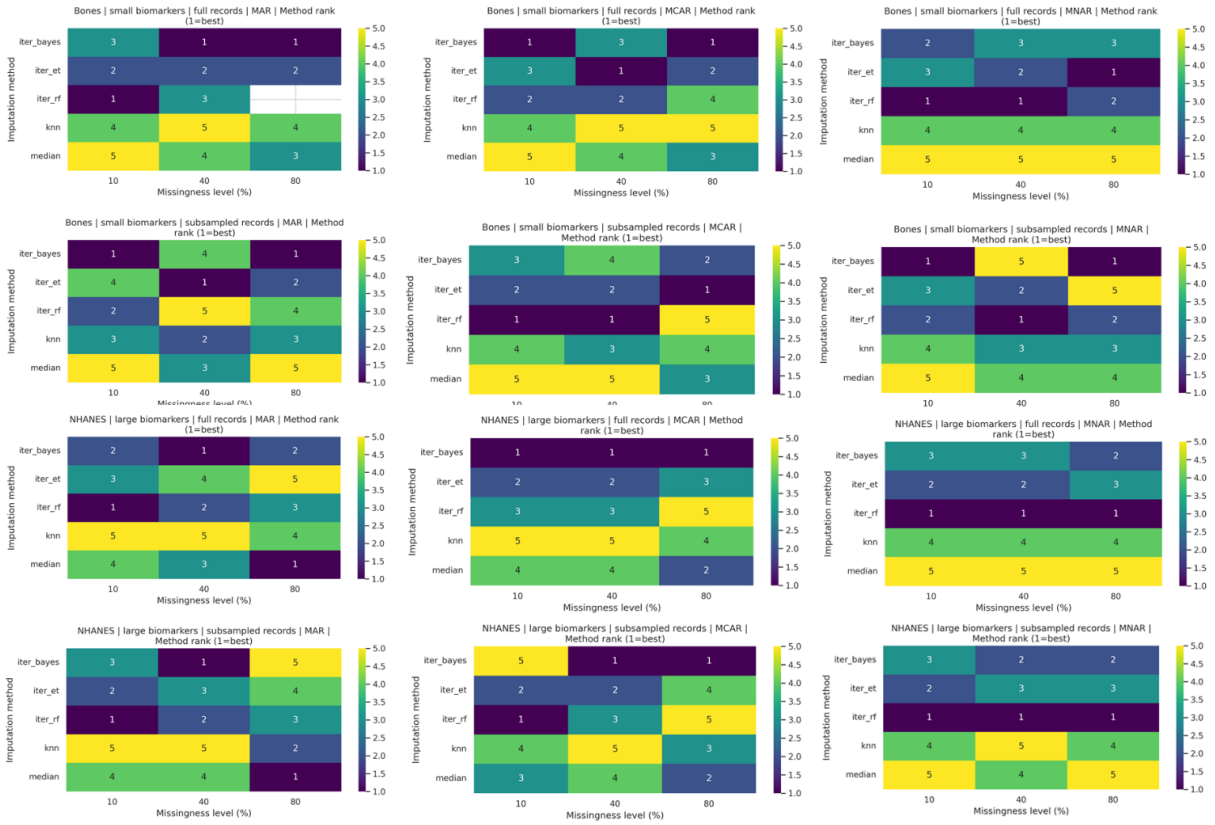

Fig. 2. Heatmaps of imputation method ranks by missingness mechanism and level of missing data

Overall, the rank heatmaps highlight that imputation performance was not uniform across datasets and missingness scenarios. Instead, the relative advantage of each method depended on the interaction between dimensionality, n/p ratio, and missingness mechanism. These observations motivated a more detailed analysis of imputation quality by mechanism and level of missing data, presented in the following subsection.

Primary Imputation Quality by Missingness Mechanism and Level of Missing Data

A summary of the results by missingness mechanism is presented in Table 4, which confirms the expected hierarchy of difficulty: MCAR (8.052) < MAR (9.691) < MNAR (11.812). The MNAR mechanism was the most challenging, with a mean error level 47% higher than that of MCAR. No universal best-performing method was identified: iter_bayes showed the best results for MCAR, iter_et was the leading method for MAR, and iter_rf achieved the lowest RMSE for MNAR.

Table 4.

Mean RMSE by missingness mechanism and distribution of feature-wise errors

Method	MCAR	MAR	MNAR	Median RMSE	IQR	% of Outliners	Best Method by Mechanism
iter_bayes	07.630	11.597	11.497	2.35	5.89	28	MCAR
iter_et	07.724	08.704	11.396	2.10	5.20	26	MAR
iter_rf	08.045	09.144	10.896	1.90	4.55	23	MNAR
knn	08.764	10.189	12.455	3.10	6.59	31	-
median	08.097	08.820	12.813	2.85	6.15	29	-

The analysis of the distribution of errors across individual features shows pronounced heterogeneity: approximately 40-50% of features can be classified as “easy” (RMSE < 5), 25-30% as “moderate” (RMSE 5-15), and a further 25-30% as “difficult” regardless of the method used. The iter_rf method demonstrated the best stability, with the lowest median RMSE (1.90) and the smallest interquartile range, whereas median imputation was characterized by the greatest heterogeneity and the most extreme error values, especially in NHANES scenarios under MNAR (max RMSE = 76.0). The substantial proportion of difficult features confirms that high feature-space dimensionality considerably complicates the imputation task.

An important observation is the nonlinear dependence of quality on the level of missing data, namely a W-shaped pattern, in which the quality at 40% missingness may be better than at 10%, followed by clear degradation at 80%. This phenomenon was observed in 55-75% of scenarios and indicates that the assumption of monotonic deterioration with increasing missingness may be incorrect. The detailed RMSE dynamics as a function of missingness level for each method are presented in Figure 3.

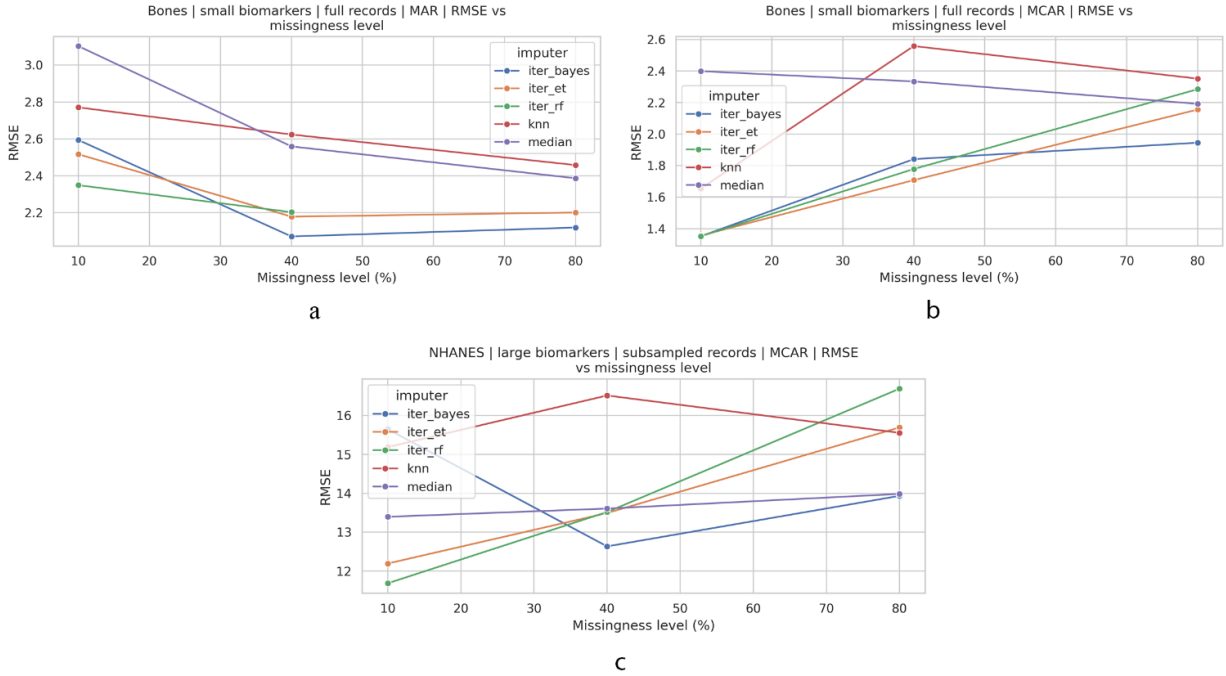

Fig. 3. Dynamics of mean RMSE as a function of missingness level for each imputation method: (a) Bones, full records, MAR; (b) Bones, full records, MCAR; (c) NHANES, subsampled records, MCAR

To assess the statistical stability of the results, bootstrap confidence intervals for RMSE were constructed using 500 resamples (Table 5).

Table 5.

Bootstrap 95% confidence intervals for RMSE (n = 500)

Method	Mean RMSE	Lower CI	Upper CI	CI Width	CV
iter_et	9.275	6.977	11.876	4.899	0.796
iter_rf	9.368	7.129	11.606	4.477	0.763
iter_bayes	10.241	7.688	13.084	5.396	1.031
median	9.910	7.432	12.701	5.269	0.776
knn	10.469	7.857	13.387	5.530	0.792

The complete overlap of the confidence intervals for iter_et, iter_rf, and iter_bayes indicates that the differences among these methods are not statistically significant at the global level. The most stable method was iter_rf (CV = 0.763), whereas iter_bayes showed the highest relative variability (CV = 1.031). The width of the confidence intervals indicates a strong influence of scenario-specific characteristics, which underscores the need for context-dependent selection of the imputation method. A consolidated visualization of method ranks across all scenarios is presented in Figure 4.

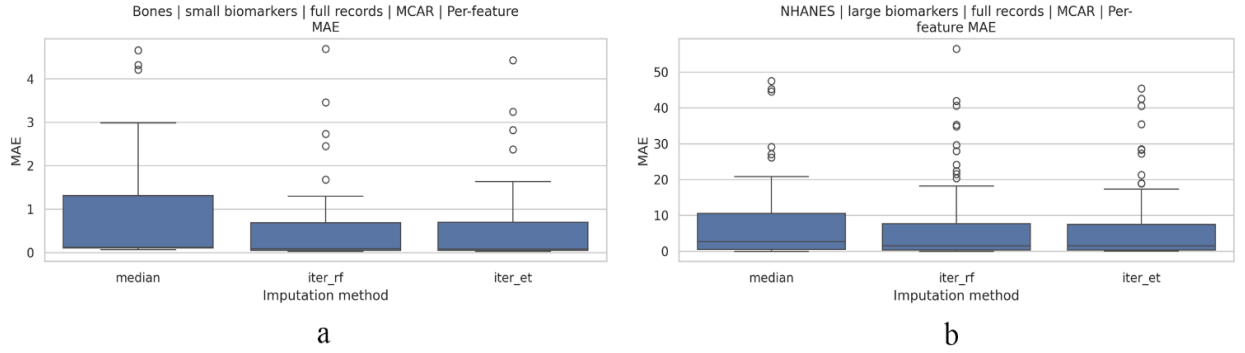

Fig. 4. Distribution of feature-wise MAE for the imputation methods under the MCAR mechanism: (a) Bones, full dataset; (b) NHANES, full dataset

Overall, the distributions shown in Figure 4 confirm that the differences among the leading methods were more strongly shaped by individual scenario characteristics than by a consistent global advantage of any single approach. This result further supports the conclusion that imputation quality should be interpreted not only through aggregate accuracy metrics, but also through its implications for subsequent analytical tasks. Therefore, the next section examines how the choice of imputation method affects biological age prediction performance.

Downstream Effect of Imputation Methods on Biological Age Prediction

The aim of this section is to analyze the influence of imputation method selection on the downstream task of biological age prediction using two approaches, namely regularized linear regression with ElasticNet and principal component analysis (PCA). The results of the comparison of the two biological age prediction models are presented in Table 6.

Table 6.

Comparison of ElasticNet and PCA for biological age prediction

Model	Mean BA RMSE	Mean Correlation	Negative Correlations	Stability
PCA	2.140	0.188	61/179 (34%)	Low
ElasticNet	5.864	0.759	0/179 (0%)	High

For the ElasticNet model, a detailed analysis of the correlations between predicted and actual biological age was conducted as a function of the imputation method. The summary results are presented in Table 7.

Table 7.

Predicted-actual biological age correlations for ElasticNet

Method	Bones	NHANES	Mean	Interpretation
iter_rf	0.798	0.763	0.780	Best downstream performance
iter_et	0.794	0.763	0.779	Stable and well-balanced
iter_bayes	0.790	0.766	0.778	Predictable performance with minimal variability across datasets (Δ=0.024)
median	0.753	0.731	0.742	Reliable baseline, fast (35 times faster than iter_rf)
knn	0.761	0.675	0.718	Worst for the downstream task

The differences among iter_rf, iter_et, and iter_bayes were minimal (0.778-0.780), whereas knn showed the lowest downstream performance, especially on NHANES (0.675). The correlations for Bones were higher (0.753-

0.798) than those for NHANES (0.675-0.766), which can be explained by the smaller number of biomarkers and the higher n/p ratio.

For the PCA-based approach, the accuracy of biological age prediction was analyzed using RMSE and MAE, and the corresponding results are presented in Table 8.

Table 8.

Accuracy of biological age prediction using PCA

Method	Mean BA RMSE	Mean BA MAE	RMSE/MAE	Rating
median	1.819	1.400	1.30	Best
knn	2.016	1.557	1.30	Good
iter_rf	2.104	1.625	1.29	Acceptable
iter_et	2.148	1.658	1.30	Acceptable
iter_bayes	2.609	2.018	1.29	Lower accuracy

Normal RMSE/MAE values (1.29-1.30) indicate the absence of substantial outliers in most scenarios. A paradoxical result is that the simplest method, median imputation, provides the best downstream accuracy for PCA. A likely explanation is that PCA itself acts as a regularizer, and the data smoothed in advance by median imputation are better aligned with the assumptions of this method.

A critical failure mode was identified for the subsampled NHANES scenario (n/p = 19, MAR, 80% missing data, iter_bayes, PCA): the correlation with age dropped to 0.045, MAE reached 2.00 years, and RMSE sharply increased to 8.8 years due to extreme outliers. At 40% missing data, the prediction quality improved; however, at 80%, PCA predicted an almost constant age. The combination of iter_bayes and PCA is therefore not recommended when n/p < 20 and the level of missing data is high. Under such conditions, ElasticNet or iter_rf are more reliable alternatives.

Analysis of the Trade-off Between Imputation Accuracy and Computational Efficiency

The aim of this task was to evaluate the trade-off between imputation quality, measured by RMSE, and computational efficiency. Pareto analysis makes it possible to identify non-dominated methods according to these two criteria. The summary results are presented in Table 9.

Only median imputation and iter_et belonged to the Pareto front: median imputation offered the fastest execution with acceptable accuracy, whereas iter_et provided the highest accuracy at the cost of increased computational demands. The other methods were either less accurate or slower, without offering substantial advantages. A visualization of the quality-time trade-off is presented in Figure 5.

Table 9.

Trade-off between imputation accuracy and execution time

Method	Mean RMSE	Mean Execution Time (s)	Speedup vs slowest	Pareto-efficient
median	9.910	0.0075	34 800×	✓
iter_bayes	10.241	2.415	108×	–
knn	10.469	7.48	35×	–
iter_et	9.275	83.12	3.1×	✓
iter_rf	9.368	260.99	1×	–

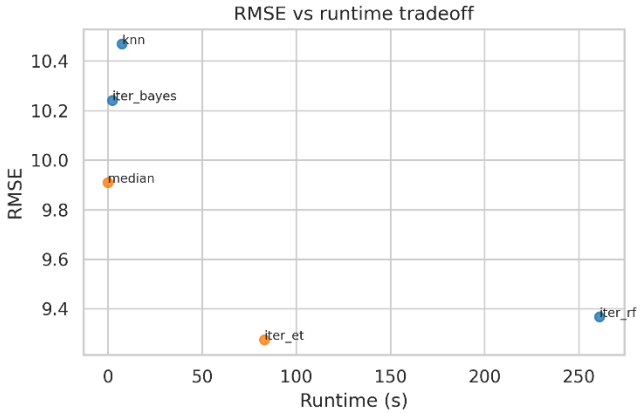

Fig. 5. Pareto front of the trade-off between mean RMSE and mean execution time of the imputation methods

As a practical guideline, iter_bayes operates 108 times faster than iter_rf and 34 times faster than iter_et, which makes it the optimal choice for interactive use. Under time constraints, iter_bayes should be preferred; when

sufficient computational resources are available and the missingness mechanisms are complex, iter_et or iter_rf should be used; if resources are limited, median imputation is the most practical option.

Practical Recommendations for Imputation Method Selection

Based on the results of all 180 experiments, a consolidated ranking of the imputation methods was developed according to the key criteria (Table 10).

Table 10.

Consolidated ranking of the imputation methods

Criterion	1st Place	2nd Place	3rd Place	4th Place	5th Place
Mean RMSE	iter_et (9.28)	iter_rf (9.37)	median (9.91)	iter_bayes (10.24)	knn (10.47)
Mean Rank	iter_bayes (2.19)	iter_rf (2.20)	iter_et (2.53)	median (3.97)	knn (4.03)
Number of Wins	iter_rf (15)	iter_bayes (15)	iter_et (4)	median (2)	knn (0)
Stability (CV)	iter_rf (0.30)	iter_et (0.33)	median (0.35)	iter_bayes (0.40)	knn (0.41)
Execution Speed	median (0.01c)	iter_bayes (2.4c)	knn (7.5c)	iter_et (83c)	iter_rf (261c)
BA Performance with PCA	median (1.82)	knn (2.02)	iter_rf (2.10)	iter_et (2.15)	iter_bayes (2.61)
BA Performance with ElasticNet	iter_rf (0.77)	iter_bayes (0.76)	iter_et (0.76)	median (0.76)	knn (0.76)
MCAR Mechanism	iter_bayes (7.63)	iter_rf (7.85)	median (7.74)	iter_et (8.15)	knn (8.76)
MAR Mechanism	iter_et (8.70)	iter_rf (8.98)	knn (9.77)	median (10.10)	iter_bayes (11.60)
MNAR Mechanism	iter_rf (10.90)	iter_et (11.40)	iter_bayes (11.50)	median (12.11)	knn (13.00)

The table confirms the absence of a universal leader. iter_rf was the most robust method, with the highest number of wins and the best downstream performance for ElasticNet, but it was also the slowest. iter_et was optimal in terms of the balance between accuracy and execution time. iter_bayes was fast and effective under MCAR. Median imputation remained a practical choice under resource constraints. KNN did not justify its use as a primary method.

For context-dependent imputation method selection, the following decision algorithm is recommended.

Step 1. Evaluate the available computational resources.

If rapid processing is required, for example in real-time settings or under limited computational resources, SimpleImputer with median imputation should be selected. If moderate resources are available, IterativeImputer with BayesianRidge may be preferred as a faster iterative alternative. If substantial computational resources are available, IterativeImputer with RandomForest or ExtraTrees can be considered.

Step 2. Assess the n/p ratio.

When $n/p < 20$, robust methods such as IterativeImputer with RandomForest or ExtraTrees should be preferred, while combinations with PCA should be avoided because of the risk of catastrophic failure in low-sample-size, high-dimensional settings. When $n/p \geq 20$, all methods may be considered depending on the remaining factors.

Step 3. Consider the missingness mechanism, if known or assumed.

Under MCAR, IterativeImputer with BayesianRidge is an effective choice, while median imputation can serve as a fast practical alternative. Under MAR, IterativeImputer with ExtraTrees is recommended. Under MNAR, IterativeImputer with RandomForest provides the best performance despite its high computational cost. If the missingness mechanism is unknown, RandomForest or ExtraTrees may be used as more robust general-purpose options.

Step 4. Account for the level of missing data.

At low levels of missing data (up to approximately 30%), all methods provide acceptable performance, and selection can be based on execution time or mechanism-specific considerations. At moderate levels (approximately 30-60%), a paradoxical W-shaped pattern may occur, where imputation quality may unexpectedly improve compared to lower missingness levels; the results should therefore be interpreted with additional caution. At high levels of missing data (above 60%), the setting becomes critical; in particular, the combination of IterativeImputer with BayesianRidge and PCA should be avoided when $n/p < 20$, while IterativeImputer with RandomForest is generally the safer choice. These thresholds are based on experiments conducted at 10%, 40%, and 80% missingness levels.

Step 5. Select the downstream model.

If PCA is used for biological age prediction, median imputation may provide favorable downstream accuracy, although the risk of negative correlations should be taken into account. If ElasticNet or another interpretable model is used, IterativeImputer with RandomForest, BayesianRidge, or ExtraTrees provides better overall downstream performance and more stable positive correlations.

Step 6. Make a compromise choice when necessary.

For the best balance between accuracy and execution time, IterativeImputer with ExtraTrees is recommended. For maximum accuracy when computation time is not a limiting factor, IterativeImputer with RandomForest is preferable. For maximum speed under limited resources, median imputation remains the most practical choice. In clinically oriented applications where interpretability is important, ElasticNet-based downstream modeling is preferable to PCA.

The logic of this algorithm is summarized in the quick selection matrix presented in Table 11.

Table 11.

Quick selection matrix for imputation method choice

Situation	n/p	Mechanism	Time Constraints?	Recommended Method	Alternative
Typical clinical setting	20–100	MAR	Yes	iter_et	iter_bayes
Cohort with MNAR	20–100	MNAR	Yes	iter_rf	iter_et
Rapid analysis	20–100	MCAR	No	iter_bayes	median
Large registry	>100	MAR	Not critical	median	iter_bayes
Critical n/p ratio	<20	Any	Not critical	iter_rf	iter_et
Downstream PCA	Any	Any	Not critical	median; use ElasticNet instead if n/p < 20	knn

The recommendations are valid for $n \geq 500$, $p \in [11, 85]$, $n/p \in [19; 299]$ and missing data levels of 10-80% under all missingness mechanisms. The combination of iter_bayes and PCA should be strictly avoided when $n/p < 20$ and the proportion of missing data exceeds 60%, because of the risk of catastrophic model failure. When $n/p < 10$, preliminary dimensionality reduction or data selection is required. To minimize the risk of incorrect method selection, a standardized validation protocol is recommended, including a masking strategy with controlled removal of a subset of values, comparison of 2-3 candidate methods, verification of the physiological plausibility of the imputed values, and sensitivity analysis with variation of the missingness level.

The choice of imputation method depends on multiple factors, and no universal solution exists. iter_rf, iter_et, and iter_bayes show similar accuracy, but each is optimal under different missingness conditions and computational resource constraints. When $n/p < 20$, robust methods are required, and PCA should be avoided because of the risk of catastrophic failure under the observed W-shaped effect. The downstream model also influences the results: PCA yields a lower RMSE, but is less stable, whereas ElasticNet is more reproducible and interpretable.

Discussion

The study confirms the advantage of IterativeImputer with RandomForest under the MNAR mechanism and the effectiveness of iter_bayes under MCAR, which is consistent with the findings of Stekhoven and Bühlmann [15], who demonstrated the superiority of ensemble based methods over parametric approaches in the presence of nonlinear relationships among variables. The lower performance of KNN, especially on the high-dimensional NHANES dataset ($n/p = 19$-130), supports the critical assessment of this method by Beretta and Santaniello [9]: the sensitivity of kNN to the "curse of dimensionality" makes it unreliable in feature spaces with a large number of variables. The effectiveness of iter_bayes under MCAR in small-sample settings is consistent with the conclusions of Antonenko et al. [11] regarding the advantages of regularized linear models in imputation tasks under limited n/p. At the same time, the paradoxical effectiveness of median imputation for the downstream PCA task has no direct analogue in the literature and requires further theoretical justification.

The W-shaped dependence of imputation quality on the level of missing data was observed in 55-75% of scenarios, which contradicts the widely accepted assumption of monotonic deterioration. Madley-Dowd et al. [17] showed that the proportion of missing data is not a reliable predictor of imputation quality; the results obtained in this study confirm and extend this conclusion by demonstrating the possibility of a nonlinear relationship even within a single scenario.

The comparison of ElasticNet and PCA indicates fundamentally different failure modes. PCA yields a lower RMSE, but is unstable because of the indeterminacy of component signs, whereas ElasticNet provides stable and interpretable results. The identified critical failure mode, namely the combination of iter_bayes and PCA when $n/p < 20$, represents a new practical finding that has not been described in previous comparative studies [3].

The main limitations of the study include the restricted parameter range ($n \geq 500$, $p \in [11; 85]$), the specificity of the two datasets, and the absence of an analysis of deep learning based approaches. At the same time, the proposed recommendations and decision algorithms have practical value, especially for avoiding critical method combinations in biomedical research.

Conclusions

This study presented a comprehensive comparison of five missing data imputation methods, namely IterativeImputer with RandomForest, ExtraTrees, and BayesianRidge estimators, KNNImputer, and SimpleImputer with median imputation, on two biomedical datasets of different dimensionality under three missingness mechanisms, namely MCAR, MAR, and MNAR, and at three levels of missing data, namely 10%, 40%, and 80%. The evaluation was performed both at the level of primary imputation quality, using RMSE, MAE, and R^2, and through the downstream effect on biological age prediction using ElasticNet and PCA.

Based on the stated objectives, the following conclusions were obtained:

1. Five imputation methods were tested on datasets with different characteristics. The iter_et method achieved the lowest average RMSE (9.275), while iter_bayes and iter_rf demonstrated the best average ranks (2.19-2.20). No single universally optimal approach was identified; therefore, method selection depends on the specific data and modeling context.

2. The hierarchy of missingness mechanism difficulty was confirmed as follows: MCAR (RMSE 8.052) < MAR (9.691) < MNAR (11.812). The best-performing methods were iter_bayes for MCAR, iter_et for MAR, iter_rf for MNAR. A W-shaped pattern in imputation quality was also observed in the 55-75% missingness range, which requires further explanation.

3. PCA achieved a lower RMSE than ElasticNet (2.14 versus 5.86), but produced 34% negative correlations. ElasticNet provided stable positive correlations (mean 0.759) and better interpretability; therefore, it appears to be the more reliable option for clinical applications. The combination of iter_bayes with PCA should be avoided when n/p < 20 and the proportion of missing data is high because of the risk of catastrophic failure.

4. The Pareto analysis showed that median imputation was extremely fast (0.0075 s) while maintaining acceptable accuracy, whereas iter_et was the most accurate among the relatively fast methods (83.12 s). The iter_rf method provided the highest accuracy under MNAR, but its very high execution time (260.99 s) substantially limits its practical applicability.

5. The choice of imputation method should be guided by the n/p ratio, the missingness mechanism, time constraints, and the downstream model. In particular, when n/p < 20, robust methods such as iter_rf and iter_et are preferable, while PCA should be avoided; under MCAR, iter_bayes or median imputation is appropriate; under MAR, iter_etis recommended; under MNAR, iter_rf is preferable when computational resources allow; for real-time applications, median imputation is the most practical option; and for clinical tasks, ElasticNet is preferable to PCA.

The scientific novelty of the study lies in the identification of the paradoxical W-shaped effect associated with increasing missingness, the systematic comparison of downstream effects for two biological age models, the identification of a critical failure mode, namely the combination of iter_bayes and PCA at low n/p ratios, and the development of a formalized algorithm for imputation method selection. The proposed recommendations are applicable to clinical and epidemiological studies with parameters $n \geq 500$, $p \in [10; 100]$, $n/p \geq 20$.

Directions for future research include the following: theoretical justification of the paradoxical quality improvement effect observed with increasing missingness; extension of the validation range to high-dimensional settings ($p \gg n$) typical of genomics and proteomics; investigation of deep learning based methods on large datasets with sufficient statistical power to train complex models; and development of adaptive hybrid approaches capable of automatically selecting the optimal method depending on data characteristics.

ADDITIONAL INFORMATION

AUTHOR CONTRIBUTIONS

Conceptualisation, V.S. and L.P.; methodology, V.S. and L.P.; software, O.V.; validation, L.P. and O.V.; formal analysis, L.P. and O.V.; investigation, L.P. and O.V.; data curation, L.P. and O.V.; writing – original draft preparation, L.P.; writing – review and editing, L.P. and O.V.; visualisation, O.V.; supervision, V.S.; project administration, V.S. All authors have read and agreed to the published version of the manuscript.

DECLARATION ON THE USE OF GENERATIVE ARTIFICIAL INTELLIGENCE TOOLS

In preparing this work, the authors used Claude 3.5 Sonnet (Anthropic) for grammar and spelling checks, paraphrasing, and improving the clarity of individual sentences. After using this tool, the authors reviewed and edited the content and take full responsibility for the content of this publication.

1. Austin P. C., White I. R., Lee D. S. et al. Missing Data in Clinical Research: A Tutorial on Multiple Imputation // Can J Cardiol. 2021. Vol. 37, No. 9. P. 1322-1331. doi: https://doi.org/10.1016/j.cjca.2020.11.010

2. Sterne J. A. C., White I. R., Carlin J. B. et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls // BMJ. 2009. Vol. 338. P. b2393. doi: https://doi.org/10.1136/bmj.b2393

3. Li J., Guo S., Ma R. et al. Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets // BMC Med Res Methodol. 2024. Vol. 24. P. 41. doi: https://doi.org/10.1186/s12874-024-02173-x

4. Donders A. R. T., van der Heijden G. J. M. G., Stijnen T. et al. Review: a gentle introduction to imputation of missing values // Journal of Clinical Epidemiology. 2006. Vol. 59, No. 10. P. 1087-1091. doi: https://doi.org/10.1016/j.jclinepi.2006.01.014

5. Jakobsen J. C., Gluud C., Wetterslev J. et al. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts // BMC Medical Research Methodology. 2017. Vol. 17, No. 1. P. 162. doi: https://doi.org/10.1186/s12874-017-0442-1

6. Salgado C. M., Azevedo C., Proença H. et al. Missing Data // Secondary Analysis of Electronic Health Records. Cham: Springer, 2016. P. 143-162. doi: https://doi.org/10.1007/978-3-319-43742-2_13

7. van Buuren S., Groothuis-Oudshoorn K. Mice: Multivariate Imputation by Chained Equations in R // Journal of Statistical Software. 2011. Vol. 45, No. 3. P. 1-67. doi: https://doi.org/10.18637/jss.v045.i03

8. Azur M. J., Stuart E. A., Frangakis C. et al. Multiple imputation by chained equations: what is it and how does it work? // International Journal of

Methods in Psychiatric Research. 2011. Vol. 20, No. 1. P. 40-49. doi: https://doi.org/10.1002/mpr.329

9. Beretta L., Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation // BMC Medical Informatics and Decision Making. 2016. Vol. 16, Suppl. 3. P. 74. doi: https://doi.org/10.1186/s12911-016-0318-z

10. Casella M., Milano N., Dolce P. et al. Transformers deep learning models for missing data imputation: an application of the ReMasker model on a psychometric scale // Front Psychol. 2024. Vol. 15. P. 1449272. doi: https://doi.org/10.3389/fpsyg.2024.1449272

11. Antonenko E., Carreño A., Read J. Autoreplicative random forests with applications to missing value imputation // Mach Learn. 2024. Vol. 113. P. 7617-7643. doi: https://doi.org/10.1007/s10994-024-06584-1

12. Jylhävä J., Pedersen N. L., Hägg S. Biological Age Predictors // EBioMedicine. 2017. Vol. 21. P. 29-36. doi: https://doi.org/10.1016/j.ebiom.2017.03.046

13. Belsky D. W., Caspi A., Houts R. et al. Quantification of biological aging in young adults // Proceedings of the National Academy of Sciences. 2015. Vol. 112, No. 30. P. E4104-E4110. doi: https://doi.org/10.1073/pnas.1506264112

14. Woods C. M., Sideridis G., Xie M. et al. Best practices for addressing missing data through multiple imputation // Infant Child Dev. 2024. Vol. 33, No. 1. P. e2407. doi: https://doi.org/10.1002/icd.2407

15. Stekhoven D. J., Bühlmann P. MissForest – non-parametric missing value imputation for mixed-type data // Bioinformatics. 2012. Vol. 28, No. 1. P. 112-118. doi: https://doi.org/10.1093/bioinformatics/btr597

16. Liang X., Liu Y., Wang T. et al. Missing Data Imputation Method Combining Random Forest and Generative Adversarial Imputation Network // Sensors. 2024. Vol. 24, No. 4. P. 1112. doi: https://doi.org/10.3390/s24041112

17. Madley-Dowd P., Hughes R., Tilling K. et al. The proportion of missing data should not be used to guide decisions on multiple imputation // Journal of Clinical Epidemiology. 2019. Vol. 110. P. 63-73. doi: https://doi.org/10.1016/j.jclinepi.2019.02.016

18. Pedregosa F., Varoquaux G., Gramfort A. et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. 2011. Vol. 12. P. 2825-2830. URL: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

19. Slipchenko V., Poliahushko L., Volkov O. et al. Hybrid imputation of biomedical data by using transformers and autoencoders for assessing human biological age // Eastern-European Journal of Enterprise Technologies. 2025. Vol. 5, No. 4(137). P. 31-40. doi: https://doi.org/10.15587/1729-4061.2025.340325

20. Slipchenko V., Grygorieva N., Poliahushko L. et al. Estimation of the bone biological age using machine learning // EUREKA: Physics and Engineering. 2025. No. 1. P. 175-186. doi: https://doi.org/10.21303/2461-4262.2025.003656

21. Centers for Disease Control and Prevention. National Health and Nutrition Examination Survey (NHANES). Atlanta: National Center for Health Statistics, 2024. URL: https://www.cdc.gov/nchs/nhanes/index.html .

Володимир СЛІПЧЕНКО, Любов ПОЛЯГУШКО, Олександр ВОЛКОВ
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

ПОРІВНЯЛЬНИЙ АНАЛІЗ МЕТОДІВ ІМПУТАЦІЇ ПРОПУЩЕНИХ ДАНИХ У БІОМЕДИЧНИХ ДОСЛІДЖЕННЯХ: ВПЛИВ НА ПЕРЕДБАЧЕННЯ БІОЛОГІЧНОГО ВІКУ

Пропущені дані залишаються однією з ключових проблем біомедичних досліджень, оскільки можуть спричиняти зміщення статистичних оцінок, знижувати точність прогнозування та послаблювати надійність наукових висновків. У роботі було виконано порівняльне оцінювання п'яти підходів до імпутації: IterativeImputer з оцінювачами RandomForest, ExtraTrees та BayesianRidge, а також KNNImputer і median-based SimpleImputer. Дослідження проведено на двох біомедичних наборах даних: Bones (3 285 записів, 11 біомаркерів) і NHANES (11 016 записів після скорочення з 55 081, 85 біомаркерів), для яких розглянуто градієнт співвідношення п/р у межах від 19 до 299. Експериментальний дизайн охоплював три механізми пропусків, а саме MCAR, MAR і MNAR, а також три рівні відсутності даних: 10%, 40% і 80%. Якість імпутації оцінювалася за показником RMSE, тоді як downstream-вплив аналізувався через прогнозування біологічного віку з використанням моделей ElasticNet і PCA. Встановлено, що IterativeImputer з ExtraTrees продемонстрував найнижче середнє значення RMSE (9.275), тоді як IterativeImputer з BayesianRidge та RandomForest показали найкращий середній ранг (2.19-2.20), що свідчить про вищу стабільність результатів у різних сценаріях. Для механізму MNAR найкращі результати було отримано для RandomForest (RMSE 10.896), тоді як для MAR найбільш ефективним виявився ExtraTrees (RMSE 8.704). Downstream-аналіз показав, що PCA забезпечував нижчий RMSE прогнозування порівняно з ElasticNet (2.14 проти 5.86), однак у 34% випадків спостерігалися негативні кореляції. Також було виявлено парадоксальний ефект покращення якості імпутації зі зростанням частки пропусків у 55-75% сценаріїв. Найшвидшим методом виявилася median-імпутація (0.0075 с), тоді як RandomForest був найповільнішим (261 с). Отримані результати дали змогу сформулювати практичні рекомендації щодо вибору методів імпутації залежно від структури даних, механізму пропусків та обчислювальних обмежень.

Ключові слова: імпутація даних, пропущені дані, біологічний вік, машинне навчання, MCAR, MAR, MNAR.