This study investigates the process of restoring missing biomedical and social data for human biological age assessment. The principal challenge is the high rate of missing values in datasets, notably NHANES – up to 40%. This complicates accurate health prediction and reduces the effectiveness of preventive interventions.

To address this issue, deep learning methods, specifically autoencoders and transformers, were employed. The autoencoder provided fast imputation (37.4 s, MAE = 7.54) but less accuracy. The transformer achieved the highest accuracy (246.3 s, MAE = 1.10) yet required substantial resources and showed overfitting risks.

A hybrid architecture has been proposed to combine the advantages of both approaches. On the NHANES dataset (55,081 records and 84 biomarkers), the model demonstrated an optimal balance (54.2 s, MAE = 5.26) and stability with up to 50% missing data. Compared to mean-value imputation, the accuracy of biological age estimation improved by 25%. The coefficient of determination reached 0.9875, and root mean squared error was 35.9, confirming strong consistency of the restored values. Sensitivity analysis revealed stable accuracy up to 55% missing data, after which degradation occurred.

A unique feature of the hybrid approach is the combination of high accuracy with moderate computational cost. This makes the model suitable for medical information systems with incomplete datasets. Practical applications include preventive medicine, biological aging monitoring, and risk group identification.

In the Ukrainian context, the model could enhance biomedical research and digital healthcare while also serving as a foundation for bioinformatics and life expectancy studies

Keywords: data imputation, composite architecture, deep learning, functional age, PhenoAge, NHANES

UDC 004.93

DOI: 10.15587/1729-4061.2025.340325

HYBRID IMPUTATION OF BIOMEDICAL DATA BY USING TRANSFORMERS AND AUTOENCODERS FOR ASSESSING HUMAN BIOLOGICAL AGE

Volodymyr Slipchenko
Doctor of Technical Sciences, Professor*
Liubov Poliahushko
Corresponding author
PhD, Associate Professor*
E-mail: liubovpoliagushko@gmail.com
Oleksandr Volkov*
Vladyslav Shatylo*
*Department of Digital Technologies in Energy
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"
Beresteiskyi ave., 37, Kyiv, Ukraine, 03056

Received 03.07.2025
Received in revised form 18.09.2025
Accepted 29.09.2025
Published 30.10.2025

How to Cite: *Slipchenko, V., Poliahushko, L., Volkov, O., Shatylo, V. (2025). Hybrid imputation of biomedical data by using transformers and autoencoders for assessing human biological age. Eastern-European Journal of Enterprise Technologies, 5 (4 (137)), 31–40. https://doi.org/10.15587/1729-4061.2025.340325*

1. Introduction

Biological age (BA) assessment is an important task in the field of health care; it allows for more accurate prediction of disease risk, determination of the effectiveness of therapeutic interventions, and planning of preventive measures [1]. Biological age is an integrated metric formed on the basis of physiological, biochemical, and metabolic indicators. These include levels of glucose, albumin, creatinine, C-reactive protein, cholesterol, alkaline phosphatase, and blood parameters (lymphocytes, RDW, MCV).

However, the accuracy of BA assessment models is significantly reduced because of the large number of missing values in medical datasets, in particular in such large representative studies as NHANES (National Health and Nutrition Examination Survey) [2]. The proportion of missing values in these data can reach 40%, which significantly complicates the application of conventional statistical and regression models. This problem is especially revealed in countries with limited access to population health data, or the lack of large-scale studies in the field of population health.

Existing methods for simple filling of missing values (mean, median) lead to the loss of complex nonlinear relationships between parameters, which significantly worsens the predictive ability of the obtained biological age models. Current studies show the advantages of deep learning. In [3], a transformer for tabular medical data was applied. In [4], the effectiveness of the modified transformer architecture ReMasker on psychometric scales under conditions of limited data was confirmed. On the other hand, improved autoencoders were used to handle non-random (non-ignorable, MNAR) omissions [5]. Taken together, these works confirm that modern deep learning architectures are able to preserve nonlinear interactions between parameters and provide better quality data recovery compared to classical approaches.

Thus, the construction of accurate and fast biomedical data imputation models, in particular hybrid ones that combine the advantages of different neural architectures, is particularly relevant. Owing to the use of deep learning-based imputation methods, the results can be applied to complex, incomplete data sets, where the relationships between different parameters are nonlinear, which is important for use on real data. By combining such incomplete medical data and environmental, sociological information about the environment, it is possible to restore missing data, take into account the expectations of the internal indicators of the

person himself/herself and the external indicators of the living environment.

This could not only improve the quality of biological age assessment but also would make such models more suitable for practical implementation in integrated medical information systems. The use of a complete dataset also opens the way to more accurate policy formation of environmental and sociological practices, exploring the interactions of indicators in complete datasets.

Therefore, research on the imputation of medical datasets is relevant. Establishing an effective method of data imputation in a dataset will allow for more efficient calculation of BA and computation of patterns among indicator values, taking into account various data components. Research on large datasets requires both resources and time, so the task to devise a more effective method for resolving a specified problem remains relevant.

2. Literature review and problem statement

Study [6] presented the results of applying Horvat's epigenetic clock, and [7] showed the PhenoAge model. It was shown that both approaches make it possible to estimate biological age with high accuracy. However, issues related to the high cost of analyses and the complexity of data collection remain unresolved. The reason for this is the expensive molecular biological procedures and the limited availability of laboratory indicators in resource-poor countries. An option to overcome these difficulties may be the search for available surrogate biomarkers and the use of incomplete data. This is the approach used in [8] but it is limited by the specificity of clinical samples. All this allows us to state that it is advisable to conduct a study on the construction of models capable of working with large population sets.

KNN imputation can improve predictions but is sensitive to scaling and metrics, degrades in high dimensionality, and is expensive to scale [9]. MICE works with MAR but omissions in auxiliary variables cause bias and instability of chain equations, increasing the variance of estimates [10]. Both approaches have limitations in accuracy and scalability in multidimensional medical tables. This motivates the transition to machine learning methods that can reproduce nonlinear interactions.

Study [11] showed the use of autoencoders for imputation of medical data but their stability was low with increasing omissions; in further work [12], the MIDA method was applied, which partially improved the accuracy but demonstrated a high demand for resources and dependence on the depth of omissions. As a result, both approaches show a disadvantage when working with significant omissions.

Another area of imputation is the use of transformer architectures. In [13], the BEHRT transformer model was tested on ECG sequences, but it did not provide for the imputation of continuous biomarkers under conditions of large MCAR-misses and was not optimized for time or resources. The transformer in [14] was applied to EEG data processing but the study was limited to synthetic misses of small depth, without checking on tabular biomarkers and PhenoAge, and without taking into account computational costs.

The system implemented in [15] is a global-local transformer for assessing brain age from MRI images, but this system does not solve the problem of imputation of tabular features at a high level of misses (35–95% MCAR) and requires significant GPU resources. It is shown that such models achieve high accuracy due to the self-attention mechanism and taking into account global patterns. However, the issues of practical applicability for tabular medical data remain unresolved. The reason for this is the high computational costs, the risk of overfitting, and the difference in the data structure between images and tables. An option to overcome these difficulties may be to build hybrid models that combine different approaches. It was in [16] that the Precious1GPT model showed the potential of language transformers for age prediction but remained difficult to explain and resource-intensive. This confirms that the construction of combined and, at the same time, interpretable models is a promising area.

In [17], the GAIN approach based on generative-competitive learning was proposed, which showed good results in restoring gaps. At the same time, the model turned out to be limited in its ability to generalize and sensitive to architecture tuning, which reduces its stability in practical scenarios. In study [18], Random Forest was used for imputation, where relatively high accuracy was demonstrated even in the presence of nonlinearities and interactions between features. However, the method was vulnerable to noise and degraded with increasing number of features, which limits its scalability. A possible way to overcome these problems is to combine classical algorithms with modern neural networks.

In this context, the ImputEHR tool [19] was proposed to visualize the imputation process, but its accuracy remained low for large data sets. This confirms the need to devise methods that combine interpretability of results with high performance and robustness to complex data structures.

In [20], a comparison of classical and modern methods was carried out. It was shown that statistical approaches are simple and interpretable but are inferior in accuracy to artificial intelligence methods. The reason for this is the ignoring of complex nonlinear patterns in the data. An option to overcome the difficulties is the design of hybrid systems that integrate the advantages of classical and modern algorithms.

In general, it can be stated that the key problem is the lack of a balanced solution that would combine the accuracy and stability of transformers with the speed and simplicity of autoencoders. Overcoming this particular contradiction is critically important for the practical application of imputation models in the tasks to estimate human biological age.

3. The aim and objectives of the study

The aim of our study is to improve the accuracy of estimating human biological age with incomplete medical data by building a hybrid imputation model that combines the advantages of autoencoders and transformers.

To achieve the goal, the following tasks were set:

– to build an imputation model based on an autoencoder and investigate its parameters;

– to build an imputation model based on a transformer and evaluate its performance;

– to design a hybrid imputation architecture and perform its parametric tuning;

– to conduct a comparative analysis of imputation models in terms of accuracy and speed and to verify the impact of imputation on the calculation of biological age.

4. The study materials and methods

4. 1. The object and hypothesis of the study

The object of our study is the process of restoring missing biomedical and social data for estimating human biological age.

The subject of the study is methods for imputing the specified data by using autoencoders, transformers, and their hybrid combination models to increase the accuracy and speed of estimating biological age.

The hypothesis of the study assumes that the hybrid architecture, which combines the advantages of autoencoders and transformers, provides a better balance between the speed of calculations and the accuracy of imputation of an incomplete data set.

The following assumptions are adopted in the work:

– data gaps are random;

– all data are numerical;

– BA is determined by the PhenoAge formula from the primary dataset;

– GPU is used to train models to speed up the study.

The following simplifications are accepted in the work:

– there is no validation on a large number of datasets;

– environmental and socioeconomic factors are partially taken into account;

– the input dataset contains more than 30% random omissions.

4. 2. Data

The study used the open biomedical dataset NHANES (National Health and Nutrition Examination Survey) for 1999–2017 [21]. The total sample size is 55,081 records. Each record includes 84 biomarkers, including physiological parameters (height, weight, blood pressure, BMI), laboratory tests (glucose levels, cholesterol, creatinine, albumin, C-reactive protein, liver enzymes, etc.), and demographic characteristics (age, gender, ethnicity).

An important characteristic of the data is the high proportion of missing values, which varies from 22 to 41% depending on the specific biomarker. Missing values are categorized [22] as follows:

– MCAR (Missing Completely At Random) – values are missing completely at random, for example because of technical errors in measurements;

– MAR (Missing At Random) – values are missing at random, for example, because of certain categories of patients who are more likely to not undergo certain tests.

Thus, the choice of NHANES is ideal for studying the effectiveness of data imputation methods.

4. 3. Computing environment

The experiments were conducted in a high-performance computing environment with the following parameters:

– CPU: 8 vCPU;

– GPU: RTX 4000 Ada (20 GB of video memory);

– RAM: 50 GB RAM;

– Framework: PyTorch 2.1;

– CUDA platform: version 11.8.

The computing environment provided fast training and efficient scaling of the models, especially the transformer, which requires significant GPU resources.

4. 4. Validation

For an objective assessment of the results, the 5-fold cross-validation method was used. The entire data set was divided into five equal parts, four of which were used for training and one for testing; the procedure was repeated five times with alternating test parts.

The quality of the models was assessed using the following metrics [23]:

– MAE (Mean Absolute Error): the average absolute deviation of the predicted values from the actual ones. A smaller value indicates a higher accuracy of the model;

– RMSE (Root Mean Squared Error): the square root of the mean square error, sensitive to large deviations and used to detect significant errors in forecasts;

– R2 (coefficient of determination) shows the proportion of the variance of the dependent variable explained by the model and characterizes the quality of generalization.

In addition, a t-test with a significance level of $\alpha = 0.05$ was used to test the statistical significance of the differences between the results obtained using different models. This allowed us to establish the reliability of the advantages of the proposed hybrid model compared to alternative approaches.

5. Results of research on imputation models for gap filling in biomedical datasets

5. 1. Construction of an autoencoder-based imputation model

For the medical data imputation task, a symmetric neural network of the Denoising Autoencoder type was developed, implemented on the basis of a Multilayer Perceptron (MLP [24]).

The proposed model has the following layer structure:

– input layer: 256 neurons (overlapping the number of biomarkers);

– hidden layers of the encoder: 128 neurons → 64 neurons;

– latent space (smallest layer): 64 neurons;

– decoder (symmetric to the encoder): 64 → 128 → 256 neurons.

The model was trained using the ReLU (Rectified Linear Unit) activation function, which provides fast and stable convergence. To reduce the risk of overtraining and increase the generalization ability of the network, the Dropout method with a probability of 0.2 was used. Parameter optimization was performed using the Adam algorithm with an initial learning rate of 0.001. The model parameters were selected experimentally based on cross-validation.

To impute missing values, a masking mechanism was introduced through element-wise multiplication. Input data containing gaps are masked with a special binary mask: $m_i \in \{0,1\}$, where 0 is missing values, 1 is present. Thus, the model learns to reconstruct only known values, minimizing losses according to the modified function

$$L_{AE}\left(W_e, b_e, W_d, b_d\right) = \frac{1}{n}\sum_{i=1}^{n} m_i \times \left(x_i - \hat{x}_i\right)^2 \to \min, \qquad (1)$$

where $W_e \in R^{h \times d}$, $b_e \in R^h$ are the encoder weights and offsets; $W_d \in R^{d \times h}$, $b_d \in R^d$ are the decoder weights and offsets; $x_i \in R_e$ is the original value; $\hat{x} \in R_d$ is the reconstructed value.

The simulation results demonstrate the following characteristics of the autoencoder: MAE is 7.543, R^2 is 0.9797, RMSE is 46.8586, and the total time for model training and imputation is 37.4 seconds.

5. 2. Construction of a transformer-based imputation model

To implement the imputation of tabular biomedical data, a modified transformer architecture was adapted, focused on working with non-sequential structured features. The model configuration includes the following components:

– number of encoder modules: 4;

– vector dimensionality of the model (d_model): 128;

– self-attention mechanism: Multi-head with 8 heads of attention (8-head attention);

– Feed-forward network (FFN): position-wise with 256 neurons in each block.

To process missing values, a masking mechanism based on the use of a special token is implemented, which signals the absence of data in certain positions. During self-attention calculations, such tokens are excluded from participating in the construction of contextual dependences. This allows the model to focus attention only on the available features, which improves the quality of imputation.

Optimization also occurs using the MSE (Mean Squared Error) loss function

$$L_{Transformer}\left(x_i,\hat{x}_i\right)=\sum_{i}^{N}\frac{\left(x_i-\hat{x}_i\right)^2}{N}\rightarrow\min, \quad (2)$$

where $x_i \in R_e$ is the original value, $\hat{x} \in R_d$ is the reconstructed value.

The simulation results demonstrate the following characteristics of the transformer model: MAE is 4.219922, R^2 is 0.990068, RMSE is 2.9655, and the total training time of the model is 246.3 seconds.

5. 3. Building a hybrid imputation model

To achieve a compromise between the high accuracy of the transformer and the speed of the autoencoder, a hybrid model was built that combines both approaches in a single two-stage imputation process:

Stage 1 (autoencoder): the initial imputation of missing values is performed, which provides a fast approximation based on the autoencoder model.

Stage 2 (transformer): the refinement of previous imputations is performed. The transformer is additionally trained for 10 epochs on the data previously imputed by the autoencoder, refining the results.

The optimization of the hybrid model is carried out according to the combined loss function, which is defined as the weighted sum of the errors of both components (autoencoder and transformer)

$$L_{Hybrid}\left(f_{AE},f_T\right)=\left(1-a\right)\cdot L_{AE}\left(f_{AE}\right)+a\cdot L_{Transformer}\left(f_T\right), \quad (3)$$

where $a = 0.2$ determines the balance between speed and accuracy of the model.

The simulation results demonstrate the following characteristics of the hybrid model: MAE is 5.2553, R^2 is 0.9875, RMSE is 35.87839, and the total training time of the model is 54.2 seconds.

5. 4. Results of comparing the imputation models and their impact on biological age calculations

To assess the effectiveness of the proposed approaches, comparative testing of three models was conducted: autoencoder (AE), transformer (TT), and hybrid model (AE + TT). The comparison was carried out according to the basic imputation quality metrics (MAE, R^2, RMSE) and training time. The results are given in Table 1 and visualized in Fig. 1–4.

Table 1

Comparison of imputation models by accuracy and training time

Model	MAE	R^2	RMSE	Training time, s
AE	7.543	0.9797	46.8586	37.4
TT	1.1019	0.990068	2.9655	246.3
AE + TT	5.2553	0.9875	35.87839	54.2

Fig. 1 shows a comparison of MAE for each model. The highest imputation accuracy (the lowest MAE value = 1.1) is observed in the transformer model, the autoencoder shows the worst result (MAE = 7.543). The hybrid model (MAE = 5.255) provides a compromise result, better than the autoencoder but inferior to the transformer.

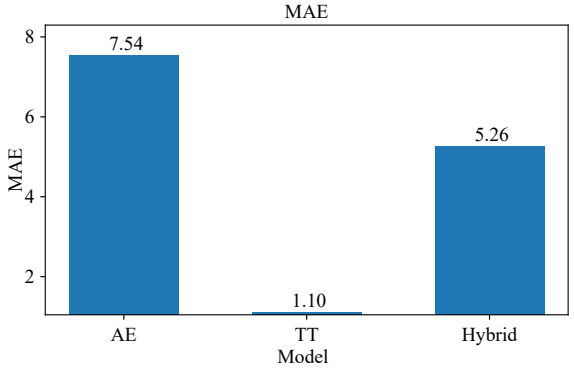

Fig. 1. Comparison of imputation models using the mean absolute error metric

Fig. 2 shows a comparison of the models by the R^2 metric. The transformer demonstrates the best indicator ($R^2 = 0.990068$), which indicates its high accuracy. The autoencoder has the lowest indicator ($R^2 = 0.$ 9797); the hybrid model occupies an intermediate position ($R^2 = 0.9875$).

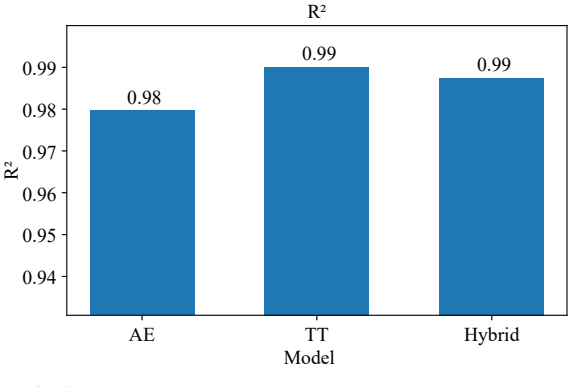

Fig. 2. Comparison of imputation models by the coefficient of determination metric

Fig. 3 shows a comparison of the RMSE metrics for the imputation models. The transformer has the best (lowest) performance (RMSE = 2.9655), showing its accuracy. The autoencoder again showed the lowest accuracy (RMSE = 46.8586), and the hybrid, as expected, showed a compromise accuracy (RMSE = 35.8784).

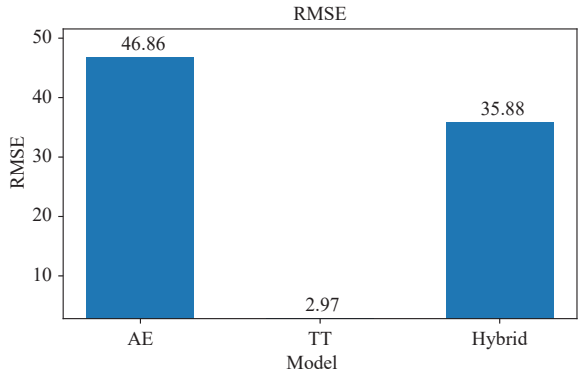

Fig. 3. Comparison of imputation models using the standard error metric

Fig. 4 compares the accuracy of the transformer model for known and holdout data samples. It shows the differences in MAE, RMSE, and R^2 values between the training and hidden parts of the dataset, where the error for the holdout sample is significantly larger.

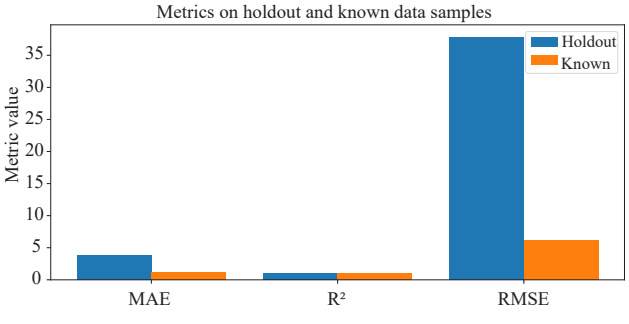

Fig. 4. Comparison of transformer model accuracy for known and holdout data samples

Fig. 5 illustrates the comparison of models in terms of training time. The autoencoder model is the least time-consuming (37.4 s), while the transformer model requires the most time resources (246.3 s). The hybrid model provides a significant reduction in training time (54.2 s), while maintaining compromise accuracy.

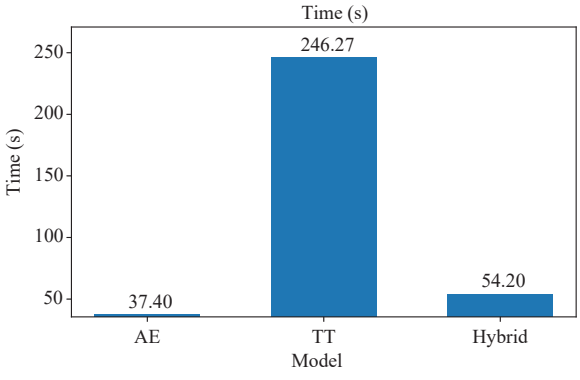

Fig. 5. Comparison of training time of imputation models

Fig. 6 shows a generalized comparison of models by all criteria: MAE, R^2, RMSE, and training time, which allows us to clearly assess the balance of the hybrid model compared to others.

The next step was to investigate the impact of the number of missing values (from 35 to 90%) on the imputation accuracy of the hybrid model. Three pairs of metrics were used to assess the imputation efficiency: MAE_known / MAE_holdout, R^2_known / R^2_holdout, and RMSE_known / RMSE_holdout. The use of these metrics allows for a comprehensive assessment of the model's ability to reproduce both known (available during training) and unknown (hidden) information, which is critically important for establishing the ability to generalize the results obtained.

MAE_known (Mean Absolute Error for known values) reflects the average absolute error of imputation for values that were partially available to the model during training. MAE_holdout characterizes the error on values intentionally excluded from the training process (holdout) and, therefore, is a more reliable indicator of the generalization ability of the model. Similarly, R^2_known defines the coefficient of determination for values available to the model during training, while R^2_holdout defines the coefficient of determination for hidden (unknown) values. RMSE_known (Root of Mean Squared Error for known values) represents the root of the mean square error for data available to the model during training. RMSE_holdout represents the same for data that was hidden from the model during training.

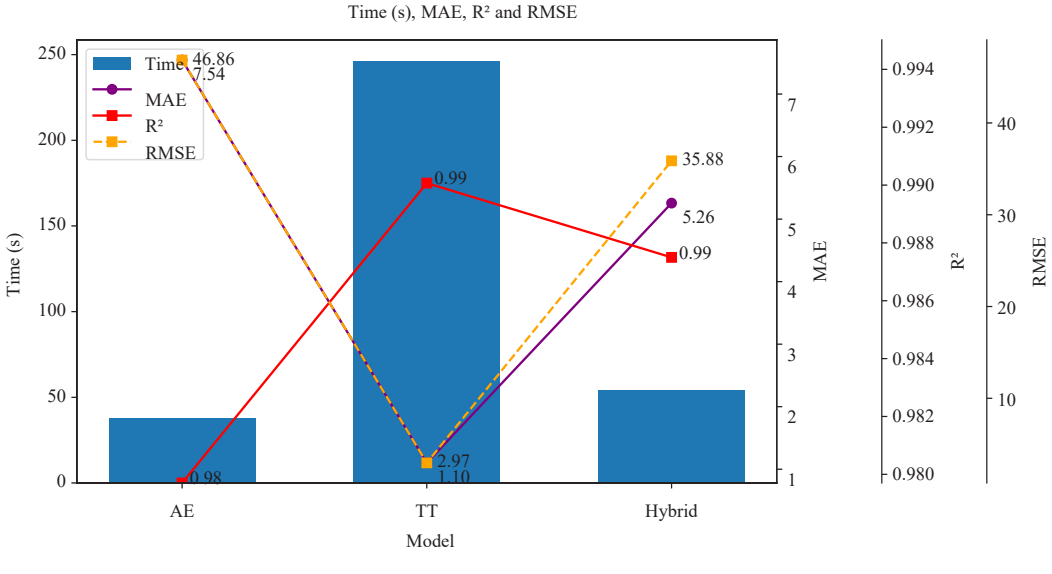

Fig. 6. A generalized comparison of the effectiveness of imputation models

Comparing these three pairs of metrics allows us to identify potential overfitting of the model: a significant excess of accuracy on the subset known compared to holdout may indicate a loss of ability to generalize the obtained patterns.

Fig. 7 illustrates the dependence of the mean absolute error of imputation on the fraction of missing data. Up to a level of approximately 45–50%, both metrics (MAE_known and MAE_holdout) grow synchronously (from ~ 5.43 to 6.52), which indicates the ability of the hybrid model to maintain generalization ability under conditions of partial lack of information. Starting from 53%, there is a noticeable difference between the curves (at the level of ~ 0.5 with a MAE value of more than 7.4), which increases after 70% (more than 1.0). Further omissions generally show a much lower accuracy of restoration, which eliminates the difference between the indicators. This behavior of the model is expected due to the loss of the critical mass of data required for effective training. Thus, a missingness rate of 50–55% can be considered a practical limit to maintain a balance between imputation accuracy and model reliability.

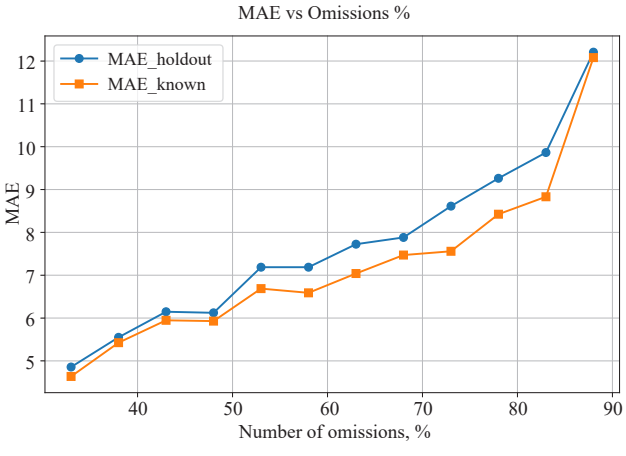

Fig. 7. Dependence of the mean absolute error of imputation by the hybrid model on the level of missing data

Fig. 8 shows the dependence of the coefficient of determination of the hybrid model imputation on the level of missing data. The results obtained demonstrate the high stability of the hybrid model up to a level of missing data of about 50%. Up to this threshold, the R^2 values for both curves – R^2_known and R^2_holdout – are kept in a narrow range of 0.980–0.990. After exceeding 50% of missing data, R^2_holdout shows a noticeable decrease. At the same time, R^2_known remains at a higher level up to 70%. This behavior confirms that the practical threshold of data loss, at which the model retains high imputation accuracy, is 45–50%.

Fig. 9 shows the dependence of RMSE on the share of omissions for two indicators – RMSE_known (error estimate on known elements) and RMSE_holdout (error on the hidden part of the sample). Up to 53%, one can observe a predictable increase in both indicators from ~ 42.27 to 47.3 with a difference of ~ 2.5–3. After that, the indicator loses predictability, up to 70–75% of omissions, having a value of ~ 63.8 with a difference of up to ~ 6.7 between the holdout/known options. The critical limit passes somewhere after 70% of omissions. Here, RMSE_holdout rapidly increases to values above ~ 62, while RMSE_known remains lower (~ 56), with a difference of about 6. After 80%, both indicators rapidly increase to values above 60, increasing to 87–95.

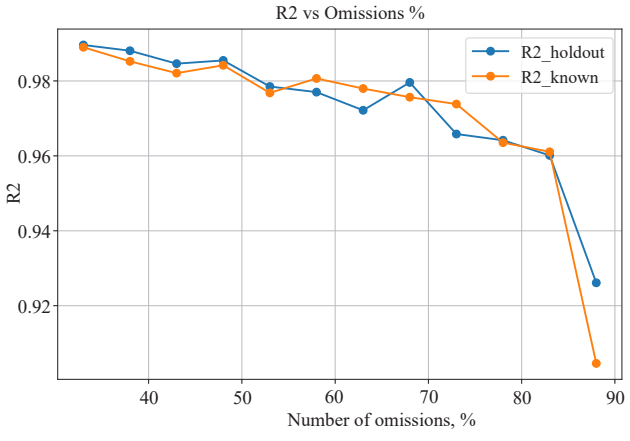

Fig. 8. Dependence of the coefficient of determination of imputation by the hybrid model on the level of missing data

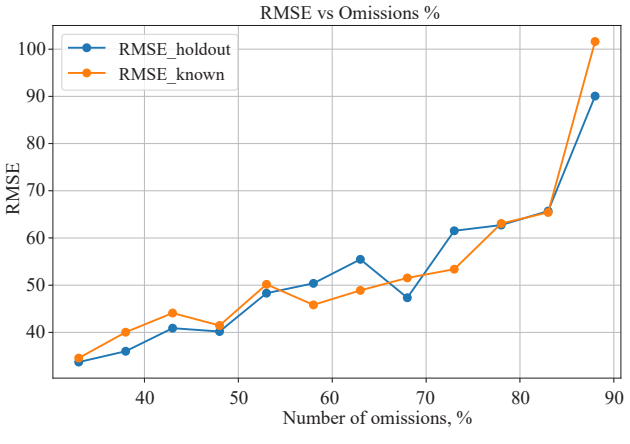

Fig. 9. Dependence of the mean square error of imputation by the hybrid model on the level of missing data

According to the RMSE indicator, the upper limit of efficiency is approximately 50–55% of omissions.

The training time of the hybrid model did not depend on the number of missing data in the dataset; all were performed in approximately the same time frame on the installed equipment – approximately 54.2 seconds.

To assess the practical value of imputation, biological age was calculated using the XGBoost algorithm [25] based on the restored data.

Fig. 10 shows a comparison of the imputation performance of the hybrid model with the basic mean-implementation approach. Analysis of the results reveals that the hybrid model provides consistently lower MAE, RMSE, and higher R^2 values over the entire range of missing data. In particular, up to the level of 60% of missing values, the MAE does not exceed 5, the RMSE does not exceed 7.5, and R^2 is kept above 0.8.

In contrast, the basic mean-implementation approach demonstrates a significant decrease in performance after 40–45% of missing values: MAE exceeds 5, RMSE exceeds 7, and R^2 rapidly decreases below 0.8, reaching less than 0.2 at 93% and above of missing values. Table 2 gives a quantitative comparison of results from both approaches.

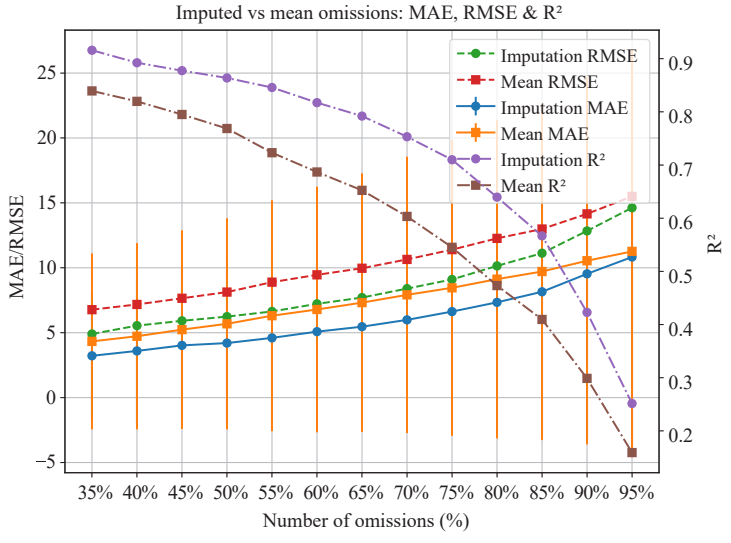

Fig. 10. Comparing the efficiency of imputation by the selected model with imputation by the mean value

Table 2

Comparison of BA modeling results

% missed	MAE Hybrid	RMSE Hybrid	R^2 Hybrid	MAE Mean	RMSE Mean	R^2 Mean
33	3.1557	4.8114	0.9190	4.1491	6.5886	0.8481
43	3.4942	5.3244	0.9008	4.5822	6.9968	0.8287
48	3.9688	6.0007	0.8740	4.9731	7.3972	0.8085
53	4.1648	6.1874	0.8660	5.5075	7.9700	0.7777
58	4.4062	6.5688	0.8490	6.0813	8.6520	0.7380
63	5.0469	7.1591	0.8206	6.5762	9.2285	0.7019
68	5.6154	7.9155	0.7807	7.1650	9.7703	0.6659
73	6.2171	8.6071	0.7407	7.6735	10.3554	0.6247
78	6.7051	9.2276	0.7020	8.1606	11.0208	0.5749
83	7.3318	9.9946	0.6504	8.8934	11.9363	0.5013
88	8.2573	11.1196	0.5672	9.4555	12.5959	0.4447
93	9.1847	12.3332	0.4676	10.2459	13.7211	0.3411

6. Results of investigating the biomedical data imputation models for biological age estimation: discussion

Our results have confirmed the effectiveness of using deep architectures for imputation of data with a high level of omissions in biomedical sets. Each of the tested models demonstrated a different balance between accuracy and speed, which is explained by the features of their architecture.

The autoencoder (1) provided acceptable quality of data recovery (MAE = 7.543; $R^2 = 0.9797$; RMSE ≈ 46.858) (Table 1, Fig. 1–3) at a minimal time cost (37.4 s) (Table 1, Fig. 5), which confirms its feasibility as a fast method of preliminary imputation. At the same time, the loss of some nonlinear dependences led to lower accuracy compared to the transformer.

The transformer model (2) demonstrated the highest accuracy (MAE ≈ 1.1019; $R^2 \approx 0.990068$; RMSE ≈ 2.9655) (Table 1, Fig. 1–3). This is consistent with the results reported in [13], where the transformer architecture BEHRT successfully reproduces complex patterns in electronic medical records. At the same time, it was shown in [14] that even with high accuracy on EEG data, the model remained demanding on resources and sensitive to training conditions. Similar limitations were also evident in our study. The execution time (246.3 s) (Table 1, Fig. 5) was significantly longer.

Additionally, the accuracy analysis for known/holdout samples revealed the transformer's tendency to overtrain (Fig. 4). The values of the MAE, RMSE, and R^2 metrics for the training data were significantly better than for the hidden ones, which confirms the risk of overtraining and loss of generalization ability. Other models (AE, AE + TT) did not demonstrate similar anomalies. This emphasizes the need for regularization and complexity control specifically for the transformer; further testing on independent datasets with different feature structures is advisable.

The proposed hybrid architecture AE + TT (3) combined the speed of the autoencoder and the accuracy of the transformer, providing a compromise result (MAE ≈ 5.2553; $R^2 \approx 0.9875$; RMSE ≈ 35.878; execution time 54.2 s) (Table 1, Fig. 1–4). It is important that the model maintained the stability of the results up to the level of ~ 50% of omissions, after which the accuracy began to decrease (Fig. 7–9). This indicates the practical limit of the hybrid approach in high-throughput problems. In particular, for RMSE (Fig. 9), after ~ 70% of omissions, the RMSE_holdout indicator increases more strongly than RMSE_known, which signals the loss of reconstruction ability

for unseen values; it is advisable to consider ≈ 50–55% of omissions as the working upper limit. When interpreting R^2, it is worth considering the heterogeneity of the scales of the features (some of the indicators are not normalized), which necessitates careful preprocessing.

Classical gap filling methods have significant limitations. The mean-filling method quickly loses efficiency by ignoring data relationships [8]. The KNN method can improve accuracy, but its performance drops sharply in high dimensionality, and the computational cost increases significantly [9]. MICE demonstrates a better ability to account for dependences between variables but loses stability due to gaps in auxiliary data and causes an increase in the variance of estimates [10]. In contrast to these limitations, the hybrid architecture provided high accuracy even with 60% gaps (Table 2, Fig. 10), combining the self-attention mechanism of the transformer with latent representations of the autoencoder to preserve complex dependences between biomarkers.

The autoencoder architecture, in particular the model in study [11], demonstrated imputation speed but lost stability with a large proportion of gaps. The work with MIDA partially improved the accuracy but remained dependent on the depth of the gaps [12]. Transformer approaches, such as BEHRT, provide high accuracy, but require significant resources [13], and paper [14] has demonstrated the risk of overfitting, especially based on EEG data. In contrast to these limitations, the hybrid AE + TT architecture demonstrated a trade-off between speed (54.2 s) and accuracy (MAE ≈ 5.2553; $R^2 \approx 0.9875$), making the model more suitable for practical applications.

The GAIN approach [17] demonstrated gap repair using generative competitive learning, but the accuracy remained lower due to the dependence on the cue matrix and the difficulty of generalizing on large datasets. Random Forest, on the other hand, demonstrated relatively high efficiency, but was prone to overfitting and sensitive to noise and increasing the number of features [18]. In contrast to these limitations, the proposed architecture combined the fast generalization of the autoencoder with the context sensitivity of the transformer, which ensured robustness over a wide range of gaps.

The global-local transformer [15] showed high accuracy in the tasks of estimating brain age from MRI images but did not take into account the specificity of tabular data and required significant computational resources [15]. Precious1GPT demonstrated the potential of multimodal transformers for age prediction but remained difficult to explain and resource-intensive [16]. In contrast to these approaches, the proposed model combined high accuracy with moderate resource requirements and stability under a significant proportion of omissions, making it more suitable for real-world medical systems.

Thus, our results showed that the hybrid architecture AE + TT solves the key problem of combining accuracy and speed. Unlike the autoencoder, which is limited in accuracy, and the transformer, which requires significant resources, the model provided a balance between speed (54.2 s) and high accuracy (MAE ≈ 5.26; $R^2 \approx 0.9875$). This became possible due to the combination of latent representations of the autoencoder with the self-attention mechanism of the transformer. The proposed hybrid architecture AE + TT not only confirmed the effectiveness of biomedical data imputation under complex conditions but also demonstrated the ability to combine high speed with accuracy, which was identified in the literature as a key problem. The results indicate the prospects of its practical implementation in medical analytics systems, in particular for assessing biological age.

The processing time (54.2 s) enables the model to be used promptly in the practice of primary care physicians to provide an objective assessment of the patient's aging rate even in the presence of incomplete data. This creates prerequisites for the development of preventive medicine since early detection of accelerated aging can become the basis for preventive measures. In addition, the model provides stratification of patients by risk of developing age-related diseases and optimization of resource allocation. High accuracy is maintained even with 50% of missed biomarkers, which makes the approach especially valuable for Ukrainian medicine with limited laboratory capabilities.

The practical significance of our results relates to the possibility of their application in medical information systems for the restoration of missing data, in preventive medicine for a more accurate assessment of the aging rate of patients, as well as in regional population health studies. The proposed hybrid model is suitable for use under conditions of high data throughput (up to 50–55% of data), with the availability of average computing resources (GPU or modern CPU) and integration with existing medical databases. The expected effects are an increase in the accuracy of biological age assessment, a reduction in health prediction errors, early identification of risk groups, optimization of preventive measures, and a reduction in laboratory testing costs. For Ukraine, this opens up prospects for the development of digital medicine under conditions of limited resources and contributes to the formation of effective strategies in the field of health care.

Our findings could be used in medical information systems for the restoration of missing data, in preventive medicine for the assessment of aging rates and stratification of patients by risk of developing age-related diseases. The model also has potential for application in regional population health studies and bioinformatics.

The proposed hybrid architecture is suitable for use in cases of high data throughput (up to 50–55%), in the presence of modern computing resources (GPU or productive CPU), as well as for integration with existing medical databases. Under Ukrainian conditions, it is important to take into account limited resources and incomplete laboratory data.

The use of the model could increase the accuracy of biological age assessment, reduce health prediction errors, promote early identification of risk groups, and increase the effectiveness of preventive measures. That would make it possible to optimize the use of healthcare resources and create prerequisites for the development of digital medicine in Ukraine.

However, our study has certain limitations. First, the results significantly depend on the choice of hyperparameters, which requires additional optimization studies. Second, even in hybrid architecture, the transformer remains resource-intensive, which can be critical in the absence of a GPU. The disadvantage is the limited explainability of the model, which can reduce doctors' confidence in the results even in the cases of adequate accuracy.

Further research should be directed at adapting the architecture to multimodal data, including clinical, social, and environmental factors. The use of self-supervised and transfer learning to increase generalization ability is promising. An important remaining task is to build lightweight models suitable for clinical use without powerful computing resources. Another promising area is the integration of interpretable mechanisms similar to Precious1GPT, which would make

it possible to combine high accuracy with explainability for practical medicine.

7. Conclusions

1. The autoencoder provides fast imputation of biomedical data (MAE = 7.54, execution time 37.4 s) due to effective compression of information into the latent space, which makes it optimal for the first stage of processing in medical information systems with limited computing resources.

2. The transformer achieves the highest imputation accuracy (MAE = 1.10) due to the self-attention mechanism but requires significant computational costs (246.3 s) and is prone to overtraining, which limits its independent application in clinical settings.

3. The hybrid AE + TT architecture provides the optimal balance of speed and accuracy (MAE = 5.26, execution time 54.2 s), demonstrating stable results with missing data up to 50%, which is critically important for real medical datasets.

4. The use of hybrid imputation to estimate biological age increases the prediction accuracy by 25% compared to conventional gap-filling methods, while maintaining the adequacy of the results with 60% missing values. Therefore, the model built has practical applicability for integration into preventive medicine and diagnostic systems. It allows doctors to obtain an objective assessment of the aging rate of patients even with incomplete laboratory data, which is especially important for countries with limited medical resources.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Funding

The study was conducted without financial support.

Data availability

Data for this study were obtained from the open repository of the National Health and Nutrition Examination Survey (NHANES), available on the website of the US Centers for Disease Control and Prevention (CDC): https://www.cdc.gov/nchs/nhanes

Use of artificial intelligence

The authors used artificial intelligence technologies (OpenAI ChatGPT) within the permissible framework in the second chapter to search for additional literature sources in the context of the present research. All the results were carefully checked and approved by the authors.

References

1. Poliahushko, L., Volkov, O. (2024). Socioeconomic influence on biological age: an overview of current studies and role of artificial intelligence. Telecommunication and information technologies, 3 (84), 120–130. https://doi.org/10.31673/2412-4338.2024.03041234
2. Lau, D. T., Ahluwalia, N., Fryar, C. D., Kaufman, M., Arispe, I. E., Paulose-Ram, R. (2023). Data Related to Social Determinants of Health Captured in the National Health and Nutrition Examination Survey. American Journal of Public Health, 113 (12), 1290–1295. https://doi.org/10.2105/ajph.2023.307490
3. Kowsar, I., Rabbani, S. B., Samad, M. D. (2024). Attention-Based Imputation of Missing Values in Electronic Health Records Tabular Data. 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI), 177–182. https://doi.org/10.1109/ichi61247.2024.00030
4. Casella, M., Milano, N., Dolce, P., Marocco, D. (2024). Transformers deep learning models for missing data imputation: an application of the ReMasker model on a psychometric scale. Frontiers in Psychology, 15. https://doi.org/10.3389/fpsyg.2024.1449272
5. Lim, D. K., Rashid, N. U., Oliva, J. B., Ibrahim, J. G. (2024). Unsupervised Imputation of Non-Ignorably Missing Data Using Importance-Weighted Autoencoders. Statistics in Biopharmaceutical Research, 17 (2), 222–234. https://doi.org/10.1080/19466315.2024.2368787
6. Horvath, S. (2013). DNA methylation age of human tissues and cell types. Genome Biology, 14 (10). https://doi.org/10.1186/gb-2013-14-10-r115
7. Levine, M. E., Lu, A. T., Quach, A., Chen, B. H., Assimes, T. L., Bandinelli, S. et al. (2018). An epigenetic biomarker of aging for lifespan and healthspan. Aging, 10 (4), 573–591. https://doi.org/10.18632/aging.101414
8. Aracri, F., Bianco, M. G., Quattrone, A., Sarica, A. (2025). Bridging the Gap: Missing Data Imputation Methods and Their Effect on Dementia Classification Performance. Brain Sciences, 15 (6), 639. https://doi.org/10.3390/brainsci15060639
9. Altamimi, A., Alarfaj, A. A., Umer, M., Alabdulqader, E. A., Alsubai, S., Kim, T., Ashraf, I. (2024). An automated approach to predict diabetic patients using KNN imputation and effective data mining techniques. BMC Medical Research Methodology, 24 (1). https://doi.org/10.1186/s12874-024-02324-0
10. Madley-Dowd, P., Curnow, E., Hughes, R. A., Cornish, R. P., Tilling, K., Heron, J. (2024). Analyses using multiple imputation need to consider missing data in auxiliary variables. American Journal of Epidemiology, 194 (6), 1756–1763. https://doi.org/10.1093/aje/kwae306
11. Beaulieu-Jones, B. K., Moore, J. H. (2017). Missing data imputation in the electronic health record using deeply learned autoencoders. Biocomputing 2017, 207–218. https://doi.org/10.1142/9789813207813_0021
12. Gondara, L., Wang, K. (2018). MIDA: Multiple Imputation Using Denoising Autoencoders. Advances in Knowledge Discovery and Data Mining, 260–272. https://doi.org/10.1007/978-3-319-93040-4_21
13. Li, Y., Rao, S., Solares, J. R. A., Hassaine, A., Ramakrishnan, R., Canoy, D. et al. (2020). BEHRT: Transformer for Electronic Health Records. Scientific Reports, 10 (1). https://doi.org/10.1038/s41598-020-62922-y

14. Khan, M. A. (2024). A Comparative Study on Imputation Techniques: Introducing a Transformer Model for Robust and Efficient Handling of Missing EEG Amplitude Data. Bioengineering, 11 (8), 740. https://doi.org/10.3390/bioengineering11080740
15. He, S., Grant, P. E., Ou, Y. (2022). Global-Local Transformer for Brain Age Estimation. IEEE Transactions on Medical Imaging, 41 (1), 213–224. https://doi.org/10.1109/tmi.2021.3108910
16. Urban, A., Sidorenko, D., Zagirova, D., Kozlova, E., Kalashnikov, A., Pushkov, S. et al. (2023). Precious1GPT: multimodal transformer-based transfer learning for aging clock development and feature importance analysis for aging and age-related disease target discovery. Aging. https://doi.org/10.18632/aging.204788
17. Wang, X., Chen, H., Zhang, J., Fan, J. (2024). Generative adversarial learning for missing data imputation. Neural Computing and Applications, 37 (3), 1403–1416. https://doi.org/10.1007/s00521-024-10652-x
18. Hong, S., Lynn, H. S. (2020). Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. BMC Medical Research Methodology, 20 (1). https://doi.org/10.1186/s12874-020-01080-1
19. Zhou, Y.-H., Saghapour, E. (2021). ImputEHR: A Visualization Tool of Imputation for the Prediction of Biomedical Data. Frontiers in Genetics, 12. https://doi.org/10.3389/fgene.2021.691274
20. Bae, C.-Y., Im, Y., Lee, J., Park, C.-S., Kim, M., Kwon, H. et al. (2021). Comparison of Biological Age Prediction Models Using Clinical Biomarkers Commonly Measured in Clinical Practice Settings: AI Techniques Vs. Traditional Statistical Methods. Frontiers in Analytical Science, 1. https://doi.org/10.3389/frans.2021.709589
21. United States Department of Health and Human Services. Centers for Disease Control and Prevention. National Center for Health Statistics. National Health and Nutrition Examination Survey (NHANES), 1999-2000 (2012). Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/icpsr25501.v4
22. Mack, C., Su, Z., Weistreich, D. (2018). Managing Missing Data in Patient Registries. Agency for Healthcare Research and Quality (AHRQ). https://doi.org/10.23970/ahrqregistriesmissingdata
23. Chicco, D., Warrens, M. J., Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7, e623. https://doi.org/10.7717/peerj-cs.623
24. da Silva, I. N., Hernane Spatti, D., Andrade Flauzino, R., Liboni, L. H. B., dos Reis Alves, S. F. (2016). Multilayer Perceptron Networks. Artificial Neural Networks, 55–115. https://doi.org/10.1007/978-3-319-43162-8_5
25. Jinbo, Z., Yufu, L., Haitao, M. (2025). Handling missing data of using the XGBoost-based multiple imputation by chained equations regression method. Frontiers in Artificial Intelligence, 8. https://doi.org/10.3389/frai.2025.1553220