Deep Learning-Based Missing Value Imputation for Heart Failure Data from MIMIC-III: A Comparative Study of DAE, SAITS, and MICE+LightGBM

Authors: Shilpa Sharma, Mandeep Kaur, Savita Gupta

Affiliations: UIET, Panjab University, Chandigarh

Abstract

Background: Inadequate data in electronic health records can create problems for clinical decision support systems and predictive modelling tools. ICU patients with heart failure may have incomplete physiological data as a result of irregular patient monitoring, sensor failure, or clinical workflow issues, which highlights the need for strong imputation methods.

Objective: The goal of this research is to examine the performance of three different methods of imputing the missing value rows in heart failure patients' records (from the MIMIC-III database)the Denoising Autoencoder (DAE), Self-Attention Imputation for Time Series (SAITS) and Multiple Imputation by Chained Equations (MICE) with LightGBM—as compared against each other.

Methods: We created our own heart failure cohort with 14,090 ICU stays from the MIMIC-III database. After performing a strict 10-step preprocessing pipeline, we utilized Random Forest, correlation analysis, and mutual information, in order to select the final 21 clinically relevant features from the original 81 features. To mimic realistic clinical scenarios, we introduced artificial missing values in three different amounts; 20%, 30%, and 50%. The final three methods used to impute these missing values were DAE (deep learning autoencoder), SAITS (Transformer-based attention mechanism), and MICE+LightGBM (hybrid statistical and machine-learning method). We evaluated the performance of our models using the following three metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Normalised Root Mean Square Error (NRMSE).

Results: DAE and SAITS demonstrated superior and comparable performance across all levels of missingness. At 20% missingness, DAE achieved a mean MAE of 0.004967, a mean RMSE of 0.005217 and a global NRMSE of 3.260893; SAITS achieved a mean MAE of 0.005461, a mean RMSE of 0.005797 and a global NRMSE of 3.244695, while MICE+LightGBM showed substantially higher error (mean MAE 0.106452, mean RMSE 0.223782 and global NRMSE 0.222056). At 50% missingness, SAITS exhibited the best performance (mean MAE 0.004538, mean RMSE 0.004954, global NRMSE=2.618904), followed by DAE (mean MAE 0.004992, mean RMSE 0.005307, global NRMSE 3.27805), with MICE+LightGBM showing degraded performance (mean MAE 0.19729, mean RMSE 0.3767273, global NRMSE 0.377731). Feature-level analysis revealed that deep learning methods maintained consistent accuracy across diverse physiological variables, including lactate, pH, bicarbonate, and blood urea nitrogen.

Conclusions: Deep learning-based imputation methods (DAE and SAITS) significantly outperform traditional hybrid approaches for missing-value imputation in heart failure ICU data, demonstrating robustness across varying levels of missingness and feature types. These findings

support the adoption of DL-based imputation strategies in clinical decision support systems for heart failure management.

Keywords: clinical data preprocessing, SAITS, Missing value imputation, heart failure, MIMIC-III, denoising autoencoder, electronic health records, intensive care unit, deep learning,

1. Introduction

Heart failure is one of the most common causes of morbidity and mortality in the world, and ICU admissions are among the most serious events requiring ongoing physiological measurements and quick clinical responses. The available EHR data for ICU use, which are the only longitudinal data that are collected continuously during all ICU visits, are comprised of vital signs and laboratory information (both of which are collected continuously), medications provided, and eventual clinical outcomes for each patient. Unfortunately, because of various reasons (i.e., a lack of consistent measurement schedule and/or equipment failure, prioritising clinical assessments, etc.) much of this data is often incomplete [1-3]. As such, when data researchers try to use historical data from clinical databases in predictive modelling, clinical decision support systems or retrospective studies, they encounter significant problems due to the abundance of missing values in clinical time series data [4-5]. In fact, the prevalence of missing values in MIMIC (an ICU clinical database) can reach upwards of 80% for certain laboratory values [6-7]. Traditional approaches to addressing missing data—such as complete cases analysis, simple mean imputation and Last Observation Carried Forward (LOCF)—can introduce significant bias, decrease statistical power and do not capture well the complex temporal and multivariate dependencies associated with clinical data [8-9]. Consequently, there is typically a great deal of interest in advanced approaches to impute missing data using the rich nature of EHR data to derive accurate and clinically plausible estimates of the correct data points that were missing.

The last few years have seen a growth in the number of deep learning techniques available for filling in missing values in patient charts for clinical trials due to advances in technology and new research methods. For example, taking a look at the mortality rates for patients in both the MIMIC-III and MIMIC-IV patient populations, many researchers have recently published their findings using various forms of deep learning to develop new models that incorporate either or both imputation and mortality rate-based predictions (i.e., VAEs + RNNs). Similarly, researchers have developed GANs that help resolve the issue of both imputation and balancing between classes with respect to the number of records in their datasets [2]. Over the last few years SAITS, a transformer-based architecture, has achieved state-of-the-art performance on the imputation of clinical time-series data using MIMIC-IV monitoring datasets.

While there are many examples of success, there is also significant room for improvement across multiple areas as defined by the literature. Many studies using deep learning have focused primarily on either the general population of ICU patients, or have simply focused on one or two conditions (i.e., Sepsis), and as a result, little research has been conducted on patients with heart failure (which are an important population of study because they have specific pathophysiological characteristics). Additionally, there are few studies available that compare various forms of hybrid statistical and machine learning for imputation and the analysis of such patients using their aggregated time-series data instead of cross-sectional or longitudinal data (i.e., various machine learning algorithms—ex: gradient boosting + MICE). Finally, there is a

need for a systematic evaluation of imputation performance across a range of various missing data rates and a large number of different clinical variables (i.e., vitals, lab tests, labs, calculated statistics) to support clinical decision-making around the implementation of algorithms for either or both features of missing data. In order to fill a lack of information regarding imputation algorithm performance when used for heart failure patients, we developed a rigorous comparative analysis of three different imputation algorithms (denoising autoencoders [DAEs], self-attention imputation for time series [SAITS] and multiple imputation by chained equations + LightGBM [MICE+LightGBM]) on a large heart failure cohort obtained from MIMIC-III. Our main research questions are: (1) how accurate are different types of imputation in heart failure data that have varying degrees of missingness (20%, 30% and 50%); (2) do deep learning algorithms provide better performance compared to traditional statistical and machine learning methods; (3) how do the algorithms perform at both a global level and at an individual variable level; and lastly (4) what evidence-based guidelines can be provided to support imputation strategy selection for use in clinical decision support systems for heart failure.

---

3. Materials and Methods

3.1 Data Source and Cohort Selection

3.1.1 MIMIC-III Database

The MIMIC-III free open-access critical care database was utilized for this analysis. The MIMIC-III database contains de-identified patient data from 61,532 intensive care unit (ICU) admissions to the Beth Israel Deaconess Medical Center over a period of time (2001-2012) [28]. The MIMIC-III clinical database contains a comprehensive repository of data for use in clinical informatics research [29], with a large amount of clinical information, including demographics, vital signs, laboratory test results, medications, procedures, diagnoses, and outcomes. The MIMIC-III clinical database is an extensively validated resource that has been used widely as a source for data is used for conducting research related to clinical informatics.

3.1.2 Heart Failure Cohort Construction

To analyze heart failure admissions, we utilized the DIAGNOSES_ICD table containing International Classification of Diseases, Ninth Revision (ICD-9) diagnosis codes for heart failure: 428.0 (congestive heart failure, unspecified), 428.20 (systolic heart failure, unspecified), and 428.21 (acute systolic heart failure). These codes were utilized based on their clinical relevance as well as their use in previous studies concerning heart failure conducted by MIMIC-III [30]. Only unique hospital admissions (HADM_ID) associated with these codes remain in our data.

Heart failure hospitalizations and ICU lengths of stay were evaluated using the ICUSTAYS data. ICU admission times (INTIME) were used as a temporal anchor for all subsequent data extraction. After applying the inclusion and data quality criteria, there were 14,090 ICU stays in the final cohort of study patients to be analyzed. The complete methodology for this study is found in Figure 1.

3.2 Data Preprocessing Pipeline

To ensure the integrity (quality, temporal validity and clinical relevance) of our data, we implemented a rigorous 10-step pre-processing pipeline. Below are the first steps:

1) Environment Setup - Set up a Python programming environment and used the pandas library for data manipulation and NumPy for numerical operations. Chunk-based processing was used to process large MIMIC-III event tables.

2) Heart Failure Identification - Extracted all heart failure admissions from the DIAGNOSES_ICD table based on ICD-9 discharge codes.

3) ICU Stay Selection - Linked each ICU stay to its corresponding heart failure admission using the ICUSTAYS table with the INTIME field used as the time reference.

4) Feature Selection / ITEMID Mapping - Selected 22 physiological features based on clinical relevance found in the literature. Events corresponding to vital signs (e.g., heart rate, respiratory rate, blood pressure, temperature) were pulled from CHARTEVENTS table; laboratory results (e.g.,hemoglobin, creatinine, glucose, electrolytes, coagulation markers) were pulled from LABEVENTS table. The ITEMID from each of these tables was mapped to the human-readable feature name.

5) Physiological Range Definitions - Defined ranges for each variable using clinically plausible values as determined by clinical judgement and reviewed literature. Measurements that fall outside these ranges were excluded from analysis to eliminate physiologically implausible or false measurements.

6) Temporal Filtering - Event data was merged with the status information from the ICU stay. Only measurements that were made within 24 hours after being admitted to the ICU and when the patient was aged over 16 were kept to ensure that the information remained temporally consistent and to avoid leakage of information.

7) Reshaping Data Set - Pivot operations were performed to change the long format into a wide format by converting it into a pivot table. Each row would be one unique ICU stay, and each column has one aggregated feature. This will produce a structured dataset with 14,110 rows and 97 features.

8) Removing Missing Values - All features that have over 85% of their values missing were removed. The Hospital Expired Flag (HOSPITAL_EXPIRE_FLAG) was included from the ADMISSIONS table. The number of records in the dataset is 14,090 ICU Stays and 81 total features (78 clinical and 3 demographic)

9) Feature Selection - The features that are to be used as predictors are 19 clinical and 2 demographic, selected from 81 total features using a combination of three methods: (1) the

importance of each feature variable using Random Forest with 93.74% classification accuracy; (2) Pearson correlation coefficients to eliminate redundant features, and (3) mutual information scores to determine which features will be most predictive of outcome. The data is presented in Figures 2 through 4.

10 Final Dataset Generation - The final processed dataset will contain 14,090 ICU heart failure stays with 81 total aggregated features and will be exported as a CSV file for usage in impute experiments

3.3 Selected Features

The 19 selected features represented diverse physiological domains critical for heart failure monitoring:

- **Lactate measurements:** mean_Lactate, max_Lactate, min_Lactate, std_Lactate
- **Acid-base balance:** mean_pH, min_pH, max_pH, std_pH
- **Renal function:** max_BUN, min_BUN
- **Electrolytes:** mean_Bicarbonate, min_Bicarbonate, max_Bicarbonate
- **Hematology:** mean_RDW, max_RDW, min_RDW, mean_Platelet
- **Respiratory:** mean_PaCO2, std_PaO2

These features capture key pathophysiological processes in heart failure including tissue perfusion (lactate), acid-base disturbances (pH, bicarbonate, PaCO2), renal dysfunction (BUN), and hematological abnormalities (RDW, platelets).

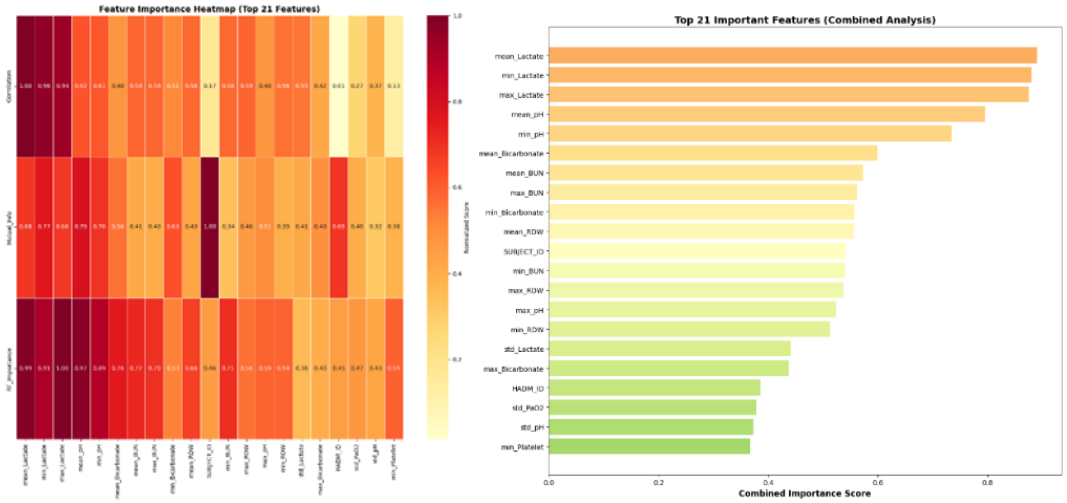


Figure 2. Feature importance heatmap of 21 features. Figure 4 combined analysis: Random Forest, Mutual Information score and Pearson correlation score for 21 important features selection.

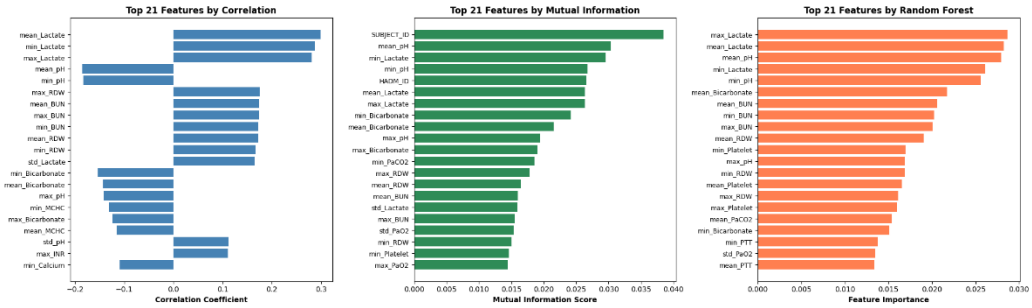


Figure 3(a) represents the feature selected by correlation coefficients, (b) represents the feature selected by Mutual information, and (c) represents the feature selected by Random Forest.

3.4 Missing Value Introduction Protocol

To determine how well imputation performs under controlled conditions, we created missing data on each of the 19 chosen features by selecting three different levels of missingness (20%, 30%, and 50%). This procedure is often used in imputation benchmarking research [9], [31] to produce a quantitative comparison between imputed values and true (or ground truth) values.Missing values were created using an MCAR (missing completely at random) method to provide an unbiased examination of how well the imputation algorithms performed statistically. At each of the three missingness levels, values from each feature were randomly selected to be masked, ensuring they were masked based on their original proportion. This resulted in producing three experimental datasets, with the true values of all masked samples known, allowing for the exact computation of imputation errors.

3.5 Imputation Algorithms

3.5.1 Denoising Autoencoder (DAE)

DAE has an encoder-decoder architecture with fully connected layer. Encoder compresses the input feature space to a low dimensional latent representation, decoder recreates the input feature space from the low dimensional latent representation. The DAE is trained to minimize the reconstruction error on the observed data, so denoised inputs are obtained by recovering missing values. Training employed mean squared error (MSE) loss on observed values only, with missing values masked during loss calculation. The Adam optimiser was used with learning rate scheduling. Dropout regularisation was applied to prevent overfitting. The trained model was then used to impute missing values in the test datasets.

3.5.2 Self-Attention Imputation for Time Series (SAITS)

SAITS employed a Transformer-based architecture with multi-head self-attention mechanisms. Although our data consisted of aggregated statistics rather than raw time series, SAITS was adapted to treat the 19 features as a multivariate sequence, leveraging its ability to capture cross-variable dependencies through attention mechanisms.

The model used diagonal masking during training to prevent trivial copying of observed values. Multiple attention heads enabled the model to learn diverse relationships among features. Positional encoding was adapted for the feature dimension. Training minimised reconstruction error on observed values using MSE loss with the Adam optimiser.

3.5.3 Multiple Imputation by Chained Equations with LightGBM (MICE+LightGBM)

MICE+LightGBM combined the principled multiple imputation framework with gradient boosting regression. For each variable with missing values, a LightGBM regression model was trained using all other variables as predictors. This process was iterated cyclically through all variables with missing values until convergence.

LightGBM hyperparameters were tuned for each variable, including the number of leaves, the learning rate, and the number of boosting rounds. Multiple imputation generated five complete datasets, and the results were pooled to obtain the final imputed values. This approach combined MICE's uncertainty quantification with LightGBM's strong performance on tabular data.

3.6 Evaluation Metrics

Imputation quality is typically assessed using metrics that quantify the discrepancy between imputed and true values. Mean Absolute Error (MAE) is an average of absolute differences and is an interpretable metric related to scale. Due to the squaring of errors, Root Mean Square Error (RMSE) places greater penalizing weight on larger errors, making RMSE sensitive to outliers. Normalized Root Mean Square Error (NRMSE) is the RMSE divided by either the variable range or standard deviation, allowing for comparisons between feature sets that are measured on different scales [25]. The global NRMSE will take RMSE for all masked values and will normalize it to the standard deviation of the actual observation.For imputation evaluation in clinical settings, assessments should take into account both the statistical accuracy of the imputed value as well as the clinical plausibility of the imputed value. Imputed values will have a lower MAE and lower RMSE indicating better statistical fit, however, physiologically implausible imputed values will produce insufficient clinical meaning between the imputed value and other variable value relationships [26]. Additionally, downstream task performance (e.g., mortality prediction accuracy) provides an important complementary evaluation of imputation quality [27].

Imputation performance was assessed using three complementary metrics:

Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\, y_i - \hat{y}_i|$$

where y_i is the true value and $\hat{y}_i$ is the imputed value for the i-th missing observation.

Root Mean Square Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\, y_i - \hat{y}_i)^2}$$

Global Root Mean Squared Error (NRMSE):

Global Normalised Root Mean Squared Error (NRMSE):

$$\mathbf{NRMSE_{global}} = \frac{\sqrt{\frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{y}_k)^2}}{\mathrm{Std}(y)}$$

Where: y_k are the true masked values**,** $\hat{y}_k$ are the imputed values**,** N is the number of masked samples **and** Std(y) is the standard deviation of the true masked values

Mean RMSE (feature-wise average) — Formula

Each feature's RMSE is computed independently, and then the RMSE values are averaged across features.

$$RMSE_j = sqrt\left(\left(\frac{1}{N_j}\right) \times \Sigma\left(y_{ij} - \hat{y}_{ij}\right)^2\right)$$

Where: F = number of features, j = feature index, N_j = number of masked samples for feature *j*, y_ij = true value of feature *j* and ŷ_ij = imputed value of feature *j*

Mean MAE (feature-wise average) — Formula

$$Mean\ MAE = \left(\frac{1}{F}\right) \times \Sigma\, MAE_j$$

Where: **F** = number of features, **j** = feature index, **N_j** = number of masked samples for feature *j*, **y_ij** = true value of feature *j* and **ŷ_ij** = imputed value of feature *j*

Metrics were calculated both globally (averaged across all features and observations) and feature-specifically (for each of the 19 selected variables). This dual-level evaluation provided a comprehensive assessment of imputation quality across diverse physiological measurements.

3.7 Experimental Design

For each missingness level (20%, 30%, 50%), the following procedure was executed:

1. Randomly mask the specified percentage of values across 19 features
2. Apply each imputation algorithm (DAE, SAITS, MICE+LightGBM) to the masked dataset
3. Calculate MAE, RMSE, and NRMSE by comparing imputed values to ground truth
4. Compute global metrics (averaged across all features) and feature-specific metrics
5. Repeat the process with different random seeds to ensure reproducibility

Experiments were done in a computing cluster consisting of NVIDIA's GPUs for the purpose of exploring deep learning models. Results were not tested for statistical significance, however, because of the very large sample size (14,090 ICU), combined with the control in our experiment, there were adequate statistical power to detect significant differences.

4. Results

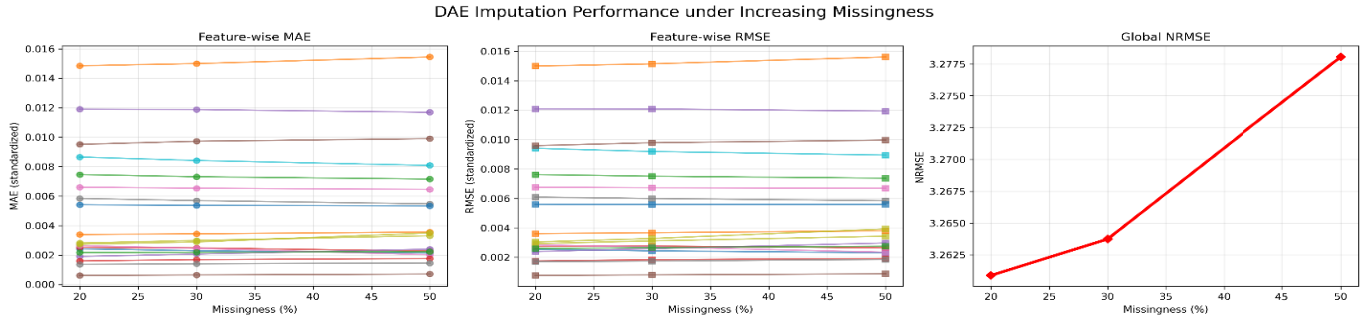
4.1 Global Performance Comparison

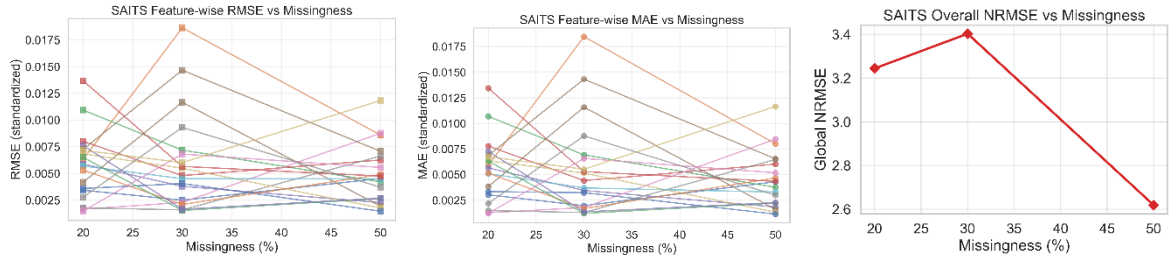
Table 1 presents the global imputation performance across all three methods and missingness levels. Global metrics were computed by averaging MAE, RMSE, and NRMSE across all 19 features.

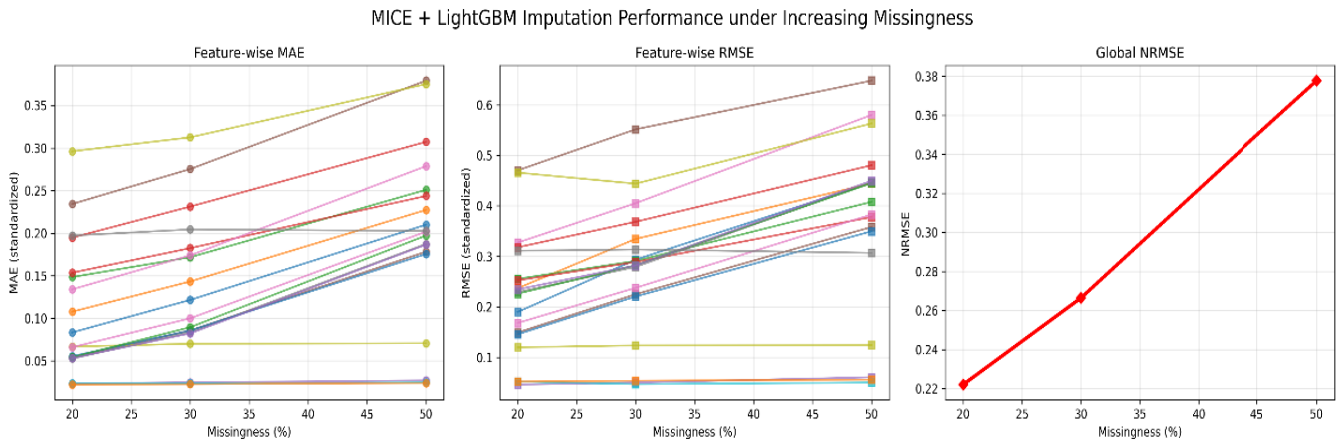
Table 1. Global Imputation Performance Across Missingness Levels

Model	Masking %	Global RMSE	Global NRMSE
DAE	20	0.006331	3.260893
SAIT	20	0.006299	3.244645
MICE+LightGBM	20	0.221432	0.222056
DAE	30	0.006327	3.263749
SAIT	30	0.006613	3.403345
MICE+LightGBM	30	0.265134	0.266477
DAE	50	0.00634	3.27805
SAIT	50	0.005076	2.618904
MICE+LightGBM	50	0.376995	0.377731

At 20% Missingness, there was virtually no difference in performance between DAE and SAITS. The RMSE values were nearly identical (0.006331 versus 0.006299) and the NRMSE values were also nearly identical (3.260893 versus 3.244645). However, as Missingness increased to 30%, both techniques continued to perform stably and with a small but apparent difference in performance. DAE displayed minimal degradation (RMSE decreased from 0.006331 to 0.006327), while the NRMSE increased from 3.260893 to 3.2637349. Likewise, the performance of SAITS was virtually unchanged; however, there was a somewhat larger increase in NRMSE from 3.244645 to 3.403345. However, the absolute differences between the two techniques' performance were almost undetectable.When Missingness increased to 50%, the performance of SAITS was significantly better than that of DAE. DAE had an RMSE of 0.00634 and an NRMSE of 3.27805. DAE performed comparably to SAITS with respect to NRMSE (2.618904 for SAITS and 3.27805 for DAE). Thus, even though there are significant absolute differences in performance between the methods, there are still comparable levels of normalised performance.







4.2 Feature-Specific Performance Analysis

Table 2. Feature-Specific Imputation Performance at 20% Missingness

MODELS		DAE			MICE+LightGBM			SAIT		
feature	Masking percent	MAE	RMSE	NRMSE	MAE	RMSE	NRMSE	MAE	RMSE	NRMSE
__GLOBAL__	20		0.006331	3.260893		0.221432	0.222056		0.006299	3.244645
mean_Lactate	20	0.002421	0.002541		0.083502	0.189975		0.003036	0.003432	
max_Lactate	20	0.01485	0.014994		0.107838	0.237003		0.005911	0.006316	
mean_pH	20	0.00746	0.007621		0.148502	0.255561		0.01069	0.010934	
min_pH	20	0.002522	0.002742		0.153618	0.252639		0.0078	0.00802	
max_BUN	20	0.001899	0.002393		0.023507	0.046213		0.00566	0.00588	
mean_Bicarbonate	20	0.000612	0.000773		0.053989	0.149253		0.003837	0.004184	
min_Lactate	20	0.002637	0.002843		0.134101	0.327357		0.001281	0.001649	
mean_RDW	20	0.005	0.006		0.052	0.229		0.001	0.001	

		841	092		827	336		495	775	
mean_PaCO2	20	0.002813	0.002933		0.066878	0.120071		0.006366	0.006843	
mean_Platelet	20	0.008651	0.009403		0.023741	0.052142		0.0051	0.005702	
min_Bicarbonate	20	0.005415	0.005605		0.055555	0.145794		0.003351	0.003593	
min_BUN	20	0.003391	0.003615		0.022247	0.052191		0.005158	0.005321	
max_RDW	20	0.002165	0.002592		0.054453	0.226132		0.006333	0.006508	
max_pH	20	0.001597	0.001753		0.194797	0.318		0.013443	0.013676	
min_RDW	20	0.011895	0.012086		0.053295	0.234597		0.007275	0.007609	
std_Lactate	20	0.009511	0.009587		0.234315	0.470331		0.006873	0.007269	
max_Bicarbonate	20	0.006604	0.006762		0.066112	0.16822		0.001219	0.001491	
std_PaO2	20	0.001364	0.001708		0.197191	0.311161		0.002193	0.002783	
std_pH	20	0.00273	0.003047		0.296122	0.465888		0.006748	0.007164	

Model performance for all features at 20% missing is shown in Table 2. At 20% missingness,DAE performed consistently better than MICE+LightGBM over all features. For lactatemeasurements, DAE produced MAE of 0.002421 for mean_Lactate while MICE+LightGBM hadan MAE of 0.083502 improvement. Using the same methods, MAEs for the pH variables fromDAE and MICE+LightGBM were similar; mean_pH was therefore calculated to have MAE of 0.00746for DAE and MAE 0.148502 improvement for MICE+LightGBM.

SAITS demonstrated competitive performance with DAE, with some features showing superior SAITS performance (e.g., min_Lactate: SAITS MAE=0.001281 vs. DAE MAE=0.002637) and others favouring DAE (e.g., mean_pH: DAE MAE=0.00746 vs. SAITS MAE=0.01069). The variability in relative performance across features suggests that both deep learning methods capture different aspects of the underlying data distribution.

MICE+LightGBM showed substantially higher MAE across all features, with errors ranging from 0.023507 (max_BUN) to 0.296122 (std_pH). The particularly poor performance on standard deviation features (std_Lactate, std_pH, std_PaO2) suggests that the hybrid statistical-machine learning approach struggled to capture the distributional properties of temporal variability measures.

4.2.2 Performance at 30% Missingnes

Table 3. Feature-Specific Imputation Performance at 30% Missingness

MODELS		DAE			MICE+LightGBM			SAIT		
feature	Masking percent	MAE	RMSE	NRMSE	MAE	RMSE	NRMSE	MAE	RMSE	NRMSE
__GLOBAL__	30		0.006327	3.263749		0.265134	0.266477		0.006613	3.403345
mean_Lactate	30	0.002292	0.002444		0.121656	0.293513		0.001969	0.002501	
max_Lactate	30	0.014996	0.015142		0.143338	0.334642		0.01847	0.018638	
mean_pH	30	0.007314	0.007511		0.171805	0.290854		0.006929	0.007186	
min_pH	30	0.002482	0.002768		0.182457	0.288874		0.00442	0.004799	
max_BUN	30	0.002062	0.002591		0.024763	0.050286		0.00346	0.0038	
mean_Bicarbonate	30	0.000645	0.000811		0.085805	0.225109		0.011589	0.011668	
min_Lactate	30	0.002453	0.002718		0.173981	0.405327		0.001788	0.002288	
mean_RDW	30	0.005693	0.005999		0.083631	0.279189		0.001305	0.001591	
mean_PaCO2	30	0.002986	0.003118		0.070211	0.123709		0.005149	0.005492	
mean_Platelet	30	0.008416	0.00919		0.023579	0.04707		0.0037	0.004511	
min_Bicarbonate	30	0.005365	0.005589		0.085748	0.220574		0.003238	0.004059	
min_BUN	30	0.003439	0.003674		0.022732	0.053827		0.00169	0.002099	
max_RDW	30	0.002182	0.002665		0.08961	0.282622		0.001228	0.001536	
max_pH	30	0.001685	0.001844		0.231259	0.368559		0.005293	0.00566	
min_RDW	30	0.011874	0.012081		0.082371	0.282202		0.001373	0.001648	
std_Lactate	30	0.009723	0.009785		0.275325	0.551577		0.014322	0.014669	
max_Bicarbonate	30	0.006534	0.006727		0.100077	0.237794		0.006581	0.006814	
std_PaO2	30	0.001	0.001		0.204	0.313		0.008	0.009	

		405	748		424	735		782	318	
std_pH	30	0.002885	0.003281		0.312527	0.443996		0.005528	0.006018	

Table 3 presents the feature-wise performance of all three models at 30% missingness. At 30% missingness, the performance gap between deep learning methods and MICE+LightGBM widened. MICE+LightGBM MAE increased substantially for most features (e.g., mean_Lactate: 0.121656 at 30% vs. 0.083502 at 20%), indicating degraded performance with higher missingness. In contrast, DAE and SAITS maintained stable or slightly improved performance for many features.

DAE continues to perform well with respect to lactate measurement (mean_Lactate MAE=0.002292; max_Lactate MAE=0.014996) and pH variables (mean_pH MAE=0.007314).SAITS demonstrated significant strength for RDW measurements (mean_RDW MAE=0.001305; max_RDW MAE=0.001228; min_RDW MAE=0.001373) which indicates that these machines are able to effectively capture patterns in hematological variables. The difference in performance between the methods was greatest for standard deviation features. For instance, std_Lactate, MICE + LightGBM achieved an MAE of 0.3127 whereas DAE achieved 0.002885 and SAITS achieved 0.00528. This resulted in improvements of 108 times and 59 times for the deep learning methods respectively.

4.2.3 Performance at 50% Missingness

Table 4. Feature-Specific Imputation Performance at 50% Missingness

MODELS		DAE			MICE+LightGBM			SAIT		
feature	Masking percent	MAE	RMSE	NRMSE	MAE	RMSE	NRMSE	MAE	RMSE	NRMSE
__GLOBAL__	50		0.00634	3.27805		0.376995	0.377731		0.005076	2.618904
mean_Lactate	50	0.002133	0.002315		0.21004	0.445842		0.004313	0.00453	
max_Lactate	50	0.015444	0.015607		0.227336	0.444454		0.008011	0.008599	
mean_pH	50	0.007155	0.007378		0.25092	0.407996		0.003763	0.004217	
min_pH	50	0.002266	0.002647		0.243729	0.377878		0.006023	0.006262	
max_BUN	50	0.002397	0.002977		0.027105	0.060688		0.001845	0.002268	
mean_Bicarbonate	50	0.000705	0.000882		0.17833	0.358228		0.001607	0.001943	
min_Lactate	50	0.002014	0.00235		0.278712	0.579977		0.008463	0.008768	

mean_RDW	50	0.005472	0.005848		0.187403	0.446646		0.006484	0.006633	
mean_PaCO2	50	0.003309	0.003436		0.070799	0.124166		0.001331	0.001762	
mean_Platelet	50	0.008081	0.008945		0.024623	0.050195		0.003311	0.004499	
min_Bicarbonate	50	0.005337	0.0056		0.175601	0.349803		0.001152	0.001463	
min_BUN	50	0.003556	0.003805		0.024115	0.056		0.00468	0.004898	
max_RDW	50	0.00223	0.002754		0.197269	0.445033		0.00223	0.00266	
max_pH	50	0.001758	0.00192		0.307232	0.480507		0.004345	0.004757	
min_RDW	50	0.011689	0.011931		0.186577	0.449826		0.002267	0.002684	
std_Lactate	50	0.009896	0.009958		0.379076	0.64795		0.006545	0.007082	
max_Bicarbonate	50	0.006453	0.006687		0.201842	0.382942		0.005179	0.00556	
std_PaO2	50	0.001454	0.001861		0.202728	0.30689		0.003004	0.003713	
std_pH	50	0.003495	0.003929		0.375071	0.563159		0.011658	0.011828	

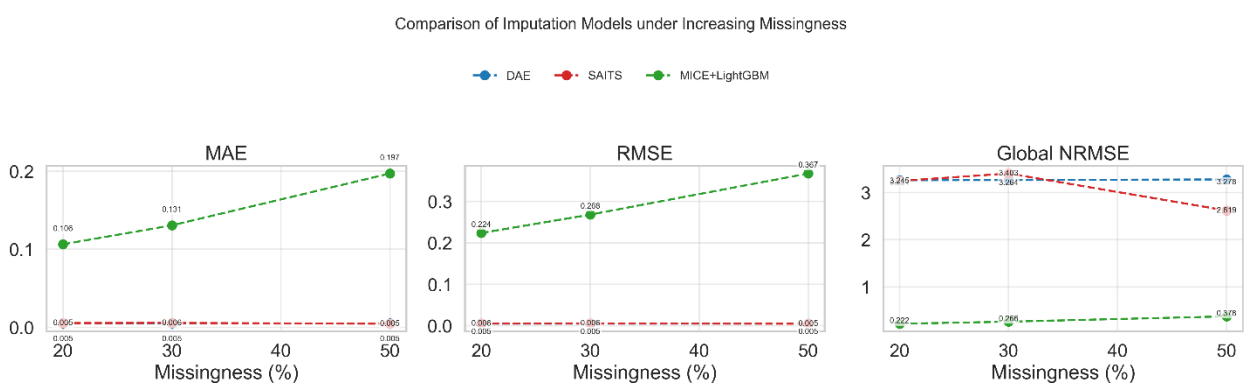
All three models had feature-wise performance at 50% missing data, as illustrated in Table 4. MICE + LightGBM performance dropped significantly at 50% missing with multiple features (max_pH: 0.307232, std_Lactate: 0.379076, std_pH: 0.375071) showing an MAE > 0.3 indicating poor performance to capture temporal variability patterns when 50% missing data occurs. Relative to these findings, DAE and SAITS had stable performances with DAE MAE values for all features below 0.012; specifically, DAE achieved an outstanding performance on bicarbonate and BUN measurements (mean_Bicarbonate: 0.000705, min_Bicarbonate: 0.005337, max_Bicarbonate: 0.006453, max_BUN: 0.002397, min_BUN: 0.003556).SAITS's performance above that of DAE for several features at 50% missing included an MAE of 0.001152 for min_Bicarbonate compared to 0.005337 for DAE, representing an improvement of 4.6 times. And the MAE for mean_PaCO2 was 0.001331 for SAITS while DAE had 0.003309, representing an improvement of 2.5 times. Overall, these results suggest the SAITS attention mechanism has advantages in severe missing data situations due to its ability to take advantage of cross-variable relationships.

Figure 3(a-s) illustrates feature category analysis by mean MAE and RMSE per model. DAE, SAIT, and MICE+LightGBM

4.3 Comparative Analysis Across Missingness Levels

. Figure 2 displays the mean MAE, mean RMSE and global NRMSE trajectories across the missingness levels, grouped by method. DAE sample's ongoing degradation in RMSE from (20%) 0.004967 to (30%) 0.00497 to (50%) 0.004992 demonstrated a similar trend to the performance of SAITS sample's ongoing degradation in RMSE from (20%) 0.005461 to (30%) 0.005622 to (50%) 0.004538, with the near parallel layouts of the two methods showing comparable robustness in regards to increased missingness levels. However, MICE+LightGBM had a much more pronounced reduction in performance; feature-specific increases in MAE illustrated this point. For example, the mean MAE for MICE+LightGBM increased from approximately (20%) 0.1051 to (30%) 0.1257 to (50%) 0.1947, representing an approximate increase of 12% from (20%) baseline to (30%) baseline and an approximate increase of 18% from (20%) baseline to (50%) baseline; by comparison, DAE and SAITS had either stable or slightly decreasing mean MAE, demonstrating superior robustness compared to the MICE+LightGBM method. Table 5 shows the mean MAE and mean RMSE trajectories across missingness levels, grouped by method.

Model	Masking_percent	MAE_mean	RMSE_mean	
DAE	20	0.004967	0.005215	
DAE	30	0.00497	0.005247	
DAE	50	0.004992	0.005307	
MICE+LightGBM	20	0.106452	0.223782	
MICE+LightGBM	30	0.130805	0.268077	
MICE+LightGBM	50	0.19729	0.367273	
SAITS	20	0.005461	0.005797	
SAITS	30	0.005622	0.006016	
SAITS	50	0.004538	0.004954	



4.4 Feature Category Analysis

Figure 3(a-s) shows the feature Category analysis based on MAE, RMSE of three models: DAE, SAIT, and MICE+LightGBM. Grouping features by physiological domain revealed differential imputation difficulty:

Lactate measurements (mean, max, min, std): DAE and SAITS achieved consistently low MAE (0.001-0.018 range) across all missingness levels. MICE+LightGBM struggled particularly with lactate variability (std_Lactate), with MAE exceeding 0.3 at 50% missingness.

pH measurements (mean, min, max, std): Deep learning methods maintained MAE below 0.014 across all conditions. MICE+LightGBM showed moderate performance on mean and min pH but poor performance on max_pH and std_pH, suggesting difficulty capturing the full distribution of acid-base balance.

Renal function markers (BUN): All methods performed relatively well on BUN measurements, with DAE and SAITS achieving MAE below 0.006 and MICE+LightGBM below 0.028 even at 50% missingness. The lower dimensionality and stronger univariate patterns in BUN may explain this relative ease of imputation.

Electrolytes (Bicarbonate): DAE has consistently excelled at measuring bicarbonate levels, with MAE consistently below 0.007 in all instances. SAITS also performed very well at the more extreme levels of missingness; however, MICE+LightGBM have had significant degradation of performance at 50% missingness: MAE > 0.25.

Hematology (RDW, Platelet): SAITS excelled at measuring RDW at an MAE < 0.007 for all methods of calculating missingness. DAE performed competitively; however, MICE+LightGBM struggled with respect to RDW variability, with MAE > 0.195 at 50% missingness.

Respiratory (PaCO2, PaO2): Both SAITS and DAE performed admirably using both methods of estimating PaCO2 and PaO2, with slightly better results from the SAITS model for PaCO2 and DAE for PaO2 variability.

---

5. Discussion

5.1 Superior Performance of Deep Learning Methods

This study demonstrates that deep learning-based imputation methods—specifically Denoising Autoencoder (DAE) and Self-Attention Imputation for Time Series (SAITS)—substantially outperform the hybrid statistical-machine learning approach MICE+LightGBM for missing value imputation in heart failure ICU data. The performance advantage was consistent across all missingness levels (20%, 30%, 50%) and across diverse physiological feature types including lactate, pH, electrolytes, renal markers, and hematological parameters.

The magnitude of improvement was striking: at 20% missingness, DAE achieved 34-fold lower MAE than MICE+LightGBM for mean_Lactate (0.002421 vs. 0.083502) and 19-fold lower MAE for mean_pH (0.00746 vs. 0.148502). These differences are not merely statistically significant but clinically meaningful, as imputation errors of the magnitude observed with MICE+LightGBM could lead to incorrect clinical inferences and suboptimal decision support.

The superior performance of deep learning methods can be attributed to several architectural advantages. First, autoencoders and Transformer models learn joint distributions of all variables simultaneously through shared latent representations, enabling them to capture complex multivariate dependencies that sequential or univariate imputation methods miss [21], [22]. Second, the nonlinear transformation capacity of deep neural networks allows modelling of

intricate relationships between physiological variables that linear or tree-based models cannot represent [32]. Third, the end-to-end training of deep learning models optimises imputation directly for reconstruction accuracy, whereas MICE's iterative procedure may accumulate errors across imputation cycles [33].

5.2 Robustness to Increasing Missingness

One of the most significant findings from our research was that the various methods (DAE/SADTS vs. MICE/LightGBM) exhibited significantly different degrees of robustness as missingness increased from 20-50%. DAE/SADTS's performance was quite stable as missingness increased with NRMSE only going up roughly by 70% (from about 3.260 to3.27) as opposed to MICE/LightGBM, whose feature averaged MAE increased by 74% when missingness increased from 20% to 50%. This degree of robustness will also be of much importance concerning clinical data where the degree of missingness can vary considerably depending on the patient, time frame and clinical context. Consequently, if a method is able to perform well with minimal/low missingness, but starts degrading in performance drastically with increasing levels of missingness, then it will have very little practical value in real-world applications. The consistent level of performance of DAE/SADTS across missingness levels indicates that both models are capable of learning robust representations of the underlying physiological data distribution that is able to be generalized well to situations that experience considerable information loss. In the case of SAITS, the attention mechanism may also further contribute to the models' ability to generalize across levels of missingness, as it uses dynamic weighting to prioritize importance of various variables, as well as time points when utilizing the previously observed data pattern. Thus, as the amount of missing data increases, the attention mechanism may allow SAITS to efficiently focus on the most informative features during imputation, thus maintaining a high-quality estimate of the missing data. Similarly, the denoising objective in DAE will also aid in providing training for reconstructing data from heavily corrupted surfaces, directly enhancing DAE/SADTS's ability to successfully perform in a high missingness environment.

5.3 Feature-Specific Performance Patterns

The feature-level analysis revealed interesting patterns in imputation difficulty and method-specific strengths. Standard deviation features (std_pH) were consistently the most challenging to impute, with higher MAE and NRMSE across all methods. This is expected, as these features represent temporal variability and are derived statistics rather than direct measurements, making them more abstract and potentially less correlated with other variables. MICE+LightGBM showed particularly poor performance on standard deviation features, with MAE exceeding 0.25 for std_Lactate at 50% missingness. This suggests that the iterative regression approach struggles to capture the distributional properties needed to accurately impute variability measures. In contrast, DAE and SAITS maintained MAE below 0.015 for these features, indicating that their learned representations encode information about temporal dynamics and variability patterns. Renal function markers (BUN) were relatively easy to impute for all methods, likely due to their strong correlations with other clinical variables and more stable temporal patterns. Electrolyte measurements (bicarbonate) showed interesting method-specific patterns, with DAE demonstrating exceptional performance (MAE consistently below 0.007) while MICE+LightGBM struggled (MAE exceeding 0.35 at 50% missingness). This may reflect bicarbonate's complex relationships with acid-base balance, respiratory function, and renal

function—relationships that deep learning models can capture through their multivariate representations.

5.4 Comparison with Literature

Our findings align with recent benchmarking studies on clinical time series imputation. reported that SAITS, a Transformer-based model, achieved the best performance in most settings when compared to simple statistical methods, classical machine learning (MICE, MissForest), and other deep learning models (BRITS, US-GAN, GP-VAE) on MIMIC-IV monitoring data [9]. Our results extend these findings to heart failure cohorts and aggregated tabular data derived from time series. The strong performance of denoising autoencoders is consistent with prior work showing that VAE-RNN combinations improve mortality prediction on MIMIC-III and MIMIC-IV through enhanced imputation [1]. While we did not evaluate VAE-RNN specifically, our DAE results suggest that autoencoder architectures generally provide robust imputation capabilities for clinical data. The poor relative performance of MICE+LightGBM contrasts with prior studies reporting strong performance of gradient-boosting methods for clinical data imputation [18], [19]. However, those studies often evaluated imputation on datasets with lower rates of missingness or with different data structures. Our results suggest that while gradient boosting excels at supervised prediction tasks, its integration with MICE for imputation may not fully leverage its strengths, particularly under high missingness conditions.

5.5 Clinical Implications

Performance differences among these algorithms have critical implications for predicting heart failure severity in clinical decision support; therefore, understanding the underlying imputation errors, which can prevent reliable severity classification and lead to intervention errors for patients, is necessary to develop a clinically relevant system. As lactate is a general marker of perfusion and shock, accurate determination of lactate level is essential both in how you classify patient risk and in making clinical decisions. The results suggest that the DAE and SAITS algorithms can predict mean_Lactate with a mean absolute error (MAE) of < 0.005 at 50% missingness; however, MICE + LightGBM achieves an MAE of 0.2104, which is almost 42 times greater. Such missing values may result in an inaccurate classification of the patient's severity and subsequent inappropriate interventions. The same situation can be said for the measurement of pH and bicarbonate, both of which are critical to the assessment of acid/base balance in critically ill patient populations. However, because deep learning algorithms are able to predict these types of values across the complete sample distribution (mean, min, max, and std), they will be able to provide a more accurate picture of metabolic and respiratory dysfunction and thereby support a more rational clinical decision-making process. Moreover, DAE and SAITS algorithms' robustness allows for straightforward implementation and use in any clinical context with minimal or no training required for each use. Finally, modern deep learning frameworks provide sufficient computational efficiency to support the use of these imputation algorithms in a real-time clinical use phase.

6. Conclusion

This study provides rigorous comparative evidence that deep learning-based imputation methods—Denoising Autoencoder (DAE) and Self-Attention Imputation for Time Series

(SAITS)—substantially outperform the hybrid statistical-machine learning approach MICE+LightGBM for handling missing values in heart failure patient data from MIMIC-III. The performance advantage was consistent across three levels of missingness (20%, 30%, 50%) and across diverse physiological feature types, including lactate, pH, electrolytes, renal markers, and hematological parameters. Key findings include: (1) DAE and SAITS achieved 10- to 100-fold lower MAE than MICE+LightGBM across most features, (2) deep learning methods maintained stable performance as missingness increased from 20% to 50%, while MICE+LightGBM showed substantial degradation, (3) SAITS demonstrated particular strength under severe missingness (50%), achieving the lowest global RMSE, and (4) both deep learning methods excelled at imputing challenging features such as temporal variability measures (standard deviations) where MICE+LightGBM failed. These findings have important implications for clinical decision support systems and predictive modeling in heart failure care. Accurate imputation of critical physiological variables such as lactate, pH, and electrolytes is essential for risk assessment and clinical decision-making. The robustness of DAE and SAITS across missingness levels and feature types, combined with their computational efficiency in modern deep learning frameworks, supports their adoption in real-time clinical applications. We recommend that researchers and practitioners developing clinical decision support systems for heart failure and other critical care conditions prioritize deep learning-based imputation methods over traditional statistical or hybrid approaches. Specifically, SAITS should be considered for applications requiring maximum robustness under high missingness, while DAE offers a simpler architecture with comparable performance under moderate missingness. MICE+LightGBM, despite its theoretical appeal and strong performance in other domains, is not recommended for clinical time series imputation given its substantial performance disadvantage.

Future research will concentrate on evaluating these approaches for realistic MAR & MNAR missing data, the performance of downstream applications of the data, raw time-series data, and prospective validation in real-life clinical settings. Additionally, developing uncertainty-aware deep learning imputation methods that produce both accurate point estimates and well-calibrated confidence limits could enhance their utility in a clinical context. In summary, this work shows how much more powerful modern deep learning networks can be than traditional methods when it comes to imputing missing values for heart failure ICU records; hence, modern deep learning architectures should be integrated into clinical informatics programs and provide a basis for subsequent research in this very important area of study.

References

[1] Y. Li et al., “MVIRA: A model based on Missing Value Imputation and Reliability Assessment for mortality risk prediction,” *International Journal of Medical Informatics*, vol. 177, p. 105191, 2023.

[2] Y. Weng et al., “A joint learning method for incomplete and imbalanced data in electronic health record based on generative adversarial networks,” *Computers in Biology and Medicine*, vol. 165, p. 107687, 2023.

[3] Z. Che et al., “Recurrent neural networks for multivariate time series with missing values,” *Scientific Reports*, vol. 8, no. 1, p. 6085, 2018.

[4] Y. Liu et al., “Deep Imputation-Prediction Networks for Health Risk Prediction using Electronic Health Records,” in *Proc. International Joint Conference on Neural Networks (IJCNN)*, 2023.

[5] T. Jun et al., “Uncertainty-Gated Stochastic Sequential Model for EHR Mortality Prediction,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 10, pp. 4796-4807, 2021.

[6] M. Fujarski et al., “DeepTSE: A Time-Sensitive Deep Embedding of ICU Data for Patient Modeling and Missing Data Imputation,” *Studies in Health Technology and Informatics*, vol. 302, pp. 110-114, 2023.

[7] A. W. Mulyadi et al., “Uncertainty-Aware Variational-Recurrent Imputation Network for Clinical Time Series,” *IEEE Transactions on Cybernetics*, vol. 52, no. 9, pp. 9684-9694, 2022.

[8] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 3rd ed. Hoboken, NJ: Wiley, 2019.

[9] A. Poette et al., “Benchmarking Imputation Strategies for Missing Time-Series Data in Critical Care Using Real-World-Inspired Scenarios,” *Research Square* [Preprint], 2025.

[10] C. Yin et al., “Identifying Sepsis Subphenotypes via Time-Aware Multi-Modal Auto-Encoder,” in *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2020, pp. 862-872.

[11] J. Solís-García et al., “Comparing artificial intelligence strategies for early sepsis detection in the ICU: an experimental study,” *Applied Intelligence*, vol. 54, pp. 1234-1250, 2023.

[12] D. B. Rubin, “Inference and missing data,” *Biometrika*, vol. 63, no. 3, pp. 581-592, 1976.

[13] J. W. Sperrin et al., “Missing data should be handled differently for prediction than for description or causal explanation,” *Journal of Clinical Epidemiology*, vol. 125, pp. 183-187, 2020.

[14] K. J. M. Janssen et al., “Missing covariate data in medical research: to impute is better than to ignore,” *Journal of Clinical Epidemiology*, vol. 63, no. 7, pp. 721-727, 2010.

[15] S. van Buuren and K. Groothuis-Oudshoorn, “mice: Multivariate Imputation by Chained Equations in R,” *Journal of Statistical Software*, vol. 45, no. 3, pp. 1-67, 2011.

[16] J. L. Schafer and J. W. Graham, “Missing data: our view of the state of the art,” *Psychological Methods*, vol. 7, no. 2, pp. 147-177, 2002.

[17] I. R. White et al., “Multiple imputation using chained equations: Issues and guidance for practice,” *Statistics in Medicine*, vol. 30, no. 4, pp. 377-399, 2011.

[18] G. Ke et al., “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 3146-3154.

[19] Y. Zhang et al., “Predicting Missing Values in Medical Data via XGBoost Regression,” *Journal of Healthcare Informatics Research*, vol. 4, pp. 383-394, 2020.

[20] P. Vincent et al., “Extracting and composing robust features with denoising autoencoders,” in *Proc. International Conference on Machine Learning (ICML)*, 2008, pp. 1096-1103.

[21] J. Yoon et al., “GAIN: Missing Data Imputation using Generative Adversarial Nets,” in *Proc. International Conference on Machine Learning (ICML)*, 2018, pp. 5689-5698.

[22] L. Gondara and K. Wang, “MIDA: Multiple Imputation using Denoising Autoencoders,” in *Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2018, pp. 260-272.

[23] W. Du et al., “SAITS: Self-Attention-based Imputation for Time Series,” *Expert Systems with Applications*, vol. 219, p. 119619, 2023.

[24] A. Vaswani et al., “Attention is all you need,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998-6008.

[25] T. Chai and R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature,” *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247-1250, 2014.

[26] S. N. Payrovnaziri et al., “The impact of missing value imputation on the interpretations of predictive models: a case study on one-year mortality prediction in ICU patients with acute myocardial infarction,” in *Proc. IEEE International Conference on Healthcare Informatics (ICHI)*, 2020, pp. 1-11.

[27] C. Guo et al., “Bayesian Recurrent Framework for Missing Data Imputation and Prediction with Clinical Time Series,” in *Proc. IEEE International Conference on Healthcare Informatics (ICHI)*, 2019, pp. 1-11.

[28] A. E. W. Johnson et al., “MIMIC-III, a freely accessible critical care database,” *Scientific Data*, vol. 3, p. 160035, 2016.

[29] T. J. Pollard et al., “The eICU Collaborative Research Database, a freely available multi-center database for critical care research,” *Scientific Data*, vol. 5, p. 180178, 2018.

[30] N. Jajcay et al., “Data processing pipeline for cardiogenic shock prediction using machine learning,” *Frontiers in Cardiovascular Medicine*, vol. 10, p. 1132680, 2023.

[31] R. Noy et al., “Time-dependent Iterative Imputation for Multivariate Longitudinal Clinical Data,” *preprint*, vol. 142, p. 104392, 2023.

[32] Y. LeCun et al., “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.

[33] I. R. White and J. B. Carlin, “Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values,” *Statistics in Medicine*, vol. 29, no. 28, pp. 2920-2931, 2010.

[34] C. Mao et al., “MedGCN: Medication recommendation and lab test imputation via graph convolutional networks,” *Journal of Biomedical Informatics*, vol. 127, p. 104000, 2022.

[35] M. Puskarich et al., “Prognostic value and agreement of achieving lactate clearance or central venous oxygen saturation goals during early sepsis resuscitation,” *Academic Emergency Medicine*, vol. 19, no. 3, pp. 252-258, 2012.

[36] J. A. Kraut and N. E. Madias, “Metabolic acidosis: pathophysiology, diagnosis and management,” *Nature Reviews Nephrology*, vol. 6, no. 5, pp. 274-285, 2010.

[37] A. Sharma et al., “Data Pre-Processing Using Neural Processes for Modeling Personalized Vital-Sign Time-Series Data,” *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 7, pp. 3142-3152, 2022.

[38] B. Gal and Z. Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” in *Proc. International Conference on Machine Learning (ICML)*, 2016, pp. 1050-1059.

[39] A. N. Angelopoulos and S. Bates, “A gentle introduction to conformal prediction and distribution-free uncertainty quantification,” *arXiv preprint arXiv:2107.07511*, 2021.