Methodological Review

Deep imputation of missing values in time series health data: A review with benchmarking

Maksims Kazijevs, Manar D. Samad *

Department of Computer Science, Tennessee State University, Nashville, TN 37209, United States

ARTICLE INFO

ABSTRACT

The imputation of missing values in multivariate time series (MTS) data is critical in ensuring data quality and producing reliable data-driven predictive models. Apart from many statistical approaches, a few recent studies have proposed state-of-the-art deep learning methods to impute missing values in MTS data. However, the evaluation of these deep methods is limited to one or two data sets, low missing rates, and completely random missing value types. This survey performs six data-centric experiments to benchmark state-of-the-art deep imputation methods on five time series health data sets. Our extensive analysis reveals that no single imputation method outperforms the others on all five data sets. The imputation performance depends on data types, individual variable statistics, missing value rates, and types. Deep learning methods that jointly perform cross-sectional (across variables) and longitudinal (across time) imputations of missing values in time series data yield statistically better data quality than traditional imputation methods. Although computationally expensive, deep learning methods are practical given the current availability of high-performance computing resources, especially when data quality and sample size are of paramount importance in healthcare informatics. Our findings highlight the importance of data-centric selection of imputation methods to optimize data-driven predictive models.

1. Introduction

The presence of missing values in real-world data sets is a major obstacle in effectively building machine learning models with reliable predictive outcomes. In the scientific literature, various arbitrary methods are applied to impute the missing values in data only to enable a machine learning-based model development. However, the selection of the imputation method regulates the input data quality and ultimately affects the accuracy and robustness of data-driven predictive models. Statisticians and biostatisticians [1] have studied the imputation of missing values in tabular data for several decades, followed by recent contributions from advanced machine and deep learning methods [2–5].

The field of missing value imputation (MVI) is primarily studied using multivariate tabular data without considering any time-varying variables. Many real-world data sets observed in health sciences and electronic sensor applications commonly include time-varying data collected for multiple variables. In electronic health records (EHR), patient follow-up data are collected over time, where missing values inevitably appear across variables (missing cross-sectional values) and time (missing longitudinal values). In addition to the variable dimensionality, the time dimension adds an extra layer of computational challenge in estimating missing values. This additional dimension requires an imputation method to capture the time dependency and variable dependency in missing data.

Several survey papers have recently been published on time series data imputation methods. Bauer et al. have benchmarked stochastic multiple imputations and deterministic spectral clustering methods for time series data imputation [6]. Shukla et al. have reviewed imputation methods for time series data based on temporal discretization, interpolation, recurrence, attention, and structural invariance [7]. Festag et al. provide a review selectively on generative adversarial networks (GAN) in forecasting and imputing biomedical time series data [8]. However, these survey articles do not perform any benchmarking of existing methods. Sun et al. have reviewed methods for imputing missing values in irregularly sampled health records data [9]. They perform mortality prediction to demonstrate the quality of imputed data sets. They compare the imputation accuracy primarily between several basic learning architectures and their variants, including GRU, LSTM, RNN, M-RNN, GRU-D, T-LSTM, and GAN-GRUI. All these imputation methods are proposed in or before 2018. Their work does not include additional experiments on missing data imputations that are required to determine the robustness of imputation algorithms. Khayati et al. evaluate matrix-based and pattern-based methods to impute time series data acquired

---

* Corresponding author.
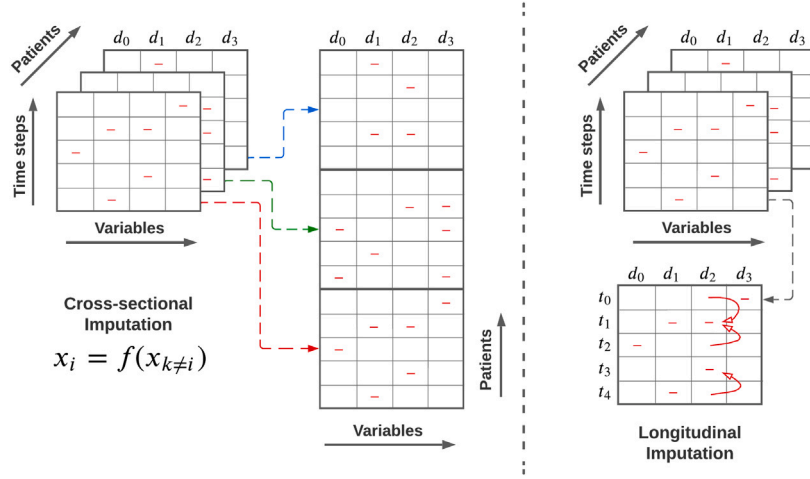 E-mail address: msamad@tnstate.edu (M.D. Samad).

Fig. 1. Cross-sectional imputation builds regression models based on column variables. Longitudinal imputation leverages time-varying values for individual variables.

from sensors [10]. They do not use deep learning methods or perform experiments under different missing value types and rates. Fang et al. have surveyed time series imputation methods based on three deep learning architectures (RNN, GAN, GRU), in addition to other statistical and machine learning-based imputation methods [11]. However, this survey does not perform any benchmarking or present experimental results to compare the imputation performance of those deep learning methods. Our paper is one of the first to comprehensively review and benchmark state-of-the-art deep missing value imputation methods on multiple time series health data.

2. Background review

The imputation of missing values is conventionally performed by modeling variables or cross-sectional data dependency. However, the cross-sectional imputation is not optimal for time series data because longitudinal data dependency across time is not considered in missing value estimation. For example, electronic health records (EHR) collect patients' follow-up treatments and clinical tests at noncontinuous and uneven time intervals. The heterogeneity in patient samples turns missing values into a stronger function of time than cross-sectional variables. This requirement necessitates methods to optimally estimate missing values in time series data using cross-sectional and longitudinal information. The sections below review the literature on cross-sectional and time series data imputation.

2.1. Cross-sectional imputation

A static data set without any time-varying component is the most common data used in the missing value imputation literature. The dependency among variables in data columns is modeled using statistical methods, which is known as the cross-sectional imputation of missing values. The most common and effective approach for cross-section imputation is based on the statistical method of multiple imputations (MI). The multiple imputations using chained equations (MICE) algorithm [12,13] is developed based on the principle of MI [1,14]. The MICE algorithm generates multiple imputed versions on the same data set, which improves accuracy by reducing uncertainty in data imputation. MICE builds linear regression models to estimate a dependent variable with missing values as a function of other variables, as shown in Fig. 1. For example, the j_{th} (x_j) variable can be estimated from the other (d-1) observed and imputed variables x_k as shown in Eq. (1).

$$x_j = f(x_k, W), k \neq j \tag{1}$$

Here, W is a vector of trainable regression parameters for the j_{th} variable, where

$$W = [w_1, w_2, w_3, \ldots, w_{d-1}]^T \tag{2}$$

A recent study has updated the MICE method by replacing linear regressors with several non-linear regressors, including gradient boosting trees (GBT) and multilayer perceptions [15]. Results suggest that combining the MI part with ensemble learning (e.g., GBT) techniques provide superior imputation accuracy performance. The MI can capture the variability of different models, while ensemble learning captures variability within a model.

Apart from statistical learning methods, deep learning has been proposed for the cross-sectional imputation of missing values. One such method trains deep autoencoders on the complete data to encode the values into high-dimensional latent space while the decoder reconstructs the original data from the latent vector to impute missing values [16]. In recent studies, deep generative methods [3,17], convolutional neural networks (CNN) [18], and recurrent neural networks (RNN) [4] are proposed for missing data imputation.

2.1.1. Missing value types

The missing value imputation literature on static data sets commonly studies three types of missing values: (1) missing completely at random (MCAR), (2) missing at random (MAR), and (3) missing not at random (MNAR). Three missing value types are statistically defined in the section below.

We can denote a data matrix, $X = \{X_{obs}, X_{miss}\}$ that can be composed of $X_{obs} \subset X$ with the observed values and $X_{mis} \subset X$ with the missing values. Matrix ($M = [m_{ij}]$) represents whether the value at row i and column j is missing or not. Entries when $m_{ij} = 0$ denote missing values and $m_{ij} = 1$ denote observed values. A conditional probability distribution of M = 0 (missingness) given X can be defined as $Pr(M = 0|X_{obs}, X_{mis})$ to model three types of missingness. First, MCAR occurs when the probability of missingness is completely independent of the observed (X_{obs}) or missing (X_{miss}) data as:

$$Pr(M = 0) = Pr(M = 0|X_{obs}, X_{mis}). \tag{3}$$

Second, MAR is defined as the missing value type when the probability of missingness depends on the observed data as follows.

$$Pr(M = 0|X_{obs}) = Pr(M = 0|X_{obs}, X_{mis}). \tag{4}$$

Third, the missing probability depends on the unobserved or missing values themselves in MNAR.

$$Pr(M = 0|X_{mis}) = Pr(M = 0|X_{obs}, X_{mis}). \tag{5}$$

However, the time series imputation methods are rarely evaluated against these missing value types. Even for cross-sectional imputation, only a few studies comprehensively analyze all three missing value types. The MCAR type is chosen as the default type in most studies [19–21]. In reality, missing values in medical data never appear completely at random because a lab test is missing due to a physician's recommendation. Physicians do not recommend a lab test when they expect normal results or no significant difference from the previous test within an interval. It has been shown that imputing missing values that are not random is far more challenging and error-prone than imputing MCAR type missing values [15]. Therefore, benchmarking missing value imputation methods for time series data requires a realistic data setting by comparing different missing value types.

2.2. Time series data imputation

The imputation of time series data models each variable as a function of time. ECG data can be acquired from one of the multiple electrodes as time-varying multivariate data. Similarly, a clinical variable (e.g., blood pressure) can be observed in the same patient over multiple visits at irregular intervals. The observed values in time can be used to build a predictive model for estimating missing values at other time points, as shown in Fig. 1.

2.2.1. Statistical and machine learning based methods

Statistical methods for imputing missing values in time series data are inspired by the multiple imputation methods developed for cross-sectional imputation. For example, the 3D-MICE method imputes missing values based on the MICE framework by leveraging cross-sectional and longitudinal dependencies in time series data [13]. Cross-sectional imputation is performed using standard MICE after flattening time series data. A single-task Gaussian process (GP) is then applied for longitudinal imputation. Both estimates are then combined using a variance-informed weighted average. The 3D-MICE method is primarily tested on one healthcare data set, and it outperforms baseline statistical methods such as mean imputation, MICE, and GP. Recently, Sun et al. have proposed multiple imputations by chained equations with data augmentation (MICE-DA) method to overcome the limitation of separate longitudinal and cross-sectional models of 3D-MICE [22]. The MICE-DA method aims to augment flattened cross-sectional data with features extracted from the longitudinal data and then apply the standard MICE method. Xu et al. has proposed the multi-directional multivariate time series (MD-MTS) method for missing value estimation [23]. They perform rigorous feature engineering to integrate temporal and cross-sectional features into a common imputation task. Zhang et al. propose a similar strategy to augment longitudinal features in their xgbooSt MIssing vaLues In timE Series (SMILES) method [24]. Technical details of the three previously mentioned methods (MICE-DA, MD-MTS, and SMILES) are addressed in Section 2.2.2. Recently, the time-aware dual-cross-visit (TA-DualCV) method has been proposed to leverage both longitudinal and cross-sectional dependencies within- and cross-patient visits [25]. The core part of the method is the dual-cross-visit imputation (DualCV), which captures multivariate and temporal dependencies in cross-visit using chained equations. DualCV consists of two chained equation-based modules: cross-visits feature perspective module (CFP) and cross-visits the temporal perspective module (CTP). Both modules utilize Gibbs sampler to impute missing values and are combined subsequently. The time-aware augmentation mechanism then captures patient-specific correlations within each time point by applying the Gaussian process (GP) on each patient visit. Lastly, the results from both components, DualCV and GP, are fused using weighted averaging. TA-DualCV is evaluated on three healthcare data sets outperforming time-aware multi-modal autoencoder (TAME) [26], 3D-MICE, and MICE methods in all experimental settings.

2.2.2. The DACMI challenge for missing value imputation

The data analytics challenge on missing data imputation (DACMI) [27] has shared de-identified EHR data with training and test partitions for benchmarking time series imputation methods. This challenge has introduced several state-of-the-art statistical and machine learning-based time series data imputation methods, including MD-MTS [23], MICE-DA [22], SMILES [24], and context-aware time series imputation (CATSI) [28].

In this MD-MTS method, Xu et al. augment longitudinal and cross-sectional features into a new feature set. The feature set includes the following variables: (1) variable values at the current time point, (2) chart time, (3) time stamps, (4) pre and post-values in 3-time stamps, and (5) min, max, and mean values. Then, a tree-based LightGBM regressor for each variable is trained using augmented features to impute missing values. The motivation behind selecting this regressor is that it is less prone to overfitting and more sensitive to outliers. Sun et al. have proposed the MICE-DA method as an improved variant of the 3D-MICE method. This method obtains local temporal and global patient similarity features for the imputation task. Local temporal features are obtained by taking all pre-and post-variable values within three-time steps. Then, the slope of the change at each time step is estimated using a one-dimensional Gaussian process (1D-GP). Global patient similarity features are obtained by calculating statistical observations from the data set, such as minimum, maximum, percentiles, mean, and standard derivations of individual patient profiles. They combine the local temporal and global patient similarity features in the standard MICE method to impute missing values. The SMILES method involves three steps: prefilling missing values, feature extraction based on window size, and imputing missing values with the XGBoost model. Local mean and soft impute strategies are used to prefill missing values. However, selecting which variables will be prefilled with the local mean versus soft impute strategy requires manual work. The XGBoost models are then trained using features extracted based on the window size to impute missing values. Each XGBoost model imputes values for a specific variable, similar to the MICE method. All these newly proposed methods outperform the baseline 3D-MICE method proposed by the organizer of DACMI challenge [13]. It is important to note that all the methods submitted to the DACMI challenge are benchmarked on only one EHR data set, namely the DACMI data set. Therefore, the generalizability of these methods across other time series data sets and data scenarios is unknown.

2.2.3. Deep time series imputation methods

Most of the state-of-the-art time series imputation methods are built upon recurrent neural networks (RNN) and more advanced RNNs such as long short-term memory (LSTM) or gated recurrent unit (GRU) methods. The bidirectional recurrent imputation for time series (BRITS) method is one of the first methods to capture both longitudinal (time) and cross-sectional (variable) dependencies for predicting missing time points [29]. BRITS performs imputation using trained LSTM models unfolded in forward and backward time directions. Thus, bidirectional learning leverages past and future trends in time series to estimate missing values. Because observations can happen at uneven time intervals, modeling the time gaps in the missing value estimation process is an important contribution of the BRITS method. On the other side, the BRITS method is computationally expensive. BRITS method is evaluated on three data sets related to air quality, health care, and human activities. It is shown that BRITS is superior to RNN-based methods such as GRU-D [30] and M-RNN [31]. Additionally, BRITS outperforms non-RNN methods, including MICE, spatio-temporal multi-view-based learning (ST-MVL) [32], imputeTS [33], matrix factorization (MF), and k-nearest neighbor (KNN) methods.

Similar to BRITS, the non-autoregressive multiresolution imputation (NAOMI) method uses bidirectional RNN for missing value imputation [34]. Furthermore, the authors of NAOMI have introduced the recursive divide and conquer principle. In this principle, two known

time steps are identified as pivots, which are used to impute the missing value that falls between them. In the next step, the newly imputed time step is used as the pivot to impute other missing values in between. This process repeats until imputing all missing values. Additionally, the NAOMI method is enhanced with adversarial training. However, NAOMI does not consider time gap information for imputing missing values. The NAOMI method is also evaluated on three data sets: a traffic data set and the other two are trajectory movement data sets. In an additional experiment, the NAOMI method has been tested for robustness on missing rates in data. In this experiment, the authors randomly remove values at a rate ranging from 10% to 90% to simulate missing values. NAOMI is compared with several baseline methods, including GRUI [35] and MaskGAN [36]. The NAOMI model outperforms these baseline methods in most of the experimental scenarios. Similar GAN-based methods have shown promising results in missing value imputation of time series data [37,38].

Yin et al. proposed the CATSI method to capture the global trends of patients. Similar to BRITS architecture, the CATSI method uses bidirectional and cross-sectional dynamics for missing value imputation. To represent the patient health state, they have introduced a "context vector" to learn global temporal dynamics, which improves the missing value imputation accuracy. Also, the CATSI method is aware of time and uses the time gap information similar to the BRITS method. To capture the time gap information between the current time point s_t and the previous time point s_{t-1}, the delta matrix δ is obtained, where d denotes the variable index, t denotes the time index, and m is the mask matrix as shown in Eq. (6). Eq. (6) shows that if the value at the previous time point is observed ($m_{t-1}^d = 1$), then the time gap at time t (δ_t^d) is the difference between the previous and current time points. When the value at the previous time point is not observed ($m_{t-1}^d = 0$), then the previous time gap value (δ_{t-1}^d) is added to the difference between the previous and current time points to obtain the delta value.

$$\delta_t^d = \begin{cases} s_t - s_{t-1} + \delta_{t-1}^d & \text{if } t > 0, m_{t-1}^d = 0 \\ s_t - s_{t-1} & \text{if } t > 0, m_{t-1}^d = 1 \\ 0 & \text{if } t = 0 \end{cases} \quad (6)$$

If a variable is not observed for an extended period, the δ value will be proportionally high. This scenario will produce a low temporal decay coefficient γ_t, as shown in Eq. (7). The data set dependent parameter γ_t is obtained via a data-driven estimation method using a trainable parameter W_γ at the time of training the imputation model.

$$\gamma_t = \exp\{-\max(0, W_\gamma \delta_t + b_\gamma)\} \quad (7)$$

The decay coefficient γ_t is used to fuse the contribution of the last observation of the variable $x_{t'}^d$ and the variable mean $\bar{x}^d$ for initializing the missing value $x_t^{,d}$, as shown in Eq. (8). The missing value will be initialized by the variable mean instead of the last observed value in the remote past.

$$x_t^{,d} = \gamma_t^d x_{t'}^d + (1 - \gamma_t^d)\bar{x}^d \quad (8)$$

Another trade-off parameter β_t is derived from γ_t and the mask matrix and is optimized via a set of trainable parameters (W_β, b_β), as shown in Eq. (9).

$$\beta_t = \sigma(W_\beta[\gamma_t; m_t] + b_\beta) \quad (9)$$

The trade-off parameter β_t regulates the contribution of cross-sectional and longitudinal components during the time series imputation model training. For example, the CATSI method uses bidirectional LSTM to obtain longitudinal estimations $\hat{x}$ and MLP for cross-sectional estimations $\hat{z}$ of the missing values. A β_t weighted fusion of these estimations yields the final estimate of the missing values y_t, as shown in Eq. (10).

$$y_t = \beta_t \odot \hat{z}_t + (1 - \beta_t) \odot \hat{x}_t \quad (10)$$

Following the DACMI challenge, several notable imputation methods have emerged, including deep imputer of missing temporal data (DETROIT) [39], Gaussian process variational autoencoder (GP-VAE) [40], time-aware multi-modal autoencoder (TAME) [26]. The DETROIT method uses a fully connected neural network with eight hidden layers. To initialize missing values, authors of the DETROIT method manually select the variables for initializing with the local mean or soft impute [41] method. To capture temporal and cross-sectional correlations of variables, the following three data values are incorporated: (1) the values of other variables at the same time step, (2) the values of all variables in the previous and following steps, and (3) the time stamp differences between the four neighboring steps. The DETROIT method is tested only with the DACMI data set, outperforming the baseline 3D-MICE method only. In the GP-VAE method, deep variational autoencoders are used to map time series data with missing values into a latent space. The Gaussian process is then used to generate the time series from the low-dimensional latent space. GP-VAE method is tested on two imaging data sets and one tabular medical data set.

Recently, Yin et al. have proposed the TAME method that combines the strengths of a time-aware attention mechanism, bidirectional LSTM architecture, and multi-modal embeddings [26]. First, patient information, such as demographics, diagnosis, medications, and tests, is concatenated and projected into a multi-modal embedding using fully connected layers of neural networks. A bidirectional LSTM architecture is then used to obtain a latent representation of multi-modal embeddings. The time-aware attention mechanism is implemented to capture longitudinal information, which maps value embedding matrices and time gap embedding matrices (the gap between the current time step and the last available time step) into a new space. Finally, missing values are imputed using a fully connected layer, where the inputs are latent vectors from bidirectional LSTM and time-aware attention outcomes. The TAME method takes advantage of time stamp availability and patient information. The method is designed for data sets where patient information, such as demographics, diagnosis, and medications, is available. However, it is also possible to evaluate the TAME method on a data set where such patient information is unavailable, for example, the DACMI data set. However, it needs to be better investigated how this method would perform when extensive patient information is unavailable on multiple benchmark data sets. The TAME method is evaluated on two selective EHR data sets, and it outperforms several recent time series data imputation methods, including 3D-MICE, CATSI, DETROIT, and BRITS.

2.3. Health records and biomedical sensor data

Medical centers follow up on patients and store their diagnostic, electronic sensor, and imaging measurements in electronic health records (EHR) for patient care management. A plethora of measurements from hundreds and thousands of patients can be an invaluable resource for retrospective research studies instead of time-consuming clinical trials. However, unlike clinical research trials, where a predefined set of parameters are systematically measured and studied over time, EHR data must be curated carefully before using such data in predictive modeling. This is because of the highly complex, incomplete, and unpredictable structure of EHR data that is far from usable in data-driven model development [42,43]. If prepared and processed appropriately, the massive EHR data acquired from a large group of the predominantly sick population may help optimize clinical decision-making and discover novel insights into patient prognosis. Notably, misdiagnosis is the third leading cause of death in the United States [44]. Therefore, large EHR data-driven tools and clinical discovery can play an important role in improving the efficacy and precision of medical services. Similarly, biomedical sensor data from electrocardiography (ECG) and electroencephalography (EEG) are widely used in medical practice.

Table 1
State-of-the-art methods for imputing missing values in time series data.

|	Method	Max. % of missing data	Number of data sets	Data domain	Irregular time intervals	Downstream classification performance
Luo et al. [13]	3D-MICE	24%	One	Healthcare	Yes	No
Cao et al. [29]	BRITS	13.3%	Three	Multiple	Yes	Yes
Liu et al. [34]	NAOMI	–	Three	Multiple	No	No
K. Yin et al. [28]	CATSI	28%	One	Healthcare	Yes	No
Yan et al. [39]	DETROIT	7%	One	Healthcare	Yes	No
Fortuin et al. [40]	GP-VAE	60%	Three	Multiple	No	Yes
C. Yin et al. [26]	TAME	90%	Two	Healthcare	Yes	No
Gao et al. [25]	TA-DualCV	90%	Three	Healthcare	Yes	Yes
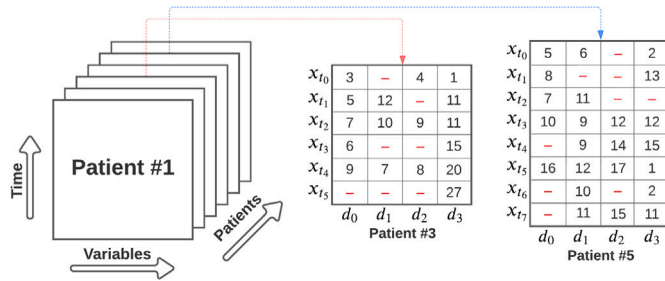
Fig. 2. Structure of time series health data with three dimensions: patient space, time-space, variable space. Each patient can have a different number of time points or follow-ups, and not all variables are measurements for all time points.

The application of EHR and biomedical sensor data is receiving growing attention in predictive modeling due to the recent advancement of deep learning in the last decade. For example, Ulloa-Cerna et al. use ~2.2 million ECG signals acquired from ~500,000 patients to predict multiple structural heart conditions using deep learning [45]. Similarly, novel machine learning methods have enabled health scientists to perform predictive analytics on EHR and medical data to answer unsolved research questions that were previously inconclusive with traditional statistical modeling [46–48]. The health science literature benefits from numerous applications of machine learning methods, including the estimation of the length of a hospital stay [49], real-time mortality prediction [50], sepsis prediction [51,52], blood and arterial pressure estimation [53–55], breathing rate estimation [56], and more. These research studies can help develop data-driven methods to discover significant clinical risk predictors and make clinical decisions faster and more accurately than those statistical methods currently used in clinical practice. However, there are three major challenges with utilizing sophisticated machine learning methods to reap the benefits of health record data, as discussed below.

2.3.1. Irregular and non-random missing values

The process of collecting EHR and patient follow-up data is unlike clinical research trials. This difference is because EHR data are meant for patient care management, not research studies. For example, patient visits and clinical variable measurements can happen irregularly depending on the physician's recommendations and the patient's health status. Therefore, the standard clinical practice collects individual patient data with varying numbers of visits (time points) and measurements taken at varying time intervals based on patient conditions, medical protocols, and administrative reasons [13,22]. See patient #3 and patient #5 in Fig. 2. On the other hand, machine learning methods, by default, expect data samples without irregular time intervals and varying numbers of time points [57,58]. In this context, Weerakody et al. present a review of gated recurrent neural networks in several prediction tasks with irregular time series data [59]. Several other reasons can be attributed to missing values in medical data, including when patients fail to visit for a specific test, a staff member fails to record the test result in the database, or the failure of medical sensors or equipment. Likewise, missing values in biomedical sensor data may occur due to broken sensors or loss of data transfer [11]. In a study on real-time sensor data acquisition for healthcare monitoring, Michalopoulos et al. mention that the data transmission process cannot be reliable and robust [60]. The reliability and robustness may not be achieved for several reasons, including time synchronization issues, interference, and transmission failures in a wireless sensor network. Therefore, the reason or patterns of missing values are never entirely random. In contrast, the proposed methods for imputing missing values in time series data assume missing values that are missing completely at random. To the best of our knowledge, the effect of missing value types (random versus not at random) has not been studied in the context of time series data imputation.

2.3.2. Predictive modeling of data with missing values

Advanced machine learning methods in default settings cannot learn from data with missing values. The selection of an imputation method in the health science literature is always arbitrary without validating or even justifying the selection. In a complete case analysis, the samples with missing entries are excluded to create a data set free of missing values, which is not an optimal way to develop predictive models [61]. Such exclusions retain samples from sicker patients who receive more clinical tests than healthier patients. Therefore, such selective population data can yield predictive models with poor generalizability due to patient selection bias [62]. The exclusion of data samples or variables can deteriorate the performance of deep learning models that are always data-hungry.

Table 1 highlights the state-of-the-art methods proposed for imputing missing values in time series data in terms of the highest missing percentages, the number of data sets used for benchmarking, data domains, irregularity in time series data, and whether downstream classification task is performed to validate the quality of imputed data. Table 1 reveals that the state-of-the-art methods lack rigorous benchmarking on multiple data sets and under various experimental scenarios. Benchmarking on only one data set to claim the superiority of a method can be misleading. The only method that performs some rigorous analysis (TA-DualCV) is not a deep learning method. However, none of these recent methods are evaluated for different missing rates and types, considering their viability in real-world settings. Therefore, this paper is inspired by the need to fairly benchmark the deep imputation methods proposed for time series data.

2.3.3. Missing rates in health data

A survey of missing value imputation in time series data suggests that the percentage of missing values is assumed constant and low (~ 10%). A study on a large EHR data set considers imputing variables up to 50% missing values [63]. In another study, Samad et al. have found that the most important echocardiographic predictor variable (tricuspid regurgitation maximum velocity - TR max vel) of mortality can be missing for more than 50% of the patients in a large EHR data set (sample size > 300,000) [64]. TR max vel is not commonly measured in patients who receive echocardiographic imaging because its importance is unknown in clinical practice.

Table 2

Summary of time series health data sets used in the experiments on missing value imputation methods. (.) represents the actual number.

Data set	Sample size	Num. of variables	Num. of time steps	Domain
DACMI	8266	13	Irregular	EHR - ICU patients
Sepsis	2164	30 (44)	20	Synthetic EHR
Hypotension	3910	9 (22)	48	Synthetic EHR
IEEEPPG	3096	5	50 (1000)	PPG, ECG, Acc. Signals
Heartbeat	409	61	50 (405)	Heart Sound

In predictive modeling, variables with more than 50% missing rates are often excluded based on the assumption that imputing such variables would introduce more impurities and eventually deteriorate the data quality. It is well known that machine learning methods require a large sample size to improve the generalizability of predictive models. The exclusion of samples or variables based on missing rates can deteriorate the overall predictive accuracy of machine learning models. Therefore, improving the imputation accuracy even at a high missing rate is imperative to retain the sample size and discover important predictor variables. The imputation of missing values in Samad et al. is conducted first by linear time interpolation and then using a vanilla MICE (missing value imputation using chained equations) method to learn the variable dependency. To the best of our knowledge, the effect of varying missing value rates on time series data imputation methods is unknown. A robust method is expected to retrain high imputation accuracy even at higher percentages of missing values.

3. Methodology

This section discusses the time series data sets, the selection of missing value imputation methods, and the justification for six experiments conducted to evaluate the state-of-the-art missing value imputation methods for time series data.

3.1. Time series health data sets

We use five health data sets to evaluate the state-of-the-art imputation methods for missing values in time series data. Table 2 summarizes the time series health data used in this paper. The DACMI data set is derived from the MIMIC-III database of electronic health records [65], and is shared publicly to conduct the DACMI challenge [27]. The DACMI data set contains 13 blood laboratory test values of 8267 patients admitted to intensive care units (ICUs). The data set comes with missing value indices and corresponding ground truths. Each patient's lab work is recorded for a specific time point. The DACMI challenge organizer has excluded patients with fewer than ten-time points or follow-up visits. Some lab values have not been measured at the time of the visit, resulting in some natively missing data in the ground-truth data. We exclude those records with natively missing values because our experiments require ground truth for evaluation.

We have benchmarked the imputation methods using two publicly available synthetic data sets (sepsis and acute hypotension) generated from the MIMIC-III database and extensively validated against real patient data by the Health Gym project [66]. The primary motivation behind using two synthetic data sets is to allow public benchmarking, whereas real patient data are protected and require extensive curation and expert annotations. These two synthetic data sets are generated from the MIMIC-III clinical database of ICU patients [65] using generative adversarial neural networks (GANs) to the point that a discriminator model cannot differentiate between synthetic and real patient samples. A three-stage validation ensures that the two synthetic data sets sufficiently mimic real patient data [67]. The first stage involves comparing the probability density functions of synthetic and real numeric variables. Binary and categorical variables are compared using side-by-side histogram plots. The second stage performed several statistical tests, including the two-sample Kolmogorov–Smirnov test,

two independent Student's t-tests, Snedecor's F-test, and the three-sigma rule test. Using Kendall's rank correlation coefficients, the third stage examines correlations between variables and trends over time.

The hypotension data set has 3910 patients measured for 22 variables at one-hour intervals over 48 h. Nine of these variables are numeric and used in this study. Non-numeric variables are excluded from our experiments on missing value imputation. The acute hypotension data set does not contain any native missing values. Like the hypotension data set, the sepsis data set is synthetically generated from the MIMIC-III database. It includes information on 2164 patients admitted to intensive care units, with 20 time points recorded for each patient. These time points are derived from 80 h of data and divided into four-hour segments. At each time point, 44 variables are observed, including 35 numeric, three binary, and six categorical variables. Only 30 numeric variables are selected for our experiments, excluding the binary and categorical variables.

The IEEEPPG data set is shared publicly to conduct the IEEE Signal Processing Cup 2015 for heart rate estimation [68]. The IEEEPPG data set contains the measurements of two-channel photoplethysmographic (PPG) signals obtained by pulse oximeters, three-axis acceleration signals, and one-channel ECG signal, all sampled at 125 Hz. The data are recorded from 12 human subjects running at a 15 km/hour peak speed. We use the modified version of the data set obtained from Monash University [69] because the original data set is no longer available. The modified version includes five variables: two PPG signals and three-axis acceleration signals only. The signals are segmented into eight seconds windows with six seconds overlap, resulting in 3096 samples with 1000 time steps per sample. We use the first 50 time points for experimental purposes in this paper.

The heartbeat data set is derived from the PhysioNet and originally shared by the Computing in Cardiology Challenge 2016 [70–72]. The heart sound recordings are collected from four areas of the heart: aortic, pulmonic, tricuspid, and mitral areas. The sound recordings are observed from healthy subjects and pathological patients with confirmed cardiac diagnoses and collected in clinical settings or non-clinical in-home visits. Each recording is truncated to 5 s, and a spectrogram of each instance is created with a window size of 0.061 s and 70% overlap, resulting in 405 time steps. The recording consists of 61 dimensions, each a frequency band from the spectrogram. The data set includes 409 samples. Each sample is truncated to the first 50 time steps in this paper.

3.2. Time series imputation methods

Our survey identifies recently published methods for multivariate time-series data imputation as of 2022. However, our method selection criteria for benchmarking purposes are as follows:

1. Deep learning-based methods only: We have excluded MD-MTS, MICE-DA, SMILES, TA-DualCV, and 3D-MICE for not involving deep neural networks;
2. Availability of source code: Excluded MD-MTS and MICE-DA methods because the source code is unavailable;
3. Recent methods only published in the last five years: Excluded GRU-D, M-RNN, and other older methods that more recent methods have consistently outperformed;

Additionally, we use two baseline methods for benchmarking: (1) the MICE algorithm as one of the most popular imputation methods in health science and (2) LSTM-based imputation as the most basic deep learning method for time series data. We identify eight state-of-the-art deep imputation methods for time series data based on the method selection criteria. First, we select the BRITS method as one of the first deep learning methods to perform longitudinal and cross-sectional imputation simultaneously. BRITS has proven its performance by outperforming baseline deep learning methods such as GRU-D and M-RNN. Second, we choose the CATSI method because of its novelty in designing a *context vector* that captures a patient's overall health state and the temporal dynamics of the time series data. Moreover, CATSI has proven its superiority by outperforming state-of-the-art imputation methods such as BRITS, 3D-MICE, and MICE. We further split the CATSI method into two parts: CATSI-LSTM and CATSI-MLP, to investigate the contribution of its longitudinal and cross-sectional imputation, respectively. The CATSI-LSTM part includes a bidirectional LSTM structure and the *context vector*. The CATSI-MLP part is a fully connected neural network with two hidden layers for cross-sectional imputation. We also compare the bidirectional CATSI-LSTM method with a unidirectional or conventional LSTM to investigate the benefits of context vectors and bidirectional learning. Third, we include the NAOMI method that utilizes a bidirectional RNN structure similar to the CATSI and BRITS methods. It is one of the first methods that use the divide and conquer principle and adversarial training in the missing value imputation problem. Fourth, we use the recently proposed TAME method that uses the bidirectional LSTM architecture, as discussed in Section 2.2.3. This method outperforms several recently proposed state-of-the-art methods, including BRITS, DETROIT, CATSI, BRNN, and 3D-MICE. The TAME method uses a complete data set with known ground truth data to create the train and test folds. We exclude ground truth for the missing values to avoid information leaks between train and test data splits. Finally, the performances of DETROIT and GP-VAE methods are evaluated in downstream classification tasks and discussed in Section 3.3.6.

3.3. Experiments

Our survey of the literature and experience in the DACMI challenge brings forward several findings and research questions. First, a simple piece-wise time interpolation performs better than the sophisticated 3D-MICE method on the DACMI data set [27,73]. It is currently unknown if the claimed superiority of state-of-the-art imputation methods will hold for multiple other data sets. Second, the robustness of time series data imputation methods for varying time lengths, missing rates, and types is currently unknown. The current time series data imputation methods are tested for default low missing rates and data with randomly introduced missing values. Therefore, the robustness and generalizability of these methods across different missing value types and rates remain unknown. Third, it is shown that the best imputation method for a given data set does not consistently perform the best for all variables [73]. That is, some variables in time series data are better imputed using cross-sectional imputation methods. This observation needs more rigorous experimentation on the existing methods. Fourth, imputation methods for time series data are evaluated on several imputation accuracy metrics against the known values. However, it is not well known whether superior imputation accuracy translates to superior predictive modeling, such as time series data classification. Model-generated synthetic values for missing entries may alter the variability and predictive information of the data set. We use several strategies to benchmark cross-sectional imputation methods in a recent study [15]. This paper performs a fair and rigorous benchmarking of state-of-the-art time series data imputation methods in six experiments using five health-related data sets.

3.3.1. Effects of initializing missing values

Missing values require some initial values for the deep learning method to start learning the final estimates. It is hypothesized that initial values can affect the final convergence of the imputation algorithm. We compare four initialization methods: (1) mean, (2) median, (3) piece-wise time interpolation, and (4) delta initialization considering the time gap between measurements as shown in Eq. (11). Here, δ is the time gap matrix for a patient. We modify the initialization method presented in CATSI and Section 2.2.3 to make it independent of the training step while retaining the general idea. For a long time gap, γ is low, and the missing value is approximated by the mean value ($\bar{x}$). Otherwise, when the time gap is shorter, the missing value will be approximated using the value of the next time point, as shown below.

$$\gamma = \exp^{-\max(0,(\frac{\delta-\mu}{\sigma}))}$$
$$x_{t\ init} = \gamma * x_{t+1} + (1 - \gamma) * \bar{x} \tag{11}$$

(4) piece-wise time interpolation, as shown in Eq. (12), where s is the time stamp, and z is the time point where the value is missing in between s_t and s_{t+1}.

$$f(z) = x_t + \frac{x_{t+1} - x_t}{s_{t+1} - s_t} * (z - s_t), \quad z \in [s_t, s_{t+1}] \tag{12}$$

3.3.2. Effects of time series length

It is known that data with longer time series can be challenging to model for recurrent neural networks. This is because the information from the remote past withers away over time during model training. Conversely, a short time length may not have sufficient temporal information for a time series model. Therefore, we will investigate the effect of time length on the performance of time series data imputation methods. A robust imputation method is expected to show stable and superior performance at shorter or longer time lengths. Each time series is truncated to varying lengths, such as 5, 10, 15, 20, and 25 time steps. If some patients have a time series shorter than any of those lengths, the remainder of the time length is padded with zeros.

3.3.3. Cross-sectional versus longitudinal versus hybrid imputation methods

Based on our preliminary results in the DACMI challenge, we hypothesize that some variables are more accurately imputed by cross-sectional than longitudinal methods despite their time-varying trends. Although integrating cross-sectional and longitudinal imputations as a hybrid method is generally superior for time series data, the superiority can vary from variable to variable. We use piece-wise time interpolation for our experiments as a baseline longitudinal imputation method, as described in Section 3.3.1 and Eq. (12). We use two cross-sectional imputation methods built on Gradient Boosting Tree (GBT) regressors and chained equations (CE): CE-GBT-I and CE-GBT-II. The CE-GBT-I method trains GBT regressors once for each variable, which are then used to iteratively update missing values in test data by replacing previously imputed values with updates in a chain of regressors. This method is proposed in [73]. In contrast, the CE-GBT-II method trains a chain of regressors multiple times by iteratively replacing missing values during the training process. Similar to CE-GBT-I, the CE-GBT-II also iteratively updates missing values.

3.3.4. Effects of percentage of missing values

The imputation methods for missing values in time series data generally assume a fixed missing rate between 5% and 10%. For example, The BRITS method is evaluated on three data sets with 10%, 13%, and 10% missing rates. Authors of the DETROIT have used the DACMI data set with 7% missing values. We hypothesize that a robust missing value imputation algorithm should also retain superior performance at higher percentages of missing values. We examine the imputation accuracy on multivariate time series data sets with up to 80% missing values.

Algorithm 1 Simulation of three missing value types.

Input: X: complete data with n samples, d variables ($V = \{x_1, x_2, ..., x_d\}$), t time steps,
      r: missing rate, τ: missing type,
Output: χ: data with missing values
$\tilde{X} \leftarrow$ reshape (X, $n * t$, d)
if missing type == missing completely at random **then**
    [i, j] $\leftarrow$ sample cell indices (r)
    $\tilde{X} \leftarrow$ remove ($\tilde{X}$, [i, j])
    $\chi \leftarrow$ reshape ($\tilde{X}$, n, t, d)
    return χ
else
    low = r / 2, high = 100 - (r / 2)
    if missing type == missing at random **then**
        while currentMissingRate < r% **do**
            sample observed variables $|V_{observed}| = 3$ from V
            for q in $V_{observed}$ **do**
                low_val (q) $\leftarrow$ Percentile ($\tilde{X}(:, q)$, low)
                high_val (q) $\leftarrow$ Percentile ($\tilde{X}(:, q)$, high)
                $i \leftarrow$ row indices (<low_val(q) $\cup$ >high_val(q))
            end for
            $V_{missing} = V - V_{osberved}$
            $\tilde{X} \leftarrow$ remove($V_{missing}$, i)
            currentMissingRate $\leftarrow \frac{\text{count missing entries}(\tilde{X})}{n*d} * 100$
        end while
    end if
    if missing type == missing not at random **then**
        for p in V **do**
            low_val (p) $\leftarrow$ Percentile ($\tilde{X}(:, p)$, low)
            high_val (p) $\leftarrow$ Percentile ($\tilde{X}(:, p)$, high)
            values $\leftarrow$ select ($\tilde{X}(:, p)$, <low_val $\cup$ >high_val)
            $\tilde{X} \leftarrow$ remove ($\tilde{X}(:, p)$, values)
        end for
    end if
    $\chi \leftarrow$ reshape ($\tilde{X}$, n, t, d)
    return χ
end if

3.3.5. Effects of missing value types

As mentioned in Section 2.1.1, three missing value types are defined statistically: (1) missing completely at random (MCAR), (2) missing at random (MAR), and (3) missing not at random (MNAR). It is noteworthy that the MCAR type is the most common in the literature because of its simplicity. However, the MCAR type is rarely observed in the real world, with the MNAR being the most practical type. Therefore, the accuracy of existing time series imputation methods adds limited practical value because the robustness of these methods across different missing value types is currently unknown. Our previous work shows that the imputation accuracy of the MNAR type is far worse and more challenging than that reported for the MCAR type [15]. Therefore, the robustness of the state-of-the-art imputation methods should be challenged using more complex missing value types. This paper simulates the three missing value types following Algorithm 1 [15]. The MCAR type is simulated by randomly selecting $r\%$ of cell indices and removing values from corresponding positions in time-varying data matrices. The MNAR type is simulated by removing values that are lower than (r/2)th percentile and higher than [100-(r/2)]th percentile of a target variable, which accounts for a total of r% missing rate. The MAR type is achieved iteratively. In each iteration, the same percentiles are obtained on three randomly selected observed variables to identify the indices corresponding to low and high values. These indices are used to remove values in the targeted missing variable set, which does not include observed variables. The iteration continues until achieving a total of r% missing rate.

3.3.6. Classification of imputed data

The imputation of missing values enables machine learning of otherwise inoperable data. However, the quality of model-generated data used to impute missing values ultimately determines the outcome of machine learning models, such as regression or classification performance. The quality of imputed data is rarely evaluated in downstream classification tasks to confirm if the imputation accuracy translates to classification performance. This experiment compares the classification accuracy of time series data completed by deep imputation methods with that of ground truth data without missing values and data with missing values without using an imputation method.

3.4. Model evaluation

All data sets are subject to 5% missing values completely at random (MCAR), except for the experiments related to the effects of the percentage of missing values (Section 3.3.4) and effects of missing value types (Section 3.3.5). We split each data set into 80%–20% train and test folds. The train fold with missing values is used to train the models. NAOMI and TAME are the two methods that require ground truth values in the training process to optimize the loss function and model parameters. Other experimental methods do not require ground truth values, as they optimize the model based on the data set's observed values. The trained model imputes missing values on the test data. The imputation accuracy is determined from the model imputed values and the corresponding ground truth on test data. The same train and test data folds are used in all experiments to ensure fair comparison and reproducibility. For evaluation, we obtain normalized root mean square deviation (NRMSD), a widely adopted metric for health data imputation [13,23,24,27,28]. In NRMSD, the absolute difference between actual (Y) and imputed values ($\hat{Y}$) is normalized to bring the variables to the same scales. The result is then multiplied by a mask matrix (M) to disregard all non-missing entries, as shown in Eq. (13), where p is the patient index, t is the time step index, d is the variable index.

$$NRMSD(d) = \sqrt{\frac{\sum_{p,t} M_{p,t,d} \left(\frac{|Y_{p,t,d} - \hat{Y}_{p,t,d}|}{\max(Y_{p,d}) - \min(Y_{p,d})} \right)^2}{\sum_{p,t} M_{p,t,d}}} \quad (13)$$

Based on the NRMSD scores, we rank the performance of the imputation or initialization methods for individual variables of a data set. The average rank of a method is obtained by averaging its ranks across all variables. Therefore, the average rank of a method cannot be lower than one or greater than the number of methods. For our experiment denoted in Section 3.3.6, we report the classification performance using a weighted F1 score, as shown in Eq. (14), where TP, FP, and FN denote true positive, false positive, and false negative predictions, respectively.

$$\text{F1 score} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})} \quad (14)$$

We choose a weighted F1 score for our classification task in Eq. (15) because it is more robust towards an imbalanced data set.

$$\text{Weighted F1 score} = \sum_{i=1}^{N} w_i * (\text{F1 score})_i, \text{ where}$$
$$w_i = \frac{\text{\# of samples in class } i}{\text{Total \# of samples}}. \quad (15)$$

Here, N denotes the total number of classes. For statistical comparisons, we perform Wilcoxon signed-rank (WSR) tests for initialization (Section 3.3.1) and classification (Section 3.3.6) experiments. The WSR test does not require normality assumption and compares matched samples (e.g., scores on the same test data fold) to find differences in mean ranks and is preferred in comparing score data (e.g., F1 scores) [74]. The WSR test has been widely used in comparing machine learning model performance scores [64,75,76].

Table 3

Effects of missing value initialization methods based on average normalized root mean squared deviation scores (NRMSD). The average NRMSD for an initialization method is obtained across seven imputation models (BRITS, CATSI, CATSI-LSTM, LSTM, CATSI-MLP, NAOMI, MICE) and all variables. The average rank of an initialization method for a data set is obtained across all variables and imputation methods.

Dataset	DACMI		Sepsis		Hypotension		IEEEPPG		Heartbeat	
Initialization	NRMSD	Avg.rank	NRMSD	Avg.rank	NRMSD	Avg.rank	NRMSD	Avg.rank	NRMSD	Avg.rank
Delta	0.383 (0.21)	2.6	0.234 (0.03)	2.4	0.211 (0.06)	2.1	0.281 (0.27)	2.0	0.204 (0.06)	2.2
Mean	0.375 (0.19)	2.9	0.233 (0.03)	1.6	0.211 (0.06)	1.5	0.287 (0.27)	3.2	0.205 (0.06)	2.4
Median	0.351 (0.17)	2.1	0.234 (0.03)	2.1	0.212 (0.06)	3.1	0.289 (0.27)	3.8	0.205 (0.06)	2.4
Piece-wise	0.394 (0.24)	2.4	0.249 (0.05)	4.0	0.215 (0.06)	3.3	0.219 (0.29)	1.0	0.207 (0.07)	3.0

Table 4

The best initialization methods for an imputation method and data set pair based on normalized root mean squared deviation (NRMSD) scores. The average rank of the best initialization method is obtained by averaging its ranks across all variables in a data set. PW = Piece-wise interpolation.

Dataset	DACMI			Sepsis			Hypotension			IEEEPPG			Heartbeat		
Method	Init.	NRMSD	Avg.rank	Init.	NRMSD	Avg.rank	Init.	NRMSD	Avg.rank	Init.	NRMSD	Avg.rank	Init.	NRMSD	Avg.rank
BRITS	Delta	0.266	3.6	PW	0.190	1.6	Mean	0.163	2.1	Delta	0.094	4.7	PW	0.194	4.2
CATSI	Delta	0.206	1.4	Delta	0.202	1.9	PW	0.158	1.3	PW	0.057	2.8	PW	0.139	1.3
CATSI-LSTM	Mean	0.223	2.6	PW	0.226	3.7	Mean	0.173	3.0	PW	0.053	1.2	PW	0.156	2.9
LSTM	PW	0.241	3.6	Mean	0.282	6.7	Mean	0.224	5.7	PW	0.080	4.3	PW	0.173	3.5
CATSI-MLP	PW	0.265	4.1	Delta	0.233	4.4	Mean	0.204	5.0	PW	0.294	6.0	Median	0.164	3.4
NAOMI	Median	0.437	6.3	Median	0.240	4.9	Delta	0.218	4.9	PW	0.054	2.0	Delta	0.245	6.1
MICE	PW	0.730	6.4	Median	0.252	4.8	PW	0.337	6.0	PW	0.901	7.0	PW	0.332	6.6

4. Experimental results

The sections below detail the results of the six experiments on state-of-the-art deep time series imputation methods.

4.1. Effects of initializing missing values

Table 3 shows the effect of four missing value initialization approaches on five data sets across all imputation methods. Table 4 presents an imputation method-specific selection of the best initialization method and corresponding scores.

The median initialization achieves the best average rank (2.1) for one (DACMI) out of five data sets with an average NRMSD score of 0.351 (0.17). Table 4 shows that the NAOMI and MICE methods yield substantially worse imputation accuracy. Therefore, excluding NAOMI and MICE, piece-wise initialization ranks as the best initialization method for the DACMI data set with an average NRMSD score of 0.241 (0.02) and an average rank of 1.2. The exclusion of the NAOMI and MICE methods turns median initialization into the worst approach, with an average NRMSD score of 0.258 (0.04) and an average rank of 3.9.

Table 3 shows that the mean initialization performs the best for the sepsis (NRMSD: 0.233 (0.03), average rank: 1.6) and hypotension (NRMSD: 0.211 (0.06), average rank: 1.5) data sets. These two data sets are synthetically generated at regular time intervals with relatively shorter time series than the other three. The sepsis and hypotension data sets have only 20 and 48 time points for all samples. In contrast to three other data sets, these factors may be important to consider mean initialization over the approaches. Table 3 shows that the piece-wise interpolation substantially outperforms all other initialization approaches for the IEEEPPG data set (NRMSD: 0.219 (0.29), average rank: 1.0). Similarly, the heartbeat data set with a long time series (405-time steps) is truncated to 50-time steps, similar to the IEEEPPG data set. Although delta is the best initialization approach for this data set (NRMSD: 0.204 (0.06), average rank: 2.2), the difference in improvement over other methods (mean, median, and piece-wise) is insignificant. Excluding the NAOMI and MICE results, piece-wise initialization ranks the best for the heartbeat data set with an average NRMSD of 0.165 (0.02) and an average rank of 1.8. No initialization method is statistically superior to other methods except for the IEEEPPG data set. The IEEEPPG data set shows that the piece-wise initialization is significantly better than mean, median, and delta initialization ($p < 0.05$). We also observe that delta initialization is significantly better than median initialization.

Table 5

Effects of time series length on the performance of missing value imputation algorithms. For this experiment, all data sets are subjected to 5% missing values of the completely-at-random type. The scores represent normalized root squared deviation (NMRSD).

Timestep	Method	DACMI	Sepsis	Hypotension	IEEEPPG	Heartbeat
5	BRITS	0.844	0.372	0.516	2.506	0.964
	CATSI	0.427	0.501	0.535	0.419	0.470
	CATSI-LSTM	0.460	0.486	0.538	0.415	0.473
	LSTM	0.471	0.466	0.454	0.792	0.686
	CATSI-MLP	0.415	0.430	0.424	0.444	0.488
	NAOMI	0.493	0.483	0.465	0.440	0.516
	TAME	0.538	0.546	0.520	0.594	0.515
10	BRITS	0.331	0.255	0.320	0.674	0.566
	CATSI	0.267	0.280	0.364	0.210	0.317
	CATSI-LSTM	0.296	0.308	0.354	0.198	0.328
	LSTM	0.352	0.345	0.338	0.419	0.438
	CATSI-MLP	0.307	0.302	0.299	0.351	0.362
	NAOMI	0.392	0.339	0.351	0.349	0.403
	TAME	0.346	0.335	0.318	0.318	0.340
15	BRITS	0.285	0.213	0.261	0.363	0.407
	CATSI	0.230	0.228	0.279	0.126	0.263
	CATSI-LSTM	0.258	0.252	0.264	0.132	0.285
	LSTM	0.325	0.305	0.294	0.438	0.344
	CATSI-MLP	0.285	0.258	0.265	0.333	0.295
	NAOMI	0.533	0.281	0.308	0.309	0.361
	TAME	0.316	0.275	0.268	0.240	0.288
20	BRITS	0.270	0.190	0.221	0.327	0.319
	CATSI	0.219	0.202	0.218	0.106	0.221
	CATSI-LSTM	0.245	0.219	0.230	0.107	0.249
	LSTM	0.317	0.282	0.274	0.359	0.304
	CATSI-MLP	0.276	0.234	0.244	0.330	0.254
	NAOMI	0.693	0.245	0.284	0.293	0.332
	TAME	0.303	0.236	0.234	0.208	0.256
25	BRITS	0.261	/	0.201	0.198	0.286
	CATSI	0.213	/	0.191	0.091	0.194
	CATSI-LSTM	0.240	/	0.211	0.091	0.211
	LSTM	0.314	/	0.258	0.366	0.265
	CATSI-MLP	0.273	/	0.231	0.323	0.225
	NAOMI	0.612	/	0.267	0.296	0.302
	TAME	0.299	/	0.217	0.176	0.237

Table A.1 in Appendix A provides comprehensive detail of the performance metrics for all initialization methods for individual imputation models and data sets. This table shows that the CATSI method with delta initialization has the best imputation performance across all other methods (average rank: 3.7) using the DACMI data set. The

Table 6

Variable specific imputation error for the DACMI data set based on normalized root mean square deviation scores. Mean initialization is used for all methods except TAME and piece-wise time interpolation.

Method	PCL	PK	PLCO2	PNA	HCT	HGB	MCV	PLT	WBC	RDW	PBUN	PCRE	PGLU	Avg.NRMSD	Avg.rank
BRITS	0.191	0.269	0.219	0.202	0.145	0.146	0.395	0.351	0.269	0.431	0.264	0.293	0.293	0.267	3.77
CATSI	0.187	0.261	0.226	0.206	0.137	0.137	0.251	0.176	0.219	0.222	0.183	0.220	0.265	0.207	1.62
CATSI-LSTM	0.210	0.260	0.231	0.233	0.218	0.215	0.254	0.177	0.219	0.216	0.184	0.221	0.263	0.223	2.85
LSTM	0.308	0.299	0.304	0.301	0.285	0.286	0.319	0.322	0.299	0.339	0.323	0.344	0.299	0.310	6.92
CATSI-MLP	0.227	0.297	0.279	0.236	0.142	0.147	0.304	0.308	0.302	0.334	0.307	0.297	0.306	0.268	4.69
NAOMI	0.940	0.428	0.378	1.114	0.453	0.504	1.269	0.343	0.328	1.222	0.353	0.354	0.334	0.617	8.77
TAME	0.262	0.319	0.291	0.267	0.224	0.228	0.332	0.278	0.311	0.300	0.331	0.331	0.339	0.294	6.46
CE-GBT-I	0.246	0.334	0.310	0.247	0.150	0.148	1.370	0.962	0.858	1.546	0.647	0.931	0.442	0.630	8.38
CE-GBT-II	0.244	0.334	0.308	0.245	0.149	0.149	1.369	0.970	0.836	1.539	0.631	0.919	0.421	0.624	7.85
Piece-wise	0.222	0.272	0.225	0.234	0.233	0.232	0.258	0.181	0.214	0.224	0.174	0.229	0.295	0.230	3.69

sepsis data set shows the best imputation rank (average rank: 5.3) using the BRITS algorithm initialized by the piece-wise time interpolation. The hypotension data set is the best imputed by the CATSI algorithm, initialized by the piece-wise time interpolation (average rank: 4.1). The same algorithm and initialization combination ranks the best for the heartbeat data set (average rank: 3.9). However, the LSTM part of the CATSI algorithm (CATSI-LSTM) with piece-wise time interpolation yields the best imputation accuracy for the IEEEPPG data set (average rank: 1.2). In other words, the missing values of the IEEEPPG data set are better predicted across time than across the variables. This observation is supported by the finding that the IEEEPPG also has the lowest correlations between variables. Overall, the CATSI method with piece-wise interpolation ranks the best (rank: 1), considering its imputation performance across all data sets. At the same time, the non-deep-learning MICE method ranks the worst for any initialization settings. These rankings suggest that advanced imputation methods benefits less from any superior approach for initializing missing values.

4.2. Effects of time series length

Table 5 compares the imputation accuracy for varying time-series lengths. Mean initialization is used, except for the TAME method, which does not explicitly initialize missing values. To achieve a constant time length for fair comparisons, longer and shorter time series data are truncated and padded with zeros, respectively. The cross-sectional imputation method (CATSI-MLP) is the best for short time series data (five time points) on the DACMI (NRMSD: 0.415) and hypotension (NRMSD: 0.424) data sets. The BRITS method performs the best for the sepsis data set for all time series lengths. The CATSI method performs the best with the heartbeat data set for all time series lengths and is superior on the IEEEPPG data set when the time series length is 15 (NRMSD: 0.126) and 20 (NRMSD: 0.106) time points. When the time series has five and ten time points, the CATSI-LSTM method shows the best imputation accuracy with NRMSD scores of 0.415 and 0.198, respectively. CATSI and CATSI-LSTM rank the best for 25 time points (NRMSD: 0.091). Similar results are observed with the DACMI data set, where CATSI outperforms other methods when the time series length is ten or more. For the hypotension data set, the results are mixed. When the time series is short (five and ten time points), the cross-sectional CATSI-MLP imputation method outperforms other methods. When the length of the time series is 15, BRITS performs the best (NRMSD: 0.261). However, the difference between the BRITS, CATSI-LSTM, CATSI-MLP, and TAME methods is insignificant. The CATSI method ranks the best when the hypotension data set has 20 (NRMSD: 0.218) and 25 (NRMSD: 0.191) time steps. Overall, the CATSI appears to be the best method when the time series has more than 15 time points.

4.3. Effect of cross-sectional versus longitudinal versus hybrid imputation methods

The sections below present data set-specific findings about the performance of different imputation strategies.

4.3.1. DACMI data set

Table 6 reveals that most variables (eight out of 13) in the DACMI data set are best imputed by a hybrid imputation method (BRITS or CATSI) that combines cross-sectional and longitudinal learning. The remaining five variables (PK, WBC, RDW, PBUN, and PGLU) are best imputed using longitudinal methods (CATSI-LSTM or piece-wise). However, two variables (HCT and HGB) correlate more with other variables than their time-series trend. A pair-wise cross-correlation analysis shows a high correlation (0.96) between HCT and HGB. The cross-sectional CATSI-MLP-based imputation method performs better than longitudinal imputation methods for these two variables. Figs. B.1(a)–(b) in Appendix B shows that higher correlation scores between variables yield better imputation accuracy (lower NRMSD) using the cross-sectional MLP method. These figures show that the imputation accuracy (NRMSD scores) is almost constant across varying correlation values when using a longitudinal imputation method (CATSI-LSTM). Overall, the cross-sectional imputation method ranks second and third for HCT and HGB variables, respectively. Similarly, the correlation between PCL and PNA time-series trends is moderate (0.70), which may not be strong enough to give the cross-sectional imputation method an edge over its longitudinal imputation counterpart. All other variable pairs have a weak correlation (<0.6), which explains the suboptimal performance of cross-sectional imputation methods for these variables. Two variables (WBC and PBUN) show the best imputation accuracy with the piece-wise time interpolation method, suggesting that these variables are more predictable from their time-series trends. The cross-sectional imputation method performs poorly for these two variables, possibly because these variables are poorly correlated with other variables. Several other variables (PK, RDW, and PGLU) are best imputed using the longitudinal part of the hybrid CATSI method (CATSI-LSTM). Therefore, the pairwise correlation between variables may provide an important guideline in selecting an appropriate imputation method.

4.3.2. Sepsis data set

The sepsis data set indicates that hybrid imputation methods are the best on 17 out of 30 variables. One variable (PaO2) shows the best imputation accuracy with the longitudinal CATSI-LSTM method (NRMSD: 0.199). Despite being part of a time series data set, the remaining 12 variables are best imputed with cross-sectional imputation methods. For example, the SysBP and meanBP variables have a 0.68 correlation, and SysBP is best imputed using the CE-GBT-II method. Intuitively, mean, systolic, and diastolic blood pressure variables can be correlated. The MeanBP and SysBP variables with a 0.76 correlation are best imputed using the CE-GBT-I and CE-GBT-II cross-sectional imputation methods, respectively. The CE-GBT-I method also performs best for several variables, including Na, Cl, CO2, BE, HCO3, SGPT, and PaCO2. The remaining variables, such as DiaBP, Creatinine, and SGOT, are best imputed by the CE-GBT-II method. Some variables that are best imputed by cross-sectional methods have a moderate correlation with each other: Na-Cl (0.65), CO2-BE (0.64), CO2-HCO3 (0.63), and SGOT-SGPT (0.79). It is important to note that the sepsis data set has a relatively short time series with 20 time steps. This may contribute to the superior performance of cross-sectional imputation methods compared to their hybrid and longitudinal counterparts.

Table 7

Variable specific imputation error for the hypotension data set based on normalized root mean square deviation scores. Mean initialization is used for all methods except TAME and piece-wise time interpolation.

Method	MAP	DBP	SBP	Urine	ALT	AST	PO2	Lactic	Serum	Avg.NRMSD	Avg.rank
BRITS	0.122	0.129	0.161	0.175	0.173	0.169	0.170	0.179	0.185	0.163	2.22
CATSI	0.124	0.141	0.161	0.161	0.162	0.163	0.171	0.180	0.160	0.158	1.89
CATSI-LSTM	0.182	0.179	0.181	0.165	0.166	0.166	0.172	0.183	0.163	0.173	3.56
LSTM	0.223	0.222	0.226	0.219	0.222	0.223	0.213	0.228	0.239	0.224	7.11
CATSI-MLP	0.133	0.157	0.190	0.211	0.216	0.231	0.226	0.229	0.240	0.204	6.67
NAOMI	0.226	0.221	0.223	0.207	0.213	0.213	0.212	0.227	0.222	0.218	6.11
TAME	0.143	0.150	0.171	0.190	0.197	0.181	0.213	0.207	0.195	0.183	4.44
CE-GBT-I	0.119	0.137	0.199	0.256	0.268	0.317	0.222	0.270	0.774	0.285	7.28
CE-GBT-II	0.119	0.137	0.199	0.256	0.268	0.317	0.222	0.270	0.774	0.285	7.28
Piece-wise	0.249	0.246	0.235	0.252	0.243	0.225	0.256	0.241	0.201	0.239	8.44

Table 8

Variable specific imputation error for the IEEEPPG data set based on normalized root mean square deviation scores. Mean initialization is used for all methods except TAME and piece-wise time interpolation.

Method	PPG1	PPG2	Acc. X	Acc. Y	Acc. Z	Avg.NRMSD	Avg.rank
BRITS	0.067	0.071	0.091	0.104	0.136	0.094	4.00
CATSI	0.067	0.060	0.065	0.065	0.068	0.065	2.00
CATSI-LSTM	0.069	0.064	0.067	0.066	0.072	0.068	3.20
LSTM	0.310	0.305	0.335	0.304	0.333	0.317	8.00
CATSI-MLP	0.297	0.292	0.293	0.298	0.293	0.295	7.00
NAOMI	0.282	0.283	0.264	0.269	0.267	0.273	6.00
TAME	0.137	0.130	0.137	0.134	0.134	0.134	4.80
CE-GBT-I	0.398	0.417	0.825	1.300	0.992	0.787	9.40
CE-GBT-II	0.396	0.426	0.829	1.338	0.945	0.787	9.60
Piece-wise	0.059	0.052	0.050	0.050	0.057	0.054	1.00

4.3.3. Hypotension data set

Table 7 shows that all features except one (MAP) are best imputed using hybrid imputation methods (BRITS and CATSI). The CE-GBT (I and II) method yields the best and second-best imputation accuracy for the MAP and DBP variables, respectively. These two variables have a 0.80 correlation, which explains the superior performance of a cross-sectional imputation method. On the other hand, a moderately correlated (0.71) variable pair (ALT and AST) is better imputed via hybrid and longitudinal imputation methods than cross-sectional. This is because all other variable pairs have a weak correlation (<0.6). Results show that several variables (ATL, AST, urine, and serum creatinine) show the second-best performance using the longitudinal part of the CATSI method, CATSI-LSTM. Overall, integrating cross-sectional and longitudinal imputation approaches in a hybrid framework improves the imputation accuracy for this data set.

4.3.4. IEEEPPG data set

None of the cross-sectional imputation methods perform well on the variables of the IEEEPPG data set, as shown in Table 8. All variables are best imputed by the piece-wise time interpolation method, outperforming hybrid methods (CATSI and BRITS). The variables have weak pairwise correlation (highest correlation: 0.31). Therefore, cross-sectional imputation methods are not expected to predict the missing values of other variables accurately. The hybrid CATSI method shows the second-best imputation performance for all variables.

4.3.5. Heartbeat data set

The heartbeat data set has 61 variables, which creates a total of 1830 unique variable pairs. Of 1830 pairs, 315 correlate greater than 0.8, and 69 correlate greater than 0.9. The CATSI method is the best for 24 variables and the second best for 20 other variables. The piece-wise time interpolation method provides the second-best imputation performance after the hybrid CATSI method. The piece-wise time interpolation offers the best imputation accuracy for 17 variables and is the second best for 12 other variables.

The CE-GBT-II cross-sectional imputation method performs the best for eight variables and the second best for five other variables. The

CE-GBT-I is the best method for imputing six variables and the second-best for the other five variables. The CATSI-LSTM longitudinal method is the best imputation method for five variables and the second-best for 13 other variables. Out of 69 strongly correlated (>0.9) variable pairs (47 unique variables), 14 variables are best imputed by cross-sectional imputation methods (CE-GBT-II: 7; CE-GBT-I: 6; CATSI-MLP: 1). In general, no single method is superior for imputing missing values of individual variables in time series data. The overall imputation accuracy can be optimized by statistical analysis of the variables and variable-specific selection between cross-sectional, longitudinal, and hybrid imputation methods.

4.4. Effects of missing values rates and types

When the missing value types are MAR (Fig. 3) and MCAR (Fig. 4), the CATSI method is the best on the DACMI data set up to 30% and 50% missing values, respectively. The BRITS method overtakes CATSI when the missing value rate is more than 50% for both MAR and MCAR types. In other words, the BRITS method is least susceptible to higher missing value rates for any missing value types. When the missing value type is MNAR, the BRITS method is the superior method for any percentage of missing values on the DACMI data set. The BRITS method is the best for the sepsis data set regardless of the missing value percentages and types. Although the TAME method is known for its high performance, it performs relatively poorly when the missing value percentage is high (Figs. 3(b), 4(b)). For the hypotension data set, the CATSI method is the best up to 10% missing values for MAR and MCAR types, while

TAME performs the best between 30% and 70% missing values. For the MNAR type, the BRITS method is generally superior to all other methods on the hypotension data set, except when the missing rate is between 40% and 60% (Fig. 5(c)). The BRITS method performs the best for the MNAR type up to 60% missing values in the heartbeat data.

In general, the CATSI method performs the best at all missing rates when the heartbeat data set has missing values of MAR and MCAR types (Figs. 3(e), 4(e). The CATSI-LSTM method is the best up to 20% MAR type missing values in the IEEEPPG data set (Fig. 3(d)). The TAME method shows the best performance for 40% or more MAR type missing values (Fig. 3(d)). Similar performance is observed with the MCAR type, where TAME is the best method for 30% and more missing rates (Fig. 4(d)). The TAME method also performs best up to 40% MNAR type missing values in the IEEEPPG dataset (Fig. 5(d)). However, the BRITS method outperforms TAME when the missing value rate is 50% or higher.

4.5. Classification of imputed data

The quality of imputed data is evaluated in downstream classification tasks. Three of five data sets (sepsis, hypotension, and heartbeat) have ground truth labels for classification. The classification label is provided for each time point for the sepsis and hypotension data sets, whereas the heartbeat data set has labels for individual samples. The DACMI and IEEEPPG data sets do not provide classification labels. We
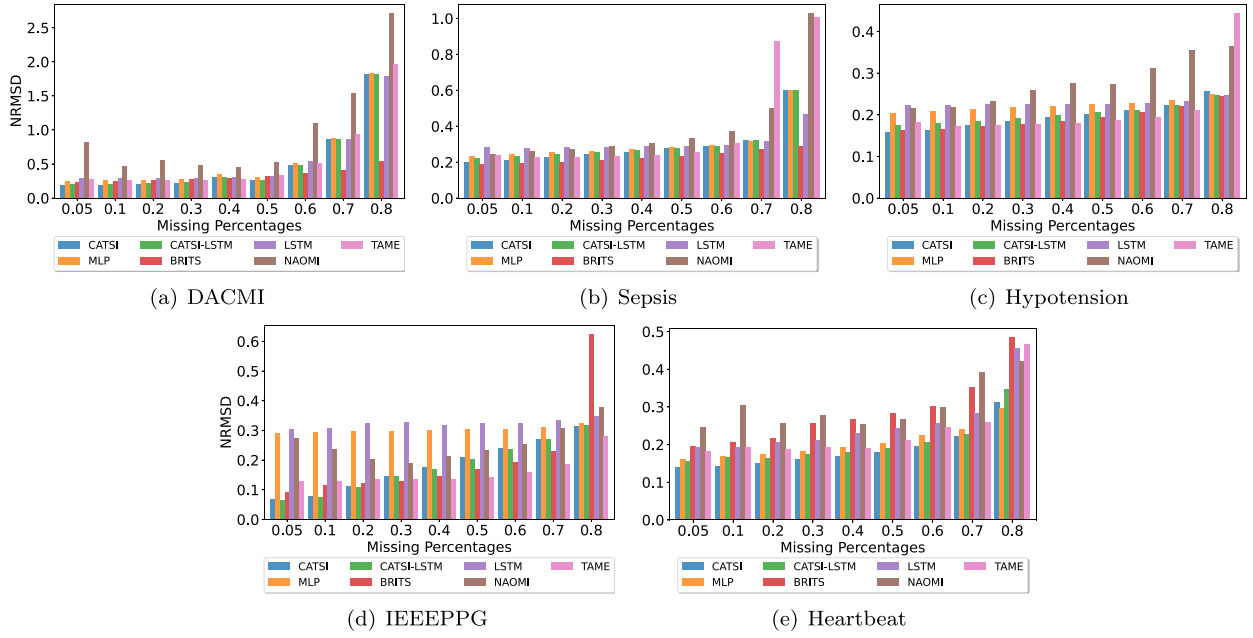
Fig. 3. Comparing normalized root mean squared deviation (NRMSD) scores of seven deep methods (BRITS, CATSI, CATSI-LSTM, LSTM, CATSI-MLP, NAOMI, TAME) for imputing missing values in time series data. The missing at random (MAR) type is used at varying missing rates.
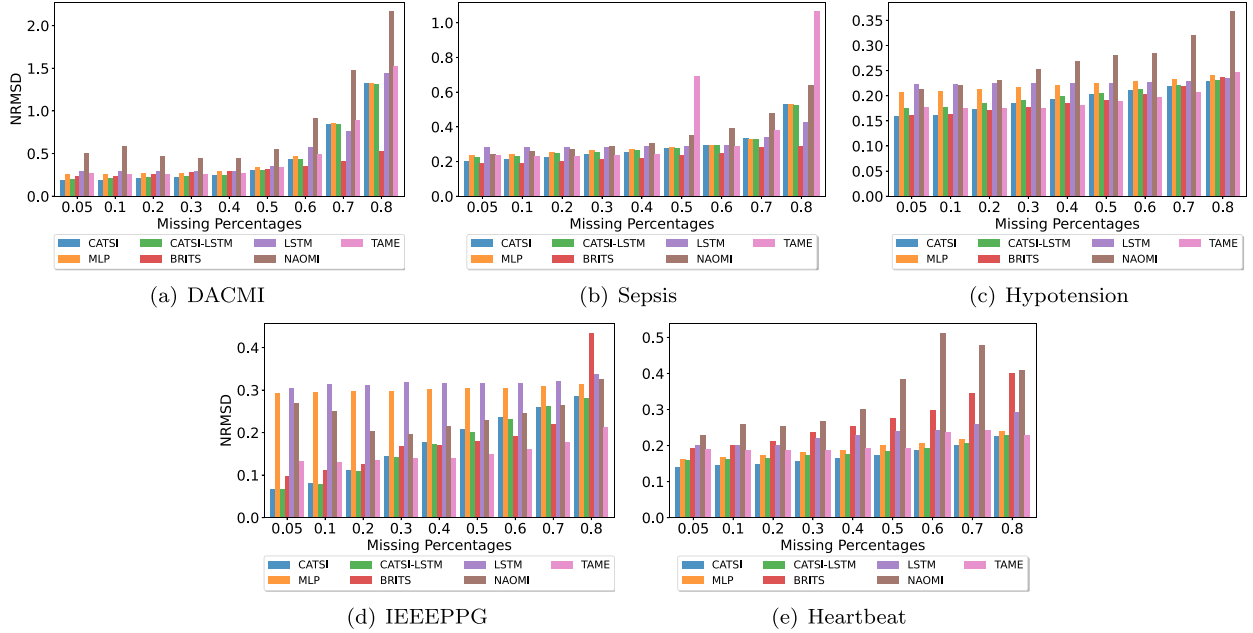

Fig. 4. Comparing normalized root mean squared deviation (NRMSD) scores of seven deep methods (BRITS, CATSI, CATSI-LSTM, LSTM, CATSI-MLP, NAOMI, TAME) for imputing missing values in time series data. The missing completely at random (MCAR) type is used at varying missing rates.

perform the following steps to obtain and compare the classification results.

First, we simulate 20% missing values of MCAR type. The data sets with missing values are then completed using the imputation methods. Second, we randomly sample 80% of the imputed data to train a classifier. The remaining 20% data are used for testing and reporting the classification accuracy. This sampling process is repeated ten times to perform statistical analysis of the imputation methods. A gradient-boosting classifier model with a default hyperparameter setting (learning rate: 0.1, number of estimators: 100, maximum tree depth: 3) is trained and tested on imputed data. Third, the quality

of imputed data is compared with two baseline scenarios: 1) when the data set is complete with no missing values (CNoMV) and (2) when no imputation method is used to handle the missing values in data [77,78], namely the no imputation method (NIM). Tomasev et al. have avoided a method to explicitly impute missing values based on a prior study [77] where imputing missing values does not constantly improve the accuracy of predictive models [78]. This observation is based on a multivariate kernel regression-based method in 2015, despite more sophisticated deep learning methods emerging over recent years. Fourth, we choose a binary *readmission* label to classify the sepsis data set. The categorical *vasopressors* label with four class categories is
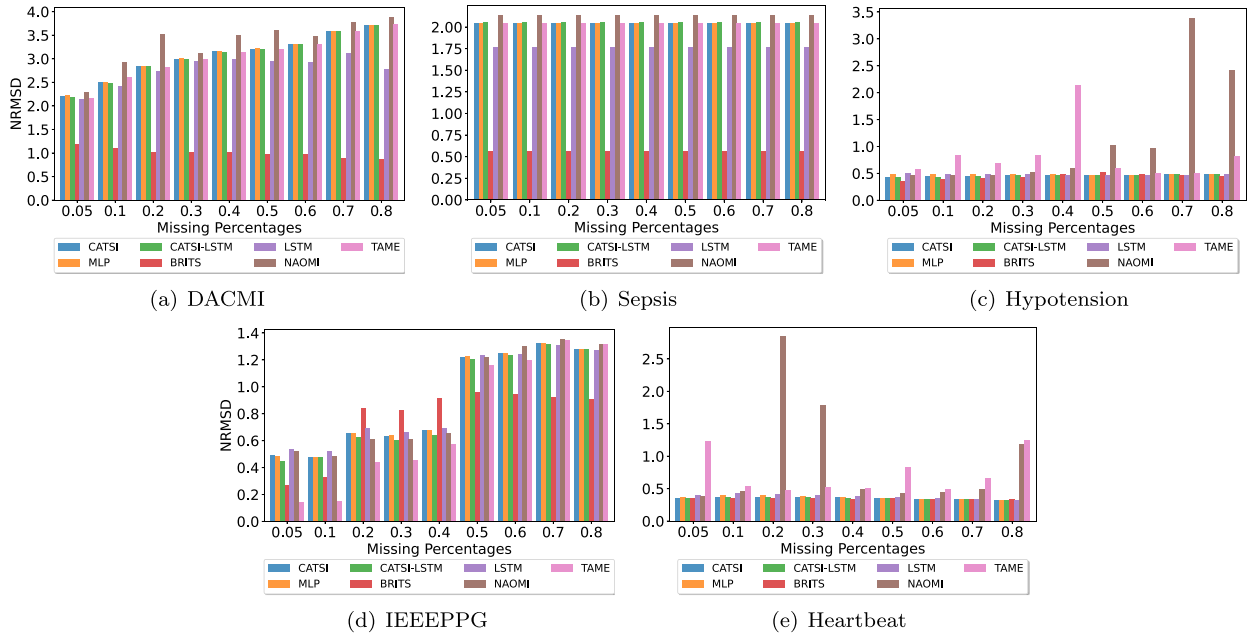
(a) DACMI (b) Sepsis (c) Hypotension

(d) IEEEPPG (e) Heartbeat

Fig. 5. Comparing normalized root mean squared deviation (NRMSD) scores of seven deep methods (BRITS, CATSI, CATSI-LSTM, LSTM, CATSI-MLP, NAOMI, TAME) for imputing missing values in time series data. The missing not-at-random (MNAR) type is used at varying missing rates.

Table 9

The quality of imputed data sets with 20% missing values of MCAR type is evaluated in downstream classification. The average F1 weighted classification scores with 95% confidence intervals are reported after sampling ten times. Missing values are initialized using mean values. CNoMV = complete with no missing values; NIM = no imputation.

Data set	BRITS	NAOMI	CATSI	CATSI-LSTM	LSTM	CATSI-MLP	TAME	DETROIT	GP-VAE	MICE	NIM	CNoMV
Sepsis	0.722 ± 0.003	0.712 ± 0.003	0.722 ± 0.003	0.724 ± 0.002	0.724 ± 0.002	0.714 ± 0.002	0.719 ± 0.002	0.721 ± 0.002	0.674 ± 0.004	0.690 ± 0.003	0.660 ± 0.002	0.735 ± 0.003
Hypotension	0.809 ± 0.001	0.805 ± 0.002	0.809 ± 0.001	0.809 ± 0.002	0.807 ± 0.001	0.805 ± 0.001	0.811 ± 0.001	0.809 ± 0.001	0.803 ± 0.001	0.800 ± 0.001	0.795 ± 0.001	0.811 ± 0.001
Heartbeat	0.614 ± 0.03	0.614 ± 0.03	0.651 ± 0.04	0.659 ± 0.04	0.638 ± 0.03	0.635 ± 0.05	0.632 ± 0.03	0.637 ± 0.02	0.604 ± 0.03	0.586 ± 0.04	0.603 ± 0.02	0.603 ± 0.03

used as the classification target for the hypotension data set. We classify the heartbeat data set using a binary label: the heartbeat as normal versus abnormal.

4.5.1. Classification accuracy

Table 9 shows the average F1 weighted classification score for each imputation method for the mean initialization setting on three data sets. The sepsis data set imputed by the CATSI-LSTM and LSTM methods shows the best classification performance (weighted F1 score: 0.724 ± 0.003. The worst classification accuracy is obtained by the no imputation method (NIM) (weighted F1 score: 0.660 ± 0.003). Intuitively, the complete data set with no missing values (CNoMV) yields the highest weighted F1 score of 0.735 ± 0.003.

For the hypotension data set, the classification accuracy on imputed test data is similar across different imputation methods. The TAME method yields the best classification performance (weighted F1 score: 0.811 ± 0.003. Complete data sets with no missing values (CNoMV) achieve the same accuracy as the data imputed by the TAME method. Data with missing values without using any explicit imputation method shows the worst classification accuracy for the hypotension data set(weighted F1 score: 0.795 ± 0.001). The heartbeat data set shows the best classification performance with CATSI-LSTM imputed data (weighted F1 score: 0.659 ± 0.04, which is better than the accuracy of the complete data set with no missing values (weighted F1 score: 0.603 ± 0.03). All other deep imputation methods have similarly shown better classification accuracy than using complete signals without missing values. This observation remains consistent in multiple experiments. Unlike EHR data (Sepsis or Hypotension), the heartbeat

data set is a collection of sensor signals, possibly with redundant frequency and noise components. It is possible that introducing missing values in such signal and then imputing can filter out noise components in the signal, improving overall signal quality. The MICE method shows the worst classification accuracy for the heartbeat data set (weighted F1 score: 0.586 ± 0.04).

4.5.2. Statistical analysis of classification accuracy

Train and test data are randomly sampled ten times from imputed data to perform the Wilcoxon signed-rank test on classification accuracies. The sepsis data set (Table 10) shows that the classification of data without missing values (CNoMV) is statistically better than that of imputed data by any method. However, all imputation methods yield statistically better classification than the case with missing values without imputation (NIH). All imputation methods are statistically better than the MICE method, except the GP-VAE, which is statistically worse than MICE. The CATSI method is better than CATSI-MLP, NAOMI, GP-VAE, and MICE. However, the CATSI method shows no statistically different performance compared to CATSI-LSTM, LSTM, TAME, and DETROIT methods.

The hypotension data set (Table 11) shows that the classification accuracy on data without missing values is statistically similar to those imputed by the CATSI, TAME, and DETROIT methods. All imputation methods are statistically better than the no imputation method (NIM). Similarly, all imputation methods are statistically better than MICE imputation. All methods except NAOMI are statistically better than GP-VAE. The classification of data imputed by NAOMI, LSTM, and CATSI-MLP methods is statistically worse than the ones imputed by

Table 10

Statistical comparison between imputation methods using the sepsis data set. Methods in the row are read as statistically better (B) or worse (W) than those in the columns. NS = not significant, CNoMV = complete with no missing values, NIM = no imputation.

Method	BRITS	NAOMI	CATSI	CATSI-LSTM	LSTM	CATSI-MLP	TAME	DETROIT	GP-VAE	MICE	NIM	CNoMV
BRITS	–	p = 0.002, B	p = 0.6, NS	p = 0.4, NS	p = 0.2, NS	p = 0.006, B	p = 0.04, B	p = 0.6, NS	p = 0.002, B	p = 0.002, B	p = 0.002, B	p = 0.002, W
NAOMI		–	p = 0.006, W	p = 0.002, W	p = 0.002, W	p = 0.4, NS	p = 0.002, W	p = 0.004, W	p = 0.002, B	p = 0.002, B	p = 0.002, B	p = 0.002, W
CATSI			–	p = 0.2, NS	p = 0.3, NS	p = 0.002, B	p = 0.08, NS	p = 0.6, NS	p = 0.002, B	p = 0.002, B	p = 0.002, B	p = 0.002, W
CATSI-LSTM				–	p = 1, NS	p = 0.002, B	p = 0.002, B	p = 0.2, NS	p = 0.002, B	p = 0.002, B	p = 0.002, B	p = 0.002, W
LSTM					–	p = 0.006, B	p = 0.01, B	p = 0.2, NS	p = 0.002, B	p = 0.002, B	p = 0.002, B	p = 0.004, W
CATSI-MLP						–	p = 0.01, W	p = 0.01, W	p = 0.002, B	p = 0.002, B	p = 0.002, B	p = 0.002, W
TAME							–	p = 0.6, NS	p = 0.002, B	p = 0.002, B	p = 0.002, B	p = 0.002, W
DETROIT								–	p = 0.002, B	p = 0.002, B	p = 0.002, B	p = 0.002, W
GP-VAE									–	p = 0.002, W	p = 0.002, B	p = 0.002, W
MICE										–	p = 0.002, B	p = 0.002, W
NIM											–	p = 0.002, W

Table 11

Statistical comparison between imputation methods using the hypotension data set. Methods in the row are read as statistically better (B) or worse (W) than those in the columns. NS = not significant, CNoMV = complete with no missing values, NIM = no imputation.

Method	BRITS	NAOMI	CATSI	CATSI-LSTM	LSTM	CATSI-MLP	TAME	DETROIT	GP-VAE	MICE	NIM	CNoMV
BRITS	–	p = 0.01, B	p = 0.6, NS	p = 1, NS	p = 0.002, B	p = 0.004, B	p = 0.2, NS	p = 0.8, NS	p = 0.002, B	p = 0.002, B	p = 0.002, B	p = 0.006, W
NAOMI		–	p = 0.04, W	p = 0.04, W	p = 0.08, NS	p = 1, NS	p = 0.02, W	p = 0.01, W	p = 0.1, NS	p = 0.002, B	p = 0.002, B	p = 0.004, W
CATSI			–	p = 0.7, NS	p = 0.08, NS	p = 0.002, B	p = 0.08, NS	p = 0.7, NS	p = 0.002, B	p = 0.002, B	p = 0.002, B	p = 0.06, NS
CATSI-LSTM				–	p = 0.2, NS	p = 0.002, B	p = 0.1, NS	p = 0.8, NS	p = 0.004, B	p = 0.002, B	p = 0.002, B	p = 0.049, W
LSTM					–	p = 0.1, NS	p = 0.01, W	p = 0.04, W	p = 0.002, B	p = 0.002, B	p = 0.002, B	p = 0.002, W
CATSI-MLP						–	p = 0.002, W	p = 0.004, W	p = 0.049, B	p = 0.004, B	p = 0.002, B	p = 0.004, W
TAME							–	p = 0.2, NS	p = 0.002, B	p = 0.002, B	p = 0.002, B	p = 0.8, NS
DETROIT								–	p = 0.002, B	p = 0.002, B	p = 0.002, B	p = 0.1, NS
GP-VAE									–	p = 0.002, B	p = 0.002, B	p = 0.002, W
MICE										–	p = 0.002, B	p = 0.002, W
NIM											–	p = 0.002, W

TAME and DETROIT methods. The BRITS method has significantly better results than NAOMI, LSTM, and CATSI-MLP methods. Furthermore, the hypotension data set imputed by the NAOMI method produces significantly worse results than CATSI and CATSI-LSTM methods. Lastly, BRITS, CATSI, and CATSI-LSTM show significantly better results than CATSI-MLP.

In general, the classification accuracies are not statistically different when comparing two imputation methods on the heartbeat data set. However, CATSI-LSTM yields statistically better accuracy than the NAOMI. An LSTM-based imputation is statistically superior to the GP-VAE, MICE, NIM, and CNoMV methods. The DETROIT method is statistically superior to the MICE and NIM methods.

5. Discussion

This paper surveys and benchmarks state-of-the-art deep learning methods for imputing missing values in time-series health data. The findings of this work can be summarized as follows. First, no single imputation method is the best for all types of health data. Hybrid imputation methods (e.g., CATSI, BRITS, DETROIT) may be preferred for EHR data (e.g., DACMI, sepsis, and hypotension data sets). Longitudinal imputation methods such as CATSI-LSTM or simple piece-wise time interpolation may be preferred for weakly correlated signal data like the IEEEPPG data set with a limited number of variables. Similar methods are also promising for strongly correlated time series signals (heartbeat data set). Second, initializing missing values with a simple piece-wise time interpolation has proven to be the best option in four of five data sets (Table 4). Specifically, piece-wise time initialization shows superiority in initializing missing values on time series signal data. Third, the CATSI imputation method is superior for time series data with a length of 15 and more (Table 5). For data with fewer than 10 time points, a cross-sectional imputation method (e.g., CATSI-MLP) may be more effective. Fourth, no single method achieves the best imputation accuracy for all variables within a specific data set. The variables highly correlated with other variables are generally better imputed by a cross-sectional imputation method than by a hybrid or longitudinal method. However, hybrid imputation methods generally show the best accuracy for imputing individual variables. Furthermore, individual variable imputation accuracy can depend on the data domain. For example, longitudinal imputation methods (e.g., CATSI-LSTM, LSTM) usually perform the best on individual variables of time series signal data (Table 8).

Fifth, the imputation performance on randomly missing values (MAR, MCAR) and non-randomly missing value types (MNAR) is substantially different. The imputation error is substantially higher on the MNAR type (Fig. 5) than on other types (MAR, MCAR) (Figs. 3, 4). This observation aligns with the previous observation that data with not-at-random missing values are more challenging to impute correctly [15]. This finding is important because missing values in EHR data do not appear randomly as the MCAR type. Therefore, most state-of-the-art imputation methods benchmarked on data with MCAR type may not demonstrate equally superior accuracy on real-world data. Sixth, the BRITS method yields the best imputation accuracy for the MNAR missing data type across varying missing rates. For the MCAR and MAR data types, BRITS also performs well on high missing rates (>50%) on several data sets (DACMI, sepsis, hypotension). The TAME method performs better at high missing rates (>50%) than at lower. In contrast, the CATSI method performs much better at lower missing rates (<50%) than higher ones. Therefore, the superiority of state-of-the-art time series imputation methods, often evaluated on low missing rates, may not work equally well on real-world data with high missing rates.

Finally, the effect of imputation on downstream patient classification tasks is mixed. Three hybrid deep imputation methods (CATSI, TAME, and DETROIT) yield statistically better classification results on EHR data sets (sepsis, hypotension) than other baseline methods, including MICE and no imputation strategies. Similar statistical significance is not observed for heartbeat sensor signal data. However, sensor signal data consistently reveal an unusual observation that imputed data sets can yield better classification accuracy than those without missing values. We attribute this observation to redundant frequency and noise components in sensor data that may be filtered out due to the removal and imputation of signal values. Therefore, the data removal and imputation process may improve the quality of sensor signals for downstream classification tasks.

Table 12

Execution times (in minutes) for training and testing of 13 imputation methods on five time series health data sets (DACMI, Sepsis, Hypotension, IEEEPPG, Heartbeat).

Dataset	BRITS	CATSI	CATSI-LSTM	LSTM	MLP	NAOMI	TAME	DETROIT	GP-VAE	MICE	CE-GBT-I	CE-GBT-II	PW
DACMI	190.5	79.7	73.3	54.6	17.6	28.9	36.4	12.5	66.4	1.3	15.3	207.9	6.1
Sepsis	22.6	8.6	8.8	6.9	3.8	4.6	9.1	9.6	1.8	0.4	48.7	681.2	3.7
Hypotension	64.5	25.9	24.8	19.2	6.9	10.8	13.8	9.5	0.9	0.6	15.9	291.8	4.8
IEEEPPG	51.9	21.4	19.7	15.1	5.1	10.1	8.2	14.8	3.2	0.3	2.1	36.6	2.1
Heartbeat	9.1	4.2	3.4	3.0	1.7	4.8	377.1	4.2	2.4	0.5	65.3	1321.5	3.3

Table A.1

Effects of missing value initialization methods on individual imputation models and data sets using average normalized root mean squared deviation scores (NRMSD) and average rank across all variables.

Dataset		DACMI		Sepsis		Hypotension		IEEEPPG		Heartbeat		Avg. perf. across all data sets	
Method	Initialization	NRMSD	Avg. rank	NRMSD	Avg. rank	NRMSD	Avg. rank	NRMSD	Avg. rank	NRMSD	Avg. rank	NRMSD	Rank
BRITS	Delta	0.266	12.2	0.190	5.9	0.163	7.0	0.094	11.6	0.199	16.4	0.182	12
	Mean	0.267	11.5	0.190	5.8	0.163	7.4	0.094	12.2	0.195	16.2	0.182	9
	Median	0.267	12.6	0.190	6.0	0.163	7.6	0.094	11.8	0.195	15.9	0.182	10
	Piece-wise	0.270	12.7	0.190	5.3	0.163	8.4	0.094	12.0	0.194	15.4	0.182	11
CATSI	Delta	0.206	3.7	0.202	6.1	0.158	5.4	0.064	5.6	0.140	4.3	0.154	2
	Mean	0.207	5.1	0.202	6.3	0.158	5.1	0.065	7.2	0.140	4.3	0.154	3
	Median	0.207	5.9	0.203	6.9	0.158	5.1	0.065	6.6	0.141	4.9	0.155	4
	Piece-wise	0.207	5.9	0.203	6.5	0.158	4.1	0.057	2.8	0.139	3.9	0.153	1
CATSI-LSTM	Delta	0.223	8.6	0.226	13.5	0.174	11.6	0.063	4.8	0.157	10.5	0.169	6
	Mean	0.223	8.5	0.227	13.7	0.173	10.0	0.068	10.2	0.159	11.7	0.170	8
	Median	0.224	9.2	0.227	13.7	0.174	11.3	0.064	6.6	0.157	10.2	0.169	7
	Piece-wise	0.223	9.0	0.226	13.0	0.173	10.9	0.053	1.2	0.156	9.8	0.166	5
LSTM	Delta	0.298	17.6	0.292	24.9	0.224	21.7	0.289	18.2	0.191	16.6	0.259	18
	Mean	0.310	19.4	0.282	22.8	0.224	21.1	0.317	24.4	0.196	17.9	0.266	19
	Median	0.323	21.7	0.288	24.3	0.228	24.1	0.319	25.0	0.194	17.1	0.270	20
	Piece-wise	0.241	12.5	0.363	26.7	0.239	24.9	0.080	10.4	0.173	12.2	0.219	13
CATSI-MLP	Delta	0.268	15.1	0.233	15.4	0.204	19.0	0.295	20.6	0.164	12.6	0.233	14
	Mean	0.268	15.5	0.234	15.7	0.204	18.1	0.295	19.8	0.164	12.4	0.233	15
	Median	0.268	15.6	0.234	15.8	0.204	18.1	0.295	20.8	0.164	12.6	0.233	16
	Piece-wise	0.265	13.6	0.238	17.3	0.205	19.4	0.294	19.2	0.164	12.7	0.233	17
NAOMI	Delta	0.688	25.2	0.242	17.4	0.218	18.8	0.259	15.0	0.245	23.1	0.330	23
	Mean	0.617	25.4	0.242	17.5	0.218	18.7	0.273	16.2	0.247	23.4	0.319	22
	Median	0.437	22.9	0.240	17.2	0.220	19.4	0.285	18.2	0.251	23.7	0.287	21
	Piece-wise	0.824	26.0	0.270	21.9	0.228	23.1	0.054	2.0	0.288	26.1	0.333	24
MICE	Delta	0.730	22.2	0.252	17.2	0.337	22.1	0.901	26.9	0.332	24.4	0.510	26
	Mean	0.733	23.8	0.252	17.2	0.337	22.1	0.901	26.6	0.332	24.5	0.511	28
	Median	0.730	23.3	0.252	16.9	0.337	22.1	0.901	27.4	0.332	25.0	0.510	27
	Piece-wise	0.730	21.9	0.252	16.9	0.337	21.0	0.901	25.1	0.332	24.1	0.510	25

5.1. Model performance and time cost trade-off

Table 12 compares the execution time for 13 imputation strategies on the five data sets. Indeed, deep learning methods are more time-consuming than MICE and other conventional baselines used in the health science literature. However, top-quality data should be a high priority to ensure reliable data-driven model outcomes, especially when cloud or local server-based high-performance computing is a reality nowadays. We outline a general performance and time-cost trade-off strategy for choosing a deep imputation method when computation time becomes a critical consideration.

The time cost of a data imputation algorithm can depend on the sample size (n), number of variables (d), number of time steps (T), or any combinations of n, d, and T parameters. We choose the five best deep imputation methods (TAME, BRITS, CATSI, CATSI-LSTM, DETROIT) with statistically similar performances (Tables 10, 11) for the performance-time cost trade-off analysis. Table 2 (summary of data statistics), in conjunction with Table 12 may explain the trade-offs based on n, d, and T parameters. The effect of sample size is evident when comparing the imputation methods on two sensor data sets (IEEEPPG and Heartbeat). The IEEEPPG data set has almost eight times more samples than the heartbeat data set, requiring substantially more execution time accordingly. Exceptions are GBT and TAME models, which are more sensitive to the number of variables (d) instead. The BRITS method is sensitive to the number of time steps (T) and

sample size (n). However, a smaller sample size can significantly reduce execution time despite having many time steps (Heartbeat). Similarly, the hypotension data set with more than double time steps requires more computation than the sepsis data set. However, DETROIT appears insensitive to differences in time steps and the number of variables. GBT and VAE methods are more sensitive to the number of variables, requiring more execution time on the sepsis data set. Overall, the DETROIT model is generally preferred considering performance and time cost trade-offs. However, when the sample size is small with many variables (heartbeat), the preferred method is CATSI-LSTM. The TAME method is preferred when the number of variables is small with moderate samples (IEEEPPG data set).

5.2. Recommendations for health scientists

Health scientists use conventional methods to handle missing values in healthcare data, including no imputation, complete case analysis by removing samples with missing values, and imputing values with mean or median. MICE-based imputation methods are arguably the most popular approach in biostatistics and appear in more recent clinical studies on healthcare data [64]. In contrast, health science often avoids superior deep learning methods due to high computational costs, availability of such methods in easy-to-use software packages (e.g., SAS) or languages (e.g., R), and poor interpretability of complex learning methods. Below, we summarize some general recommendations for selecting missing value imputation methods.

- The data quality following imputation is critically important for the reliability of data-driven outcomes. Therefore, imputation methods should be considered instead of choosing an arbitrary strategy to fill in the missing values.
- A major benefit of deep learning is incremental learning on future data, unlike statistical models (MICE). When new data are available in the future, statistical learning methods would require retraining on the entire data set (past and future data) from scratch.
- We recommend high-performance computing, including cloud-based or local GPU servers, to effectively reap the fruits of deep learning. Therefore, the selection of an imputation method should be based on missing rates in data or missing types (MCAR versus MNAR) instead of execution time to ensure the highest quality data.
- Health data most commonly appear as multivariate time series, whereas MICE methods model relationships between variables, ignoring essential temporal relationships in time series. Therefore, the most basic imputation strategy should involve piece-wise time interpolation followed by MICE-based cross-sectional imputation.
- Hybrid methods that learn cross-sectional and longitudinal relationships in data are preferred to ensure the highest quality data. When sufficient computing resources are available, we recommend the CATSI or BRITS methods because the TAME and DETROIT methods require ground truth values for model training, which may not be available in practice.
- Hybrid deep imputation methods generally show imputation and classification accuracy statistically superior to conventional baselines. However, the improvement may not appear statistically significant in all data scenarios. Despite statistical insignificance, a 1% relative improvement in classification accuracy infers correctly prognosticating 100 additional patients in 10,000 samples, which is valuable in clinical practice. Therefore, statistical tests should not always dissuade researchers from adopting advanced imputation methods to ensure the highest data quality.
- Implementing open-source deep imputation methods in Python remains a challenge for researchers trained in R programming and statistical software packages. The productive transfer of knowledge and tools is an open problem and would entail interdisciplinary efforts and research support.

6. Conclusions

This study presents one of the first surveys to benchmark state-of-the-art deep learning methods for imputing missing values in time series data. Our six data-centric experiments compare the effectiveness of imputation methods across different data types, variable statistics, time series length, missing value rates and types, and in downstream classification tasks. The experimental results have identified several important insights into method selection for imputing time series data and shared recommendations for health scientists. Therefore, we emphasize the importance of considering data-centric factors over an arbitrary selection of an imputation strategy to ensure data quality and the reliability of data-driven outcomes. In the future, the interactions between the six experimental conditions may be an important direction of study for optimizing the data-centric selection of imputation methods.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
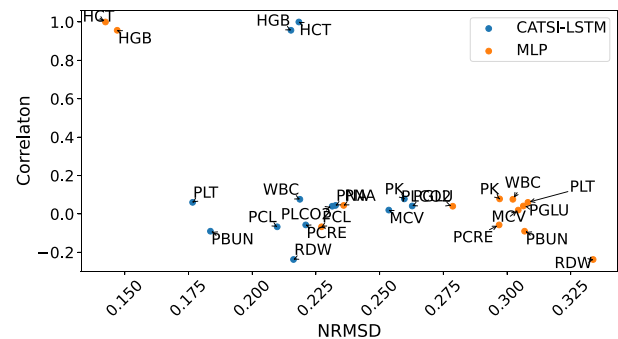
(a) DACMI - HCT variable
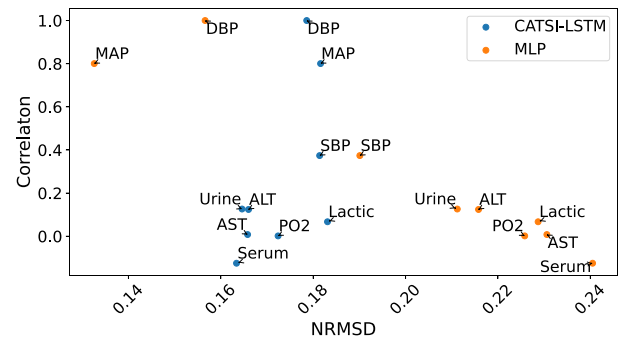
(b) Hypotension - DBP variable

Fig. B.1. (a) Correlation between HCT and other variables in the DACMI data set versus the NRMSD scores for individual variables, (b) the same plot for the DBP variable in the Hypotension data set.

Acknowledgments

Appendix A. Effects of missing value initialization methods

See Table A.1.

Appendix B. Correlation between variables versus imputation accuracy (NRMSD scores) to demonstrate the selection of cross-sectional (MLP) and longitudinal (CATSI-LSTM) imputation methods for individual variables.

See Fig. B.1.

References

[1] S. Van Buuren, Flexible Imputation of Missing Data, CRC Press, 2018.
[2] F. Biessmann, D. Salinas, S. Schelter, P. Schmidt, D. Lange, "Deep" learning for missing value imputation in tables with non-numerical data, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 2017–2025, http://dx.doi.org/10.1145/3269206.3272005.
[3] H. Zhang, P. Xie, E. Xing, Missing value imputation based on deep generative models, 2018, arXiv preprint arXiv:1808.01684.
[4] M. Sangeetha, M.S. Kumaran, Deep learning-based data imputation on time-variant data using recurrent neural network, Soft Comput. 24 (2020) 13369–13380, http://dx.doi.org/10.1007/s00500-020-04755-5.
[5] G. Madhu, B. Bharadwaj, G. Nagachandrika, K. Vardhan, A novel algorithm for missing data imputation on machine learning, in: 2019 International Conference on Smart Systems and Inventive Technology, ICSSIT, 2019, pp. 173–177, http://dx.doi.org/10.1109/ICSSIT46314.2019.8987895.

[6] J. Bauer, O. Angelini, A. Denev, Imputation of multivariate time series data - Performance benchmarks for multiple imputation and spectral techniques, Article SSRN Electron. J. XXI (2017) 1–5, http://dx.doi.org/10.2139/ssrn.2996611.

[7] S.N. Shukla, B.M. Marlin, A survey on principles, models and methods for learning from irregularly sampled time series, 2020, http://dx.doi.org/10.48550/arxiv.2012.00168, arXiv.

[8] S. Festag, J. Denzler, C. Spreckelsen, Generative adversarial networks for biomedical time series forecasting and imputation, J. Biomed. Inform. 129 (2022) 104058, http://dx.doi.org/10.1016/j.jbi.2022.104058.

[9] C. Sun, S. Hong, M. Song, H. Li, A review of deep learning methods for irregularly sampled medical time series data, 2020, http://dx.doi.org/10.48550/arxiv.2010.12493, arXiv.

[10] M. Khayati, A. Lerner, Z. Tymchenko, P. Cudré-Mauroux, Mind the gap: An experimental evaluation of imputation of missing values techniques in time series, Proc. VLDB Endow. 13 (5) (2020) 768–782, http://dx.doi.org/10.14778/3377369.3377383.

[11] C. Fang, C. Wang, Time series data imputation: A survey on deep learning approaches, 2020, http://dx.doi.org/10.48550/arxiv.2011.11347, arXiv.

[12] M. Resche-Rigon, I.R. White, Multiple imputation by chained equations for systematically and sporadically missing multilevel data, Stat. Methods Med. Res. 27 (6) (2018) 1634–1649, http://dx.doi.org/10.1177/0962280216666564.

[13] Y. Luo, P. Szolovits, A.S. Dighe, J.M. Baron, 3D-MICE: Integration of cross-sectional and longitudinal imputation for multi-analyte longitudinal clinical data, J. Am. Med. Inf. Assoc. 25 (6) (2017) 645–653, http://dx.doi.org/10.1093/jamia/ocx133.

[14] V. Nassiri, G. Molenberghs, G. Verbeke, J. Barbosa-Breda, Iterative multiple imputation: A framework to determine the number of imputed datasets, Amer. Statist. 74 (2) (2020) 125–136, http://dx.doi.org/10.1080/00031305.2018.1543615.

[15] M.D. Samad, S. Abrar, N. Diawara, Missing value estimation using clustering and deep learning within multiple imputation framework, Knowl.-Based Syst. 249 (2022) 108968, http://dx.doi.org/10.1016/j.knosys.2022.108968.

[16] S.J. Choudhury, N.R. Pal, Imputation of missing data with neural networks for classification, Knowl.-Based Syst. 182 (2019) 104838.

[17] R.D. Camino, C.A. Hammerschmidt, R. State, Improving missing data imputation with deep generative models, 2019, http://dx.doi.org/10.48550/ARXIV.1902.10666, arXiv.

[18] Y. Zhuang, R. Ke, Y. Wang, Innovative method for traffic data imputation based on convolutional neural network, IET Intell. Transp. Syst. 13 (4) (2019) 605–613.

[19] S. Nikfalazar, C.H. Yeh, S. Bedingfield, H.A. Khorshidi, Missing data imputation using decision trees and fuzzy clustering with iterative learning, Knowl. Inf. Syst. 62 (2020) 2419–2437, http://dx.doi.org/10.1007/s10115-019-01427-1.

[20] A. Gonzalez-Vidal, P. Rathore, A.S. Rao, J. Mendoza-Bernal, M. Palaniswami, A.F. Skarmeta-Gomez, Missing data imputation with Bayesian maximum entropy for internet of things applications, IEEE Internet Things J. 4662 (2020) 1, http://dx.doi.org/10.1109/jiot.2020.2987979.

[21] H. Hegde, N. Shimpi, A. Panny, I. Glurich, P. Christie, A. Acharya, MICE vs PPCA: Missing data imputation in healthcare, Inf. Med. Unlocked 17 (2019) 100275, http://dx.doi.org/10.1016/j.imu.2019.100275.

[22] P. Sun, MICE-DA: A MICE method with data augmentation for missing data imputation in IEEE ichi 2019 DACMI challenge, in: 2019 IEEE International Conference on Healthcare Informatics, ICHI 2019, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 1–3, http://dx.doi.org/10.1109/ichi.2019.8904724.

[23] X. Xu, X. Liu, Y. Kang, X. Xu, J. Wang, Y. Sun, Q. Chen, X. Jia, X. Ma, X. Meng, X. Li, G. Xie, A multi-directional approach for missing value estimation in multivariate time series clinical data, J. Healthc. Inf. Res. 4 (2020) 365–382, http://dx.doi.org/10.1007/s41666-020-00076-2.

[24] X. Zhang, C. Yan, C. Gao, B.A. Malin, Y. Chen, Predicting missing values in medical data via XGBoost regression, J. Healthc. Inf. Res. 4 (4) (2020) 383–394.

[25] G. Gao, F. Khoshnevisan, M. Chi, Reconstructing missing EHRs using time-aware within- and cross-visit information for septic shock early prediction, in: 2022 IEEE 10th International Conference on Healthcare Informatics, ICHI, 2022, pp. 151–162, http://dx.doi.org/10.1109/ichi54592.2022.00034.

[26] C. Yin, R. Liu, D. Zhang, P. Zhang, Identifying sepsis subphenotypes via time-aware multi-modal auto-encoder, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 862–872.

[27] Y. Luo, Evaluating the state of the art in missing data imputation for clinical data, Brief. Bioinform. 23 (2022) http://dx.doi.org/10.1093/BIB/BBAB489.

[28] K. Yin, L. Feng, W.K. Cheung, Context-aware time series imputation for multi-analyte clinical data, J. Healthc. Inf. Res. 4 (2020) 411–426, http://dx.doi.org/10.1007/s41666-020-00075-3.

[29] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, Y. Li, BRITS: Bidirectional recurrent imputation for time series, Adv. Neural Inf. Process. Syst. 31 (2018).

[30] Z. Che, S. Purushotham, K. Cho, D. Sontag, Y. Liu, Recurrent neural networks for multivariate time series with missing values, Sci. Rep. 8 (1) (2018) 1–12.

[31] J. Yoon, W.R. Zame, M. van der Schaar, Multi-directional recurrent neural networks: A novel method for estimating missing data, in: Time Series Workshop in International Conference on Machine Learning, 2017.

[32] X. Yi, Y. Zheng, J. Zhang, T. Li, ST-MVL: Filling missing values in geo-sensory time series data, in: Proceedings of the 25th International Joint Conference on Artificial Intelligence, 2016.

[33] S. Moritz, T. Bartz-Beielstein, ImputeTS: Time series missing value imputation in R, R J. 9 (1) (2017) 207.

[34] Y. Liu, R. Yu, S. Zheng, E. Zhan, Y. Yue, NAOMI: Non-autoregressive multiresolution sequence imputation, Adv. Neural Inf. Process. Syst. 32 (2019) 1–11.

[35] Y. Luo, X. Cai, Y. Zhang, J. Xu, et al., Multivariate time series imputation with generative adversarial networks, Adv. Neural Inf. Process. Syst. 31 (2018).

[36] W. Fedus, I. Goodfellow, A.M. Dai, Maskgan: Better text generation via filling in the, in: International Conference on Learning Representations, 2018, https://openreview.nt/forum?id=ByOExmWAb.

[37] Y. Zhang, B. Zhou, X. Cai, W. Guo, X. Ding, X. Yuan, Missing value imputation in multivariate time series with end-to-end generative adversarial networks, Inform. Sci. 551 (2021) 67–82, http://dx.doi.org/10.1016/j.ins.2020.11.035.

[38] Z. Guo, Y. Wan, H. Ye, A data imputation method for multivariate time series based on generative adversarial network, Neurocomputing 360 (2019) 185–197, http://dx.doi.org/10.1016/j.neucom.2019.06.007.

[39] C. Yan, C. Gao, X. Zhang, Y. Chen, B. Malin, Deep imputation of temporal data, in: 2019 IEEE International Conference on Healthcare Informatics, ICHI, 2019, pp. 1–3, http://dx.doi.org/10.1109/ichi.2019.8904776.

[40] V. Fortuin, D. Baranchuk, G. Raetsch, S. Mandt, GP-VAE: Deep probabilistic time series imputation, in: Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, in: Proceedings of Machine Learning Research, vol. 108, PMLR, 2020, pp. 1651–1661.

[41] R. Mazumder, T. Hastie, R. Tibshirani, Spectral regularization algorithms for learning large incomplete matrices, J. Mach. Learn. Res. 11 (2010) 2287–2322.

[42] B.K. Beaulieu-Jones, J.H. Moore, Missing data imputation in the electronic health record using deeply learned autoencoders, in: Pacific Symposium on Biocomputing 2017, World Scientific, 2017, pp. 207–218.

[43] E. Zhang, R. Robinson, B. Pfahringer, Deep holistic representation learning from EHR, in: 2018 12th International Symposium on Medical Information and Communication Technology, ISMICT, 2018, pp. 1–6, http://dx.doi.org/10.1109/ISMICT.2018.8573698.

[44] M.A. Makary, M. Daniel, Medical error-the third leading cause of death in the US, BMJ (Clin. Res. Ed.) 353 (2016) i2139, http://dx.doi.org/10.1136/bmj.i2139.

[45] A.E. Ulloa-Cerna, L. Jing, J.M. Pfeifer, S. Raghunath, J.A. Ruhl, D.B. Rocha, J.B. Leader, N. Zimmerman, G. Lee, S.R. Steinhubl, et al., RECHOmmend: An ECG-based machine learning approach for identifying patients at increased risk of undiagnosed structural heart disease detectable by echocardiography, Circulation 146 (1) (2022) 36–47.

[46] R.E. Sherman, S.A. Anderson, G.J.D. Pan, G.W. Gray, T. Gross, N.L. Hunter, L. LaVange, D. Marinac-Dabic, P.W. Marks, M.A. Robb, J. Shuren, R. Temple, J. Woodcock, L.Q. Yue, R.M. Califf, Real-world evidence — What is it and what can it tell us? N. Engl. J. Med. 375 (2016) 2293–2297, http://dx.doi.org/10.1056/NEJMsb1609216.

[47] L. Jing, G.J. Wehner, J.D. Suever, R.J. Charnigo, S. Alhadad, E. Stearns, D. Mojsejenko, C.M. Haggerty, K. Hickey, A.M. Valente, et al., Left and right ventricular dyssynchrony and strains from cardiovascular magnetic resonance feature tracking do not predict deterioration of ventricular function in patients with repaired tetralogy of Fallot, J. Cardiovasc. Magn. Reson. 18 (1) (2016) 1–10, http://dx.doi.org/10.1186/s12968-016-0268-8.

[48] M.D. Samad, G.J. Wehner, M.R. Arbabshirani, L. Jing, A.J. Powell, T. Geva, C.M. Haggerty, B.K. Fornwalt, Predicting deterioration of ventricular function in patients with repaired tetralogy of Fallot using machine learning, Eur. Heart J. - Cardiovasc. Imaging 19 (7) (2018) 730–738, http://dx.doi.org/10.1093/ehjci/jey003.

[49] T. Gentimis, A.J. Alnaser, A. Durante, K. Cook, R. Steele, Predicting hospital length of stay using neural networks on MIMIC III data, in: 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress, DASC/PiCom/DataCom/CyberSciTech, 2017, pp. 1194–1201, http://dx.doi.org/10.1109/DASC-PICom-DataCom-CyberSciTec.2017.191.

[50] Z. Nowroozilarki, A. Pakbin, J. Royalty, D.K. Lee, B.J. Mortazavi, Real-time mortality prediction using MIMIC-IV ICU data via boosted nonparametric hazards, in: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI, 2021, pp. 1–4, http://dx.doi.org/10.1109/BHI50953.2021.9508537.

[51] M. Scherpf, F. Gräßer, H. Malberg, S. Zaunseder, Predicting sepsis with a recurrent neural network using the MIMIC III database, Comput. Biol. Med. 113 (2019) 103395, http://dx.doi.org/10.1016/j.compbiomed.2019.103395.

[52] S. Zhou, Z. Zeng, H. Wei, T. Sha, S. An, Early combination of albumin with crystalloids administration might be beneficial for the survival of septic patients: A retrospective analysis from MIMIC-IV database, Ann. Intens. Care 11 (2021) 1–10, http://dx.doi.org/10.1186/s13613-021-00830-8.

[53] J.L. Vincent, N.D. Nielsen, N.I. Shapiro, M.E. Gerbasi, A. Grossman, R. Doroff, F. Zeng, P.J. Young, J.A. Russell, Mean arterial pressure and mortality in patients with distributive shock: A retrospective analysis of the MIMIC-III database, Ann. Intens. Care 8 (2018) 1–10, http://dx.doi.org/10.1186/s13613-018-0448-9.

[54] L. Wang, W. Zhou, Y. Xing, X. Zhou, A novel neural network model for blood pressure estimation using photoplethesmography without electrocardiogram, J. Healthc. Eng. 2018 (2018) http://dx.doi.org/10.1155/2018/7804243.

[55] X. Xing, M. Sun, Optical blood pressure estimation with photoplethysmography and FFT-based neural networks, Biomed. Opt. Express 7 (8) (2016) 3007–3020, http://dx.doi.org/10.1364/BOE.7.003007.

[56] P.H. Charlton, D.A. Birrenkott, T. Bonnici, M.A. Pimentel, A.E. Johnson, J. Alastruey, L. Tarassenko, P.J. Watkinson, R. Beale, D.A. Clifton, Breathing rate estimation from the electrocardiogram and photoplethysmogram: A review, IEEE Rev. Biomed. Eng. 11 (2018) 2–20, http://dx.doi.org/10.1109/RBME.2017.2763681.

[57] A.K. Waljee, A. Mukherjee, A.G. Singal, Y. Zhang, J. Warren, U. Balis, J. Marrero, J. Zhu, P.D. Higgins, Comparison of imputation methods for missing laboratory data in medicine, BMJ Open 3 (8) (2013) http://dx.doi.org/10.1136/bmjopen-2013-002847.

[58] S. Cheng, X. Li, B.M. Marlin, Learning from irregularly-sampled time series: A missing data perspective, in: International Conference on Machine Learning, 2020, pp. 5937–5946.

[59] P.B. Weerakody, K.W. Wong, G. Wang, W. Ela, A review of irregular time series data handling with gated recurrent neural networks, Neurocomputing 441 (2021) 161–178, http://dx.doi.org/10.1016/j.neucom.2021.02.046.

[60] M. Michalopoulos, C. Anagnostopoulos, C. Doukas, I. Maglogiannis, S. Hadjiefthymiades, Optimizing pervasive sensor data acquisition utilizing missing values substitution, in: Proceedings of the 3rd International Conference on Pervasive Technologies Related to Assistive Environments, 2010, pp. 1–8.

[61] G.E. Batista, M.C. Monard, et al., A study of K-nearest neighbour as an imputation method, His 87 (251–260) (2002) 48.

[62] G.M. Weber, W.G. Adams, E.V. Bernstam, J.P. Bickel, K.P. Fox, K. Marsolo, V.A. Raghavan, A. Turchin, X. Zhou, S.N. Murphy, et al., Biases introduced by filtering electronic health records for patients with "complete data", J. Am. Med. Inf. Assoc. 24 (6) (2017) 1134–1141.

[63] B.K. Beaulieu-Jones, D.R. Lavage, J.W. Snyder, J.H. Moore, S.A. Pendergrass, C.R. Bauer, Characterizing and managing missing structured data in electronic health records: Data analysis, JMIR Med. Inf. 6 (2018) http://dx.doi.org/10.2196/medinform.8960.

[64] M.D. Samad, A. Ulloa, G.J. Wehner, L. Jing, D. Hartzel, C.W. Good, B.A. Williams, C.M. Haggerty, B.K. Fornwalt, Predicting survival from large echocardiography and electronic health record datasets, JACC: Cardiovasc. Imaging 12 (2018) 681–689, http://dx.doi.org/10.1016/j.jcmg.2018.04.026.

[65] A.E. Johnson, T.J. Pollard, L. Shen, L.-w.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data 3 (1) (2016) 1–9.

[66] N.I. Kuo, M. Polizzotto, S. Finfer, L. Jorm, S. Barbieri, et al., Synthetic acute hypotension and sepsis datasets based on MIMIC-III and published as part of the health gym project, 2021, arXiv preprint arXiv:2112.03914.

[67] N.I.-H. Kuo, M.N. Polizzotto, S. Finfer, F. Garcia, A. Sönnerborg, M. Zazzi, M. Böhm, R. Kaiser, L. Jorm, S. Barbieri, The health Gym: Synthetic health-related datasets for the development of reinforcement learning algorithms, Sci. Data 9 (1) (2022) 693.

[68] Z. Zhang, Z. Pi, B. Liu, TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise, IEEE Trans. Biomed. Eng. 62 (2) (2014) 522–531.

[69] C.W. Tan, C. Bergmeir, F. Petitjean, G.I. Webb, Monash university, UEA, UCR time series extrinsic regression archive, 2020, arXiv preprint arXiv:2006.10996.

[70] C. Liu, D. Springer, Q. Li, B. Moody, R.A. Juan, F.J. Chorro, F. Castells, J.M. Roig, I. Silva, A.E.W. Johnson, Z. Syed, S.E. Schmidt, C.D. Papadaniil, L. Hadjileontiadis, H. Naseri, A. Moukadem, A. Dieterlen, C. Brandt, H. Tang, M. Samieinasab, M.R. Samieinasab, R. Sameni, R.G. Mark, G.D. Clifford, An open access database for the evaluation of heart sound algorithms, Physiol. Meas. 37 (12) (2016) 2181, http://dx.doi.org/10.1088/0967-3334/37/12/2181.

[71] A.L. Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000) e215–e220.

[72] A. Bagnall, H.A. Dau, J. Lines, M. Flynn, J. Large, A. Bostrom, P. Southam, E. Keogh, The UEA multivariate time series classification archive, 2018, 2018, arXiv preprint arXiv:1811.00075.

[73] M.D. Samad, L. Yin, Non-linear regression models for imputing longitudinal missing data, in: 2019 IEEE International Conference on Healthcare Informatics, ICHI, 2019, pp. 1–3, http://dx.doi.org/10.1109/ichi.2019.8904528.

[74] S.W. Scheff, Fundamental Statistical Principles for the Neurobiologist: A Survival Guide, Elsevier, 2016, pp. 1–216, http://dx.doi.org/10.1016/C2015-0-02471-6, URL http://www.sciencedirect.com:5070/book/9780128047538/fundamental-statistical-principles-for-the-neurobiologist.

[75] S. Daberdaku, E. Tavazzi, B. Di Camillo, A combined interpolation and weighted K-nearest neighbours approach for the imputation of longitudinal icu laboratory data, J. Healthc. Inf. Res. 4 (2) (2020) 174–188, http://dx.doi.org/10.1007/S41666-020-00069-1/METRICS, URL https://link.springer.com/article/10.1007/s41666-020-00069-1.

[76] S.C. Nistor, G. Czibula, IntelliSwAS: Optimizing deep neural network architectures using a particle swarm-based approach, Expert Syst. Appl. 187 (2022) 115945, http://dx.doi.org/10.1016/j.eswa.2021.115945, URL https://www.sciencedirect.com/science/article/pii/S0957417421012987.

[77] N. Razavian, D. Sontag, Temporal convolutional neural networks for diagnosis from lab tests, 2015, arXiv preprint arXiv:1511.07938.

[78] N. Tomašev, X. Glorot, J.W. Rae, M. Zielinski, H. Askham, A. Saraiva, A. Mottram, C. Meyer, S. Ravuri, I. Protsyuk, et al., A clinically applicable approach to continuous prediction of future acute kidney injury, Nature 572 (7767) (2019) 116–119.