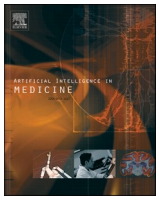
Artificial Intelligence In Medicine 142 (2023) 102587

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed



Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques

Mingxuan Liu [a,1], Siqi Li [a,1], Han Yuan [a], Marcus Eng Hock Ong [b,c], Yilin Ning [a], Feng Xie [a,b], Seyed Ehsan Saffari [a,b], Yuqing Shang [a], Victor Volovici [d], Bibhas Chakraborty [a,b,e,f], Nan Liu [a,b,g,h,*]

[a] Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore
[b] Programme in Health Services and Systems Research, Duke-NUS Medical School, Singapore
[c] Department of Emergency Medicine, Singapore General Hospital, Singapore
[d] Department of Neurosurgery, Erasmus MC University Medical Center, Rotterdam, the Netherlands
[e] Department of Statistics and Data Science, National University of Singapore, Singapore
[f] Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA
[g] SingHealth AI Office, Singapore Health Services, Singapore
[h] Institute of Data Science, National University of Singapore, Singapore

ARTICLE INFO

Keywords:
Missing value
Imputation
Deep learning
Neural networks
Healthcare

ABSTRACT

Objective: The proper handling of missing values is critical to delivering reliable estimates and decisions, especially in high-stakes fields such as clinical research. In response to the increasing diversity and complexity of data, many researchers have developed deep learning (DL)-based imputation techniques. We conducted a systematic review to evaluate the use of these techniques, with a particular focus on the types of data, intending to assist healthcare researchers from various disciplines in dealing with missing data.

Materials and methods: We searched five databases (MEDLINE, Web of Science, Embase, CINAHL, and Scopus) for articles published prior to February 8, 2023 that described the use of DL-based models for imputation. We examined selected articles from four perspectives: data types, model backbones (i.e., main architectures), imputation strategies, and comparisons with non-DL-based methods. Based on data types, we created an evidence map to illustrate the adoption of DL models.

Results: Out of 1822 articles, a total of 111 were included, of which tabular static data (29 %, 32/111) and temporal data (40 %, 44/111) were the most frequently investigated. Our findings revealed a discernible pattern in the choice of model backbones and data types, for example, the dominance of autoencoder and recurrent neural networks for tabular temporal data. The discrepancy in imputation strategy usage among data types was also observed. The "integrated" imputation strategy, which solves the imputation task simultaneously with downstream tasks, was most popular for tabular temporal data (52 %, 23/44) and multi-modal data (56 %, 5/9). Moreover, DL-based imputation methods yielded a higher level of imputation accuracy than non-DL methods in most studies.

Conclusion: The DL-based imputation models are a family of techniques, with diverse network structures. Their designation in healthcare is usually tailored to data types with different characteristics. Although DL-based imputation models may not be superior to conventional approaches across all datasets, it is highly possible for them to achieve satisfactory results for a particular data type or dataset. There are, however, still issues with regard to portability, interpretability, and fairness associated with current DL-based imputation models.

Abbreviations: DL, Deep Learning; AE, Auto Encoder; DAE, Deep Auto Encoder; GAN, Generative Adversarial Network; LSTM, Long Short-term Memory; MLP, Multilayer Perceptron; RNN, Recurrent Neural Networks.

* Corresponding author at: Centre for Quantitative Medicine, Duke-NUS Medical School, 8 College Road, Singapore 169857, Singapore.
E-mail address: liu.nan@duke-nus.edu.sg (N. Liu).

[1] These authors contributed equally.

https://doi.org/10.1016/j.artmed.2023.102587
Received 17 October 2022; Received in revised form 8 April 2023; Accepted 16 May 2023
Available online 22 May 2023

1. Introduction

Healthcare data has emerged in diverse formats in the era of big data. Personalized health monitoring devices, for instance, enable the collection of data tailored to an individual's daily activities. Likewise, rapidly evolving laboratory techniques generate vast amounts of sequencing data. However, these new data formats are more susceptible to the problem of missing values than traditional tabular clinical data collected from prospective observational or randomized trials.

Missing values cast a shadow on data analysis: they can reduce prediction power and result in bias in downstream decision-making [1,2], which is particularly problematic in high-fidelity decision-making situations, such as those in healthcare. Complete data analysis or simple imputation (mean, median, or mode) may resolve missingness for tabular static data, but such strategies may not be adequate for a variety of data types and architectures, ranging from static to temporal, tabular to imaging and sequencing data. Therefore, advanced approaches are necessary to ensure the quality and robustness of models.

As described by Little and Rubin [3,4], missing data can be categorized into three types: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Prior to the widespread adoption of deep learning (DL), traditional statistical and machine learning approaches, such as interpolation methods, k-nearest neighbor ($k-$NN) [5], multiple imputation by chained equations (MICE) [6], and random forest (RF)-based models like MissForest [7], have been used to impute missing values. However, these methods may be restricted to certain types of missing data; for example, MICE generally assumes that the missing data type is MAR [6]. When applied to complex healthcare data, these non-DL-based imputation strategies may exhibit low accuracy [8,9], especially when the mechanism of missingness cannot be determined.

In recent years, DL-based methods have increasingly been used to solve missing value problems and shown to enhance imputation accuracy [10,11]. As well, DL-based models can be customized to handle complex missing patterns and data structures, such as time-series data with unique sequential structures and image data with spatial patterns [12,13]. Due to the superior performance and designation flexibility, DL-based imputation models have gained popularity in a wide range of applications, such as in-patient mortality prediction [14,15] and early detection of Alzheimer's Disease (AD) [12,16].

In spite of the presence of several existing reviews on missing value imputation, most of them either focus on non-DL-based methods [17–19], or treat the neural network as a single type of method [20–24]. Due to the lack of specificity, these articles cannot adequately assist prospective researchers contemplating the application of DL-based models to their own data. To our knowledge, no systematic review has been conducted regarding DL-based missing value imputation methods for diverse types of healthcare data. Toward bridging this gap, we present an evidence map analysis [25] that examines model-use by data type and provide guidance for researchers using DL-based methodologies to manage missing values.

2. Materials and methods

2.1. Search strategies

In this study, we undertook a systematic literature search to identify relevant research articles. We searched five databases (MEDLINE, Web of Science, Embase, CINAHL, and Scopus) using a combination of search phrases "missing value", "imputation", "machine learning", "deep learning", and "healthcare". Detailed search strategies are provided in eTable 1.

2.2. Exclusion criteria

We conducted the study according to the Preferred Reporting Items for Systematic Reviews (PRISMA) guidelines [26]. The following reasons were considered to exclude studies: the study was not in the medical or clinical domain, the imputation model used was not DL-based or was not specified, the study was not published as a research article (e.g., a conference poster, conference abstract, or book chapter), or the article was not written in English.

2.3. Selection procedure and data extraction

Two reviewers (ML and SL) independently screened the titles and abstracts between 6 August and 11 September 2021, and 8 February and 23 February 2023 in accordance with the eligibility criteria. The discrepancies were resolved through discussions with a third reviewer (HY). For full-text screening and information extraction, ML and SL separately accessed the documents between 12 September and 22 October 2021, and 24 February and 7 March 2023. In the event of disagreement, they consulted with HY between these dates. Four aspects of information were gathered from the included articles: data types, model backbones (i.e., main architectures), imputation strategies, and comparisons with non-DL-based methods.

2.4. Data analysis

To generate an evidence map [25] that illustrates the application of DL-based imputation models across various data types, we classified the types of data involved in imputation into six categories: tabular static data, tabular temporal data, genetic and genomic data, image data, signal data, and multi-modal data. While both tabular static and tabular temporal data contain observations as rows and features as columns, only tabular temporal data include the time factor. Genetic and genomic data encompasses both DNA data for organisms and personal genetic information. Image data and signal data refer to the information generated by specific medical devices, such as magnetic resonance imaging (MRI) and electrocardiogram (ECG). Multi-modal data refers to the use of serval types of data in performing a single imputation task.

We then categorized the articles according to the "backbones" of the imputation model: 1) multi-layer perceptron (MLP) [27]; 2) recurrent neural network (RNN) [28], including vanilla RNN, long short-term memory (LSTM), and gated recurrent unit (GRU); 3) the framework of autoencoder (AE) [29] which includes vanilla autoencoder, denoising autoencoder (DAE), and variational autoencoder (VAE); 4) the framework of generative adversarial network (GAN) [30]; 5) a hybrid of the four backbones mentioned above; and 6) other less common models, such as self-organizing map (SOM), graphical network (GNN), convolutional neural network (CNN) and Transformer component. Detailed definitions of these models and their corresponding general imputation mechanisms are provided in Fig. 1. Our assessment of imputation strategies was divided into two categories: separated and integrated. As the name suggests, "separated" means that the imputation process is separated from downstream tasks such as disease classification and risk prediction, whereas "integrated", also known as "end-to-end", refers to the imputation process being undertaken concurrently with downstream tasks.

We presented the evidence map based on the cross-tabulation of model backbones and data types. Additionally, a bar plot was created to illustrate the distribution of imputation strategies. Python version 3.8.3 (Python Software Foundation, Delaware, USA) and R version 4.0.2 (The R Foundation for Statistical Computing) were used for data analysis.

3. Results

Our search of five databases yielded 1822 studies, of which 111 were included for analysis. Fig. 2 illustrates the selection procedure in detail. A summary of the included studies is presented in Table 1. Fig. 3 depicts the evidence map between the "backbones" (i.e., main architectures) of DL-based imputation models and the types of healthcare data. Among

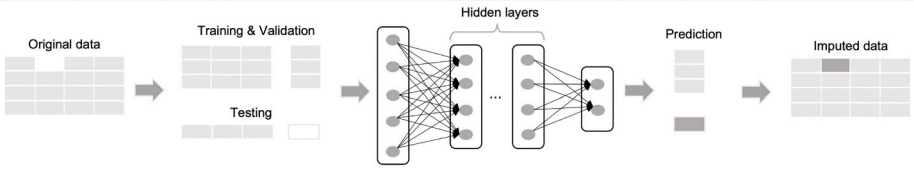
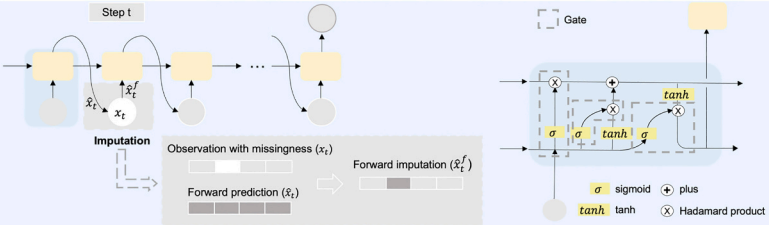
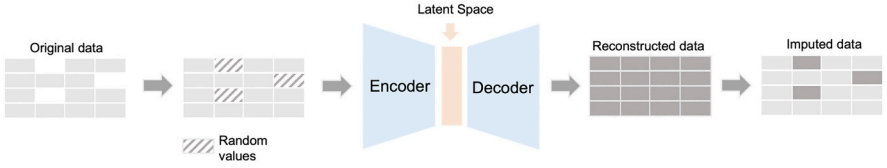
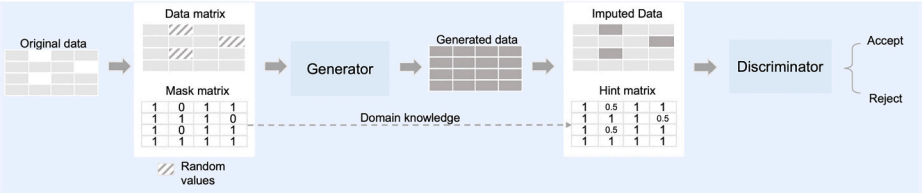
Model	Definition	Imputation Mechanism[1]
Multi-layer perceptron (MLP)	It consists of at least three layers of neurons (input, hidden, and output), each of which is fed a non-linear activation function to capture the patterns.	
Recurrent neural network (RNN)[2]	It uses the hidden states (memory) to process sequences of inputs. At each step t, the input includes the current observation and the previous hidden state.	
Autoencoder (AE)[3]	It is a framework[4] that co-trains two modules: an encoder that maps the input data into a latent embedding, and a decoder that reconstructs the input from the latent embedding.	
Generative adversarial network (GAN)	It is a framework[4] in an adversarial process, co-training two modules: a generator that captures the data distribution, and a discriminator that detects whether a sample came from the training data rather than the generator.	

Fig. 1. Definitions of models and the corresponding imputation mechanisms.

[1]Imputation mechanisms: MLP models can be trained on the complete observations to predict the missing values; RNN models can predict the missing values based on the previous hidden state (forward imputation [69]); Autoencoder can maintain the whole data structure in a good manner and reveal the missing values in its output; GAN can use the generator to capture the data distribution, impute the missing values with the generated data, and apply the discriminator to decide the rightness of the imputation with the assistance of domain knowledge, if applicable. The adversarial process allows for precise data distribution capturing. Based on these general ideas, applications and variants are discussed in Subsections 3.1–3.6.

[2]Long short-term memory (LSTM) and Gated recurrent unit (GRU) are two main branches of RNN. Compared with vanilla RNN, they have an additional mechanism of "gates" to control the contribution of "memory", i.e., sequence-dependencies. GRU with two gates is simpler than LSTM with three gates, but performs similarly in many scenarios.

[3]Denoising autoencoder (DAE) and variational autoencoder (VAE) hold the same fundamental structure as vanilla autoencoder. DAE receives corrupted data points as input and is trained to predict the uncorrupted data points as its output [141]. Considering missingness as one of the forms of corruption, DAE can be more robust to missing values than vanilla AE. Variational Autoencoder (VAE) utilizes the technique of variational inference in statistics, which introduces probabilistic modeling in latent space to better approximate the true data [100].

[4]Framework refers to the fact that the modules (encoder/decoder and generator/discriminator) which respectively shaped the structures of AE and GAN, can embed with various models based on the data input, for example, convolutional neural network (CNN) to tackle images.

the 111 studies, 32 presented missing value imputation models for tabular static data, 44 for tabular temporal data, 15 for genetic and genomic data, six for image data, six for signal data, and nine for multi-modal data. It is important to note that these numbers are not mutually exclusive, as a single study may take into account multiple data types and impute them each individually with a single model, or alternatively, multiple DL-based imputation models may be applied to and compared for a single data type.

According to Fig. 4, most studies (68 %, 76/111) adopted the "separated" strategy. The "integrated" strategy was popular among tabular temporal and multi-modal data, but less used for tabular static data and genetic and genomic data, and rarely applied to image or signal data. Moreover, 61 out of the 111 selected studies investigated the type of data missingness, while the remaining addressed imputation directly without examining this specification. Table 2 indicates that explanation methods are rarely considered among the included studies (6 %, 7/111). In addition, we have compiled a summary of code sources in eTable 2, which provides more detailed information about the models used in the included studies. The following section presents DL-based imputation techniques based on different types of health data.

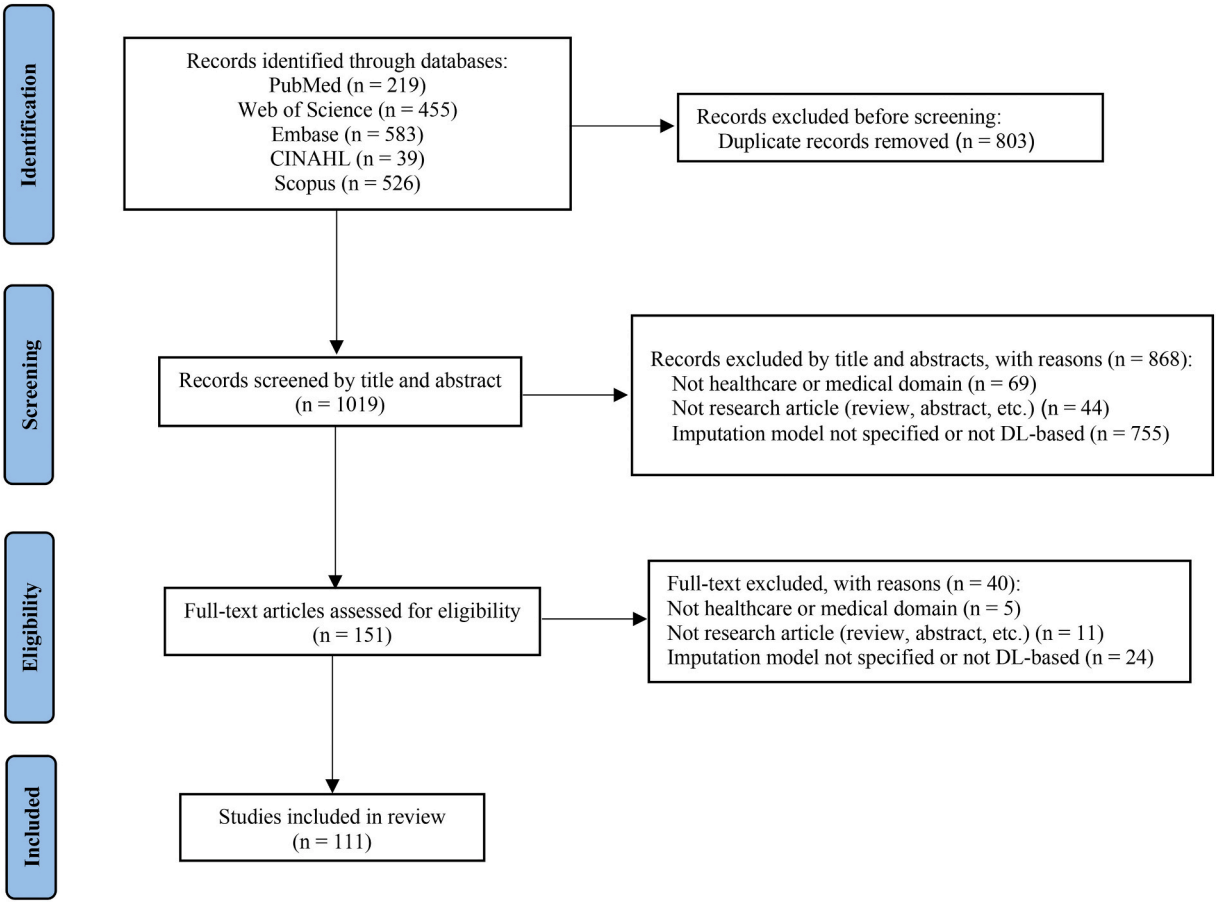


Fig. 2. Preferred Reporting Items for Systematic Reviews (PRISMA) flow diagram.

3.1. Tabular static data

A total of 32 studies [10,31–61] used tabular static data in this review. Most (81 %, 26/32) of them provided non-DL-based imputation methods as baselines for comparison with DL-based methods in terms of imputation accuracy (if complete data is available) and prediction performance, with simple imputation (mean/median/mode imputation, 44 %, 14/32), k-NN (44 %, 14/32), MICE (34 %, 11/32) and RF-based methods (31 %, 10/32) being the most common options. Most of these studies demonstrated the superiority of DL-based approaches.

MLP-oriented models dominate the structures of DL-based imputation methods for tabular static data. Ten studies [33,34,40–42,46,48,49,57,60] used MLP models directly, 12 [35–37,43,47,51–55,58,59] created autoencoder models employing MLP modules as encoders and decoders, and five studies [31,45,47,56,59] using GANs also implemented MLP modules as generators and discriminators. Studies conducted on one dataset and involving both autoencoders and GANs demonstrated the superiority of DL-based models in comparison to non-DL-based models [47,59]. Two studies [51,59] attempted to improve imputation accuracy by using k-NN for pre-imputation and Khan et al. [54] applied GAN for data augmentation prior to DAE-based imputation.

There are several alternatives to MLP. A dynamic layered RNN was applied to repeat the imputation process in order to enhance the accuracy of imputation [38]. Peralta et al. [10] embedded a residual connection in their autoencoder to allow for capturing non-linearities. To improve the robustness of imputation with respect to varying missing rates, Hallaji et al. developed an autoencoder with a ladder network structure [44] and a hybrid model incorporating a DAE module within the GAN framework [32]. This hybrid model differed from the method in [54], in which DAE and GAN were separated. In their research, Feng et al. [50] incorporated a transformer module within the generator of GAN to capture spatial correlations of the population health data. Traynor et al. [61] also used the transformer module, and specifically the model "TabNet" for imputation.

The handling of mixed variable types is challenging for tabular data (both static and temporal) due to the assumptions about data distribution and limitations of some non-DL-based methods [36,54]. However, DL-based imputation methods could accommodate a mixture of variable types with proper encoding and activation functions [35].

3.2. Tabular temporal data

Among the 111 included studies, 44 [8,9,11,13–16,44,62–97] addressed missing value imputation for tabular temporal data. This type of missing data is usually caused by factors that are less controllable, such as patients dropping out or using different assessment patterns for different patient subgroups [13,62]. As a result, informative missingness accompanied by time dynamics, heterogeneity, and high missing rates pose challenges for imputation.

AE-based (34 %, 15/44) and RNN-based (48 %, 21/44) methods were commonly employed to impute tabular temporal data, with "separated" imputation strategy being used more often in conjunction with the former (93 %, 14/15 for AE-based versus 24 %, 5/21 for RNN-based). The development of the "separated" imputation strategy relied on ground truth, that is, artificially composing missingness by dropping values at random and comparing the imputed values to the dropped values. Thirty out of 44 studies reported missing rates, with 13 of these studies assessing the robustness of imputation with different rates ranging from 5 % to 90 %.

Table 1
Summary of included studies.

Author	Health data type	Missing mechanism	Model backbone	Imputation Strategy	Non-DL Baselines
Ennett et al. (2008) [42]	Tabular Static Data	MNAR	MLP	Separated	Mean, Random
Hernandez-Pereira et al. (2015) [39]	Tabular Static Data	MNAR	Other (SOM)	Separated	Mean, Mode, k-NN, Multiple Linear Regression, hot-deck
Seffens et al. (2015) [41]	Tabular Static Data	MAR, MNAR	MLP	Separated	/
Bektaş et al. (2018) [46]	Tabular Static Data	/	MLP	Separated	Mean, Delete, Kmeans
Turabieh et al. (2018) [38]	Tabular Static Data	/	RNN	Separated	/
Huang et al. (2018) [40]	Tabular Static Data	MCAR	MLP	Separated	k-NN, SVM
Abiri et al. (2019) [35]	Tabular Static Data	MAR	Autoencoder	Separated	Mean, k-NN, MICE, RF
Miok et al. (2019) [36]	Tabular Static Data	MCAR	VAE	Integrated	/
Phung et al. (2019) [37]	Tabular Static Data	/	DAE	Separated	Mean, Median, Iterative SVD, k-NN, MF, SoftImpute
Cheng et al. (2020) [33]	Tabular Static Data	/	MLP	Separated	/
Vrbaski et al. (2020) [34]	Tabular Static Data	MCAR, MNAR, MAR	MLP	Separated	PMM, SLR, RF, Mean
Kachuee et al. (2020) [45]	Tabular Static Data	MCAR	GAN	Separated	MICE
Huang et al. (2020) [51]	Tabular Static Data	MAR	VAE	Separated	k-NN
Dong et al. (2021) [31]	Tabular Static Data	MAR	GAN	Separated	MICE, MissForest
Hallaji et al. (2021) [32]	Tabular Static Data	MAR	Hybrid (GAN, DAE)	Integrated	MICE, MissForest, k-NN, EM
Macias et al. (2021) [55]	Tabular Static Data	MAR	Autoencoder	Separated	Mean
Chen et al. (2021) [43]	Tabular Static Data	/	Autoencoder	Separated	MICE, MissForest, Matrix Completion
Kalweit et al. (2021) [53]	Tabular Static Data	MAR	Autoencoder	Separated	Zero, Mean, k-NN
Peralta et al. (2021) [10]	Tabular Static Data	MCAR	Autoencoder	Integrated	Pairwise Correlation PCA, Iterative PCA
Traynor et al. (2022) [61]	Tabular Static Data	MNAR	Other (Transformer)	Separated	EM, PMM with MICE, MIPCA, RF
Boursalie et al. (2022) [47]	Tabular Static Data	MNAR, MCAR, MAR	DAE, GAN	Separated	/
Bram et al. (2022) [48]	Tabular Static Data	MCAR, MAR	MLP	Separated	Mean, PMM, NORM, RF
Chang et al. (2022) [49]	Tabular Static Data	MNAR	MLP	Integrated	Mean, Median, Mode, k-NN, MICE
Feng et al. (2022) [50]	Tabular Static Data	/	GAN	Separated	Mean, Median, k-NN
Kabir et al. (2022) [52]	Tabular Static Data	/	Autoencoder	Separated	Iterative, k-NN, SVD, Mean
Khan et al. (2022) [54]	Tabular Static Data	MCAR	DAE	Separated	MissForest, MICE
Neves et al. (2022) [56]	Tabular Static Data	MCAR	GAN	Separated	/
Pan et al. (2022) [57]	Tabular Static Data	MAR, MCAR, MNAR	MLP	Separated	Mode, Random, Hot-deck, k-NN
Pereira et al. (2022) [58]	Tabular Static Data	MNAR	VAE	Integrated	Mean, MICE, k-NN, SoftImpute
Psychogyios et al. (2022) [59]	Tabular Static Data	MNAR	GAN, Autoencoder	Separated	Mean, Mode, k-NN, MissForest
Samad et al. (2022) [60]	Tabular Static Data	MAR, MCAR, MNAR	MLP	Separated	MICE, Iterative SVD, MF, k-NN
Beaulieu-Jones et al. (2016) [71]	Tabular Temporal Data	MCAR, MNAR	Autoencoder	Separated	Iterative SVD, k-NN, SoftImpute, Mean, Median, MICE
Bianchi et al. (2018) [72]	Tabular Temporal Data	MAR	Autoencoder	Integrated	Mean, LOCF
Che et al. (2018) [63]	Tabular Temporal Data	MCAR	GRU	Integrated	Mean, Forward, Concatenating, SoftImpute, k-NN, Cubic Spline, MICE, MF, MissForest
de Jong et al. (2019) [64]	Tabular Temporal Data	MAR, MNAR	Hybrid (LSTM, DAE)	Integrated	/
Ghazi et al. (2019) [62]	Tabular Temporal Data	MAR	LSTM	Integrated	Mean, Forward
Jun et al. (2019) [76]	Tabular Temporal Data	/	VAE	Separated	Zero, SoftImpute, k-NN, MICE
Jung et al. (2019) [68]	Tabular Temporal Data	/	RNN	Integrated	Mean, Forward
Park et al. (2019) [9]	Tabular Temporal Data	/	GAN	Separated	User-Avg, k-NN
Codella et al. (2019) [73]	Tabular Temporal Data	/	RNN	Separated	3D-MICE
Yoon et al. (2019) [8]	Tabular Temporal Data	MAR	RNN	Separated	Cubic Spline, MICE, MissForest, EM, Matrix Completion, MCMC
Fortuin et al. (2020) [74]	Tabular Temporal Data	MNAR	VAE	Separated	Forward, Mean, GP
Ma et al. (2020) [69]	Tabular Temporal Data	MCAR	Hybrid (GAN, RNN)	Integrated	Zero, RegEM, DynaMMo, TRMF
Tao et al. (2020) [79]	Tabular Temporal Data	/	DAE	Separated	Mean, Auto-regression
Xu et al. (2020) [11]	Tabular Temporal Data	MNAR, MAR	Autoencoder	Separated	SoftImpute, k-NN
Habiba et al. (2020) [65]	Tabular Temporal Data	MNAR, MAR	GRU	Integrated	/
Lin et al. (2020) [77]	Tabular Temporal Data	MNAR	Autoencoder	Separated	Interpolation, EWMA, k-NN, Kalman smoothing, LOCF
Zhao et al. (2020) [67]	Tabular Temporal Data	/	GRU	Separated	Ridge Regression
Yin et al. (2020) [78]	Tabular Temporal Data	/	LSTM	Integrated	Mean, k-NN, 3D-MICE, T-LGBM
Tsiligkaridis et al. (2020) [66]	Tabular Temporal Data	MNAR	LSTM	Integrated	/
Jun et al. (2020) [75]	Tabular Temporal Data	/	RNN	Integrated	Zero, Mean, k-NN
Jung et al. (2021) [16]	Tabular Temporal Data	/	LSTM	Integrated	Mean, Forward, Zero
Mulyadi et al. (2021) [70]	Tabular Temporal Data	/	Hybrid (VAE, RNN)	Integrated	/
Xu et al. (2021) [13]	Tabular Temporal Data	/	Autoencoder	Separated	SoftImpute, MICE, k-NN, MissForest, MiceForest
Gordon et al. (2021) [84]	Tabular Temporal Data	MCAR	Other (GNN)	Separated	MICE, k-NN, Mean, MissForest, interpolation
Liang et al. (2021) [89]	Tabular Temporal Data	/	LSTM	Integrated	/
Ramch et al. (2021) [14]	Tabular Temporal Data	/	VAE	Separated	/
Wang et al. (2021) [94]	Tabular Temporal Data	/	RNN	Separated	Mean, k-NN, Matrix Factorization(MF), MICE

(*continued on next page*)

Table 1 (*continued*)

Author	Health data type	Missing mechanism	Model backbone	Imputation Strategy	Non-DL Baselines
Zamanzadeh et al. (2021) [96]	Tabular Temporal Data	MAR, MCAR, MNAR	Autoencoder	Separated	k-NN, MICE, Mean/Mode
Zhang et al. (2021) [97]	Tabular temporal data	/	Hybrid (Autoencoder, GRU)	Integrated	/
Chen et al. (2022) [80]	Tabular Temporal Data	/	GAN	Integrated	Mean, k-NN, MICE, EM
Deshmukh et al. (2022) [81]	Tabular Temporal Data	MAR	VAE	Separated	MICE, MF, k-NN
Farrell et al. (2022) [82]	Tabular Temporal Data	/	VAE	Separated	Mean, fixed value, GLM
Haliduola et al. (2022) [85]	Tabular Temporal Data	MAR, MCAR	RNN	Integrated	/
Ho et al. (2022) [86]	Tabular Temporal Data	MAR	RNN	Integrated	/
Lee et al. (2022) [87]	Tabular Temporal Data	MNAR	Other (Transformer)	Integrated	/
Li et al. (2022) [88]	Tabular Temporal Data	MAR	GRU	Integrated	k-NN, MF, MICE
Liu et al. (2022) [90]	Tabular Temporal Data	/	GRU	Integrated	Mean, k-NN, 3D-MICE, Concatenating
Liu et al. (2022) [91]	Tabular Temporal Data	/	GRU	Integrated	Mean, k-NN, MICE, Concatenating
Porta et al. (2022) [93]	Tabular Temporal Data	MAR	LSTM	Integrated	SoftImpute, ST-MVL
Rasmy et al. (2022) [15]	Tabular Temporal Data	/	GRU	Integrated	/
Yildiz et al. (2022) [95]	Tabular Temporal Data	/	Autoencoder	Separated	/
Getz et al. (2023) [83]	Tabular Temporal Data	MAR, MNAR	VAE	Separated	RF, MICE
Lu et al. (2023) [92]	Tabular Temporal Data	/	LSTM	Separated	Cubic Spline, k-NN
Hallaji et al. (2020) [44]	Tabular Static Data, Tabular Temporal Data	MNAR, MCAP, MAR	Autoencoder	Separated	EM, MissForest, MICE
Sun et al. (2008) [103]	Genetic and Genomic Data	MAR	MLP	Separated	fastPHASE, EM
Badsha et al. (2019) [101]	Genetic and Genomic Data	MAR	Autoencoder	Separated	scVI, SAVER, MAGIC, ALRA, scImpute
Kinalis et al. (2019) [102]	Genetic and Genomic Data	/	Autoencoder	Integrated	/
Chen et al. (2019) [98]	Genetic and Genomic Data	/	MLP	Separated	Mean, SVD, FactoMineR, fastICA, Bayesian-based NMF, gradient-based NMF
Qiu et al. (2020) [100]	Genetic and Genomic Data	MCAR, MNAR	VAE	Separated	Mean, k-NN, Iterative SVD
Mongia et al. (2020) [99]	Genetic and Genomic Data	/	MLP	Separated	scImpute, drImpute, MAGIC, SAVER
Tian et al. (2021) [104]	Genetic and Genomic Data	/	Autoencoder	Separated	scImpute, SAVER, MAGIC
Dai et al. (2021) [107]	Genetic and Genomic Data	MAR	GAN	Separated	MICE, SoftImpute, Sinkhorn, Linear RR
Zhang et al. (2021) [110]	Genetic and Genomic Data	/	Other (CNN)	Separated	MAGIC, SAVER, scImpute
Chen et al. (2022) [106]	Genetic and Genomic Data	MAR	MLP	Separated	Mean, PMM, NORM
Mahbub et al. (2022) [108]	Genetic and Genomic Data	MCAR	Autoencoder	Separated	/
Peacock et al. (2022) [109]	Genetic and Genomic Data	/	VAE	Separated	/
Zhang et al. (2022) [111]	Genetic and Genomic Data	MCAR	Autoencoder	Separated	ALRA, SAVER, scImpute, DrImpute, MAGIC, EnImpute , VIPER
Zhou et al. (2022) [112]	Genetic and Genomic Data	/	Autoencoder	Integrated	/
Chen et al. (2023) [105]	Genetic and Genomic Data	/	Autoencoder	Separated	SAVER, scImpute, VIPER, bayNorm, scRecover, ALRA, SCRABBLE
Pan et al. (2018) [114]	Image	MCAR, MNAR, MAR	GAN	Separated	/
Lee et al. (2019) [115]	Image	/	GAN	Separated	/
Pan et al. (2020) [116]	Image	/	GAN	Separated	/
Xia et al. (2021) [113]	Image	/	GAN	Separated	Interpolation, Mean
Gao et al. (2021) [12]	Image	/	GAN	Separated	/
Peng et al. (2021) [117]	Image	/	GAN	Separated	/
Miller et al. (2018) [122]	Signal	/	Autoencoder	Separated	/
Saeed et al. (2018) [120]	Signal	/	Hybrid (Autoencoder, GAN)	Separated	Mean, Median, filling with −1, PCA
Feng et al. (2019) [121]	Signal	MAR	RNN	Separated	Mean, k-NN, SoftImpute, BiScaler, MICE
Jang et al. (2020) [119]	Signal	/	DAE	Separated	Mean, MICE
Calhas et al. (2020) [124]	Signal	MAR	GRU	Separated	k-NN, Mean, Barycenter, MICE
Lee et al. (2021) [123]	Signal	/	GAN	Separated	Random
Thung et al. (2017) [128]	Multi-modal Data	MNAR	Other	Integrated	LRMC, iMSF
Jabason et al. (2018) [130]	Multi-modal Data	MNAR	Autoencoder	Integrated	Mean, k-NN
Kim et al. (2020) [127]	Multi-modal Data	MAR, MNAR	MLP	Integrated	/
Kim et al. (2020) [126]	Multi-modal Data	/	DAE	Separated	k-NN, SVD, Mean
Akramifard et al. (2020) [125]	Multi-modal Data	/	Autoencoder	Separated	Mean
Fan et al. (2021) [133]	Multi-modal Data	MAR, MCAR	MLP	Separated	MF
Vivar et al. (2021) [129]	Multi-modal Data	MAR	Hybrid (LSTM, GCN)	Integrated	k-NN, PPCA, MICE-LR
Li et al. (2021) [131]	Multi-modal Data	/	RNN	Integrated	Mean, k-NN, MICE
Xu et al. (2022) [132]	Multi-modal Data	/	VAE	Separated	Mean, k-NN, MissForest, SoftImpute

SAVER: Single-cell Analysis Via Expression Recovery; MAGIC: Markov Affinity-based; Graph Imputation of Cells; k-NN: k-Nearest Neighbor Imputation; PPCA: Probabilistic Principal Component Analysis; MICE: Multiple Imputation by Chained Equations; MICE-LR: Multiple Imputation by Chained Equations with Linear Regression; EM: Expectation Maximization; PMM: Predictive Mean Matching; SLR: Stochastic Linear Regression; RF: Random Forest; ALRA: Adaptively thresholded Low-Rank Approximation; EWMA: Exponentially Weighted Moving Average; LOCF: Last Observation Carried Forward; GP: Gaussian Process; SVD: Singular Value

Decomposition; MCMC: Markov Chain Monte Carlo; LRMC: Low-Rank Matrix Completion with sparse feature selection; iMSF: incomplete Multi-Source joint Feature learning; SVM: Support Vector Machine; RegEM: Regularized Expectation Maximization; TRMF: Temporal Regularized Matrix Factorization; NMF: Non-negative Matrix Factorization; T-LGBM: Light Gradient Boosting Machine on Temporal and Cross-variable Features; VIPER: Variability-Preserving ImPutation for Expression Recovery; MF: Matrix Factorization; GLM: Generalized linear model; ST-MVL: Spatio-Temporal Multi- View-based Learning; MIPCA: Multiple Imputation with Principal Component Analysis.

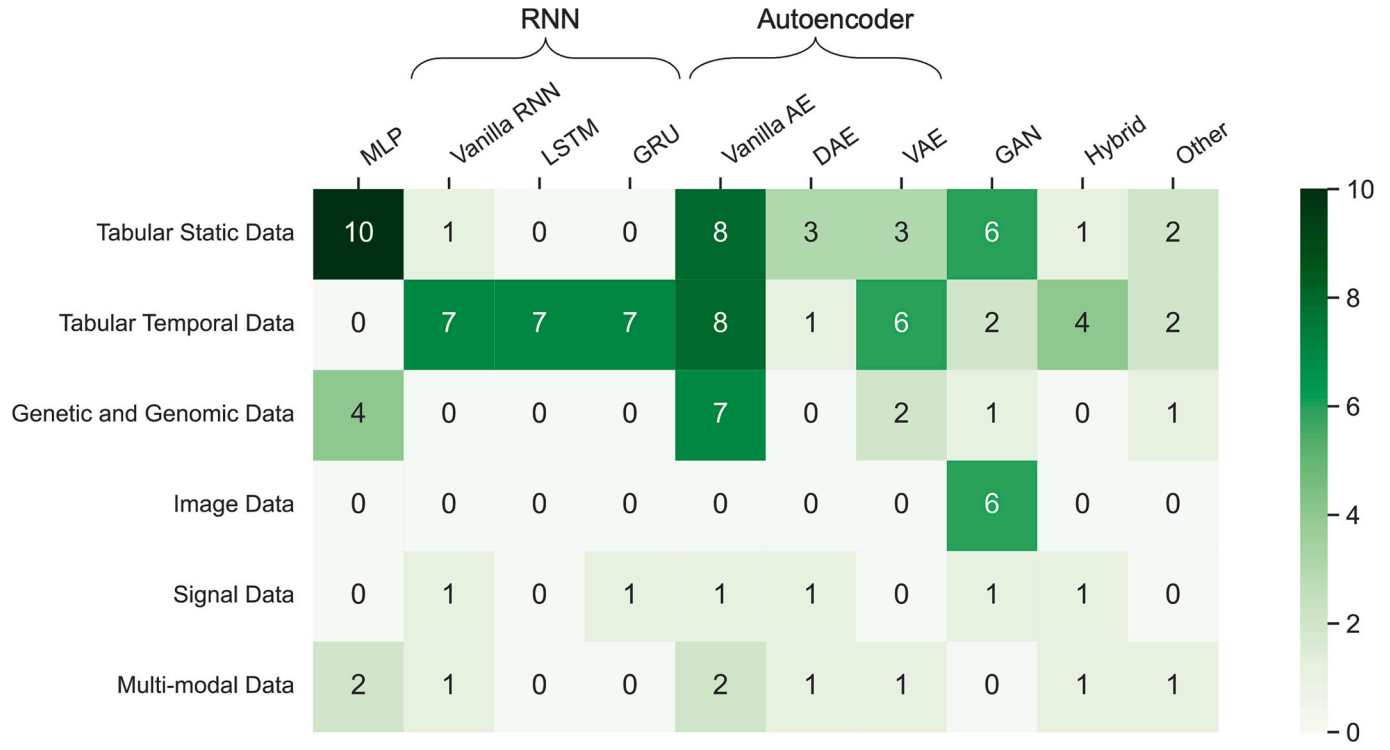


Fig. 3. The evidence map between "backbones" (main architectures) of model and data type "Backbones" are classified into ten categories: MLP (multi-layer perceptron), RNN (recurrent neural network), LSTM (Long short-term memory), GRU (gated recurrent unit), AE (autoencoder), DAE (denoising autoencoder), VAE (variational autoencoder), GAN (generative adversarial network) and Other, which includes less frequently used models such as SOM (self-organizing map). Data types are categorized into seven categories: tabular static, tabular temporal, genetic and genomic, image, signal, and multi-modal data. The numbers are non-exclusive.

The framework of autoencoder provides great flexibility when handling the complex characteristics of tabular temporal data during imputation [13]. Eight studies [11,13,44,71,72,77,95,96] designed vanilla autoencoders, some of which were specifically customized to fit the data. The customizations included adding an extra encoder to deal with patient heterogeneity, implementing a ladder network to tackle both spatial and temporal relationships, and incorporating a transformer module to capture long-term dependencies. Tao et al. [79] developed a DAE method to denoise missing data. Six studies [14,74,76,81–83] adopted VAE models to reflect correlations over time based on variational posteriors. These models are based on statistical knowledge such as Gaussian process and Bayesian inference, which permit a robust and accurate representation of tabular temporal data.

Modeling sequence data with RNN-based methods allows for capturing missing patterns related to time dynamics. Among the 21 studies that applied RNN-based methods, 14 [15,16,62,63,65–67,78,88–93] developed LSTM or GRU models using the gate mechanism to control the information flow along the sequence. In particular, three studies [63,65,67] applied or developed variants of GRU-D – a model that uses a specific parameter to characterize the decay of effects over time. Among the seven studies [8,68,73,75,85,86,94] that utilized vanilla RNN models, Jun et al. [75] specifically employed the variational posterior to capture uncertainty. Other than one-direction RNN-based models, some [8,73,78,86,88,92,94] designed their models with bi (multi)-directions to incorporate both past and future information for imputation.

Moreover, three studies [64,70,97] developed hybrid RNN-based and AE-based methods in which the AE component was added after initial imputation by RNN. Three other studies used the GAN framework, where adversarial learning (either alone [9] or in conjunction with an additional transformer module to encode the missingness parallelly [80] or in combination with RNN [69]) can help prevent error propagation from imputation to downstream tasks. Besides the aforementioned models, Gorden et al. [84] developed a GNN model based on a joint bipartite graph, and Lee et al. [87] applied a hierarchical transformer model to accommodate irregular time sequences.

3.3. Genetic and genomic data

In this review, 15 studies [98–112] dealt with genetic and genomic data, including single-cell RNA sequencing data [99,102,104,105,110,111], gene expression data [101,103,107,108], and combinations of several data formats such as DNA methylation, mRNA, and microRNA data [98,100,106,109,112]. Data obtained from single-cell sequencing may contain around 50 % zero-count observations [101,104], some of which are "false zeros" or "false negatives", i.e., missing values due to inadequate sequencing input [99]. Furthermore, the high dimension property complicates the imputation of genetic and genomic data [106,107].

A total of nine studies [100–102,104,105,108,109,111,112] used AE-based models (vanilla AE and VAE), four studies [98,99,103,106] were based on MLP, and another two studies employed GAN [107] and

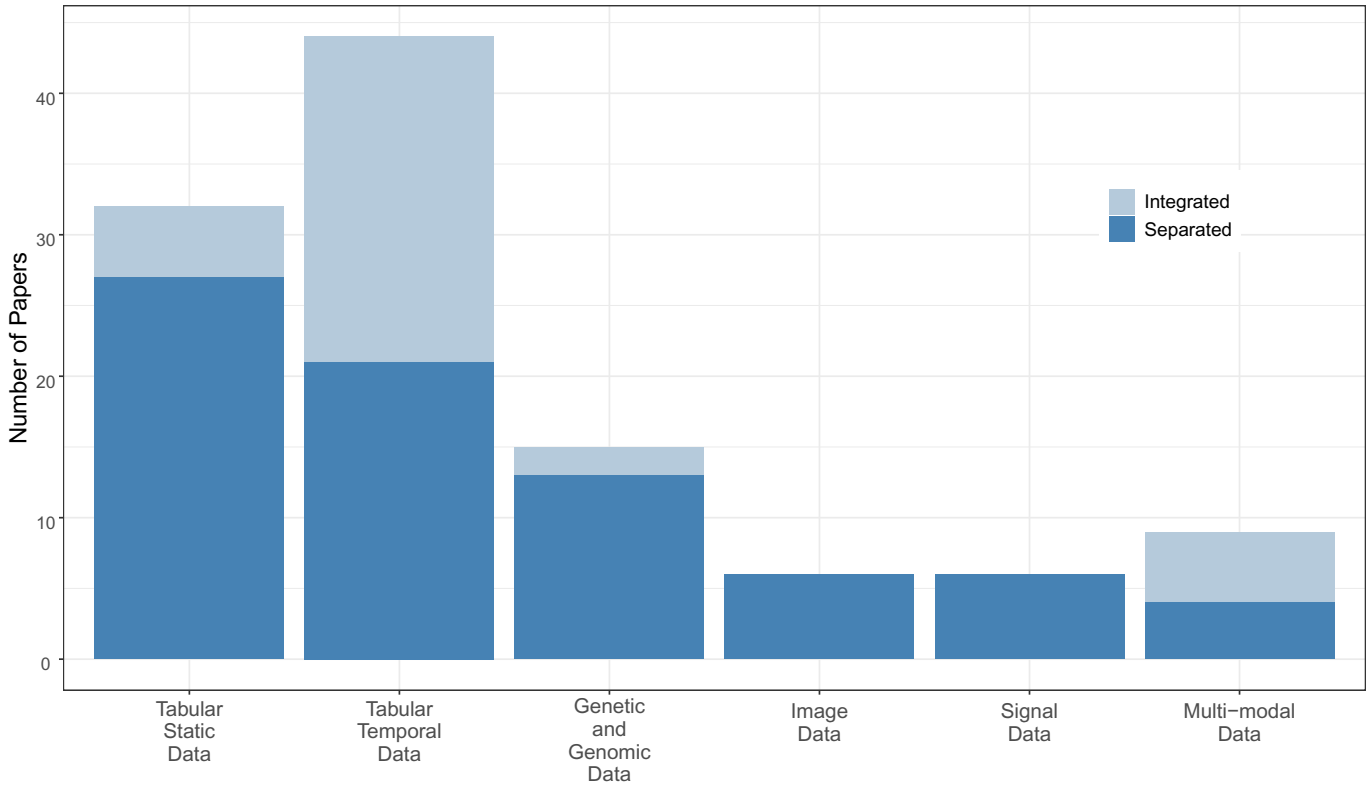


Fig. 4. The distribution of imputation frameworks (integrated or separated) by data type.

Table 2
Explanation of missing value imputation models.

Author	Health data type	Model backbone	Explanation type	Explanation method
Huang et al. (2020) [51]	Tabular Static Data	VAE	Post hoc	RF
Park et al. (2019) [9]	Tabular Temporal Data	GAN	Post hoc	Attention
Zhang et al. (2021) [97]	Tabular Temporal Data	Hybrid (Autoencoder, GRU)	Post hoc	Attention
Chen et al. (2022) [80]	Tabular Temporal Data	GAN	Post hoc	Attention
Ho et al. (2022) [86]	Tabular Temporal Data	RNN	Post hoc	SHAP
Lee et al. (2022) [87]	Tabular Temporal Data	Other (Transformer)	Post hoc	LRP
Rasmy et al. (2022) [15]	Tabular Temporal Data	GRU	Post hoc	Integrated gradient

RF: Random Forest; SHAP: SHapley Additive exPlanations; LRP: Layer-wise Relevance Propagation.

CNN [110] models. Non-DL methods such as "scImpute" (40 %, 6/15) were commonly used for comparison. By using AE-based models and transfer learning, Badsha et al. [101] extracted prior knowledge about gene-gene relationships and learnt the dependence structure within the reference panel. Kinalis et al. [102] suggested that the architecture of AE and VAE facilitates more interpretable imputation procedures. Upon proper training, the autoencoder can be interpreted as a combination of biologically meaningful modules. In their study, Chen et al. [105] constrained their autoencoder by bulk sequencing data when imputing single-cell sequencing data. Aside from AE's ability to reduce the dimension in latent space, MLP may also contribute to dimensionality reduction for imputation [98,99]. In one example, Chen et al. [98] deciphered two low-dimensional hidden representations from the original high-dimensional data to explain a molecular relationship and a sample-level connection. Moreover, instead of using MLPs to handle high-dimensional missingness, Dai et al. [107] used GAN and Zhang et al. [110] used CNN models to examine expression patterns.

3.4. Image data

There were six studies [12,113–117] that focused on image data, five of which analyzed neuroimages, including magnetic resonance imaging (MRI) and positron emission tomography (PET); these images have been widely adopted for computer-aided diagnosis of AD and mild cognitive impairment (MCI) [12,114–117]. Another study [113] examined cardiac magnetic resonance (CMR) images, which are often considered the gold standard for many cardiovascular medicine analyses. Insufficient image quality and acquisition or storage errors are two of the most common causes of missing pixels in image data [12,113]. In practice, PET scans may be rejected by many MRI patients due to high cost and radioactive exposure [116], resulting in the absence of the entire image.

All six studies constructed GAN frameworks for missing value imputation. Only one study [113] established non-DL-based imputation baselines for comparison (mean imputation and an interpolation approach), while the other five did not. Moreover, only one study [114] clarified the missing mechanism (i.e., MCAR, MAR, or MNAR). By utilizing label information in a conditional GAN, missing pixels in an image can be imputed or a new image can be synthesized when the whole image is missing [113]. Imputing PET data based on the corresponding MRI images was also investigated for brain diseases diagnosis [12,118]. To impute the entire image, a task-induced GAN can be developed with two tasks designed for the discriminator: whether the image is true and

whether it indicates disease [12]. This task was addressed by Cycle-GAN and Colla-GAN [114,115], similar to Pan et al. [116] that connected MRI and PET by designing two generators and two discriminators in their GAN framework. Another study by Peng et al. [117] combined voxel-wise reconstruction loss with perceptual loss to maintain the consistency of disease details.

3.5. Signal data

A variety of signal data was discussed in six studies, including actigraphy device data [119], smartphone applications data [120], wearable sensor data [121], medical waveforms (e.g. ECG, EEG) [122,123], and time-series signals measured at fMRI [124]. In two studies, the missing mechanism was identified as MAR [121,124], while no claim was made in four other studies. In contrast to tabular temporal data, signal data usually has a high sampling rate and is susceptible to noise. As a result of intermittent disconnections, body movement, and firmware malfunctions, there is a large amount of missing data collected during the movement [121], with blocks of features being lost simultaneously. It may be necessary to characterize the missing interval by analyzing it as a period of continuous repetition of zero values [119]. For medical waveform data, even a low missing rate in real-time operational systems, such as removing the lead for a few seconds, can significantly impact the prediction performance [122]. Also, due to the long recording duration and lack of frequent monitoring, missingness tends to occur repeatedly until the cause is identified [123]. Similarly to image modality, fMRI is prone to noise artifacts, resulting in missing values for signal modality. This presents challenges to the imputation process, where the spatiotemporal nature of the data can be helpful [124].

The non-DL-based imputation methods, simple imputation (mean/median imputation, 67 %, 4/6) and MICE (50 %, 3/6), are common baselines in these six studies. Three studies used customized AE-based models: denoising autoencoder that treated missingness as a type of noise [119], adversarial autoencoder where the encoder contributed to feature representation and followed by the discriminator of GAN [120], and 1-D convolutional and corresponding deconvolutional modules [122]. One study applied a GAN imputation model with a prediction loss to preserve the contextual information about features [123]. In the other two studies [121,124], RNN models were applied with the time factor taken into account. Signal imputation in fMRI data is accomplished by first filling in the missing values using spatial information, and then regularizing the time domain using a GRU layer [124].

3.6. Multi-modal data

A total of nine studies [125–133] investigated imputation methods for multi-modal data, where information fusion procedures were essential to connect modality-specific models. Five studies employed autoencoder models and built linkages between encoders and decoders to concatenate different modalities [125,126,130–132]. For example, Kim et al. [126] developed a stacked DAE model with a merged hidden layer that served as a linkage. They also created a collaborative layer to connect MLP models across different modalities [127]. Moreover, Xu et al. designed a product-of-expert module to discover the intrinsic correlations between different data mortalities [132]. Rather than fusing information across modalities, Li et al. [131] applied a sequence structure for linking modality-specific models. The task-specific layers designed by Thung et al. [128] could enable iterative communication between modality-specific layers, thereby facilitating the exchange of cross-modal information. In their end-to-end framework, Vivar et al. [129] aggregated recurrent graph convolutional models through the self-attention process. Using this architecture, missing value imputation was transformed into the completion of a geometric matrix. Solving this geometric problem has been demonstrated to be effective when graph convolutional models are coupled with LSTMs.

4. Discussion

With this systematic review, we contribute to a comprehensive summary of knowledge regarding the efficacy of DL models in missing value imputation for healthcare data. We found that DL models are superior to non-DL-based methods in that they are customized to take into account the data type as well as the missing patterns, thereby improving the quality of data imputation. Besides, the "integrated" imputation strategy could enhance the performance of both imputation and downstream analysis, and its usage varied across data types, highlighting the advantages of imputation based on the characteristics of the data type. Our investigation also revealed a lack of attention toward the issues related to method practicability, interpretability, and fairness concerns.

4.1. The mapping of DL-based imputation methods with data types

Data-type-oriented DL-based imputation models are both beneficial for the imputation process and downstream tasks. As illustrated in Fig. 3, DL-based imputation models are associated with data types. The MLP-dominated models (MLP, autoencoders, and MLP-based GANs) are widely used to determine the feature relevance of tabular static data. With tabular temporal data, informative missingness and high missing rates make it difficult to characterize time dynamics [11,13]. RNN-based models, such as bi-directional RNN [8], and autoencoders with statistical modeling, such as VAE [74], are commonly used for capturing complex time patterns.

In the case of genetic and genomic data, the biological characteristics and associated biological knowledge, such as gene-gene relationships, can be effectively incorporated into the imputation process [98,99,101,102]. In the context of image imputation, GAN-based models are commonly used. Providing additional information, such as labeling (Co-GAN [134]) and relevant images (Cycle-GAN [135] or Colla-GAN [115]), can enhance the performance of imputation. The use of CNNs, residual networks, and attention blocks is prevalent in addressing deep spatial information contained in image data [12,113–117]. There is a wide range of imputation procedures for signal data, partly because different signal types have different causes of missingness.

The fusion of mode-specific models is essential when encoding multi-model data. Currently, most operations are focused on the layer level, for example, stacking and self-attention mechanism [125,126,130–132]. Some researchers use the term "multi-modality" when describing datasets collected from different sources but of the same type (e.g., image data of MRI and PET [12,114], RNA and methylation data [109,112]).

An opening exists in the imputation approach for medical text data. There may be an explanation for this: since techniques in natural language processing (e.g., BERT [136]) inherently learn representation through masking, i.e., considering some language tokens as missing on purpose, so the actual absence of tokens will not be an issue. Medical text data can be analyzed using customized biomedical research models, such as BioBERT [137] and MedBERT [138].

4.2. The benefits of "integrated" imputation strategy

The block-building logic enables DL-based models to adopt an "integrated" strategy, i.e., co-training imputation and downstream tasks, which is advantageous for several reasons. First, the interaction between these two tasks can be mutually beneficial, reducing the bias in imputation, and providing prior information for downstream modeling [64,69,87,129,131]. Additionally, the "integrated" strategy is more practical since it avoids the difficulties in defining imputation accuracy when the missing rate of the original dataset is high, indicating limited ground truth for imputation quality checking [75,78,87,139]. Moreover, when multiple data types are involved in an "integrated" framework, the fusion of latent information during the imputation process can

directly be used for downstream tasks, thereby preventing redundant training efforts [75,128,130]. This is in line with the relative prevalence of the "integrated" strategy when working with tabular temporal and multi-modal data, as shown in Fig. 4.

In contrast to the "separated" strategy, the "integrated" strategy does not emphasize the selection of the optimal combination of imputation and downstream models [129]. The "separated" strategy, in which the best imputation model is determined first and then downstream models are chosen, may not be effective given the belief that imputation accuracy does not directly affect downstream tasks [34,129]. The "integrated" strategy can resolve these practical difficulties by imputing missing data together with the downstream models being developed. It should be noted, however, the "integrated" strategy results in greater model complexity, which explains its limited application in current studies (Fig. 4).

4.3. Comparison with non-DL-based imputation models

When both non-DL-based and DL-based imputation models are available, the former may be preferred for its simplicity of implementation; some non-DL-based methods (such as MICE, XGboost, LightGBM, etc.) could produce effective imputation when paired with carefully constructed data presented in a tabular or temporal format [17]. The ease of application is, however, dependent upon restrictive statistical assumptions about data, which can be difficult to identify in real-world scenarios [85,140]. Furthermore, feature engineering requires a substantial amount of time and effort [37], diverting researchers from their primary research objectives. Other concerns, such as high data dimension [11,79] and low time efficiency [11,126], also pose obstacles for non-DL-based methods. For healthcare data in complex formats, DL-based models seem ideal, as statistical assumptions and feature engineering are relatively less needed, and they do not suffer from the curse of dimensionality. Additionally, pre-trained DL-based models can reduce computational costs at the evaluation stage [100].

4.4. Drawbacks and future directions of DL-based imputation models

Several potential concerns have been raised based on this review, which can influence the adoption of DL-based models for missing value imputation on a large scale. A high degree of portability is essential considering the heavy burden placed on healthcare systems. When dealing with complex healthcare data, researchers may easily fall victim to model stacking. Models should be carefully and efficiently designed to better capture missing patterns and take advantage of module interaction [69,129], rather than stacking for novelty. Besides, clinical practitioners lacking deep learning expertise may find it challenging to implement DL-based imputation models.

The interpretability of DL-based models is fundamental to bridging the gap between clinicians and algorithm developers. Nevertheless, this aspect has only been mentioned in a few studies [9,15,51,80,86]. Although full transparency is a difficult goal to attain, model interpretability can still be achieved through post-hoc methods such as SHapley Additive exPlanations (SHAP), or by using the attention mechanism for explanations (Table 2). As such, explanations like feature importance ranking can contribute to objective variable selection and model evaluation; consequently, it can not only improve the practicability of DL-based models, but will also enhance clinicians' confidence and trust in complex models.

Moreover, researchers should also pay attention to the fairness in the imputation process, which has not been adequately addressed at the moment, and there is a lack of discussion on social bias, or discrimination against certain groups or individuals [1,13]. Using imputed data influenced by such bias may adversely affect the subsequent analysis and result in unjustified decision-making and medical inequality.

4.5. Limitations

This study has several limitations. First, the scope of our review was limited to clinical and translational research; however, some DL-based imputation techniques may be published in other research fields. Second, Transformer and CNN were viewed as additive modules, rather than model "backbones" because they were commonly used within autoencoder and GAN frameworks as opposed to individual models, which may differ slightly from the usual usage. Third, we focused primarily on data types and their corresponding imputation strategies. Lastly, we did not provide data type-specific experimental comparisons since a comprehensive and quantitative evaluation is beyond the scope of this study.

5. Conclusions

Our study fills a gap in the existing literature by systematically reviewing and evaluating DL-based methods for missing value imputation. In contrast to conventional imputation techniques like k-NN and MICE, DL-based imputation models represent a family of techniques. The design of DL-based imputation models in healthcare should be tailored to data types and characteristics. As with non-DL models, there is no universally ideal DL-based imputation model, but achieving satisfactory performance with respect to a specific data type or dataset is highly feasible. Research in the future may focus on the portability, interpretability, and fairness of DL-based imputation models.

Declaration of Competing Interest

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.artmed.2023.102587.

References

[1] Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. JAMA 2019;322(24):2377–8. https://doi.org/10.1001/jama.2019.18058.
[2] Kang H. The prevention and handling of the missing data. Korean J Anesthesiol 2013;64(5):402–6. https://doi.org/10.4097/kjae.2013.64.5.402.
[3] Little R, Little RJA, Rubin DB. Statistical analysis with missing data. Wiley; 2019.
[4] Rubin DB. Inference and missing data. Biometrika 1976;63(3):581–92. https://doi.org/10.1093/biomet/63.3.581.
[5] Batista GEAPA, Monard MC. An analysis of four missing data treatment methods for supervised learning. Appl Artif Intell 2003;17(5–6):519–33. https://doi.org/10.1080/713827181.
[6] van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. J Stat Softw 2011;45(3):1–67. https://doi.org/10.18637/jss.v045.i03.
[7] Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics 2011;28(1):112–8. https://doi.org/10.1093/bioinformatics/btr597.
[8] Yoon J, Zame WR, van der Schaar M, Yoon J, Zame WR, van der Schaar M. Estimating missing data in temporal data streams using multi-directional recurrent neural networks. IEEE Trans Biomed Eng 2019;66(5):1477–90. https://doi.org/10.1109/TBME.2018.2874712.
[9] Park S, et al. Learning sleep quality from daily logs. In: KDD'19: Proceedings of the 25th ACM SIGKDD international conferencce on knowledge discovery and data mining; 2019. p. 2421–9.
[10] Peralta M, Jannin P, Haegelen C, Baxter JSH. Data imputation and compression for Parkinson's disease clinical questionnaires. Artif Intell Med 2021;114:102051. https://doi.org/10.1016/j.artmed.2021.102051.
[11] Xu D, et al. A deep learning-based, unsupervised method to impute missing values in electronic health records for improved patient management. J Biomed Inform 2020:111. https://doi.org/10.1016/j.jbi.2020.103576.
[12] Gao X, Shi F, Shen D, Liu M. Task-induced pyramid and attention GAN for multimodal brain image imputation and classification in Alzheimers disease. IEEE J Biomed Health Inform 2021;26(1):36–43. https://doi.org/10.1109/JBHI.2021.3097721.
[13] Xu D, et al. A deep learning-based unsupervised method to impute missing values in patient records for improved management of cardiovascular patients. IEEE J

Biomed Health Inform 2021;25(6):2260–72. https://doi.org/10.1109/JBHI.2020.3033323.
[14] Ramchandran S, Tikhonov G, Kujanpaa K, Koskinen M, Lahdesmaki H. Longitudinal variational autoencoder. In: Presented at the 24th international conference on artificial intelligence and statistics (AISTATS); 2021.
[15] Rasmy L, et al. Recurrent neural network models (CovRNN) for predicting outcomes of patients with COVID-19 on admission to hospital: model development and validation using electronic health record data. Lancet Digital Health 2022;4(6):e415–25. https://doi.org/10.1016/S2589-7500(22)00049-8.
[16] Jung W, et al. Deep recurrent model for individualized prediction of Alzheimer's disease progression. NEUROIMAGE 2021;237:118143.
[17] Luo Y. Evaluating the state of the art in missing data imputation for clinical data. Brief Bioinform 2021. https://doi.org/10.1093/bib/bbab489.
[18] Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. BMC Med Res Methodol 2017;17(1):162. https://doi.org/10.1186/s12874-017-0442-1.
[19] Bell ML, Fiero M, Horton NJ, Hsu C-H. Handling missing data in RCTs; a review of the top medical journals. BMC Med Res Methodol 2014;14(1):118. https://doi.org/10.1186/1471-2288-14-118.
[20] Thomas T, Rajabi E. A systematic review of machine learning-based missing value imputation techniques. Data Technol Appl 2021;55(4):558–85. https://doi.org/10.1108/DTA-12-2020-0298.
[21] Jäger S, Allhorn A, Bießmann F. A benchmark for data imputation methods. Front Big Data 2021;4. https://doi.org/10.3389/fdata.2021.693674.
[22] Ismail AR, Abidin NZ, Maen MK. Systematic review on missing data imputation techniques with machine learning algorithms for healthcare. Review; Missing Data Imputation; Machine Learning. Healthcare 2022;3(2):10. https://doi.org/10.18196/jrc.v3i2.13133.
[23] Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O. A survey on missing data in machine learning. J Big Data 2021;8(1):140. https://doi.org/10.1186/s40537-021-00516-9.
[24] Alabadla M, et al. Systematic review of using machine learning in imputing missing values. IEEE Access 2022;10:44483–502. https://doi.org/10.1109/ACCESS.2022.3160841.
[25] Liu N, et al. coronavirus disease 2019 (COVID-19): an evidence map of medical literature. BMC Med Res Methodol Jul 2 2020;20(1):177. https://doi.org/10.1186/s12874-020-01059-y.
[26] Page MJ, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. https://doi.org/10.1136/bmj.n71.
[27] Huang K, Xiao C, Glass LM, Critchlow CW, Gibson G, Sun J. Machine learning applications for therapeutic tasks with genomics data. Patterns 2021;2(10):100328. https://doi.org/10.1016/j.patter.2021.100328.
[28] David ER, James LM. Learning internal representations by error propagation. In: Parallel distributed processing: explorations in the microstructure of cognition: foundations. MIT Press; 1987. p. 318–62.
[29] Kramer MA. Nonlinear principal component analysis using autoassociative neural networks. AIChE J 1991;37(2):233–43. https://doi.org/10.1002/aic.690370209.
[30] Goodfellow I, et al. Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ, editors. Advances in neural information processing systems. 27. Curran Associates, Inc; 2014 [Online]. Available, https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.
[31] Dong W, et al. Generative adversarial networks for imputing missing data for big data clinical research. BMC Med Res Methodol 2021;21(1). https://doi.org/10.1186/s12874-021-01272-3.
[32] Hallaji E, Razavi-Far R, Palade V, Saif M. Adversarial learning on incomplete and imbalanced medical data for robust survival prediction of liver transplant patients. IEEE Access 2021;9:73641–50. https://doi.org/10.1109/ACCESS.2021.3081040.
[33] Cheng CY, et al. A deep learning approach for missing data imputation of rating scales assessing attention-deficit hyperactivity disorder. Front Psych 2020:11. https://doi.org/10.3389/fpsyt.2020.00673.
[34] Vrbaski D, et al. Missing data imputation in cardiometabolic risk assessment: a solution based on Artificial neural networks. Comput Sci Inf Syst 2020;17(2):379–401. https://doi.org/10.2298/CSIS190710003V.
[35] Abiri N, et al. Establishing strong imputation performance of a denoising autoencoder in a wide range of missing data problems. Neurocomputing 2019;365:137–46. https://doi.org/10.1016/j.neucom.2019.07.065.
[36] Miok K, et al. Multiple imputation for biomedical data using monte carlo dropout autoencoders. In: 2019 E-Health and Bioengineering Conference (EHB); 2019.
[37] Phung S, Kumar A, Kim J. A deep learning technique for imputing missing healthcare data. In: 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC); 2019. p. 6513–6.
[38] Turabieh H, Abu Salem A, Abu-El-Rub N, Turabieh H, Abu Salem A, Abu-El-Rub N. Dynamic L-RNN recovery of missing data in IoMT applications. Future Gen Comput Syst Int J Esci 2018;89:575–83. https://doi.org/10.1016/j.future.2018.07.006.
[39] Hernandez-Pereira EM, Alvarez-Estevez D, Moret-Bonillo V, Hernandez-Pereira EM, Alvarez-Estevez D, Moret-Bonillo V. Automatic classification of respiratory patterns involving missing data imputation techniques. Biosyst Eng 2015;138:65–76. https://doi.org/10.1016/j.biosystemseng.2015.06.011.
[40] Huang MW, Lin WC, Tsai CF. Outlier removal in model-based missing value imputation for medical datasets. J Healthcare Eng 2018;2018:1817479. https://doi.org/10.1155/2018/1817479.
[41] Seffens W, Evans C, Taylor H. Machine learning data imputation and classification in a multicohort hypertension clinical study. Bioinform Biol Insight 2015;9:43–54. https://doi.org/10.4137/BBI.S29473.
[42] Ennett CM, Frize M, Walker C. Imputation of missing values by integrating neural networks and case-based reasoning. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2008. IEEE engineering in medicine and biology society. Annual international conference; 2008. p. 4337–41. https://doi.org/10.1109/IEMBS.2008.4650170.
[43] Chen Z, Cao B, Edwards A, Deng H, Zhang K. A deep imputation and inference framework for estimating personalized and race-specific causal effects of genomic alterations on PSA. J Bioinform Comput Biol 2021:2150016. https://doi.org/10.1142/S0219720021500165.
[44] Hallaji E, Razavi-Far R, Saif M. DLIN: deep ladder imputation network. IEEE Trans Cybern 2021;52(9):8629–41. https://doi.org/10.1109/TCYB.2021.3054878.
[45] Kachuee M, Karkkainen K, Goldstein O, Darabi S, Sarrafzadeh M. Generative imputation and stochastic prediction. IEEE Trans Pattern Anal Mach Intell 2020;44(3):1278–88. https://doi.org/10.1109/TPAMI.2020.3022383.
[46] Bektaş J, Ibrikçi T, Özcan İT. The impact of imputation procedures with machine learning methods on the performance of classifiers: an application to coronary artery disease data including missing values. Biomed Res 2018;29(13):2780–5. https://doi.org/10.4066/biomedicalresearch.29-18-199.
[47] Boursalie O, Samavi R, Doyle TE. Evaluation methodology for deep learning imputation models. Exp Biol Med NOV 2022;247(22):1972–87. https://doi.org/10.1177/15353702221121602.
[48] Bram DS, Nahum U, Atkinson A, Koch G, Pfister M. Evaluation of machine learning methods for covariate data imputation in pharmacometrics. CPT-Pharm Syst Pharmacol DEC 2022;11(12):1638–48. https://doi.org/10.1002/psp4.12874.
[49] Chang YW, et al. Neural network training with highly incomplete medical datasets. Mach Learn Sci Technol SEP 1 2022;3(3). https://doi.org/10.1088/2632-2153/ac7b69.
[50] Feng Y, Wang J, Wang Y, Chu X. Spatial-attention and demographic-augmented generative adversarial imputation network for population health data reconstruction. IEEE Trans Big Data 2022:1–14. https://doi.org/10.1109/TBDATA.2022.3227089.
[51] Huang XX, Cui GS, Wu D, Li Y. A semi-supervised approach for early identifying the abnormal carotid arteries using a modified variational autoencoder. In: Presented at the 2020 IEEE international conference on bioinformatics and biomedicine; 2020.
[52] Kabir S, Farrokhvar L. Non-linear missing data imputation for healthcare data via index-aware autoencoders. Health Care Manag Sci SEP 2022;25(3):484–97. https://doi.org/10.1007/s10729-022-09597-1.
[53] Kalweit M, Kalweit G, Boedecker J. AnyNets: Adaptive deep neural networks for medical data with missing values. In: CEUR workshop proceedings. 2926; 2021. p. 12–21 [Online]. Available, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85113823808&partnerID=40&md5=2c4893e2874af4411511ff655061e57c.
[54] Khan W, et al. Mixed data imputation using generative adversarial networks. IEEE Acces 2022;10:124475–90. https://doi.org/10.1109/ACCESS.2022.3218067.
[55] Macias E, Serrano J, Vicario JL, Morell A. Novel imputation method using average code from autoencoders in clinical data. In: European signal processing conference. 2021-January; 2021. p. 1576–9. https://doi.org/10.23919/Eusipco47968.2020.9287343.
[56] Neves DT, Alves J, Naik MG, Proenca AJ, Prasser F. From missing data imputation to data generation. J Comput Sci MAY 2022;61. https://doi.org/10.1016/j.jocs.2022.101640.
[57] Pan H, et al. "Discrete missing data imputation using multilayer perceptron and momentum gradient descent," (in English). Sensors (Basel, Switzerland) 2022;22(15). https://doi.org/10.3390/s22155645.
[58] Pereira RC, Abreu PH, Rodrigues PP. Partial multiple imputation with variational autoencoders: tackling not at randomness in healthcare data. IEEE J Biomed Health Inform 2022;26(8):4218–27. https://doi.org/10.1109/JBHI.2022.3172656.
[59] Psychogyios K, Ilias L, Askounis D, Ieee. Comparison of missing data imputation methods using the Framingham Heart study dataset. In: Presented at the 2022 IEEE-EMBS international conference on biomedical and health informatics (BHI) jointly organised with thE IEEE-EMBS international conference on wearable and implantable body sensor networks (BSN'22); 2022.
[60] Samad MD, Abrar S, Diawara N. Missing value estimation using clustering and deep learning within multiple imputation framework. Knowledge-Based Syst AUG 5 2022;249. https://doi.org/10.1016/j.knosys.2022.108968.
[61] Traynor C, Sahota T, Tomkinson H, Gonzalez-Garcia I, Evans N, Chappell M. Imputing biomarker status from RWE datasets-a comparative study. J Personal Med DEC 2021;11(12). https://doi.org/10.3390/jpm11121356.
[62] Ghazi MM, et al. Training recurrent neural networks robust to incomplete data: application to Alzheimer's disease progression modeling. Med Image Anal 2019;53:39–46. https://doi.org/10.1016/j.media.2019.01.004.
[63] Che ZP, et al. Recurrent neural networks for multivariate time series with missing values. Sci Rep 2018;8. https://doi.org/10.1038/s41598-018-24271-9.
[64] de Jong J, et al. Deep learning for clustering of multivariate clinical patient trajectories with missing values. GigaScience 2019;8(11). https://doi.org/10.1093/gigascience/giz134.
[65] Habiba M, Pearlmutter BA. Neural ODEs for informative missingess in multivariate time series. In: 2020 31st Irish Signals and Systems Conference (ISSC); 2020. p. 1–6.

[66] Tsiligkaridis T, Sloboda J. A multi-task LSTM framework for improved early sepsis prediction. 2020. p. 49–58.
[67] Zhao Y, Berretta M, Wang T, Chitnis T. GRU-DF: A temporal model with dynamic imputation for missing target values in longitudinal patient data. In: 2020 IEEE international conference on healthcare informatics, ICHI 2020; 2020.
[68] Jung W, Mulyadi AW, Suk HI, Jung W, Mulyadi AW, Suk H-I. Unified modeling of imputation, forecasting, and prediction for AD progression. In: Medical image computing and computer assisted intervention - MICCAI 2019, PT IV. vol. 11767; 2019. p. 168–76.
[69] Ma Q, Li S, Cottrell G. Adversarial joint-learning recurrent neural network for incomplete time series classification. IEEE Trans Pattern Anal Mach Intell 2020; 44(4):1765–76. https://doi.org/10.1109/TPAMI.2020.3027975.
[70] Mulyadi AW, Jun E, Suk H. Uncertainty-aware variational-recurrent imputation network for clinical time series. IEEE Trans Cybernet 2021;52(9):9684–94. https://doi.org/10.1109/TCYB.2021.3053599.
[71] Beaulieu-Jones BK, Moore JH. Missing data imputation in the electronic health record using deeply learned autoencoders. Pac Symp Biocomput 2016;22:207–18. https://doi.org/10.1142/9789813207813_0021.
[72] Bianchi FM, et al. Learning representations of multivariate time series with missing data. Pattern Recog 2019:96. https://doi.org/10.1016/j.patcog.2019.106973.
[73] Codella J, Sarker H, Chakraborty P, Ghalwash M, Yao Z, Sow D. EXITs: an ensemble approach for imputing missing EHR data. IEEE International Conference on Healthcare Informatics; 2019. p. 1–3. https://doi.org/10.1109/ICHI.2019.8904779.
[74] Fortuin V, et al. GP-VAE: deep probabilistic multivariate time series imputation.. International conference on artificial intelligence and statistics. 2020. p. 1651–61.
[75] Jun E, Mulyadi AW, Choi J, Suk H. Uncertainty-gated stochastic sequential model for EHR mortality prediction. IEEE Trans Neural Networks Learn Syst 2020;32(9): 4052–62. https://doi.org/10.1109/TNNLS.2020.3016670.
[76] Jun E, Mulyadi AW, Suk HI. Stochastic imputation and uncertainty-aware attention to EHR for mortality prediction. In: 2019 international joint conference on neural networks (IJCNN); 2019. p. 1–7. https://doi.org/10.1109/IJCNN.2019.8852132.
[77] Lin SW, et al. Filling missing values on wearable-sensory time series data. In: Proceedings of the 2020 SIAM international conference on data mining (SDM); 2020. p. 46–54.
[78] Yin C, Liu R, Zhang D, Zhang P. Identifying sepsis subphenotypes via time-aware multi-modal auto-encoder. In: Presented at the proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. CA, USA: Virtual Event; 2020. https://doi.org/10.1145/3394486.3403129.
[79] Tao HM, Deng QQ, Xiao SZ, Huamin T, Qiuqun D, Shanzhu X. Reconstruction of time series with missing value using 2D representation-based denoising autoencoder. J Syst Eng Electron 2020;31(6):1087–96. https://doi.org/10.23919/JSEE.2020.000081.
[80] Chen YP, Huang CH, Lo YH, Chen YY, Lai F. Combining attention with spectrum to handle missing values on time series data without imputation. Inform Sci 2022; 609:1271–87. https://doi.org/10.1016/j.ins.2022.07.124.
[81] Deshmukh A, Choudhary J, Singh DP. Multi kernel scaled deep time series imputation. In: 8th international conference on advanced computing and communication systems, ICACCS 2022; 2022. p. 829–34. https://doi.org/10.1109/ICACCS54159.2022.9784998.
[82] Farrell S, Mitnitski A, Rockwood K, Rutenberg A. Interpretable machine learning for high-dimensional trajectories of aging health. PLoS Comput Biol JAN 2022;18 (1). https://doi.org/10.1371/journal.pcbi.1009746.
[83] Getz K, Hubbard RA, Linn KA. Performance of multiple imputation using modern machine learning methods in electronic health records data. Epidemiology Mar 1 2023;34(2):206–15. https://doi.org/10.1097/ede.0000000000001578.
[84] Gordon D, Petousis P, Zheng H, Zamanzadeh D, Bui AAT. TSI-GNN: extending graph neural networks to handle missing data in temporal settings. Front Big Data SEP 15 2021;4. https://doi.org/10.3389/fdata.2021.693869.
[85] Haliduola HN, Bretz F, Mansmann U. Missing data imputation in clinical trials using recurrent neural network facilitated by clustering and oversampling. Biom J JUN 2022;64(5):863–82. https://doi.org/10.1002/bimj.202000393.
[86] Ho NH, Yang HJ, Kim J, Dao DP, Park HR, Pant S. Predicting progression of Alzheimer's disease using forward-to-backward bi-directional network with integrative imputation. Neural Netw 2022;150:422–39. https://doi.org/10.1016/j.neunet.2022.03.016.
[87] Lee Y, Jun E, Choi J, Suk HI. Multi-view integrative attention-based deep representation learning for irregular clinical time-series data. IEEE J Biomed Health Inform AUG 2022;26(8):4270–80. https://doi.org/10.1109/JBHI.2022.3172549.
[88] Li B, Shi Y, Cheng L, Yan Z, Wang X, Li H. MTSSP: Missing value imputation in multivariate time series for survival prediction. In: Proceedings of the international joint conference on neural networks. vol. 2022-July; 2022. https://doi.org/10.1109/IJCNN55064.2022.9892806.
[89] Liang W, Zhang K, Cao P, Liu X, Yang J, Zaiane O. Rethinking modeling Alzheimer's disease progression from a multi-task learning perspective with deep recurrent neural network. Comput Biol Med 2021;138. https://doi.org/10.1016/j.compbiomed.2021.104935.
[90] Liu Y, Qin S, Yepes AJ, Shao W, Zhang Z, Salim FD. Integrated convolutional and recurrent neural networks for health risk prediction using patient journey data with many missing values. In: Proceedings - 2022 IEEE international conference on bioinformatics and biomedicine, BIBM 2022; 2022. p. 1658–63. https://doi.org/10.1109/BIBM55620.2022.9995048.
[91] Liu Y, Qin S, Zhang Z, Shao W. Compound density networks for risk prediction using electronic health records. In: Proceedings - 2022 IEEE international conference on bioinformatics and biomedicine, BIBM 2022; 2022. p. 1078–85. https://doi.org/10.1109/BIBM55620.2022.9995587.
[92] Lu X, Yuan L, Li R, Xing Z, Yao N, Yu Y. An improved Bi-LSTM-based missing value imputation approach for pregnancy examination data. Algorithms 2023;16 (1). https://doi.org/10.3390/a16010012.
[93] Porta JI, Domínguez MA, Tamarit F. Automatic data imputation in time series processing using neural networks for industry and medical datasets. In: Communications in computer and information science. vol. 1577. CCIS; 2022. p. 3–16. https://doi.org/10.1007/978-3-031-04447-2_1.
[94] Wang Q, Ren S, Xia Y, Cao L. BiCMTS: Bidirectional coupled multivariate learning of irregular time series with missing values. In: International conference on information and knowledge management, proceedings; 2021. p. 3493–7. https://doi.org/10.1145/3459637.3482064.
[95] Yildiz AY, Koc E, Koc A. Multivariate time series imputation with transformers. IEEE Signal Process Lett 2022;29:2517–21. https://doi.org/10.1109/LSP.2022.3224880.
[96] Zamanzadeh DJ, et al. Autopopulus: A novel framework for autoencoder imputation on large clinical datasets. In: Proceedings of the annual international conference of the IEEE engineering in medicine and biology society, EMBS; 2021. p. 2303–9. https://doi.org/10.1109/EMBC46164.2021.9630135.
[97] Zhang K, Jiang X, Madadi M, Chen L, Savitz S, Shams S. DBNet: A novel deep learning framework for mechanical ventilation prediction using electronic health records. In: Proceedings of the 12th ACM conference on bioinformatics, computational biology, and health informatics, BCB 2021; 2021. https://doi.org/10.1145/3459930.3469551.
[98] Chen L, Xu J, Li SC. DeepMF: deciphering the latent patterns in omics profiles with a deep learning method. BMC Bioinforma 2019;20(23):1–13. https://doi.org/10.1186/s12859-019-3291-6.
[99] Mongia A, Sengupta D, Majumdar A. deepMc: deep matrix completion for imputation of single-cell RNA-seq data. J Comput Biol 2020;27(7):1011–9. https://doi.org/10.1089/cmb.2019.0278.
[100] Qiu YL, Zheng H, Gevaert O. Genomic data imputation with variational auto-encoders. GigaScience 2020;9(8). https://doi.org/10.1093/gigascience/giaa082.
[101] Badsha MB, et al. Imputation of single-cell gene expression with an autoencoder neural network. Quant Biol 2020;8(1):78–94. https://doi.org/10.1007/s40484-019-0192-7.
[102] Kinalis S, Nielsen FC, Winther O, Bagger FO. Deconvolution of autoencoders to learn biological regulatory modules from single cell mRNA sequencing data. BMC Bioinforma 2019;20:1–9. https://doi.org/10.1186/s12859-019-2952-9.
[103] Sun YV, Kardia SLR. Imputing missing genotypic data of single-nucleotide polymorphisms using neural networks. Eur J Hum Genet 2008;16(4):487–95. https://doi.org/10.1038/sj.ejhg.5201988.
[104] Tian T, Min MR, Wei Z, Tian T, Min MR, Wei Z. Model-based autoencoders for imputing discrete single-cell RNA-seq data. Methods 2021:192. https://doi.org/10.1016/j.ymeth.2020.09.010.
[105] Chen S, Yan X, Zheng R, Li M. Bubble: a fast single-cell RNA-seq imputation using an autoencoder constrained by bulk RNA-seq data. Brief Bioinform 2023;24(1). https://doi.org/10.1093/bib/bbac580.
[106] Chen SX, Xu C. Handling high-dimensional data with missing values by modern machine learning techniques. J Appl Stat 2023. https://doi.org/10.1080/02664763.2022.2068514.
[107] Dai ZY, Bu ZQ, Long Q. Multiple imputation via generative adversarial network for high-dimensional blockwise missing value problems. In: Presented at the 20th IEEE international conference on machine learning and applications (ICMLA 2021); 2021.
[108] Mahbub S, Sawmya S, Saha A, Reaz R, Rahman MS, Bayzid MS. Quartet based gene tree imputation using deep learning improves phylogenomic analyses despite missing data. J Comput Biol 2022;29(11):1156–72. https://doi.org/10.1089/cmb.2022.0212.
[109] Peacock S, Jacob E, Burlutskiy N. Coupling deep imputation with multitask learning for downstream tasks on omics data. In: Proceedings of the international joint conference on neural networks. vol. 2022-July; 2022. https://doi.org/10.1109/IJCNN55064.2022.9892749.
[110] Zhang WJ, Yang W, Talburt J, Weissman S, Yang MQ, Ieee. Missing Value Recovery for Single Cell RNA Sequencing Data. In: Presented at the 2021 international conference on computational science and computational intelligence (CSCI 2021); 2021.
[111] Zhang X, et al. NISC: neural network-imputation for single-cell RNA sequencing and cell type clustering. Front Genet MAY 3 2022;13. https://doi.org/10.3389/fgene.2022.847112.
[112] Zhou KY, Kottoori BS, Munj SA, Zhang ZW, Draghici S, Arslanturk S. Integration of multimodal data from disparate sources for identifying disease subtypes. Biology-Basel MAR 2022;11(3). https://doi.org/10.3390/biology11030360.
[113] Xia Y, et al. Recovering from missing data in population imaging - Cardiac MR image imputation via conditional generative adversarial nets. Med Image Anal 2021;67. https://doi.org/10.1016/j.media.2020.101812.
[114] Pan YS, et al. Synthesizing missing PET from MRI with cycle-consistent generative adversarial networks for Alzheimer's disease diagnosis. In: Medical image computing and computer assisted intervention, PT III. vol. 11072; 2018. p. 455–63.
[115] Lee D, Kim J, Moon WJ, Ye JC. Collagan: Collaborative gan for missing image data imputation. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition. vol. 2019; 2019. p. 2482–91.

[116] Pan YS, et al. Spatially-constrained fisher representation for brain disease identification with incomplete multi-modal neuroimages. IEEE Trans Med Imaging 2020;39(9):2965–75. https://doi.org/10.1109/TMI.2020.2983085.
[117] Peng L, et al. Longitudinal prediction of infant MR images with multi-contrast perceptual adversarial learning. Front Neurosci 2021;15. https://doi.org/10.3389/fnins.2021.653213.
[118] James BD, Leurgans SE, Hebert LE, Scherr PA, Yaffe K, Bennett DA. Contribution of Alzheimer disease to mortality in the United States. Neurology Mar 25 2014;82(12):1045–50. https://doi.org/10.1212/wnl.0000000000000240.
[119] Jang JH, et al. Deep learning approach for imputation of missing values in actigraphy data: algorithm development study. JMIR Mhealth Uhealth 2020;8(7). https://doi.org/10.2196/16113.
[120] Saeed A, Ozcelebi T, Lukkien J. Synthesizing and reconstructing missing sensory modalities in behavioral context recognition. Sensors (Basel) 2018;18(9). https://doi.org/10.3390/s18092967.
[121] Feng T, Narayanan S. Imputing missing data in large-scale multivariate biomedical wearable recordings using bidirectional recurrent neural networks with temporal activation regularization. In: 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC), 23–27 July 2019; 2019. p. 2529–34. https://doi.org/10.1109/EMBC.2019.8856966.
[122] Miller D, et al. Physiological waveform imputation of missing data using convolutional autoencoders. In: 2018 IEEE 20th international conference on E-health networking, applications and services (HEALTHCOM); 2018.
[123] Lee W, Lee J, Kim Y. Contextual imputation with missing sequence of EEG signals using generative adversarial networks. IEEE Acces 2021;9:151753–65. https://doi.org/10.1109/ACCESS.2021.3126345.
[124] Calhas D, Henriques R. fMRI multiple missing values imputation regularized by a recurrent denoiser. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). vol. 12721; 2021. p. 25–35.
[125] Akramifard H, Balafar MA, Razavi SN, Ramli AR. Early detection of Alzheimer's disease based on clinical trials, three-dimensional imaging data, and personal information using autoencoders. J Med Signals Sens Apr-Jun 2021;11(2):120–30. https://doi.org/10.4103/jmss.JMSS_11_20.
[126] Kim JC, Chung K, Kim J-C, Chung K. Multi-modal stacked denoising autoencoder for handling missing data in healthcare big data. IEEE Acces 2020;8:104933–43. https://doi.org/10.1109/ACCESS.2020.2997255.
[127] Kim JC, Chung K, Kim J-C, Chung K. Hybrid multi-modal deep learning using collaborative concat layer in health bigdata. IEEE Acces 2020;8:192469–80. https://doi.org/10.1109/ACCESS.2020.3031762.
[128] Thung KH, Yap PT, Shen DG, Thung K-H, Yap P-T, Shen D. Multi-stage diagnosis of Alzheimer's disease with incomplete Multimodal data via multi-task deep learning. In: Deep learning in medical image analysis and multimodal learning for clinical DECISION support. vol. 10553; 2017. p. 160–8.
[129] Vivar G, et al. Simultaneous imputation and classification using Multigraph Geometric Matrix Completion (MGMC): application to neurodegenerative disease classification. Artif Intell Med 2021;117. https://doi.org/10.1016/j.artmed.2021.102097.
[130] Jabason E, Ahmad MO, Swamy MNS. Missing structural and clinical features imputation for semi-supervised Alzheimer's disease classification using stacked sparse autoencoder. In: 2018 IEEE biomedical circuits and systems conference (BioCAS); 2018. p. 1–4. https://doi.org/10.1109/BIOCAS.2018.8584844.
[131] Li D, Lyons P, Klaus J, Gage B, Kollef M, Lu C. Integrating static and time-series data in deep recurrent models for oncology early warning systems. In: International conference on information and knowledge management, proceedings; 2021. p. 913–22. https://doi.org/10.1145/3459637.3482441.
[132] Xu YM, et al. Explainable dynamic multimodal variational autoencoder for the prediction of patients with suspected central precocious puberty. IEEE J Biomed Health Inform MAR 2022;26(3):1362–73. https://doi.org/10.1109/JBHI.2021.3103271.
[133] Fan M, et al. A deep matrix completion method for imputing missing histological data in breast cancer by integrating DCE-MRI radiomics. Med Phys 2021;48(12):7685–97. https://doi.org/10.1002/mp.15316.
[134] Mirza M, Osindero S. Conditional generative adversarial nets. arXiv e-prints, pp. arXiv:1411.1784-arXiv:1411.1784. 2014.
[135] Zhu J-Y, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE International Conference on Computer Vision 2017:2223–32.
[136] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv e-prints, pp. arXiv:1810.04805-arXiv:1810.04805. 2018.
[137] Lee J, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2019;36(4):1234–40. https://doi.org/10.1093/bioinformatics/btz682.
[138] Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digital Med 2021;4(1):86. https://doi.org/10.1038/s41746-021-00455-y.
[139] Luo Y, Cai X, Zhang Y, Xu J, Xiaojie Y. Multivariate time series imputation with generative adversarial networks. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in neural information processing systems. vol. 31. Curran Associates, Inc; 2018 [Online]. Available, https://proceedings.neurips.cc/paper/2018/file/96b9bff013acedfb1d140579e2fbeb63-Paper.pdf.
[140] Lee CH, Yoon H-J. Medical big data: promise and challenges. Kidney Res Clin Pract 2017;36(1):3–11. https://doi.org/10.23876/j.krcp.2017.36.1.3.
[141] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press; 2016.