How Deep is your Guess? A Fresh Perspective on Deep Learning for Medical Time-Series Imputation

Linglong Qian 1,4,5† Tao Wang 1,4 Jun Wang 2 Hugh Logan Ellis 1 Robin Mitra 3 Richard Dobson 1,6 Zina Ibrahim 1,4,5,6†

- ¹ Department of Biostatistics and Health Informatics, King's College London
- Department of Computer Science, University of Warwick
 Department of Statistics, University College London
 King's AI Institute
 PyPOTS Research
- ⁶ Institute of Health Inforantics, Unviversity College London

Abstract

We introduce a novel classification framework for time-series imputation using deep learning, with a particular focus on clinical data. By identifying conceptual gaps in the literature and existing reviews, we devise a taxonomy grounded on the inductive bias of neural imputation frameworks, resulting in a classification of existing deep imputation strategies based on their suitability for specific imputation scenarios and dataspecific properties. Our review further examines the existing methodologies employed to benchmark deep imputation models, evaluating their effectiveness in capturing the missingness scenarios found in clinical data and emphasising the importance of reconciling mathematical abstraction with clinical insights. Our classification aims to serve as a guide for researchers to facilitate the selection of appropriate deep learning imputation techniques tailored to their specific clinical data. Our novel perspective also highlights the significance of bridging the gap between computational methodologies and medical insights to achieve clinically sound imputation models.

1 Introduction

In key areas such as finance [78], healthcare [25] and weather forecasting [26], predictive analytics frequently rely on extensive, diverse, and multimodal time series data. These datasets are complex, exhibiting characteristics such as

skewed and long-tailed distributions and typically comprise a large number of multimodal and interrelated variables sampled at varying frequencies. The complexity is further compounded by the prevalence of non-random missingness, which is informative of the underlying data structure and highly influences the quality of predictive models performing downstream tasks.

Recent advances in deep learning techniques have outperformed traditional statistical and machine learning imputation methods when applied to large and heterogeneous time-series datasets [91]. These models, herein referred to as *deep imputers*, do not make strong assumptions about the underlying distribution of the data and learn directly from the data itself, allowing them to capture complex patterns and relationships without being constrained by predefined distributions. Such flexibility is particularly valuable in dealing with large and heterogeneous time-series datasets, where the underlying distribution may be unknown or highly variable.

The literature contains a wide variety of deep imputers based on various architectures including convolutional neural networks (CNNs) [35], recurrent neural networks (RNNs) [77], and multi-layer perceptrons (MLPs) [2]. Several recent reviews exist, categorising existing methods in a number of ways. For example, [27] classifies deep imputers by the type of missingness handled (e.g., missing completely at random, missing at random or missing not at random) and the neural architectures employed. In contrast, [91] classifies imputers based on their abilities to model and quantify stochasticity. Furthermore, because of unique characteristics and challenges in healthcare datasets, some literature reviews have focused on surveys specific to medical time series such as [44] and [48]; particularly on benchmarking performance across multiple datasets and exploring how different deep imputers perform across different data types and levels of correlations among the variables modelled.

Examining the literature on deep imputers, particularly those benchmarking model performance using medical datasets, often shows disparities in reported performance depending on the task, architecture, implementation approach and training data [24]. Our review has also uncovered a variety of data processing and masking techniques used to simulate missingness during experimental evaluation; those are yet to be systematically studied, leaving a critical gap in our understanding of the factors leading to the performance metrics reported in existing literature reviews. More importantly, the current level of scrutiny of the various types of deep imputers is insufficient to make an informed choice of the appropriateness of a given model for a specific task or dataset. To our knowledge, no comprehensive review thoroughly examines deep imputers beyond their architecture and general performance metrics.

We have therefore endeavoured to ask the following questions to support the assessment of the suitability of a given deep imputer for a given task: a) what are the characteristics of a given multivariate time series that make a particular category of deep imputers more suitable than others? b) what is the effect of the data processing steps adopted by a given model on the rigor of evaluation and subsequent performance? c) what are the distinct properties of medical time-series data and how effectively does the current paradigm of deep imputers account for them?

Our review builds on the latest review of deep imputers in the context of electronic health records (EHRs) [48] by offering a road map that connects the characteristics of deep imputers to the mechanisms by which they influence model performance for a given dataset with different types of missingness. We present a taxonomy grounded in the concept of **inductive bias** [6], which refers to the preferences, priors or assumptions a deep learning model inherently makes about the data, guiding its generalisation from training data to unseen test data. Moreover, we examine the extent to which data processing and experimental methods reflect real-world missingness in healthcare datasets and investigate current gaps in the benchmarking techniques found in the literature. Our goal is to enhance the understanding of deep imputation mechanisms, improve model and architecture selection, and refine evaluation methodologies for data with specific attributes, thereby addressing the current shortcomings in deep learning imputation research for EHR data.

2 Background

2.1 Characteristics of EHR Data

The intended use of EHR data to support clinical care and inform treatment decisions introduces a unique set of characteristics that directly impact the imputation process. Specifically, EHR time series are inherently multimodal, capturing patient health dynamics through **continuous measurements** (e.g., heart rate), **discrete events** (e.g., medication changes) and **ordinal data** (e.g., cancer stages). These multimodal variables are recorded asynchronously at intervals which vary according to clinical needs. For example, heart rate is sampled more frequently than capillary blood glucose, and laboratory tests are conducted when clinicians anticipate a need [42].

As the variables recorded in EHRs jointly capture a patient's treatment journey, they exhibit multiple forms of dependencies. Many variables are **cross-sectionally correlated**, which can introduce redundancy by overemphasising certain features. For example, clinical practice often dictates ordering test panels and laboratory tests are seldom requested in isolation; ordering electrolyte tests typically includes kidney function tests, and calcium tests require concurrent albumin measurements to adjust calcium levels [68]. The correlations are particularly problematic given that clinical insights often lie within the extreme values of **skewed distributions** of clinical variables, such as abnormally high or low blood sugar. An imputation model, therefore, needs to accurately discern those informative outliers which mark physiological events, from noise.

The dependencies among EHR time-series also extend across the temporal dimension, encompassing **short-term and long-term temporal dependencies** that hold significant meaning within a patient's trajectory. Immediate physiological responses reveal acute body reactions, while prolonged interventions or chronic conditions manifest long-term effects. For example, a heart

attack will result in immediate-term changes in blood pressure and heart rate, intermediate-term changes in renal function, and long-term varying effects on many features of interest if it leads to heart failure. Accurately modelling these temporal dependencies requires understanding the sequential patterns in EHRs, and balancing the impact of historical health events against recent ones. Moreover, temporal locality, which refers to physiological events occurring closely together in time, plays significant roles in clinical diagnosis. For instance, recurring patterns of irregular heartbeats persisting over short periods are characteristic of atrial fibrillation. Finally, many variables adhere to the notion of temporal invariance, whereby certain physiological patterns maintain clinical relevance regardless of the specific moment in the patient's timeline and regardless of temporal shift. For example, the presence of elevated troponin levels in blood tests consistently indicates heart injury, regardless of when the test is conducted within the patient's clinical timeline. Here, the imputation must maintain the salience of the variables adhering to temporal invariance.

Finally, medical time-series distributions are often highly-skewed; they exhibit a natural class imbalance where the amount of training data available for a given outcome of interest is generally low. To illustrate, consider a warning system for in-hospital cardiac arrest, which requires training on patient records whose hospital stays culminate in a cardiac arrest. Cardiac arrest incidence is estimated to be as low as 2.3% of intensive care unit admissions [3], which makes the target population a minority with much less training data available compared to the majority (no cardiac arrest) class. Class imbalance is further amplified by the variability in clinical presentations of patients with the same disease [59], making the group of interest (those with a given clinical presentation) only a minority in any patient population.

2.2 EHR Missingness Beyond Traditional Frameworks

The complex characteristics of EHR data raise the question: how can one effectively model missingness and develop imputation strategies that preserve the dependencies and clinical meaning of the data? Following Rubin's classification [75], missing data mechanisms are traditionally classified into Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR), reflecting the relationship between missingness in both observed and unobserved data. While Rubin's framework remains fundamental in missing data analysis and continues to be used for classifying imputation models in the latest literature reviews [48,91], it fails to fully capture the complexities of EHR data. In EHRs, missingness often results from the documentation practices involved in routine patient care and may patient-specific insights [7]. For example, a lab test might be missing because it was not ordered for a given patient, or because normal results are not typically recorded. Such practices blur the distinction between missing and observed variables [33,90] and complicate the practicality of the categorical distinctions of MCAR, MAR and MNAR [33].

A more pertinent issue in EHRs is **structured missingness** [62], where the large volumes of heterogeneous and multimodal data, along with specific modes

of data collection, cause missing values to exhibit associations and structural patterns. This non-random, multivariate associative pattern of missing values fundamentally hinders large-scale machine learning. In EHRs, structured missingness naturally arises from the asynchronous and decision-driven nature of healthcare data collection [90]. Over 60% of EHR data is missing not at random [54] due to irregular sampling intervals dictated by clinical decisions and carry meaningful insights.

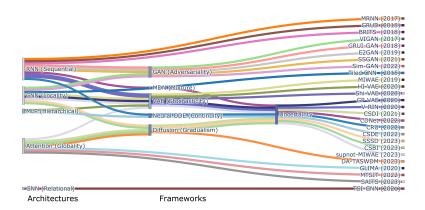


Figure 1: Taxonomy of Deep Imputers. On the left, the foundational neural network architectures are listed: RNN, GNN, Attention mechanisms, CNN, and MLP. The middle section classifies these architectures into broader categories based on their imputation frameworks, such as GAN, MDN, Neural OD, VAE and Diffusion models. Architectures and frameworks are labeled accordingly. The right section of the diagram lists deep imputation models developed using these methodologies, like TSI-GNN, MRNN, BRITS, Sim-GAN, and GP-VAE, among others.

3 Taxonomy of Deep Imputers

Our multidimensional exploration of the literature on deep imputers is depicted in Figure 1 and further detailed in Table 1. Our taxonomy is based on the following principles:

1) The effectiveness of an imputation model in interpreting the complexity and missingness patterns of a given dataset relies on the fundamental connection between the dataset's characteristics and the model's **inductive bias** [32]. Inductive bias refers to the set of assumptions, preferences, or constraints that

guide the learning process to reduce the space of possible solutions by prioritising one solution over another, independent of the observed data. This significantly influences the model's behaviour, generalisation capabilities, robustness, and ultimately shapes the resulting imputations. Inductive bias can take several forms related to the complexity of the learned representation, the underlying data distribution, or preferences regarding the learned parameters [32]. Since our review focuses on understanding the link between existing deep imputers and specific data properties and tasks, we highlight two types of inductive bias: preference bias, which dictates a model's assumptions in hypothesis selection by prioritising certain functions over others, effectively shaping its preferences for certain types, patterns, or relationships, and uncertainty bias, which dictates how a model accounts for uncertainty in its produced output.

2) Modern deep imputers are a combination of neural architectures and frameworks. A network's architecture dictates the physical structure and design of the neural network itself, including the arrangement of layers, the type of neurons used, and how these neurons are connected, e.g. convolutional networks. On the other hand, a framework dictates how this structure is employed and trained to perform the imputation task. Frameworks are higher-level constructs that define the algorithmic approach that leverages a given neural architecture, e.g. a recurrent architecture could be trained within an encoder-decoder or a generative adversarial network (GAN) framework. Each architecture and framework has its own inductive bias, shaping the model's approach to handling missing data and influencing its performance in different scenarios.

We first examine existing imputation **architectures**, focusing on the inductive biases that influence their generalisation capabilities and effectiveness in addressing various types of missingness, data structures, and task complexities. We then distinguish between architecture and the conceptual or mathematical **frameworks** that utilise a neural network for imputation. Here, inductive bias dictates how a framework approaches specific challenges within the imputation task, shaping the effectiveness of different models across varied scenarios. Finally, different frameworks vary in their approaches to **handling uncertainty** in the resulting imputations. Here, a model's inductive bias reflects the methods employed to incorporate uncertainty into its imputation outputs, for example, whether uncertainty can be directly modelled through probabilistic means or variability in the model's outputs.

In our subsequent discussion, we outline the general inductive biases associated with each class of architectures and frameworks. For each class, we detail the deep imputation models found in the literature, discussing their specific inductive biases. We also highlight the latest developments aimed at enhancing these models to handle data that falls outside their inherent biases, as well as known areas for improvement and ongoing research. Our granular distinction aims to illuminate the strengths and limitations of existing methods for specific datasets, providing a roadmap to assess a model's suitability for various imputation contexts.

3.1 Neural Network Architectures

3.1.1 Recurrent Neural Networks (RNNs)

are intuitively fit for handling temporal sequences [60]. Their inherent inductive bias favours learning temporal dependencies between variables over time, as a result of the recurrent nature of their architecture, which allows them to maintain an internal state across time steps. This bias aligns with the sequential and dynamic nature of medical event sequences, enabling RNNs to effectively capture patient trajectories.

Vanilla RNNs are especially suited for capturing short-term temporal correlations in EHRs, but struggle with modeling long-term dependencies, especially when faced with known EHR issues such as irregular sampling and diverse record lengths. To address these, modified RNN architectures have been introduced in imputers, e.g., GRUD [11] which incorporates temporal decay into its architecture effectively managing time-sensitive missing data. Subsequent models, including MRNN [103] and BRITS [10] integrate solutions to handle irregular sampling into their architectures for more accurate and contextually relevant data imputation. Despite these advances, RNN imputers have yet to accommodate mixed variable types and multi-dimensional time-series [101]. Developing more sophisticated RNN variants that can seamlessly model the complex interplay between continuous and discrete variables within medical time series data remains a crucial next step.

3.1.2 Convolutional Neural Networks (CNNs)

possess an inductive bias towards capturing local patterns, embodying the principle of temporal locality for detecting acute physiological changes that can mark important clinical events, e.g. tachycardia. One-dimensional (1D) CNNs excel in identifying these pivotal moments by focusing on short-term local variations within the data [46], thus facilitating prompt clinical intervention and effective patient monitoring.

The exploration of two-dimensional (2D) CNNs extends the utility of CNNs to capture complex spatial relations across variables [37]. Innovations such as TimesNet [95] and Tiled CNNs [93] employ techniques like Gramian Angular Fields to encode multiple time series as images and leverage spatial correlations. This approach allows CNNs to unearth patterns that span multiple time points and variables, effectively revealing relationships characteristic of physiological signals. For example, multi-channel 2D CNNs can concurrently analyse ECG readings, respiratory rates, and oxygen saturation levels, offering a comprehensive representation of patient health status. However, while these models provide significant insights, they also introduce challenges related to the complexity of transforming and interpreting time series in 2D spaces, potentially obscuring the temporal sequence and causality inherent in the data. Therefore, making the transition from 1D to 2D CNNs requires careful consideration [85] to ensure that temporal information is preserved and accurately represented across time-series.

3.1.3 Transformers

deviate from traditional sequence processing methods by capturing long-range dependencies through self-attention, enabling the representation of global contextual relationships within data [88]. The intrinsic inductive bias of Transformers towards comprehensive contextuality aligns well with the multifaceted, long-sequence semantics of medical data, enhancing the recognition of complex, multivariate temporal patterns of patient trajectories across extended time frames. The global perspective of Transformers is particularly advantageous in identifying subtle, yet clinically significant patterns that might be overlooked by models with a narrower focus. However, adapting the Transformer to medical time series data requires bespoke modifications to preserve the strict sequential integrity that defines these datasets [13, 104] and to capture equally-useful short-term temporal associations [21]. Models such as DeepMVI [4], NRTSI [79], SAITS [23], GLIMA [84], and MTSIT [102] and CrosFormer [106] incorporate temporal encoding and locality-enhanced attention to maintain sequential coherence while leveraging global dependencies. However, the success of these sophisticated Transformer-based approaches heavily relies on the breadth and depth of available data. The scarcity of comprehensive, publicly accessible medical time series datasets remains a significant barrier, underlining the necessity for expanded data resources for healthcare analytics.

3.1.4 Graph Neural Networks (GNNs)

[107] stand out for their ability to model the complex relational structures prevalent in medical data, e.g., modeling the dependencies between health indicators such as heart rate, blood pressure, and laboratory results. The foundational inductive bias of GNNs towards capturing spatial relationships allows these models to represent health variables and their interdependencies as nodes and edges within a graph. This representation not only facilitates the preservation of temporal sequences but also enriches the representation to a broader context derived from interconnected health parameters.

Several works have demonstrated the application of GNNs to address the challenges of medical time series imputation. Models like GACN [100] interweave Graph Attention Networks (GAT) [89] and temporal convolution layers to model spatio-temporal dependencies within the imputation process. Similarly, SPIN [57] and GRIN [19] leverage multi-layered attention mechanisms and graph recurrent imputation networks to enhance initial time series representations by using complex relational insights before imputation. Advanced architectures such as AGRN [14] and MDGCN [47] utilise bidirectional recurrent structures, integrating graph convolution with RNN cells to capture spatio-temporal relations. Here, the translation of time series into effective graph structures presents considerable challenges, notably the absence of natural graph representations in traditional time series datasets [105]. Although recent models like TSI-GNN [31] explored incorporating temporal information into bipartite graphs through an extension of graph representation learning, constructing and interpreting these

graph models require significant computational efforts and domain knowledge, limiting their scalability and practicality. Furthermore, seamlessly incorporating temporal information into the GNN framework has proven difficult [96], especially aligning the static nature of graphs with the dynamic progression of medical time series [43].

Table 1: Chronological overview of deep learning models specialized for medical time series imputation.

Model	Year	Architecture	Framework	Inductive bias	Uncertainty Quantification	Missingness Mechanisms
MRNN [103]	2017	RNN	-	Sequential	×	MAR
GRUD [11]	2018	RNN	-	Sequential	×	MCAR, MNAR
BRITS [10]	2018	RNN	-	Sequential	×	MAR
Tiled CNN [93]	2015	CNN	-	Locality	×	MAR
GLIMA [84]	2020	Attention	-	Globality	×	MCAR, MAR
MTSIT [102]	2022	Attention	-	Globality	×	MCAR, MAR
SAITS [23]	2023	Attention	-	Globality	×	MCAR
TSI-GNN [31]	2021	GNN	-	Relational	×	MCAR
MIWAE [58]	2019	CNN	VAE	Locality, Stochasticity	×	MCAR, MAR
GP-VAE [28]	2020	CNN	VAE	Locality, Stochasticity	<i>V</i>	MCAR, MAR, MNAR
V-RIN [63]	2020	RNN	VAE	Sequential, Stochasticity	<i>V</i>	MCAR, MAR
HI-VAE [64]	2020	MLP	VAE	Stochasticity	×	MCAR
Shi-VAE [5]	2022	RNN	VAE	Sequential, Stochasticity	×	MAR
supnot-MIWAE [45]	2023	CNN, Attention	VAE	Locality, Globality, Stochasticity	~	MNAR
CDNet [51]	2022	RNN	MDN	Sequential, Mixture	V	=
VIGAN [80]	2017	CNN	GAN	Locality, Adversariality	×	-
GRUI-GAN [52]	2018	RNN	GAN	Sequential, Adversariality	×	=
E^2 GAN [53]	2019	RNN	GAN	Sequential, Adversariality	×	=
SSGAN [61]	2021	RNN	GAN	Sequential, Adversariality	×	MCAR
Sim-GAN [67]	2022	CNN	GAN	Locality, Adversariality	×	-
CSDI [86]	2021	Attention	Diffusion	Globality, Gradualism	V	MCAR, MAR, MNAR
SSSD [1]	2023	CNN	Diffusion	Locality, Gradualism	✓	MCAR, MAR, MNAR
CSBI [15]	2023	CNN, Attention	Diffusion	Locality, Globality, Gradualism	✓	MAR
DA-TASWDM [97]	2023	Attention	Diffusion	Globality, Gradualism	×	MAR
CRU [76]	2022	RNN	Neural ODE	Sequential, Continuity	V	MAR
CSDE [66]	2022	MLP	Neural ODE	Continuity	✓	=

3.2 Learning Frameworks

In modern deep imputers, a network architecture provides the learning mechanisms for a learning framework, which guides the imputation process towards plausible generalisations of the data to capture and replicate complex data distributions. This synergy ensures that the generated data reflects realistic and clinically relevant patterns. Additionally, neural frameworks offer various approaches to quantifying confidence in the resulting imputation, crucial given that accurate modeling of EHR data directly impacts subsequent downstream tasks and clinical decision-making. Different neural frameworks employ varied paradigms for representing and handling uncertainty, which we include in our discussion to explore the diverse strategies for managing the inherent unpredictability of healthcare data.

3.2.1 Variational Autoencoders (VAEs)

consist of an encoder and a decoder network. The encoder maps input data to a latent space distribution, while the decoder reconstructs data from this distribution. During training, VAEs optimise a variational lower bound on the log-likelihood of the data, ensuring the latent space captures key features of the input distribution. Using an expressive neural network architecture, VAEs

effectively learn representations that reflect common properties of EHRs, such as skewness and multimodality.

Inductive Bias & Handling Uncertainty: VAEs assume that data is generated from a latent space with a known, usually Gaussian, distribution [56] and aim to learn a distribution which captures the underlying structure of the data. As such, VAEs are inherently probabilistic models and provide a measure of uncertainty by learning the probability distribution of the latent space.

State of the Art: VAE variants such as the Heterogeneous Incomplete Variational Autoencoder (HI-VAE) [64], Mixed VAE (VAEM) [55], and MIWAE [58] handle missingness in diverse data types. GP-VAE [28] is a known early imputation model based on a Gaussian process. GP-VAE's performance plummets with heterogeneity in observations and extended missingness. V-RIN [63] aims to bypass GP-VAE's distribution-related imputation bias by incorporating an uncertainty-aware Gated Recurrent Unit (GRU) to blend temporal dynamics with the imputed data. Supnot-MIVAE [45] extends this approach by introducing an additional classifier to refine the evidence lower bounds, enhancing imputation accuracy for classification tasks. Shi-VAE [5] further expands these capabilities by including LSTMs for better temporal structure handling and effectively addressing missing data episodes.

Limitations: The success of VAEs depends on their ability to create meaningful data representations that align with their assumed distribution. With EHR data, this may lead to oversimplifications which obscure vital clinical subtleties. While hybrid VAE models such as V-RIN and Shi-VA bypass the distribution problem by incorporating temporal dynamics, these models face significant challenges in producing interpretable, clinically relevant outputs. Furthermore, the computational intensity for training VAEs, especially when integrating temporal dynamics, remains a barrier for their wide adoption for large medical datasets.

3.2.2 Mixture Density Networks (MDNs)

[9] combine the predictive power of deep neural networks with the probabilistic precision of mixture models to model the conditional probability distribution of targets based on inputs. MDNs comprise a neural network architecture that outputs parameters of a mixture model (e.g., mean and variance, mixture weights) conditioned on the input data. This enables MDNs to predict a range of possible outcomes or data paths.

Inductive Bias & Handling Uncertainty: MDNs assume that the data is generated from a mixture of probability distributions. This bias towards probabilistic rather than singular outcome representations aligns with the high variance in outcomes and treatment responses observed in medical time series. Consequently, MDNs can directly capture uncertainty through a mixture of weights and variances in the assumed distributions.

State of the Art: A highly-performing example is CDNet [51], which effectively models imputed feature distributions and addresses the heterogeneity and irregularity of EHR data by integrating an MDN with a GRU and a

Regularized Attention Network (RAN). In CDNet, the GRU captures time dependencies, while the MDN handles latent variable sampling through a mix of neural networks and distributions. This setup allows CDNet to potentially capture complex relationships within EHR data.

Limitations: A key challenge in all MDN models is the ability to optimally configure the model to fully exploit its theoretical potential for capturing complex relationships while avoiding overfitting and maintaining the interpretability and clinical relevance of the output.

3.2.3 Generative Adversarial Networks (GANs)

[17] establish a framework where two neural networks, a generator and a discriminator, compete in a zero-sum minimax game. The generator is tasked with replicating real data distributions to produce synthetic data samples that are indistinguishable from real data, while the discriminator learns to differentiate between real and synthetic samples. The training process involves iteratively updating the generator and discriminator networks to improve the quality of generated samples.

Inductive Bias & Handling Uncertainty: GANs have an inductive bias towards generating realistically diverse data distributions, as the generator aims to fool the discriminator. This bias enables the adversarial model to effectively generate and impute incomplete multivariate time-series. However, GANs inherently lack direct mechanisms to quantify uncertainty within the imputations, and their application to establishing confidence in the generated data is still in its early stages [65].

State of the Art: Two prominent GAN examples found in the literature are GRUI-GAN [52] and E^2 GAN [53]. GRUI-GAN employs a modified GRU to account for the temporal irregularity of incomplete time series. The model adapts the GRU in both the discriminator and generator to learn the distribution of the entire dataset, the implicit relationships between observations, and the temporal information of the dataset. In the second phase, the input 'noise' of the GAN's generator is trained so that the generated time series closely resembles the original incomplete time series, increasing the likelihood of highquality generated data. However, optimising the noise vector of GRUI-GAN has proven difficult [53]. To address this, E^2 GAN integrates the GAN structure with a denoising autoencoder, streamlining the imputation by bypassing direct noise optimisation. Advances continue with frameworks like NAOMI [50], which adopts a non-autoregressive approach to minimise cumulative errors in extended sequences, offering a more robust solution for datasets with significant missingness. Additionally, SSGAN [61] enhances the GAN paradigm by incorporating elements such as a temporal reminder matrix and additional classification layers to improve imputation quality.

Limitation: GANs have the potential to greatly enhance medical research by creating diverse and comprehensive datasets, including those representing unrepresented conditions and groups. Nevertheless, their inability to quantify confidence in the generated data necessitates additional methods, adding

complexity to their practical use. The adversarial training process, though innovative, can induce instabilities like "mode collapse" observed in NAOMI and Sim-GAN, where the generator produces limited and repetitive outputs instead of capturing the full diversity of the data. Such limitations are especially significant in clinical settings, where the reliability of data and predictions is paramount.

3.2.4 Diffusion Models

[39,82] employ a stochastic process to gradually generate synthetic data, progressing from random noise towards distributions that mimic the observed data. During training, diffusion models learn to reverse this process by denoising synthetic samples to match observed data.

Inductive Bias & Handling Uncertainty: Diffusion models assume that data evolves over time according to a diffusion process, inherently biasing them towards solutions that mimic this generating process and away from those that do not. Although diffusion models do not provide an explicit mechanism to quantify uncertainty, they iteratively incorporate noise into the diffusion process, allowing for the generation of stochastic samples that reflect the variability in the data. Additionally, diffusion models can estimate uncertainty by measuring the divergence between predicted and observed data distributions at each time step.

Examples of diffusion models include NETRATE State of the Art: [30] and MedDiff [38], which integrate temporal dynamics and domain-specific knowledge into the diffusion process to better handle time-dependent variations and complex medical scenarios. CSDI [86] leverages observed data subsets to guide the generation process but encounters scalability issues due to the quadratic complexity induced by the transformer-based architecture [81]. To tackle this, SSSD [1] substitutes transformers with structured state-space models [34], while CSBI [15] and MIDM [92] enhance efficiency and accuracy by modelling the diffusion process as a Schrödinger bridge problem [18] and sampling noise from the conditional distribution of observed representations. PriSTI [49] and DA-TASWDM [97] further push the boundaries by integrating spatio-temporal dependencies and dynamic temporal relationships. SPD [8] represents a paradigm shift by modelling time series from a continuous perspective, better aligning with the stochastic and irregular nature of medical data timelines.

Limitations: In most diffusion models, computational efficiency, clinical relevance, and accuracy remain challenges. While models like PriSTI and DATASWDM mark significant advances in personalised and contextually relevant imputations, the lack of straightforward and explicit mechanisms to quantify and communicate uncertainty significantly hampers their practical utility.

3.2.5 Neural Ordinary Differential Equations (Neural ODEs)

embed a neural network modeled by some function ${\bf f}$ into an ODE framework [12], which describes the temporal evolution of the data using differential equations. Once trained, an ODE model uses the learned function to simulate the dynamics of the data over time or make predictions about future states based on current observations.

Inductive Bias & Handling Uncertainty: ODE models assume that the underlying temporal evolution of the data data can be described by a system of ordinary differential equations. This bias favours continuous data transitions aligning with the functions learned by the model [66,76], enabling Neural ODEs to capture gradual transitions within patient timelines, potentially overcoming the issue of irregular sampling of EHR data [73]. These models do not directly facilitate uncertainty measurement, but uncertainty can be incorporated as a stochastic process into the differential equations, allowing for the propagation of uncertainty through time. Additionally, ODE models can estimate uncertainty by comparing model predictions with observed data and adjusting model parameters to minimise the discrepancy.

State of the Art: The utility of Neural ODEs have been demonstrated through various extensions and applications. ODE-GRU-D [36] extends GRU-D by using an ODE solver to precisely decipher the decay dynamics within time series, thus refining control over decay rates compared to the original GRU-D. Additionally, CRU [76] offers a probabilistic recurrent framework for irregularly sampled time series, utilising a linear stochastic differential equation (SDE) [87] within a latent space structure, thereby incorporating the analytic solutions of continuous-discrete Kalman filter [94] formulations with medical time series analysis. Similarly, CSDE [66] presents a novel probabilistic framework that overcomes the limitations of traditional dynamic models by incorporating Markov dynamic programming [40] and multi-conditional forward-backward losses, facilitating rigorous training and ensuring theoretical optimality.

Limitations: Solving the differential equations of Neural ODEs is computationally demanding and is sensitive to the initial set conditions. Furthermore, the robustness and domain relevance of Neural ODE-based imputations is critically reliant on the model's capability to accurately capture the underlying dynamics from medical datasets, which may be compounded by data sparsity. The sophisticated mathematical underpinnings of Neural ODEs can deter clinical applicability due to the abstract nature of their outputs, making it challenging for clinicians to derive clear, actionable insights.

4 Mind the Gaps

Having established our taxonomy of the state-of-the-art deep imputers, we now critically appraise the practical aspects of the current paradigm, particularly issues that have a direct impact on clinical utility. We first explore experimental design, identify issues that contribute to inconsistencies in model evaluation,

and highlight visible gaps between the capabilities of existing deep imputers and the specific requirements of the medical domain. We then turn our discussion to model reliability, particularly our ability to quantify one's uncertainty in the resulting imputation. Here, we highlight the importance of post-hoc uncertainty quantification methods, particularly for models based on deterministic architectures. Finally, this section highlights the potential for integrating clinical insight with the imputation process to ensure that the generated values align with clinical protocols and are both statistically plausible and clinically meaningful. Our discussions underscore the need for more rigorous, transparent, and comprehensive approaches to performance and reliability evaluation, particularly in handling complex missingness patterns and data distributions.

4.1 Mind the Masking Gap

During experimental evaluation of a deep imputer, *masking* is used to simulate incomplete data conditions by designating certain data points as missing during training and evaluation. Masking provides a controlled way to test how an algorithm handles incomplete datasets and is thus essential for performance evaluation. Our examination of the literature identified a wide discrepancy in the preprocessing steps employed by different models and potential misalignments between the masking techniques used and the missingness assumptions the models are designed to handle. Our observations are summarised below.

Misalignment with Missingness Assumptions: As shown in Table 1, deep imputers have been designed to recognise different flavours of missingness (MCAR, MAR, MNAR). During experimental evaluation, however, all models shown in Table 1 use random masking (Figure 2 (a)) to generate missing datasets, predominantly producing MCAR scenarios. This approach oversimplifies the correlations embedded within the EHR time-series, which are reflected in complex missingness patterns across time and cross-sectionally. As discussed in section 2.1, these patterns arise from the underlying physiological processes and recording practices of clinical workers.

Interestingly, the literature contains masking techniques that can capture spatio-temporal MNAR missingness patterns of medical datasets [20], including temporal masking (Figure 2 (b)), which captures missingness patterns over time, spatial masking (Figure 2 (c)), which captures cross-sectional missingness and block masking (Figure 2 (d)), which combines the two to concurrently capture different flavours of temporal and cross-sectional correlations and dependencies. Despite their direct applicability to biomedical domains, they are rarely used to evaluate imputation models operating on medical datasets. The only examples using spatio-temporal masking of time-series come from the traffic domain [47, 99].

The above problem is exasperated by the lack of information in published work. With the exception of BRITS and CSDI, the use of random masking is not mentioned in the experimental design, and one must examine the accompanying code to discern it. While the use of random masking facilitates model evaluation, it contrasts with the complex and informative MNAR patterns ob-

served in real-world EHRs [29] which many deep imputers have been designed to address. The discrepancy between the theoretical model and experimental evaluation technique therefore highly undermines a deep imputer's capacity, leaving it under-evaluated.

Under-reporting of Masking Pipelines: There are significant discrepancies and under-reporting of when masking is introduced during experimental evaluation. Data could be pre-masked before being ingested by the model or masked dynamically during the training phase. Traditional pre-masking methods, while more straightforward, limit the model's training to incomplete datasets, reducing its ability to learn from the entire range of clinical features and associated dependencies. In contrast, adopting in-mini-batch masking strategies promises a more dynamic approach by iteratively masking different subsets of the same dataset across training epochs. However, this approach risks overfitting, as the model may become too focused on the artificial missing patterns and fail to recognise the original data structures. Therefore, the decision of when masking is introduced can have a profound impact on the model's capacity to interpret the diverse missing patterns found in a given dataset [70]. Despite the potential impact on the results, this aspect of the experimental design is not reported in most deep imputers discussed in this survey, except for BRITS and GRUD, which mask before training, and CSDI and STAITS, which use in-mini-batch masking during training.

Overlooked Design Decisions There are other decisions that highly influence the resulting masked data but are not discussed in most of the deep imputation literature. An important issue is the methodology used to implement masking. Generally, masking can be implemented using overlaying [22] or augmenting [16] as shown in Figures 2 e-f. Overlaying involves adding artificial missingness in addition to the original missingness the dataset contains, while augmentations only mask complete data, separating the artificial missingness generated from the original missingness. Choosing the type of masking has consequences during model training and evaluation. Although overlaying exposes the model to a broader array of missing data scenarios leading to more robust training and effective imputation strategies, it requires complex evaluation processes and increases the risk of overfitting. On the other hand, augmenting simplifies the model's learning process by allowing it to learn from the artificially introduced missingness without interference from the original missing patterns, but may not fully equip the model to handle the intricate missingness patterns in real-world data. In addition to the above, there is growing evidence that multiple masking, which refers to repeatedly applying masking operations to generate diverse examples during model training, is an effective strategy for improving imputation performance [72]. However, it is unclear how any of the deep imputers implement masking, creating a big gap in our understanding of the rigour of the evaluation techniques, especially in models designed to accommodate non-random EHR missingness.

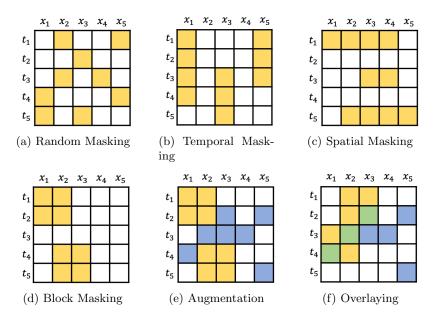


Figure 2: Masking techniques and approaches demonstrated over a time-series of five features (x_1-x_5) and five time points (t_1-t_5) : (a) random masking, (b) temporal masking, (c) spatial masking, (d) block masking. The yellow cells indicate those labeled as missing via masking. In (e) augmentation and (f) overlaying, the blue cells indicate cells that are missing within the original data. In (e), the masked (yellow) cells have no overlap with the original missingness in the data. Green: masked data coming from both the original missingness and artificial missingness. In (f), overlaying masks cells from either the original missingness or simulates artificial missingness from non-missing data.

4.2 Mind the Uncertainty Gap

Accurate imputation is crucial for downstream prediction tasks. In addition, the ability to quantify one's confidence in the resulting imputation enables understanding the limitations and reliability of the generated data. The correlation between imputation accuracy and uncertainty is subtle and model-dependent and has been shown to weaken with increased data diversity [41]. Therefore, a dedicated component for quantifying imputation uncertainty is particularly critical for high-stake medical downstream tasks. For example, [11] demonstrated that accurate imputation in GRUD significantly improves the prediction of patient outcomes in the ICU. Similarly, [28] showed that the ability of GP-VAE to quantify uncertainty improves the interpretability and trustworthiness of imputed laboratory tests by clinical staff.

The Current State of Uncertainty Quantification: The importance of establishing one's confidence in the resulting imputation is not currently reflected in the state-of-the-art of medical deep imputers. Neural architectures such as RNNs, CNNs, and transformers are different flavours of deterministic imputers with no inherent capabilities to quantify uncertainty. Highly performing models such as GRUD [11] and BRITS [10] effectively handle temporal dependencies and irregular sampling, but do not measure uncertainty. Although generative frameworks such as VAEs and MDNs are naturally probabilistic and can provide measures of confidence in the resulting imputation, they are computationally complex. Moreover, their uncertainty quantification mechanisms rely on distribution-specific assumptions and are highly dependent on the models' inductive bias, limiting the ability to generate interpretable insights that are directly applicable to the complex missingness patterns observed in medical time series.

Need for Post-Hoc Uncertainty Quantification: Given the diversity of deep imputers and the distribution-dependent nature of non-deterministic models, there is a need for general post-hoc uncertainty quantification mechanisms that can be utilised regardless of the imputer's underlying architecture or framework. Such independent components will allow uncertainty quantification after training, avoiding potential performance degradation and associated model complexities. These methods can employ model-agnostic uncertainty estimates independent of the inductive bias of the imputation model, enhancing imputation robustness and reliability across different models and datasets.

There is a small but growing number of work proposing post-hoc uncertainty quantification for efficient and effective deep imputation architectures. A prominent example is DEARI [69], which extends BRITS by integrating a self-attention mechanism to enhance imputation accuracy and employs a post-hoc Bayesian marginalization strategy to provide reliable uncertainty bounds. CF-RNN [83] adapts the inductive conformal prediction framework to time-series forecasting, which constructs prediction intervals that may potentially undercover the response when conditioned on certain missing patterns. Another approach which remains unexplored is to adopt a multiple imputation framework [74], which involves creating multiple imputed datasets from the pre-

dictive distribution conditional on the observed data and combining the results to account for uncertainty. Applying the principles of multiple imputations in conjunction with deep imputers can enhance post-hoc uncertainty quantification by generating diverse imputation scenarios and ensuring robust estimates. This approach, though not immediately obvious in its application to deep learning models, merits investigation to improve the scalability and reliability of imputation models.

4.3 Mind the Knowledge Gap

While medicine is data-rich, it is also knowledge-rich and acknowledging the existing body of clinical knowledge can have a great impact on the reliability and interpretability of the imputed data. The incorporation of domain knowledge into neural network architectures can have a direct influence on alleviating the possible discrepancies between the imputer's inductive bias and the patterns embedded within the data. For instance, RNNs can be fine-tuned with clinical temporal patterns to capture treatment effects or disease progression, while CNNs can be adapted to embed clinical significance into spatial-temporal relationships. This idea is recognised by a few recent imputation attempts such as [71] where conditional statements derived from clinical guidelines are used to guide the imputation process, ensuring that the generated values are not only statistically plausible but also clinically meaningful. Another example is [98] where signal temporal logic (STL) is used to define clinically meaningful temporal patterns used as a dictionary to guide the training process, greatly improving the alignment between established clinical protocols and model output. These efforts, however, are a minority and the importance of domain knowledge has not made its mark on the deep imputation models most cited in the literature.

Knowledge to Enhance Performance in Skewed Distributions: Incorporating medical knowledge into the imputation process can also alleviate the issue of *skewed distributions* and class imbalance prevalent in medical timeseries. By embedding domain knowledge, imputation models can ensure that imputed values for infrequent events, such as in-hospital cardiac arrests, are clinically plausible and contextually appropriate. This approach reduces the risk of biased imputations that favor the majority class, ultimately leading to more reliable and clinically relevant predictions. Incorporating medical knowledge can also improve the model's ability to handle the variability in clinical presentations, ensuring that imputed data maintains its integrity across diverse patient populations.

5 Conclusion

In this review, we have examined deep learning imputation approaches for EHR data, highlighting the relationship between the inductive biases of these models and the distinct characteristics of medical time-series. By analysing these biases, we emphasise the need to balance mathematical abstraction with clinical

insights to improve the interpretability and applicability of imputation techniques. We identify issues in data masking, timing, implementation, and the need for post-hoc uncertainty quantification and highlight the importance of incorporating domain and expert knowledge in our current paradigm, revealing gaps in our current research and identifying avenues for further research. This study highlights the challenges and potential innovations of bridging data science and clinical practice. Our goal is to develop models that better align with clinical complexities, promoting advancements that are both methodologically robust and clinically relevant.

Acknowledgments

This paper represents independent research funded by the NIHR Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust, the EPSRC Centre for Doctoral Training in Data-Driven Health (DRIVE-Health) and King's College London. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. The work of LQ is supported by the Kings-China Scholarship Council PhD Scholarship Programme (K-CSC) under Grant CSC202008060096. ZI is supported by Innovate UK under grant 10104845. TW is supported by the NIHR Maudsley Biomedical Research Centre at South London, Maudsley Charity and Early Career Research Award from the Institute of Psychiatry, Psychology & Neuroscience. ZI and RD were supported by in part by the NIHR Biomedical Research Centre at SLaM, in part by Kings College London, London, U.K., and in part by the NIHR University College London Hospitals Biomedical Research Centre. H.L.E is supported by a Research Fellowship Award from the Department of Medicine at Dalhousie University and DRIVE-Health. RM's research on Structured Missingness has been supported by the Turing-Roche Strategic Partnership.

References

- [1] Juan Lopez Alcaraz and Nils Strodthoff. Diffusion-based time series imputation and forecasting with structured state space models. *Transactions on Machine Learning Research*, 2022.
- [2] Silvia Arber, John J Hunter, John Ross Jr, Minoru Hongo, Gilles Sansig, Jacques Borg, Jean-Claude Perriard, Kenneth R Chien, and Pico Caroni. Mlp-deficient mice exhibit a disruption of cardiac cytoarchitectural organization, dilated cardiomyopathy, and heart failure. Cell, 88(3):393–403, 1997.
- [3] Richard A Armstrong, Caroline Kane, Fiona Oglesby, Katie Barnard, Jasmeet Soar, and Matt Thomas. The incidence of cardiac arrest in the in-

- tensive care unit: A systematic review and meta-analysis. *Journal of the Intensive Care Society*, 20(2):144–154, 2019.
- [4] Parikshit Bansal, Prathamesh Deshpande, and Sunita Sarawagi. Missing value imputation on multidimensional time series. *Proceedings of the VLDB Endowment*, 14(11):2533–2545, 2021.
- [5] Daniel Barrejón, Pablo M Olmos, and Antonio Artés-Rodríguez. Medical data wrangling with sequential variational autoencoders. *IEEE Journal* of Biomedical and Health Informatics, 26(6):2737–2745, 2021.
- [6] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261, 2018.
- [7] Jeroen Berrevoets, Fergus Imrie, Trent Kyono, James Jordon, and Mihaela Van der Schaar. To impute or not to impute? missing data in treatment effect estimation. In *International Conference on Artificial Intelligence* and Statistics, pages 3568–3590. PMLR, 2023.
- [8] Marin Biloš, Kashif Rasul, Anderson Schneider, Yuriy Nevmyvaka, and Stephan Günnemann. Modeling temporal data as continuous functions with stochastic process diffusion. In *International Conference on Machine Learning*, pages 2452–2470. PMLR, 2023.
- [9] Christopher M Bishop. Mixture density networks. 1994.
- [10] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems*, 31, 2018.
- [11] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):1–12, 2018.
- [12] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. Advances in neural information processing systems, 31, 2018.
- [13] Si-An Chen, Chun-Liang Li, Nate Yoder, Sercan O Arik, and Tomas Pfister. Tsmixer: An all-mlp architecture for time series forecasting. arXiv preprint arXiv:2303.06053, 2023.
- [14] Yakun Chen, Zihao Li, Chao Yang, Xianzhi Wang, Guodong Long, and Guandong Xu. Adaptive graph recurrent network for multivariate time series imputation. In *International Conference on Neural Information Processing*, pages 64–73. Springer, 2022.

- [15] Yu Chen, Wei Deng, Shikai Fang, Fengpei Li, Nicole Tianjiao Yang, Yikai Zhang, Kashif Rasul, Shandian Zhe, Anderson Schneider, and Yuriy Nevmyvaka. Provably convergent schrödinger bridge with applications to probabilistic time series imputation. In *International Conference on Ma*chine Learning, pages 4485–4513. PMLR, 2023.
- [16] Tae-Min Choi, Ji-Su Kang, and Jong-Hwan Kim. Rdis: Random drop imputation with self-training for incomplete time series data. *IEEE Access*, 2023.
- [17] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.
- [18] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. Advances in Neural Information Processing Systems, 34:17695– 17709, 2021.
- [19] Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4027–4035, 2021.
- [20] Min Deng, Zide Fan, Qiliang Liu, and Jianya Gong. A hybrid method for interpolating missing data in heterogeneous spatio-temporal datasets. *ISPRS International Journal of Geo-Information*, 5(2), 2016.
- [21] Shengdong Du, Tianrui Li, Yan Yang, and Shi-Jinn Horng. Multivariate time series forecasting via attention-based encoder—decoder framework. Neurocomputing, 388:269–279, 2020.
- [22] Wenjie Du. Pypots: A python toolbox for data mining on partially-observed time series. arXiv preprint arXiv:2305.18811, 2023.
- [23] Wenjie Du, David Côté, and Yan Liu. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, 2023.
- [24] Wenjie Du, Jun Wang, Linglong Qian, Yiyuan Yang, Fanxing Liu, Zepu Wang, Zina Ibrahim, Haoxin Liu, Zhiyuan Zhao, Yingjie Zhou, et al. Tsi-bench: Benchmarking time series imputation. arXiv preprint arXiv:2406.12747, 2024.
- [25] Jose Roberto et al. Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *Journal of Biomedical Informatics*, 101:103337, 2020.
- [26] Martin Schultz et al. Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A*, 379(2194):20200097, 2021.

- [27] Chenguang Fang and Chen Wang. Time series data imputation: A survey on deep learning approaches. arXiv preprint arXiv:2011.11347, 2020.
- [28] Vincent Fortuin, Dmitry Baranchuk, Gunnar Rätsch, and Stephan Mandt. Gp-vae: Deep probabilistic time series imputation. In *International conference on artificial intelligence and statistics*, pages 1651–1661. PMLR, 2020.
- [29] Pedro J García-Laencina, José-Luis Sancho-Gómez, and Aníbal R Figueiras-Vidal. Pattern classification with missing data: a review. Neural Computing and Applications, 19:263–282, 2010.
- [30] Manuel Gomez-Rodriguez, David Balduzzi, and Bernhard Schölkopf. Uncovering the temporal dynamics of diffusion networks. In Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 July 2, 2011, 2011.
- [31] David Gordon, Panayiotis Petousis, Henry Zheng, Davina Zamanzadeh, and Alex AT Bui. Tsi-gnn: Extending graph neural networks to handle missing data in temporal settings. *Frontiers in big Data*, 4:693869, 2021.
- [32] A Goyal and Bengiom Y. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society*, 478:20210068, 2022.
- [33] John W Graham and John W Graham. Analysis of missing data. *Missing data: Analysis and design*, pages 47–69, 2012.
- [34] Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2021.
- [35] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern recognition*, 77:354–377, 2018.
- [36] Mansura Habiba and Barak A Pearlmutter. Neural odes for informative missingess in multivariate time series. In 2020 31st Irish Signals and Systems Conference (ISSC), pages 1–6. IEEE, 2020.
- [37] Nima Hatami, Yann Gavet, and Johan Debayle. Classification of timeseries images using deep convolutional neural networks. In *Tenth interna*tional conference on machine vision (ICMV 2017), volume 10696, pages 242–249. SPIE, 2018.
- [38] Huan He, Shifan Zhao, Yuanzhe Xi, and Joyce C Ho. Meddiff: Generating electronic health records using accelerated denoising diffusion model. arXiv preprint arXiv:2302.04355, 2023.

- [39] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [40] Ronald A Howard. Dynamic programming and markov processes. 1960.
- [41] Gabriele Incorvaia, Darryl Hond, and Hamid Asgari. Uncertainty quantification of machine learning model performance via anomaly-based dataset dissimilarity measures. *Electronics*, 13(5), 2024.
- [42] Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.
- [43] Ming Jin, Huan Yee Koh, Qingsong Wen, Daniele Zambon, Cesare Alippi, Geoffrey I Webb, Irwin King, and Shirui Pan. A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection. arXiv preprint arXiv:2307.03759, 2023.
- [44] Maksims Kazijevs and Manar D. Samad. Deep imputation of missing values in time series health data: A review with benchmarking. *Journal* of *Biomedical Informatics*, 144:104440, 2023.
- [45] SeungHyun Kim, Hyunsu Kim, Eunggu Yun, Hwangrae Lee, Jaehun Lee, and Juho Lee. Probabilistic imputation for time-series classification with missing data. In *International Conference on Machine Learning*, pages 16654–16667. PMLR, 2023.
- [46] Moshe Kravchik and Asaf Shabtai. Detecting cyber attacks in industrial control systems using convolutional neural networks. In *Proceedings of the 2018 workshop on cyber-physical systems security and privacy*, pages 72–83, 2018.
- [47] Yuebing Liang, Zhan Zhao, and Lijun Sun. Memory-augmented dynamic graph convolution networks for traffic data imputation with diverse missing patterns. *Transportation Research Part C: Emerging Technologies*, 143:103826, 2022.
- [48] Mingxuan Liu, Siqi Li, Han Yuan, Marcus Eng Hock Ong, Yilin Ning, Feng Xie, Seyed Ehsan Saffari, Yuqing Shang, Victor Volovici, Bibhas Chakraborty, and Nan Liu. Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. Artificial Intelligence in Medicine, 142:102587, 2023.
- [49] Mingzhe Liu, Han Huang, Hao Feng, Leilei Sun, Bowen Du, and Yanjie Fu. Pristi: A conditional diffusion framework for spatiotemporal imputation. In 2023 IEEE 39th International Conference on Data Engineering (ICDE), pages 1927–1939. IEEE, 2023.

- [50] Yukai Liu, Rose Yu, Stephan Zheng, Eric Zhan, and Yisong Yue. Naomi: Non-autoregressive multiresolution sequence imputation. Advances in neural information processing systems, 32, 2019.
- [51] Yuxi Liu, Shaowen Qin, Zhenhao Zhang, and Wei Shao. Compound density networks for risk prediction using electronic health records. In 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 1078–1085. IEEE, 2022.
- [52] Yonghong Luo, Xiangrui Cai, Ying Zhang, Jun Xu, et al. Multivariate time series imputation with generative adversarial networks. *Advances in neural information processing systems*, 31, 2018.
- [53] Yonghong Luo, Ying Zhang, Xiangrui Cai, and Xiaojie Yuan. E2gan: Endto-end generative adversarial network for multivariate time series imputation. In *Proceedings of the 28th international joint conference on artificial intelligence*, pages 3094–3100. AAAI Press Palo Alto, CA, USA, 2019.
- [54] Yuan Luo. Evaluating the state of the art in missing data imputation for clinical data. *Briefings in Bioinformatics*, 23(1):bbab489, 2022.
- [55] Chao Ma, Sebastian Tschiatschek, Richard Turner, José Miguel Hernández-Lobato, and Cheng Zhang. Vaem: a deep generative model for heterogeneous mixed type data. Advances in Neural Information Processing Systems, 33:11237–11247, 2020.
- [56] David JC MacKay et al. Introduction to gaussian processes. NATO ASI series F computer and systems sciences, 168:133–166, 1998.
- [57] Ivan Marisca, Andrea Cini, and Cesare Alippi. Learning to reconstruct missing data from spatiotemporal graphs with sparse observations. Advances in Neural Information Processing Systems, 35:32069–32082, 2022.
- [58] Pierre-Alexandre Mattei and Jes Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International conference* on machine learning, pages 4413–4423. PMLR, 2019.
- [59] Maciej A Mazurowski, Piotr A Habas, Jacek M Zurada, Joseph Y Lo, Jay A Baker, and Georgia D Tourassi. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural networks, 21(2-3):427–436, 2008.
- [60] Larry R Medsker and LC Jain. Recurrent neural networks. Design and Applications, 5(64-67):2, 2001.
- [61] Xiaoye Miao, Yangyang Wu, Jun Wang, Yunjun Gao, Xudong Mao, and Jianwei Yin. Generative semi-supervised learning for multivariate time series imputation. In *Proceedings of the AAAI conference on artificial* intelligence, volume 35, pages 8983–8991, 2021.

- [62] Robin Mitra, Sarah F McGough, Tapabrata Chakraborti, Chris Holmes, Ryan Copping, Niels Hagenbuch, Stefanie Biedermann, Jack Noonan, Brieuc Lehmann, Aditi Shenvi, et al. Learning from data with structured missingness. *Nature Machine Intelligence*, 5(1):13–23, 2023.
- [63] Ahmad Wisnu Mulyadi, Eunji Jun, and Heung-Il Suk. Uncertainty-aware variational-recurrent imputation network for clinical time series. *IEEE Transactions on Cybernetics*, 52(9):9684–9694, 2021.
- [64] Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, 107:107501, 2020.
- [65] Philipp Oberdiek, Gernot Fink, and Matthias Rottmann. Uqgan: A unified model for uncertainty quantification of deep classifiers trained via conditional gans. Advances in Neural Information Processing Systems, 35:21371–21385, 2022.
- [66] Sung Woo Park, Kyungjae Lee, and Junseok Kwon. Neural markov controlled sde: Stochastic optimization for continuous-time data. In *International Conference on Learning Representations*, 2021.
- [67] Soumen Kumar Pati, Manan Kumar Gupta, Rinita Shai, Ayan Banerjee, and Arijit Ghosh. Missing value estimation of microarray data using simgan. Knowledge and Information Systems, 64(10):2661–2687, 2022.
- [68] Rimma Pivovarov, David J Albers, Jorge L Sepulveda, and Noémie Elhadad. Identifying and mitigating biases in ehr laboratory tests. *Journal* of biomedical informatics, 51:24–34, 2014.
- [69] Linglong Qian, Zina Ibrahim, and Richard Dobson. Uncertainty-aware deep attention recurrent neural network for heterogeneous time series imputation. arXiv preprint arXiv:2401.02258, 2024.
- [70] Linglong Qian, Zina Ibrahim, Wenjie Du, Yiyuan Yang, and Richard JB Dobson. Unveiling the secrets: How masking strategies shape time series imputation. arXiv preprint arXiv:2405.17508, 2024.
- [71] Linglong Qian, Zina Ibrahim, Hugh Logan Ellis, Ao Zhang, Yuezhou Zhang, Tao Wang, and Richard Dobson. Knowledge enhanced conditional imputation for healthcare time-series, 2024.
- [72] Linglong Qian, Zina M. Ibrahim, Ao Zhang, and Richard J. B. Dobson. Addressing class imbalance in electronic health records data imputation. In Proceedings of the 6th International Workshop on Knowledge Discovery from Healthcare Data co-located with 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023), Macao, China, August 20, 2023, volume 3479 of CEUR Workshop Proceedings. CEUR-WS.org, 2023.

- [73] Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.
- [74] DB Rubin. Multiple imputation for nonresponse in surveys1987john wiley & sons. *New York*, 1987.
- [75] Donald B Rubin. Inference and missing data. Biometrika, 63(3):581–592, 1976.
- [76] Mona Schirmer, Mazin Eltayeb, Stefan Lessmann, and Maja Rudolph. Modeling irregular time series with continuous recurrent units. In *International Conference on Machine Learning*, pages 19388–19405. PMLR, 2022.
- [77] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.
- [78] Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. Applied Soft Computing, 90:106181, 2020.
- [79] Siyuan Shan, Yang Li, and Junier B Oliva. Nrtsi: Non-recurrent time series imputation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [80] Chao Shang, Aaron Palmer, Jiangwen Sun, Ko-Shin Chen, Jin Lu, and Jinbo Bi. Vigan: Missing view imputation with generative adversarial networks. In 2017 IEEE International conference on big data (Big Data), pages 766–775. IEEE, 2017.
- [81] Lifeng Shen and James Kwok. Non-autoregressive conditional diffusion models for time series prediction. arXiv preprint arXiv:2306.05043, 2023.
- [82] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.
- [83] Kamile Stankeviciute, Ahmed M Alaa, and Mihaela van der Schaar. Conformal time-series forecasting. Advances in neural information processing systems, 34:6216–6228, 2021.
- [84] Qiuling Suo, Weida Zhong, Guangxu Xun, Jianhui Sun, Changyou Chen, and Aidong Zhang. Glima: Global and local time series imputation with multi-directional attention learning. In 2020 IEEE International Conference on Biq Data (Biq Data), pages 798–807. IEEE, 2020.

- [85] Wensi Tang, Guodong Long, Lu Liu, Tianyi Zhou, Jing Jiang, and Michael Blumenstein. Rethinking 1d-cnn for time series classification: A stronger baseline. arXiv preprint arXiv:2002.10061, pages 1–7, 2020.
- [86] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. Advances in Neural Information Processing Systems, 34:24804– 24816, 2021.
- [87] Nicolaas G Van Kampen. Stochastic differential equations. Physics reports, 24(3):171–228, 1976.
- [88] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.
- [89] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.
- [90] Brian Wahl, Aline Cossy-Gantner, Stefan Germann, and Nina R Schwalbe. Artificial intelligence (ai) and global health: how can ai contribute to health in resource-poor settings? *BMJ global health*, 3(4), 2018.
- [91] Jun Wang, Wenjie Du, Wei Cao, Keli Zhang, Wenjia Wang, Yuxuan Liang, and Qingsong Wen. Deep learning for multivariate time series imputation: A survey. arXiv preprint arXiv:2402.04059, 2024.
- [92] Xu Wang, Hongbo Zhang, Pengkun Wang, Yudong Zhang, Binwu Wang, Zhengyang Zhou, and Yang Wang. An observed value consistent diffusion model for imputing missing values in multivariate time series. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 2409–2418, 2023.
- [93] Zhiguang Wang and Tim Oates. Imaging time-series to improve classification and imputation. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 3939–3945, 2015.
- [94] Greg Welch, Gary Bishop, et al. An introduction to the kalman filter. 1995.
- [95] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *The Eleventh International Conference on Learning Representations*, 2023.
- [96] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD*

- international conference on knowledge discovery \mathcal{C} data mining, pages 753–763, 2020.
- [97] Jingwen Xu, Fei Lyu, and Pong C Yuen. Density-aware temporal attentive step-wise diffusion model for medical time series imputation. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, pages 2836–2845, 2023.
- [98] Ruixuan Yan, Tengfei Ma, Achille Fokoue, Maria Chang, and Agung Julius. Neuro-symbolic models for interpretable time series classification using temporal logic description. In 2022 IEEE International Conference on Data Mining (ICDM), pages 618–627. IEEE, 2022.
- [99] Yongchao Ye, Shiyao Zhang, and James J. Q. Yu. Spatial-temporal traffic data imputation via graph attention convolutional network. In Igor Farkaš, Paolo Masulli, Sebastian Otte, and Stefan Wermter, editors, Artificial Neural Networks and Machine Learning – ICANN 2021, pages 241–252, 2021.
- [100] Yongchao Ye, Shiyao Zhang, and James JQ Yu. Spatial-temporal traffic data imputation via graph attention convolutional network. In *Interna*tional Conference on Artificial Neural Networks, pages 241–252. Springer, 2021.
- [101] Joonyoung Yi, Juhyuk Lee, Kwang Joon Kim, Sung Ju Hwang, and Eunho Yang. Why not to use zero imputation? correcting sparsity bias in training neural networks. In *Eighth International Conference on Learning Representations*, *ICLR 2020*. International Conference on Learning Representations, 2020.
- [102] A Yarkın Yıldız, Emirhan Koç, and Aykut Koç. Multivariate time series imputation with transformers. *IEEE Signal Processing Letters*, 29:2517– 2521, 2022.
- [103] Jinsung Yoon, William R Zame, and Mihaela van der Schaar. Multidirectional recurrent neural networks: A novel method for estimating missing data. In *Time series workshop in international conference on machine learning*, 2017.
- [104] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2023.
- [105] Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. In International Conference on Learning Representations, 2021.
- [106] Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*, 2022.

[107] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. AI open, 1:57– 81, 2020.