Addressing Class Imbalance in Electronic Health Records Data Imputation

Linglong Qian[1,*], Zina Ibrahim[1,*], Ao Zhang[2] and Richard JB Dobson[1,3,4,5]

[1]Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom
[2]Faculty of Natural, Mathematical & Engineering Sciences, King's College London, London, United Kingdom
[3]University College London, London, UK
[4]South London and Maudsley NHS Foundation Trust, London, UK
[5]Health Data Research UK London, University College London, London, UK

Abstract

Imputing missing values in imbalanced datasets remains an open challenge. Most methods assume data are missing at random or follow a standard distribution, lacking robustness for complex real-world data. Electronic health records exhibit severe class imbalance with non-random missingness, hindering model performance. We propose M^3-BRITS for greater scalability and flexibility, modeling temporal and cross-feature correlations to impute missing data, by optimizing sample similarity with deep metric learning for self-supervised learning. Evaluating imputation alone avoids reduced diversity and model bias from joint downstream tasks. Our model achieves superior performance to all baseline methods on four real-world datasets. This shows promise for increasing model scalability and flexibility to handle complex real-world data.

Keywords

Missingness, Imbalance, Imputation, Self-supervised learning

1. Introduction

The development of successful machine learning models, leveraging data procured from Electronic Health Records (EHRs), necessitates confronting the prevalent issue of data missingness inherent in this extensive information repository. Nevertheless, the assorted data types encapsulated within EHRs are gathered at variable time intervals, mirroring both clinical and administrative decisions enacted by healthcare practitioners to bolster patient care. Consequently, the data amalgamates both static and irregularly-sampled temporal data. For instance, vital signs like heart rate are often subject to regular monitoring, while the recording of white blood cell (WBC) count is not universally consistent across patients. This is primarily because it is infrequently ordered for clinically stable patients who are minimally suspected of harbouring an infection [1].

The complex phenomenon of missingness in EHRs poses significant challenges to imputation algorithms. Beyond the high rates of missing data, over 50% of EHR information is missing non-randomly [2, 3], making the volume of available training data for a specific outcome of interest typically sparse, irrespective of the task in question. This availability is greatly influenced by the accuracy of class labels.

To elucidate, consider a predictive system for in-hospital cardiac arrests, which necessitates training on patient records that ultimately result in a cardiac arrest. The incidence of cardiac arrest is estimated to be as low as 2.3% of intensive care unit admissions [4], thereby designating the target population as a minority, with significantly less training data available compared to the majority class (patients without a cardiac arrest).

Similarly, the clinical manifestations of patients diagnosed with the same disease can vary substantially [5], rendering the population of interest (those with a specific clinical presentation) a minority within any patient set. As such, class imbalance is intrinsic to tasks focused on EHRs [6], and each target group exhibits unique missing data patterns that reflect the divergences of groups and individuals from the standard distributions of measured variables.

These divergences carry profound clinical implications that must be considered to preserve the integrity of the imputed data as well as the associated task [7]. Consequently, the analysis and treatment of missingness in EHRs require meticulous attention to ensure accurate and meaningful machine learning model development.

The aforementioned complexities are well-documented in retrospective studies evaluating long-term missingness patterns in multi-center medical data. In these analyses, significant variation was

THE 6TH INTERNATIONAL WORKSHOP ON KNOWLEDGE DISCOVERY FROM HEALTHCARE DATA

[*]Corresponding author.
linglong.qian@kcl.ac.uk (L. Qian); zina.ibrahim@kcl.ac.uk (Z. Ibrahim); ao.zhang@kcl.ac.uk (A. Zhang); richard.j.dobson@kcl.ac.uk (R. J. Dobson)
0000-0003-3930-045X (L. Qian); 0000-0001-6203-2727 (Z. Ibrahim); 0000-0002-0825-1264 (A. Zhang); 0000-0003-4224-9245 (R. J. Dobson)

observed across tasks, variables, and time [3]. Such sophisticated attributes undoubtedly constrain the applicability of traditional statistical and machine learning imputation methods [8, 9]. These methods typically make strong assumptions about the data's originating distributions, markedly limiting the generalizability of imputation algorithms.

Deep learning (DL) models, encompassing convolutional neural networks (CNN) [10], recurrent neural networks (RNN) [11], and multi-layer perceptions (MLP) [12], have demonstrated success in estimating non-randomly missing values in temporal medical data. Notwithstanding, in all successful imputations, the network components (either recurrent or convolutional) were trained in tandem with the classification/regression component [13]. Consequently, the imputation task has been intimately linked with the downstream task, making it challenging to determine the imputer's contribution to the predictor's final performance.

Our study aims to address these gaps in managing missing data in Electronic Health Records (EHRs) by extending the RNN model BRITS [13]. BRITS captures non-random missingness patterns both historically and across features. We restructure imputation as a scalable, self-supervised learning task using multiple masking metric learning. This approach helps overcome issues related to skewed distributions while enabling scalability for complex EHR data imputation, accommodating non-random missingness, mixed data types, and uneven sampling.

Our model, M^3-BRITS, employs BRITS as a backbone and adaptively learns properties intrinsic to clinical time-series data through self-supervised deep metric learning. M^3-BRITS surpasses all existing imputation methods, including BRITS, in realistic complex scenarios.

2. Related Work

Efforts to impute multivariate time series data have resulted in numerous strategies. Among these, the GRUD model [14] incorporates temporal decay for missing data imputation, and its extensions, MRNN [15] and BRITS [13], capture temporal dynamics and missingness patterns across multiple features utilizing bidirectional RNNs. However, the MRNN model's limitation lies in treating imputed values as constants without sufficient updates during iterations. In contrast, BRITS, free from specific data assumptions, has exhibited superior performance across domains, indicating a need for enhanced approaches such as ours.

Existing temporal imputation methods struggle in managing class imbalances and disparate missingness distributions, often evident in clinical data [16, 17]. Strategies such as resampling [18], cost-sensitive [19], and ensemble learning [20] are deployed to address class imbalance, yet they inadequately handle uneven missingness distributions. This inadequacy leads us to consider deep metric learning [21], which enables intrinsic distribution recalibration via sample distance learning.

Deep metric learning, necessitating indications of similarity or dissimilarity, has shown success in handling complex data types and structures, as demonstrated by MeLIM [22] and DECADE [23]. Unlike traditional deep classification methods, this approach doesn't require class balance within a minibatch, offering a promising solution to data imputation challenges.

Models such as GRUU [24], V-RIN [25], and BRITS illustrate the simultaneous execution of imputation and downstream tasks within a single neural network, although the results are not consistently satisfactory. We argue that this integration might introduce model bias due to potential divergence in classification and imputation focuses. Consequently, our study concentrates solely on the imputation process, aiming to minimize such biases and enhance data restoration efficiency.

3. Terminology and Background

For a temporal interval observed over T discrete time-steps, we represent a multivariate time series as a matrix $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_T}\}$, composed of T observations. Each observation, denoted by $\boldsymbol{x_t} \in \mathbb{R}^D$, is a vector of D features. These features encapsulate various modalities, namely numerical (D_{num}), categorical (D_{cat}), static (D_{sta}), and dynamic (D_{dyn}) variables. It is crucial to note that $\mathbb{R}^D$ is heterogeneous, encompassing structured data types that extend beyond purely numerical features. This configuration allows for a comprehensive representation of the diverse data elements inherent in complex multivariate time series.

Information related to missing values is encapsulated within two derived matrices (see Fig. 1). The mask matrix $\boldsymbol{M} \in \mathbb{R}^{T \times D}$ indicates whether each element of $\boldsymbol{X}$ is observed or missing:

$$m_t^d = \begin{cases} 0, & \text{if } x_t^d \text{ is missing} \\ 1, & \text{otherwise} \end{cases} \tag{1}$$

Additionally, given that the time elapsed between consecutive observations can vary across the interval, we denote the time gaps at each time step t as δ_t. For features observed over time steps, δ_T represents the gap between the current time step (e.g., s_t) and the last observed value. Given the potential for non-uniform sampling across features in the data $\boldsymbol{X}$, there is a corresponding variability in δ_t. The $\delta \in \mathbb{R}^{N \times D}$ encodes the time gap between two successive observed values for each feature d, providing an additional indicator of temporal context to the dataset.

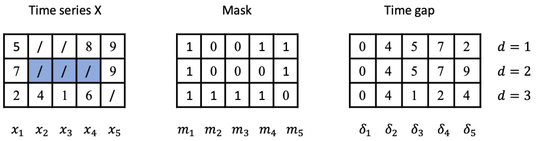

Figure 1: An example of multivariate time-series. x_{1-5}: observations in time steps t_1, ..., t_5 with corresponding time-stamps $s_{1-5} = 0, 4, 5, 7, 9$. Feature d_2 was missing during t_{2-4}, the last observation took place at s_1. Hence, $\delta_5^2 = t_5 - t_1 = 9 - 0 = 9$.

The definition of this indicator follows:

$$\delta_t^d = \begin{cases} s_t - s_{t-1} + \delta_{t-1}^d & \text{if } t > 1,\ m_t^d = 0 \\ s_t - s_{t-1} & \text{if } t > 1,\ m_t^d = 1 \\ 0 & \text{if } t = 1 \end{cases} \tag{2}$$

3.1. Overview of the BRITS Backbone

Our work extends BRITS's assumptions, emphasizing temporal and feature correlations embodied in $\boldsymbol{X}$. BRITS architecture, combining a fully-connected regression module and a recurrent component, applies temporal decay (Eq. (3)) and a decay factor to handle temporal correlations and adjust influence based on temporal distance. Missing values within an observation $\boldsymbol{x_t}$ are managed via a historical representation $\boldsymbol{\hat{x}_t}$ and a masking vector $\boldsymbol{m_t}$, producing a complement vector $\boldsymbol{x_t^{hc}}$ that accounts for missingness patterns (Eq. (3)-(6)).

$$\gamma_{th} = \exp\left(-\max(0, \boldsymbol{W_{\gamma h}}\delta_t + \boldsymbol{b_{\gamma h}})\right) \tag{3}$$

$$\boldsymbol{\hat{h}_{t-1}} = \boldsymbol{h_{t-1}} \odot \gamma_{th} \tag{4}$$

$$\boldsymbol{\hat{x}_t} = \boldsymbol{W_x}\hat{h}_{t-1} + \boldsymbol{b_x} \tag{5}$$

$$\boldsymbol{x_t^{hc}} = \boldsymbol{m_t} \odot \boldsymbol{x_t} + (1 - \boldsymbol{m_t}) \odot \boldsymbol{\hat{x}_t} \tag{6}$$

BRITS explores intra-observation correlations through a fully-connected layer, generating $\boldsymbol{x_t^{fc}}$, a feature-wise approximation of missing values (Eq. (7)). The concept of decay extends to feature space, resulting in a learnable factor, $\hat{\beta}t$, considering both temporal decay and the masking vector (Eq. (8)-(9)). This integration produces the imputed matrix $\boldsymbol{C_t}$, effectively combining observed and imputed data (Eq. (10)-(11)).

$$\boldsymbol{x_t^{fc}} = \boldsymbol{W_z}\boldsymbol{x_t^{hc}} + \boldsymbol{b_z} \tag{7}$$

$$\gamma_{tf} = \exp\left(-\max(0, \boldsymbol{W_{\gamma f}}\delta_t + \boldsymbol{b_{\gamma f}})\right) \tag{8}$$

$$\hat{\beta}_t = \sigma(\boldsymbol{W_\beta}[\gamma_{tf} \circ \boldsymbol{m_t}] + \boldsymbol{b_\beta}) \tag{9}$$

$$\boldsymbol{x_t^c} = \beta_t \odot \boldsymbol{x_t^{fc}} + (1 - \beta_t) \odot \boldsymbol{x_t^{hc}} \tag{10}$$

$$\boldsymbol{C_t} = \boldsymbol{m_t} \odot \boldsymbol{x_t} + (1 - \boldsymbol{m_t}) \odot \boldsymbol{x_t^c} \tag{11}$$

$$\boldsymbol{h_t} = \sigma(\boldsymbol{W_t}\boldsymbol{\hat{h}_{t-1}} + \boldsymbol{U_h}[\boldsymbol{C_t} \circ \boldsymbol{m_t}] + \boldsymbol{b_h}) \tag{12}$$

$$\boldsymbol{C_t^*} = \frac{{\boldsymbol{C_{t_F}} + \boldsymbol{C_{t_B}}}^T}{2} \tag{13}$$

The final step (Eq. (12)) updates the hidden state via Recurrent Neural Networks (RNNs), leveraging various indicators to learn functions of past observations. The bidirectional recurrent dynamics approach integrates backward information to tackle slow convergence, providing paired outputs $\boldsymbol{C_{t_F}}; \boldsymbol{C_{t_B}}$ and $\boldsymbol{h_{t_F}}; \boldsymbol{h_{t_B}}$ (Eq. (13)).

In essence, BRITS exploits temporal and feature correlations in multivariate time series data, employing decay factors, a regression module, and a bidirectional RNN for imputing missing values. The final hidden states are updated using imputations and corresponding masks, with the integrated processes visualized in Figure 2.

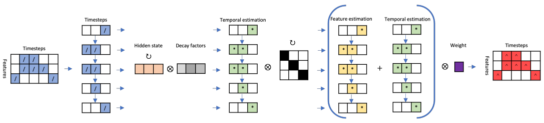

Figure 2: The backbone processes of BRITS

4. Methodology

We introduce the two components that facilitate scalable and flexible imputation through the use of M^3-BRITS, as illustrated in Figure 3. In real-world temporal data, the degree of heterogeneity varies across different domains. For instance, healthcare data often demonstrate significant class imbalance. Furthermore, the time-varying statistics of various types of data present substantial challenges for interpretation within a single location, a complexity that amplifies when extended to multiple locations. This makes the incorporation of all relevant prior knowledge into the analytical model impractical. Traditional metric learning techniques [26, 27] are designed to construct task-specific distance metrics from data automatically. To navigate this complexity, we extend the base model by applying Self-Supervised Deep Metric Learning to diminish task bias while facilitating the exploration of individual properties.

Enhancing Self-Supervised Learning Performance in Imbalanced Datasets The key advantage of self-supervised learning is its inherent ability to handle imbalanced datasets. By exploiting its capacity to learn representative features from a vast corpus of unlabeled data, self-supervised learning can effectively engage with underrepresented classes that might otherwise remain underexplored [28, 29]. By employing repeated masking

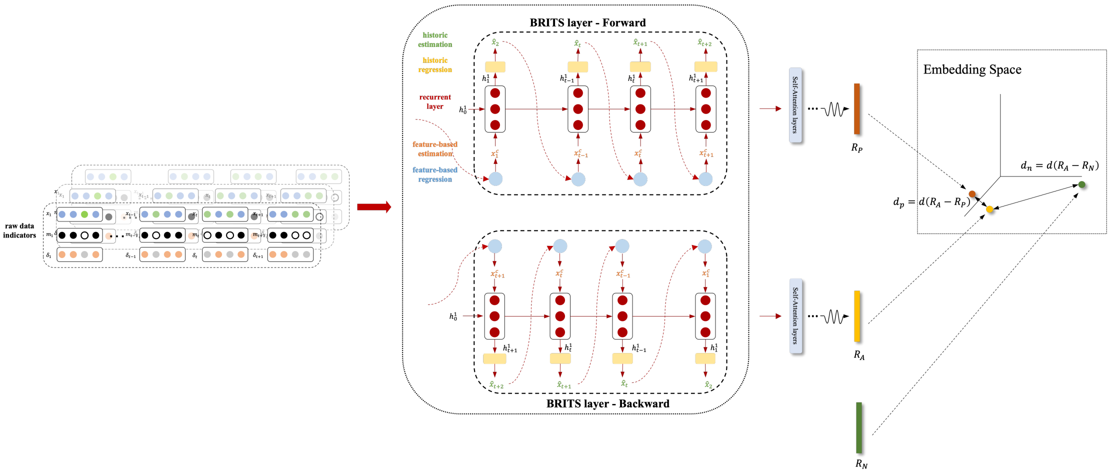

Figure 3: The structure of the proposed methods.

and generating a diverse array of examples, the model can discern valuable features distributed throughout the entire data, inclusive of those from minority classes, thereby enhancing the overall representational richness. This strategy can yield considerable benefits, especially under conditions where labeled data is either scarce or imbalanced. Furthermore, self-supervised learning can provide assistance in alleviating issues associated with overfitting, which are frequently encountered when models are trained on imbalanced datasets. Notwithstanding these advantages, challenges might emerge, including the identification of suitable auxiliary tasks and the risk of learning non-discriminative features [30].

Online Hard Triplets Mining The directionality of representations, whether originating from single-layer BRITS or our M^3-BRITS model, is employed to guide the construction of triplets. The dual directional indicators double the mini-batch size. This strategy applies to both labeled and unlabeled data, ensuring each triplet is balanced with one positive and one negative sample pair. However, this paper primarily focuses on labeled data in an effort to address the imbalance problem. Specifically, one representation is assigned as the anchor (denoted as $\boldsymbol{R_A}$), and another randomly chosen representation from the same group is defined as the positive sample (denoted as $\boldsymbol{R_P}$). Concurrently, representations from a different group within the same mini-batch are classified as negative samples (denoted as $\boldsymbol{R_N}$).

Incorporating Self-Attention for Optimal Representation Learning In order to optimally learn representations in each direction, we employ a methodology analogous to BERT [31]. We initially prepend a learnable embedding, the [CLS] token, to the hidden states as delineated in Equation 14. Each transformer encoder layer comprises two sublayers: (a) a multi-headed self-attention mechanism (MSA, as defined in Equation 15), and (b) a feed-forward network (FFN, as defined in Equation 16). Residual connections [32] are utilized around each of the sublayers in both the MSA and the FFN, followed by layer normalization (LN). The [CLS] embedding serves as $\boldsymbol{R_A, R_P, and R_N}$.

$$\hat{\boldsymbol{h}} = [\text{CLS}; \boldsymbol{h}_0; \boldsymbol{h}_1; ...; \boldsymbol{h}_t] \tag{14}$$

$$\hat{\boldsymbol{h}}' = \text{LN}(\text{MSA}(\hat{\boldsymbol{h}}) + \hat{\boldsymbol{h}}) \tag{15}$$

$$\hat{\boldsymbol{h}}^{*} = \text{LN}(FFN(\hat{\boldsymbol{h}}') + \hat{\boldsymbol{h}}') \tag{16}$$

$$\begin{aligned} \boldsymbol{R_A}, \boldsymbol{R_P} &\in [f_r(f_{SA}(\boldsymbol{H}_{nF})),\ f_r(f_{SA}(\boldsymbol{H}_{nB}))] \\ \boldsymbol{R_N} &\in [f_r(f_{SA}(\boldsymbol{H}_{mF})),\ f_r f_{SA}(\boldsymbol{H}_{mB}))], m \neq n \end{aligned} \tag{17}$$

where the $\boldsymbol{H}_{nF}$ represents the forward hidden states of sample n, $\boldsymbol{H}_{nB}$ for backward. Among the constructed triplets, we select all triplets that violate the following condition:

$$\| \boldsymbol{R_A} - \boldsymbol{R_P} \|_2 + \lambda \ < \| \boldsymbol{R_A} - \boldsymbol{R_N} \|_2 \tag{18}$$

λ is a pre-set margin, meaning we only consider samples that are easily confused with the anchor sample by a margin λ.

Loss Function This study utilizes the state-of-the-art Multi-Similarity loss (MS loss) objective function, a renowned metric learning objective [33]. The MS loss function prioritizes the significance of the samples by exploring the similarities within positive pairs and between negative pairs. The pairs that are most informative are often indistinguishable and hence contribute more substantially during the training process.

$$\mathscr{L} = \frac{1}{\mid B \mid} \sum_{i \in B} \left\{ \frac{1}{\alpha} \log[1 + \sum_{n \in \mathcal{N}_i} e^{\alpha(\mathrm{S}_{in} - \epsilon)}] \right. \\ \left. + \frac{1}{\beta} \log[1 + \sum_{p \in \mathscr{P}_i} e^{\beta(\mathrm{S}_{ip} - \epsilon)}] \right\} \tag{19}$$

where α, β, ϵ are hyperparameters; B denotes mini-batch samples.

Table 1
The mean absolute error (MAE) and mean relative error (MRE) for all datasets.

Models	Metric	Air	MIMIC-III (89)	Traffic	PhysioNet
M^3-BRITS	MAE	/	**0.297±0.006**	/	**0.249±0.004**
	MRE	/	0.462±0.005	/	0.351±0.008
M^2-BRITS	MAE	**0.106±0.002**	0.301±0.013	**0.065±0.006**	0.252±0.003
	MRE	0.145±0.002	0.466±0.011	0.236±0.016	0.354±0.003
M-BRITS	MAE	/	0.302±0.005	/	0.257±0.005
	MRE	/	0.468±0.004	/	0.363±0.003
BRITS	MAE	0.120±0.003	0.305±0.014	0.073±0.012	0.263±0.010
	MRE	0.165±0.004	0.473±0.012	0.266±0.032	0.371±0.007
MRNN	MAE	0.292±0.009	0.519±0.015	0.151±0.012	0.555±0.012
	MRE	0.400±0.004	0.804±0.004	0.548±0.022	0.783±0.013
GRUD	MAE	6.139±0.151	2.653±0.228	0.133±0.013	0.496±0.016
	MRE	8.414±0.317	4.112±0.380	0.484±0.028	0.700±0.006
V-RIN-full	MAE	0.166±0.003	0.303±0.012	0.140±0.012	0.273±0.013
	MRE	0.227±0.001	0.470±0.008	0.507±0.023	0.386±0.012

5. Experiments

In this section, we carefully assess and analyze the performance of M^3-BRITS by comparing it with state-of-the-art models using four real-world datasets in the domains of healthcare, environment, and traffic. The models against which M^3-BRITS is compared include BRITS, GRUD, V-RIN(full), and MRNN. For all experiments, only the best model from each study is employed for comparison. While a comparison with the E^2GAN model would have been intriguing, especially considering that [34] does not provide a quantitative comparison with BRITS, its inclusion in our study was precluded due to the obsolescence of the publicly-available code based on TensorFlow 1.7 and Python 2.7, rendering it incompatible with our accessible GPU hardware.

5.1. Datasets

The four datasets chosen for experimental evaluation each exhibit distinctive data distributions, with the MIMIC-III database particularly noted for its high rate of missing data. In our replication of the benchmarking studies for these publicly accessible datasets, we opted to bypass any steps that involved the removal of all-NaN samples to preserve the original missingness inherent in the data.

5.2. Implementation Details

In this study, we utilized the Adam optimizer across all models, with the number of RNN hidden units fixed at 108. The batch size was determined based on the specific dataset; MIMIC-Mortality and Air Quality were allocated a batch size of 128, while Traffic and MIMIC-Physionet Challenge were assigned a batch size of 64. To promote stable training, each dataset was normalized to have zero mean and unit variance. Randomly, we selected 10% of each dataset for validation and another 10% for testing, training the models on the remaining data. For the imputation task, we randomly masked 10% of observations in each dataset to serve as the ground truth, which was used as validation data. A 5-fold cross-validation method was implemented to evaluate the models. Imputation performance was gauged using the Mean Absolute Error (MAE) and Mean Relative Error (MRE). As for the Multiple Masking strategy, we restricted multiple masking operations to the training set alone.

5.3. Experimental Results

In the conducted experiments, we utilized three variants of our BRITS model, namely M^3-BRITS, M^2-BRITS, and M-BRITS. The M^3-BRITS model represents Multiple Masking Metric Learning. Similarly, the M^2-BRITS denotes Multiple Masking Learning, and the M-BRITS variant only encompasses Metric Learning. These different versions allowed us to evaluate the effectiveness and contribution of each component in a variety of real-world datasets.

Table 1 presents the Mean Absolute Error (MAE) and Mean Relative Error (MRE) across all evaluated datasets. In the MIMIC-III (89) dataset, our M^3-BRITS model achieved the lowest MAE of 0.297, surpassing the performance of all other models. Similarly, for the PhysioNet dataset, M^3-BRITS consistently demonstrated superior performance, with an MAE of 0.249. Regarding the unlabeled Air Quality and Traffic datasets, M^2-BRITS outperformed the other models, achieving the lowest MAE values of 0.106 and 0.065, respectively. This suggests that models with better generalization across datasets are likely to capture properties that are translatable to various data types.

Considering the MRE, no single model significantly outperformed the others across all datasets. However, it is noteworthy that, for each dataset, the MREs of our proposed M^3-BRITS and M^2-BRITS models were generally equivalent to, or better than, those of the other competing models. Baseline models, namely MRNN, GRUD, and V-RIN-full, did not outperform our proposed models in terms of either MAE or MRE across all datasets. This underlines the efficacy of the M^3-BRITS and M^2-BRITS methodologies in processing these real-world datasets.

6. Discussion and Conclusions

We propose M^3-BRITS, an innovative architecture that incorporates deep metric learning into the BRITS model for imputing missing values in multivariate time series characterized by non-random missingness. Our empirical findings indicate that M^3-BRITS delivers state-of-the-art performance across various datasets in the realm of time series imputation.

The M^3-BRITS model advances the BRITS architecture by optimizing sample similarity through self-supervised deep metric learning. This approach allows for circumventing prevalent issues such as class imbalance, thereby averting the imposition of artificial data distribution. By assessing imputation performance in isolation, as opposed to jointly with downstream tasks, M^3-BRITS prevents the diminution of diversity and model bias that can emerge from varying data focuses.

The experimental results substantiate that M^3-BRITS outperforms all baseline models. These outcomes underscore how augmenting an imputation model with deep metric learning and self-supervised learning empowers it to manage more intricate data with non-random missingness, an area where other techniques might falter. The M^3-BRITS model presents a promising prospect for achieving superior scalability and flexibility to address challenges encountered in real-world time series data.

In forthcoming research, we aspire to evaluate the performance of M^3-BRITS on additional complex healthcare datasets, particularly those typified by high missing rates, class imbalance, and mixed variable types. We will also explore potential enhancements to M^3-BRITS, inclusive of the incorporation of other self-supervised learning methodologies.

Acknowledgments

This paper represents independent research funded by the NIHR Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust 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 Linglong Qian was supported by the Kings-China Scholarship Council PhD Scholarship Programme (K-CSC) under Grant CSC202008060096. All models are implemented on CREATE HPC.[1]

References

[1] Z. Hu, et al., Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record, Journal of biomedical informatics 68 (2017) 112–120.

[2] Z. C. Lipton, D. Kale, R. Wetzel, Directly modeling missing data in sequences with rnns: Improved classification of clinical time series, in: Machine learning for healthcare conference, PMLR, 2016, pp. 253–270.

[3] B. J. Wells, K. M. Chagin, A. S. Nowacki, M. W. Kattan, Strategies for handling missing data in electronic health record derived data, Egems 1 (2013).

[4] R. Armstrong, et al., The incidence of cardiac arrest in the intensive care unit: A systematic review and meta-analysis, Journal of the Intensive Care Society 20 (2019) 144–154.

[5] M. Mazurowski, et al., Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance, Neural networks 21 (2008) 427–436.

[6] J. Wu, J. He, Y. Liu, Imverde: Vertex-diminished random walk for learning imbalanced network representation, in: 2018 IEEE International Conference on Big Data (Big Data), IEEE, 2018, pp. 871–880.

[7] L. E. Zarate, B. M. Nogueira, T. R. Santos, M. A. Song, Techniques for missing value recovering in imbalanced databases: Application in a marketing database with massive missing data, in: 2006 IEEE International Conference on Systems, Man and Cybernetics, volume 3, IEEE, 2006, pp. 2658–2664.

[8] P. Yadav, M. Steinbach, V. Kumar, G. Simon, Mining electronic health records (ehrs) a survey, ACM Computing Surveys (CSUR) 50 (2018) 1–40.

[9] P. B. Jensen, L. J. Jensen, S. Brunak, Mining electronic health records: towards better research applications and clinical care, Nature Reviews Genetics 13 (2012) 395–405.

[10] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, et al., Recent advances in convolutional neural networks, Pattern recognition 77 (2018) 354–377.

[11] M. Schuster, K. K. Paliwal, Bidirectional recurrent neural networks, IEEE transactions on Signal Processing 45 (1997) 2673–2681.

[12] S. Arber, J. J. Hunter, J. Ross Jr, M. Hongo, G. Sansig, J. Borg, J.-C. Perriard, K. R. Chien, P. Caroni, Mlp-deficient mice exhibit a disruption of cardiac cytoarchitectural organization, dilated cardiomyopathy, and heart failure, Cell 88 (1997) 393–403.

[13] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, Y. Li, Brits: Bidirectional recurrent imputation for time series, Advances in neural information processing systems 31 (2018).

[1] King's College London. (2022). King's Computational Research, Engineering and Technology Environment (CREATE). Retrieved March 2, 2022, from https://doi.org/10.18742/rnvf-m076

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

[15] J. Yoon, W. R. Zame, M. van der Schaar, Multi-directional recurrent neural networks: A novel method for estimating missing data, in: Time series workshop in international conference on machine learning, 2017.

[16] R. J. Little, D. B. Rubin, Statistical analysis with missing data, volume 793, John Wiley & Sons, 2019.

[17] H. He, E. A. Garcia, Learning from imbalanced data, IEEE Transactions on knowledge and data engineering 21 (2009) 1263–1284.

[18] O. Loyola-González, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, M. García-Borroto, Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases, Neurocomputing 175 (2016) 935–947.

[19] M. Galar, A. Fernández, E. Barrenechea, F. Herrera, Eusboost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling, Pattern recognition 46 (2013) 3460–3471.

[20] B. Krawczyk, M. Woźniak, G. Schaefer, Cost-sensitive decision tree ensembles for effective imbalanced classification, Applied Soft Computing 14 (2014) 554–562.

[21] M. Kaya, H. Ş. Bilge, Deep metric learning: A survey, Symmetry 11 (2019) 1066.

[22] Q. Suo, W. Zhong, F. Ma, Y. Yuan, J. Gao, A. Zhang, Metric learning on healthcare data with incomplete modalities., in: IJCAI, 2019, pp. 3534–3540.

[23] Z. Che, X. He, K. Xu, Y. Liu, Decade: a deep metric learning model for multivariate time series, in: KDD workshop on mining and learning from time series, sn, 2017.

[24] E. Jun, A. W. Mulyadi, J. Choi, H.-I. Suk, Uncertainty-gated stochastic sequential model for ehr mortality prediction, IEEE Transactions on Neural Networks and Learning Systems 32 (2020) 4052–4062.

[25] A. W. Mulyadi, E. Jun, H.-I. Suk, Uncertainty-aware variational-recurrent imputation network for clinical time series, IEEE Transactions on Cybernetics (2021).

[26] B. Kulis, et al., Metric learning: A survey, Foundations and Trends® in Machine Learning 5 (2013) 287–364.

[27] L. Yang, R. Jin, Distance metric learning: A comprehensive survey, Michigan State Universiy 2 (2006) 4.

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

[29] Q. Xie, Z. Dai, E. Hovy, T. Luong, Q. Le, Unsupervised data augmentation for consistency training, Advances in neural information processing systems 33 (2020) 6256–6268.

[30] A. Y. Ng, et al., Preventing" overfitting" of cross-validation data, in: ICML, volume 97, Citeseer, 1997, pp. 245–253.

[31] J. Devlin, M.-W. Chang, K. Lee, K. N. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL: https://arxiv.org/abs/1810.04805.

[32] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[33] X. Wang, X. Han, W. Huang, D. Dong, M. R. Scott, Multi-similarity loss with general pair weighting for deep metric learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5022–5030.

[34] Y. Luo, Y. Zhang, X. Cai, X. Yuan, E²gan: End-to-end generative adversarial network for multivariate time series imputation, in: International Joint Conference on Artificial Intelligence (IJCAI), 2019, pp. 3094–3100. doi:10.24963/ijcai.2019/429.

A. Datasets

Traffic Dataset. The Metro Interstate Traffic Volume Data Set, presents the hourly volume of traffic on the interstate highway I-94 in Minneapolis-St Paul, MN, USA.[2]

Air Quality Dataset. The Beijing Multi-Site Air-Quality Data offers hourly records of air pollutants.[3]

MIMIC-III Dataset. The Medical Information Mart for Intensive Care III (MIMIC-III) constitutes an extensive, freely accessible database, encompassing over 40,000 critical care patients.[4]

PhysioNet Challenge 2012 Dataset. The Predicting Mortality of ICU Patients: The PhysioNet/Computing in Cardiology Challenge 2012 is a publicly available medical benchmarking dataset.[5]

[2] https://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume

[3] https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data#

[4] https://physionet.org/content/mimiciii/1.4/

[5] https://physionet.org/content/challenge-2012/1.0.0/