Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 10, No. 1, 2024

Masked Self-Attention Dynamic Imputation Model for Physiological Indicators

Shaofeng Wu

College of Software Engineering, Chengdu University of Information Technology, Chengdu, China

Abstract: Accurate imputation of missing and outlier values in intraoperative physiological indicators can assist doctors in swiftly taking measures to prevent risks during surgery, which in turn helps improve postoperative prognosis for patients. Traditional data imputation methods have focused too much on the current data itself, neglecting the contextual information provided by surrounding data points. Addressing this issue, the imputation process for intraoperative physiological indicator data obtained through monitoring has been enhanced with a model that employs a self-attention mechanism focusing on the data context for dynamic weighting during imputation. Experimental results indicate that the proposed model achieves a MAE of 2.06%, a RMSE of 7.08%, and a MRE of 2.8%, outperforming other comparative models.

Keywords: Intraoperative physiological indicators, data imputation, dynamic weighting, self-attention mechanism.

1. Introduction

The monitoring of intraoperative physiological indicators serves multiple roles. Doctors can ensure patient safety, guide anesthesia management [1,2], prevent complications [3,4], and assist in surgical decision-making by observing changes in intraoperative physiological indicators. During surgery, accurate physiological indicators can provide clinical practitioners with the information needed to proactively take measures to prevent risks, thereby improving postoperative prognosis for patients.

However, due to equipment limitations or practical issues in clinical operations, some indicators may experience data loss; during patient positioning adjustments or equipment maintenance, monitoring may be temporarily interrupted, leading to informational gaps; outlier data may also occur during surgery, such as sporadic reading distortions caused by sensor detachment or technical failures. These situations can impact the quality of surgical decisions and patient care. As shown in Figure 1, the missing and outlier values for systolic pressure are displayed.

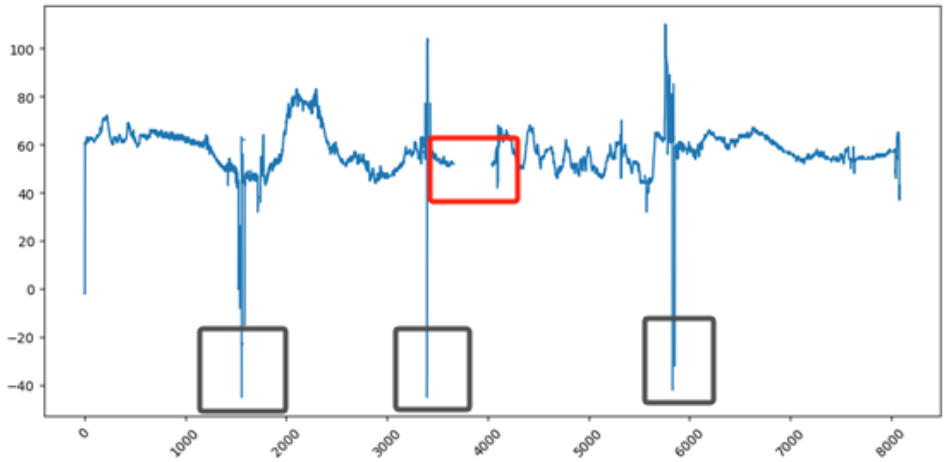

Figure 1. Missing and Outlier Systolic Pressure Data. The missing values in the data are highlighted in the red box, while the outliers are indicated in the gray box.

Currently, the handling of intraoperative physiological indicators is one of the hot research topics. There are mainly two methods for dealing with incomplete data: one is to delete the data that is missing or erroneous [5], and the other is to perform data imputation [6]. Choe et al. [7] excluded arterial waveform data that was clearly outside physiological ranges during their analysis. This included removing segments with missing values, data where blood pressure readings were higher than 200 millimeters of mercury (mmHg) or lower than 20 mmHg, waveforms where the difference between the maximum and minimum blood pressure values was less than 20 mmHg, and instances where the difference between adjacent readings exceeded 30 mmHg. Deleting missing data can lead to the loss of important information, and when the missingness is not random, this practice may introduce bias that undermines the validity of the analysis. Therefore, a more reasonable approach is to use advanced data imputation techniques to retain as much information as possible and improve the utilization of data. Imputation can be applied not only to missing data but also to some outlier values, which can be treated as missing values for the purpose of imputation.

The imputation of intraoperative physiological indicators

can be categorized into statistical learning imputation and machine learning imputation. Ribeiro et al. [8] compared methods such as special value imputation, mean imputation, and median imputation to choose the appropriate imputation method based on the characteristics of the data and the mechanism of missingness, and demonstrated the practical application and effectiveness of different imputation methods. Ahn et al. [9] showed the effectiveness of the K-Nearest Neighbors (KNN) imputation method in multivariate time series datasets by finding the K closest neighbors of a missing data point in a multidimensional feature space and using their values to estimate the missing ones.

In recent years, as deep learning has been validated in numerous fields for imputation tasks, capable of capturing complex feature transformations to yield superior imputed data[10,11]. Cao et al. [12] proposed BRITS (Bidirectional Recurrent Imputation for Time series), a recurrent dynamic model for estimating missing values. The closer a missing value is to the target value, the greater the influence it has on the generation process. To handle missing values, a bidirectional Recurrent Neural Network (RNN) model was designed to input and predict missing values.

With the success of self-attention mechanisms in natural language processing [13,14], they have also been introduced into time series models. Wu et al. [15] proposed a self-correlation mechanism by improving the self-attention mechanism, which identifies the similarity between subsequences at the same phase positions across different periods based on the periodicity of time series. This mechanism aggregates information from the identified similar subsequences, thereby reducing the computational complexity of the self-attention mechanism. Shukla et al. [16] utilized the self-attention mechanism from transformers, proposing learnable time embeddings as positional encodings and using temporally distributed latent representations to better capture the local structures of time series data, albeit the structure was too simplistic. Du et al. [17] introduced a Diagonal-Masked Self-Attention (DMSA) mechanism to eliminate the influence of sequence elements on themselves by applying a zero matrix where diagonal elements are masked on the attention weights, modifying the process and imputing missing values with dynamic weights, though the integration of dynamic weights was not comprehensive.

To address the aforementioned issues, this paper models the joint representation of multiple intraoperative physiological monitoring data of patients, uses two DMSA blocks to obtain learned representations, and dynamically integrates the weights from the two DMSA blocks, allowing for a more comprehensive learning for the data that requires imputation.

2. Methods

2.1. Problem Definition

Given a collection of multivariate time series (MTS) with T time steps and D dimensions, it is denoted as $\boldsymbol{X} = (x^1, x^2, ..., x^d ..., x^D)$. where the d-th step $\boldsymbol{x}^d = (x_1^d, x_2^d, ..., x_t^d, ..., x_T^D)$, x_t^d represents the d-th dimension variable of the t-th step in X. To enable the model to learn the location of the data that needs to be imputed within the input multivariate time series, a missingness matrix M is introduced, where $M = (m^1, m^2, ..., m^d)$ is missing, and $\boldsymbol{m}^d = (m_1^d, m_2^d, ..., m_t^d)$ otherwise. The calculation formula is as follows:

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

2.2. Diagonal-Masked Self-Attention Mechanism

The DMSA mechanism is an extension of the self-attention mechanism, adding a unique feature within the basic framework of self-attention to enhance the model's ability to capture temporal relationships in sequential data. The DMSA mechanism is shown in Figure 2. In this mechanism, the query, key, and value vectors are still obtained by mapping the given sequence data, maintaining consistency with the original self-attention mechanism. The difference lies in the introduction of a diagonal masking module in the DMSA, specifically designed to strengthen the model's understanding of the relationships between each element in the sequence and other elements, especially those temporally adjacent. This diagonal masking sets the diagonal elements of the attention score matrix to negative infinity (or in practice, a very small number, such as $-\infty$), ensuring that the attention weights for these positions are close to zero after the SoftMax normalization step. The calculation formula is as follows:

$$\left[\text{DiagMask}\left(x\right)\right]\left(i,j\right) = \begin{Bmatrix} -\infty & i = j \\ x\left(i,j\right) & i \neq j \end{Bmatrix} \tag{2}$$

$$\text{DMSA}\left(Q,K,V\right) = \text{Softmax}\left(\text{DiagMask}\left(\frac{QK^{\mathsf{T}}}{\sqrt{d_k}}\right)\right)V \tag{3}$$

Q, K, and V represent the query vector, key vector, and value vector, respectively, and d_k is the dimension of these three vectors. As a result, the model, when computing the representation of each element, is forced to ignore itself and focus on other relevant elements.

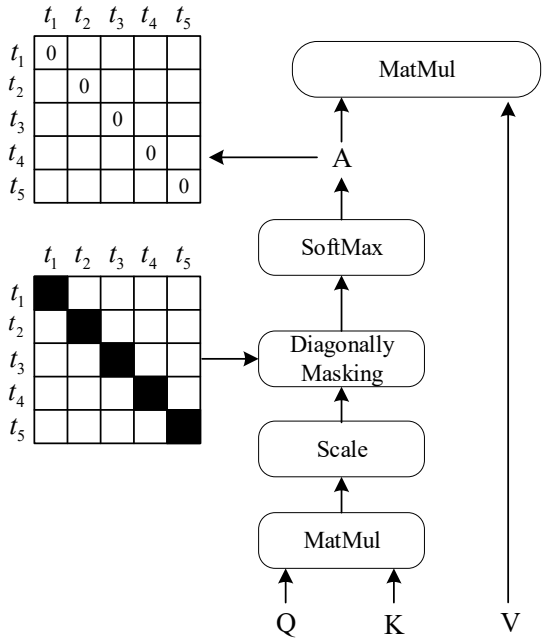

Figure 2. Diagonal-Masked Self-Attention mechanism

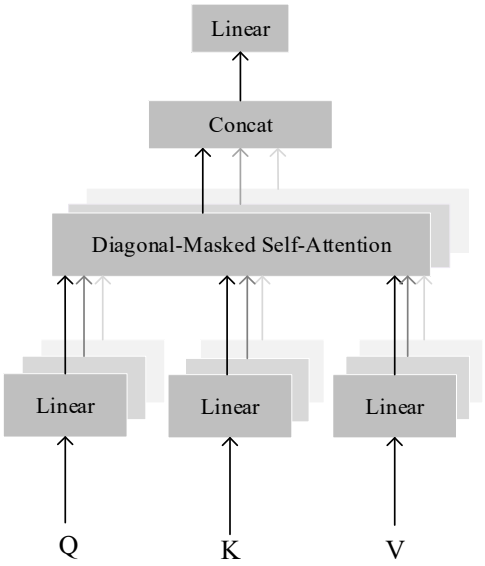

Figure 3. multi-head Diagonal-Masked Self-Attention mechanism

The multi-head attention mechanism first maps the input vector to multiple sets of queries, keys, and values vectors through different linear transformations, with each mapping corresponding to an attention "head". As each head has a different attention distribution, the model can flexibly focus on different parts of the input data, which is particularly valuable when handling complex dependencies. In this paper, we utilize the multi-head Diagonal-Masked Self-Attention mechanism (MHDMSA), as shown in Figure 3.

2.3. BiDMSA-DWI

To solve the problem of missing physiological data during surgery, this paper has developed a Bi-Diagonal-Masked Self-Attention Dynamic Weight Imputation model (BiDMSA-DWI), which aims to make full use of the interrelations among physiological indicator data to enhance imputation accuracy. The structure of the model is illustrated in Figure 4. The BiDMSA-DWI is primarily composed of three parts: the first DMSA module, the second DMSA module, and the dynamic weighting module.

The actual input feature vector is concatenated with a mask vector that represents the missing data situation. This combined vector is then fed into the model and through the first DMSA block, it obtains the first learned representation and imputed data. Next, the imputed data is concatenated with the mask vector and this combined vector is fed into the model again, passing through the second DMSA block to obtain the second learned representation. Using the output weights from both the first and second DMSA blocks, concatenated with the mask vector, a combined weight is obtained. This weight is then used to merge the first and second learned representations to obtain a third learned representation, ultimately resulting in the imputed data.

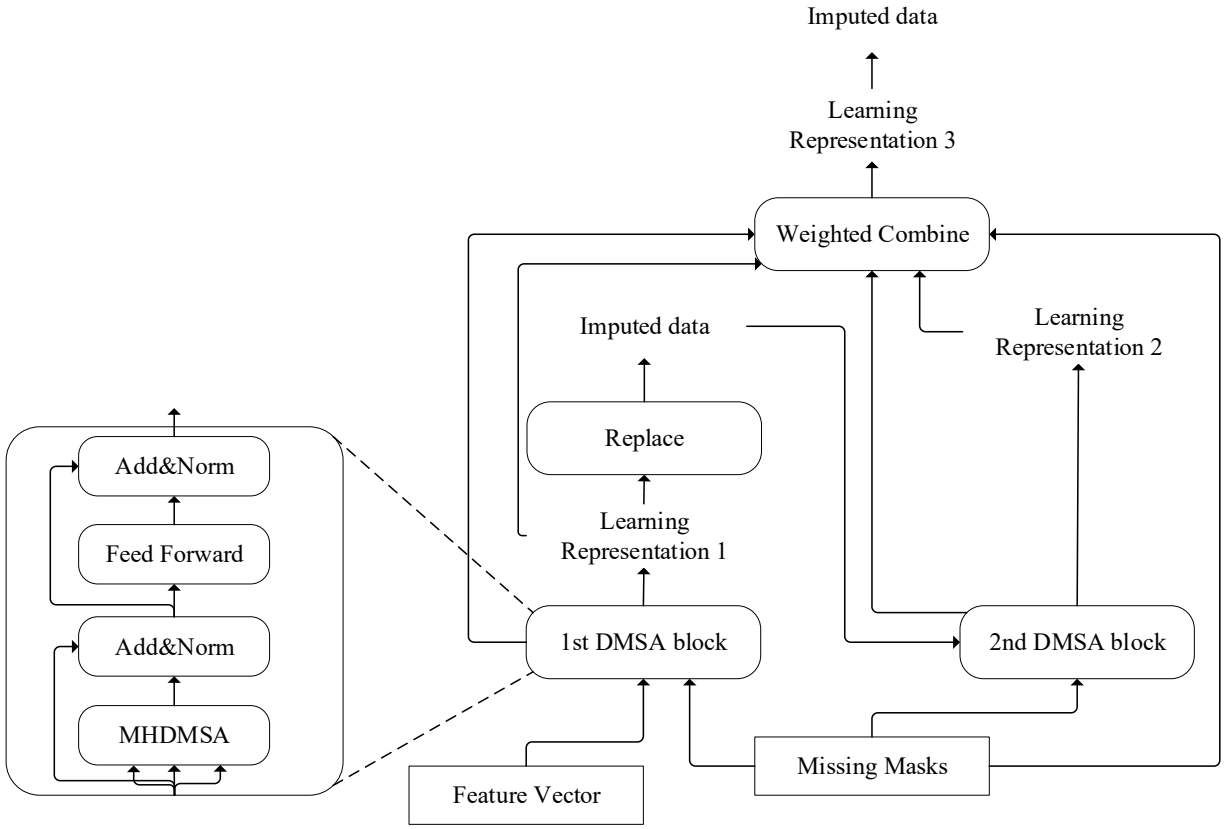

Figure 4. Bi-Diagonal-Masked Self-Attention Dynamic Weight Imputation model

3. Experiment

3.1. Dataset

The physiological monitoring data used in this study were all sourced from a tertiary hospital in Chengdu, and all materials have been anonymized to ensure patient privacy and security. Data from April to May 2022 was used as the training and test sets. The dataset was segmented using a sliding window method, with each window covering a 20-minute span and a step interval of one minute. Out of 77 detected physiological indicators, we selected 11 of the most widely applied indicators for our research, and set corresponding thresholds for these indicators, as shown in Table 1. The threshold is set lower because anesthetic drugs

themselves can affect physiological indicators, and each patient's condition is unique, thus making the threshold settings more universally applicable. The normal minimum heart rate and pulse are 60, but it may be lower in athletes. Therefore, the threshold for heart rate and pulse is set at 37. During surgery, due to blood loss, a patient's blood pressure is indeed lower than that of a healthy individual, hence the thresholds for diastolic pressure, systolic pressure, and mean arterial pressure are set at 45, 25, and 30, respectively. The thresholds for oxygen saturation and perfusion index are set at 73 and 0.01, which are already far below the safe lower limit. The fraction of inspired oxygen threshold of 20 is below the oxygen level in ambient air. End-tidal carbon dioxide, end-tidal oxygen, and the respiratory rate carbon dioxide reaching 13, 10, and 1 indicate that the patient may be experiencing cardiac arrest, so the threshold is set at these values.

Table 1. Physiological Indicator Thresholds

Abbreviation	Full Name	Threshold
HR	Heart Rate	37
PR	Pulse Rate	37
DBP	Diastolic Blood Pressure	45
SBP	Systolic Blood Pressure	25
MAP	Mean Arterial Pressure	30
SpO2	Oxygen saturation	73
PI	Perfusion Index	0.01
FiO2	Fraction of Inspired Oxygen	20
EtCO2	End-Tidal Carbon Dioxide	13
EtO2	End-Tidal Oxygen	10
RRCo2	Respiratory Rate Carbon Dioxide	1

3.2. Experimental Settings

In the deep learning model, the Adam optimizer was employed to update the parameters, with the initial learning rate set to 0.0006, batch size set to 16, and the number of iterations set for 500 epochs. An early stopping mechanism was implemented; it halts training if there is no decrease in the loss value of the training set within 20 epochs. Dropout was used to prevent overfitting, with a dropout probability value of 0.1. The model's hidden dimension was 256, the hidden size of the feedforward layer was 512, the number of heads in the attention mechanism was 8, and in the attention mechanism, the dimension of each head's key and value vectors was 32.

3.3. Evaluation Metrics

In the experiments of this chapter, the common evaluation metrics for data imputation tasks are adopted: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Relative Error (MRE). The mathematical expressions for these evaluation metrics are provided below. It should be noted that the calculation of the error is based solely on the missing values specified by the corresponding mask in the formula input. The calculation formula is as follows:

$$\mathrm{MAE} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T} |(estimation - target)\odot mask|_t^d}{\sum_{d=1}^{D}\sum_{t=1}^{T} mask_t^d} \tag{4}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{d=1}^{D}\sum_{t=1}^{T} (((estimation - target)\odot mask)^2)_t^d}{\sum_{d=1}^{D}\sum_{t=1}^{T} mask_t^d}} \tag{5}$$

$$\mathrm{MRE} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T} |(estimation - target)\odot mask|_t^d}{\sum_{d=1}^{D}\sum_{t=1}^{T} |target\odot mask|_t^d} \tag{6}$$

"estimation" represents the predicted value estimate, "target" represents the actual value data, and "mask" indicates the missing positions in the data.

3.4. Comparative Experimental Analysis

To demonstrate the cutting-edge performance of the BiDMSA-DWI model adopted in this paper for handling complex data imputation tasks, this section of the experiment has selected a range of comparison models for a comprehensive performance evaluation. These comparative models include not only traditional machine learning algorithms, such as KNN but also models based on recurrent neural networks like BRITS, as well as those using attention mechanisms such as NRTSI. The term "our" is used to refer to the model presented in this paper.

Table 2. Comparative Experiment of Imputation Models

Model	MAE	RMSE	MRE
KNN	5.08%	13.36%	7.13%
BRITS	3.91%	8.5%	4.94%
NRTSI	4.87%	12.46%	6.46%
Our	2.06%	7.08%	2.8%

By observing the experimental results in Table 2, it can be seen that the deep learning-based methods, especially the BiDMSA-DWI model proposed in this paper, significantly outperform the traditional machine learning algorithm KNN across all evaluation metrics. This outcome suggests that when it comes to imputing missing values in irregular multivariate data, deep learning, with its powerful feature extraction capabilities, is superior to traditional statistical machine learning methods.

The BiDMSA-DWI model also demonstrates exceptional performance when compared with other advanced deep-learning models, such as BRITS and NRTSI, which further confirms the superiority of our model in the field of multivariate data imputation. The BiDMSA-DWI utilizes a bidirectional masked self-attention mechanism that effectively captures the dependencies between data points. The dynamic weighting of the two bidirectional masked self-attention modules enables more accurate predictions in data imputation tasks.

3.5. Attention Ablation Experiment Analysis

To further explore the potential of the bidirectional masked self-attention mechanism in enhancing the predictive performance of the model, this study employed an innovative experimental design. The experiment involved substituting the bidirectional masked self-attention module in the BiDMSA-DWI model with a standard self-attention mechanism, thus creating a control model named "Self."

Table 3. Ablation Experiment of the Attention Mechanism

Model	MAE	RMSE	MRE
Self	2.13%	7.21%	2.9%
Our	2.06%	7.08%	2.8%

From the results in Table 3, we can observe that the model employing bidirectional masked self-attention outperformed the model that only used the self-attention mechanism across all evaluation metrics. This indicates that the bidirectional masked self-attention mechanism, by restricting the model to rely solely on information from other elements rather than its own information when making predictions, can indeed enhance the model's predictive accuracy.

The bidirectional masked self-attention mechanism introduces an efficient way of encoding contextual information, ensuring that each element in the sequence, when integrating information, is not distracted by its own signal. This allows for accurately reflecting the internal dynamics of the sequence.

3.6. Dynamic Weighting Ablation Experiment Analysis

To further validate the enhancing effect of the dynamic weighting module on model prediction, and to reveal how the inputs of two different DMSA weights affect the performance of the data imputation model, this section has designed an ablation experiment targeting the weighting component. These experiments aim to compare the model prediction performance between the average weight weighting strategy and the models weighted individually by either the first or the second DMSA module's weights. The experimental results are recorded in Table 3, where -att2 represents the removal of the weighting from the second DMSA module, -att1 signifies the removal of the weighting from the first DMSA module, and -att1-att2 indicates the adoption of the average weighting strategy.

By observing Table 4, it can be found that compared to the model that did not use dynamic weighting of the two attention module weights, the model proposed in this study reduced the MAE by 0.2%, RMSE by 0.46%, and MRE by 0.27%, demonstrating the effectiveness of the dynamic weighting mechanism in data imputation.

Table 4. Three Scheme comparing

Model	MAE	RMSE	MRE
Our	**2.06%**	**7.08%**	**2.8%**
Our-att2	2.27%	7.48%	3.09%
Our-att1	2.1%	7.26%	2.86%
Our-att1-att2	2.26%	7.54%	3.07%

Further analysis reveals that both the first and the second bidirectional masked self-attention modules play a positive role in the predictive performance of the model. However, when the second module is removed, the performance loss is more significant, which suggests that the second module holds a more crucial position in the model. When both modules are removed, the performance loss is greatest, which further confirms that these two modules play a synergistic, mutually enhancing role in the model, which is vital for achieving optimal predictive performance.

4. Conclusion

This paper introduces the BiDMSA-DWI model for imputing missing and anomalous values in intraoperative physiological indicators, employing the DMSA mechanism. It constructs two DMSA blocks and dynamically weights the learned representations and output weights obtained from these two DMSA blocks to achieve the imputed data. Experimental analysis has shown that the performance of this model is significantly superior to comparative models in clinical data, and ablation experiments have proven the effectiveness of the DMSA and dynamic weighting imputation. However, since the model does not account for individual differences, there is still room for performance improvement. In the future, more data will be collected for further experimentation and clinical application validation to explore the effects and value of this model in practical applications.

References

[1] Joosten A, Lucidi V, Ickx B, et al. Intraoperative hypotension during liver transplant surgery is associated with postopera-tive acute kidney injury: a historical cohort study[J]. BMC anesthesiology, 2021, 21(1): 1-10.

[2] Salmasi V, Maheshwari K, Yang D, Mascha EJ, Singh A, Sessler DI, Kurz A. Relationship between Intraoperative Hypotension, Defined by Either Reduction from Baseline or Absolute Thresholds, and Acute Kidney and Myocardial Injury after Noncardiac Surgery: A Retrospective Cohort Analysis. Anesthesiology. 2017 Jan;126(1):47-65.

[3] Lundberg S M, Nair B, Vavilala M S, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery[J]. Nature biomedical engineering, 2018, 2(10): 749-760.

[4] Healy M A, Mullard A J, Campbell D A, et al. Hospital and payer costs associated with surgical complications[J]. JAMA surgery, 2016, 151(9): 823-830.

[5] Choe, Sooho, et al. "Short-term event prediction in the operating room (STEP-OP) of five-minute intraoperative hypotension using hybrid deep learning: retrospective observational study and model development." JMIR Medical Informatics 9.9 (2021): e31311.

[6] Moore L, Hanley J A, Lavoie A, et al. Evaluating the validity of multiple imputation for missing physiological data in the national trauma data bank[J]. Journal of emergencies, trauma, and shock, 2009, 2(2): 73-79.

[7] Choe, Sooho, et al. "Short-term event prediction in the operating room (STEP-OP) of five-minute intraoperative hypotension using hybrid deep learning: retrospective observational study and model development." JMIR Medical Informatics 9.9 (2021): e31311.

[8] Ribeiro S M, de Castro C L. Missing data in time series: A review of imputation methods and case study[J]. Learning and Nonlinear Models-Revista Da Sociedade Brasileira De Redes Neurais-Special Issue: Time Series Analysis and Forecasting Using Computational Intelligence, 2021, 19(2).

[9] Ahn H, Sun K, Kim K P. Comparison of missing data imputation methods in time series forecasting[J]. Computers, Materials & Continua, 2022, 70(1): 767-779.

[10] Fang C, Wang C. Time series data imputation: A survey on deep learning approaches[J]. arXiv preprint arXiv:2011.11347, 2020.

[11] Wang J, Du W, Cao W, et al. Deep Learning for Multivariate Time Series Imputation: A Survey[J]. arXiv preprint arXiv:2402.04059, 2024.

[12] Cao W , Wang D , Li J ,et al.BRITS: Bidirectional Recurrent Imputation for Time Series[J]. 2018.DOI:10.48550/arXiv.1805.10572.

[13] Chaudhari S, Mithal V, Polatkan G, et al. An attentive survey of attention models[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2021, 12(5): 1-32.

[14] Galassi A, Lippi M, Torroni P. Attention in natural language processing[J]. IEEE transactions on neural networks and learning systems, 2020, 32(10): 4291-4308.

[15] Wu H, Xu J, Wang J, et al. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting[J]. Advances in Neural Information Processing Systems, 2021, 34: 22419-22430.

[16] Shukla S N, Marlin B M. Multi-time attention networks for irregularly sampled time series[J]. arXiv preprint arXiv:2101.10318, 2021.

[17] Du W, Côté D, Liu Y. Saits: Self-attention-based imputation for time series[J]. Expert Systems with Applications, 2023, 219: 119619.