Information Fusion 105 (2024) 102207

Contents lists available at ScienceDirect

Information Fusion

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

Multi-task oriented diffusion model for mortality prediction in shock patients with incomplete data

Weijie Zhao [a,b], Zihang Chen [b], Puguang Xie [b], Jinyang Liu [b], Siyu Hou [c], Liang Xu [c], Yuan Qiu [d], Dongdong Wu [e,f], Jingjing Xiao [c,*], Kunlun He [f,*]

[a] Department of Cardiovascular Surgery, Xinqiao Hospital, Chongqing, Chongqing, 40037, China
[b] Bioengineering College, Chongqing University, Chongqing, 400044, China
[c] Bio-Med Informatics Research Centre & Clinical Research Centre, The Second Affiliated Hospital of the Army Medical University, Chongqing, 400037, China
[d] Department of General Surgery, The Second Affiliated Hospital of the Army Medical University, Chongqing, 400037, China
[e] Department of Information, Daping Hospital, Army Medical University, Chongqing, 400042, China
[f] Medical Big Data Research Center, Medical Innovation Research Division, Chinese PLA General Hospital, Beijing, 100853, China

ARTICLE INFO

Dataset link: https://eicu-crd.mit.edu/, https://mimic.mit.edu/

Keywords:
Mortality prediction
Incomplete data
Multi-task oriented diffusion model

ABSTRACT

Mortality prediction based on electronic medical records is crucial for treatment decisions of shock patients in the ICU. Although clinical data on such patients often contain many missing values, the multi-view property of medical data could compensate for such missing information. Traditionally, mortality prediction models are built as two-stage approaches with additional data imputation steps, leading to issues in which the local optimal model at each step may not necessarily obtain a globally optimal solution. To overcome this problem, we conducted a multi-centre study using real-world data and aimed to develop an end-to-end mortality prediction model for shock patients. A Multi-task Oriented Diffusion Model (MODM) is proposed to fill in missing values and predict mortality simultaneously. Specifically, the model incorporates label information from different tasks to guide the optimal direction and effectively reduce uncertainty in the diffusion process. In addition, we propose a self-adjusting training strategy that balances the convergence rates among different tasks. The two largest well-known ICU datasets were used in this study, where 14,278 shock patients from eICU-CRD (2018) were included in the internal experiment, and 5,310 shock patients from MIMIC-IV (2012) were used as an external test. Compared with 14 state-of-the-art methods, the proposed model achieved the best performance with an AUC of 0.7998 in mortality prediction and notably good performance in terms of RMSE (0.0058, 0.0034, 0.0030, 0.0027) and MAE (0.3959, 0.4358, 0.4975, 0.5435) at random missing rates (10%, 30%, 50%, 70%, respectively) in the data imputation stage. The experimental results indicate the superiority of the proposed end-to-end MODM algorithm in handling real-world data. We released our code at https://github.com/zha0wj/MODM.

1. Introduction

Hospitals generate large amounts of biomedical data through daily inspections, including those of a significant proportion of ICU patients with shock symptoms [1]. In such cases, accurate mortality prediction is crucial for treatment planning. However, the effective management and use of these rapidly accumulating data for clinical purposes remains a major challenge [2]. With advancements in artificial intelligence, various diagnostic and prognostic models have been developed in the medical domain [3,4]. Generally, complete data are essential for conducting experiments and obtaining accurate and consistent results [5]. Nonetheless, inherent factors, such as differences in medical instruments, physicians' perceptions, and patients' contraindications, often cause incompleteness and poor consistency of clinical data [6]. This results in a limited clinically useful sample size along with strong noise in the model design, creating difficulties in generalising prognostic models across different clinical centres. To overcome this challenge, a thorough consideration of missing data handling is essential, as such data loss might eliminate critical features.

Owing to the complementarity and correlation between different features, biomedical data are often regarded as multi-view data. Machine learning-based algorithms can exploit these features to improve their overall performance [7]. However, existing models have mostly been developed using complete and well-structured medical datasets, which are difficult to obtain in real-world scenarios. Additional efforts

* Corresponding authors.
E-mail addresses: shine636363@sina.com (J. Xiao), kunlunhe@plagh.org (K. He).

https://doi.org/10.1016/j.inffus.2023.102207
Received 2 July 2023; Received in revised form 14 November 2023; Accepted 15 December 2023
Available online 30 December 2023

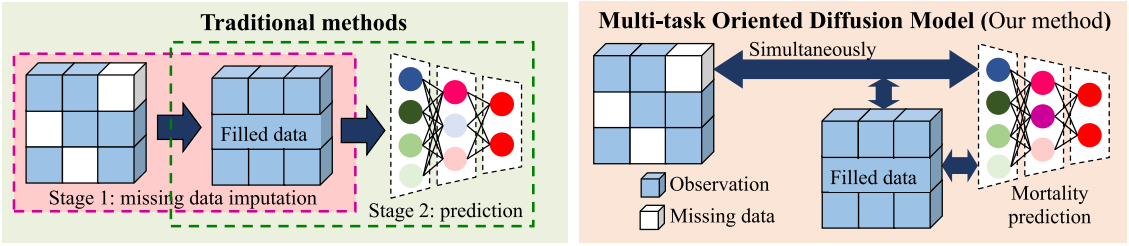


Fig. 1. Comparison of the proposed and other traditional methods. The traditional methods are those with two independent stages: data imputation (area in red dashed line) and prediction (area in green dashed line). Our method (Multi-task Oriented Diffusion Model) is an end-to-end framework that simultaneously fills in incomplete data and predicts the mortality of shock patients using mutual influence between these two tasks.

are required when missing values are present in the model development or validation stages. Typically, traditional models are built in two independent stages (Fig. 1). The first stage, called data imputation, fills in the missing data with commonly used generative models, such as the Variational AutoEncoder (VAE) [8], Generative Adversarial Network (GAN) [9], Continuous Normalising Flow (CNF) [10], and Diffusion Models (DM) [11]. In the second stage, machine learning-based methods are often adopted for building prediction or classification models, such as Support Vector Machine (SVM) [12], Random Forest [13], Boosting [14], and Convolutional Neural Network (CNN) [15]. While researchers aim to improve the accuracy of each stage, they may overlook the overall performance of the tasks. In other words, due to the presence of unknown noise, the best data imputation models may not be applicable for building the best prediction models. A locally optimal model at each step may not necessarily yield a globally optimal solution.

Therefore, the main objective of this study was to build an end-to-end mortality prediction model for multi-centre shock patients based on real-world data (Fig. 1). To achieve this, we propose a new diffusion model that simultaneously fills in missing data and predicts mortality by capturing the intrinsic correlations among the data. The auxiliary task of data imputation can effectively reduce the domain gap when applying the built prediction model to other dataset. Specifically, this model incorporates label information from different tasks during the diffusion process to guide the optimal direction and overcome the problem of high randomness in traditional diffusion models. Additionally, we propose a self-adjusting training strategy that balances the convergence rates of the data imputation and prediction tasks, resulting in better performance and higher accuracy of the overall model.

To evaluate the performance of the proposed method, we used known databases from multiple centres worldwide, including patients with shock from eICU-CRD [16] and MIMIC-IV [17] who underwent examination within 24 h. MIMIC-IV was developed by a team of computer scientists and physicians at Massachusetts Institute of Technology. It covers clinical data from about 300,000 ICU patients in the Boston area, allowing for the study and improvement of healthcare practices. Similarly, eICU-CRD, developed by Philips Healthcare, includes medical records from more than 200 institutions and data from more than 200,000 ICU patients. The database is intended for research purposes and aims to enhance healthcare and medical decision making.

Based on the specific inclusion and exclusion criteria for shock patients, 14,278 patients from the eICU-CRD were included in the internal experiment, while the external test of the MIMIC-IV involved 5,310 patients. In our study, we compared two-stage and end-to-end approaches. For the two-stage approaches, we first evaluated the performance of various commonly used data imputation methods, such as zero-value imputation, mean-value imputation [18], and K-Nearest Neighbour (KNN) [19], as well as other generative filling methods like Generative Adversarial Nets (GAIN) [20], diffusion models like Conditional Score-based Diffusion models (CSDI) [21], and Masked Autoencoding (ReMasker) [22], on internal datasets using metrics like RMSE and MAE. We then compared the performance of machine learning models on the mortality prediction task, such as Linear Support Vector Machine (SVM) [12], Random Forest (RF) [13], and deep learning models like ResNet34 [23], Gated Recurrent Neural Network (GRU) [24], Feature Tokenizer Transformer (FTT) [25], TabNet [26], TabTransformer [27], and TabAttention [28], using completed datasets (with filled data). We also compared the modified ResNet34 [23], GRU [24], TabNet [26], and TabTransformer [27] models in our designed end-to-end manner. The main contributions of this study are as follows:

(1) The designed multi-task oriented diffusion model simultaneously fills in the missing data and predicts the mortality of shock patients via an end-to-end approach. The domain gap between different datasets decreases in the data imputation task (auxiliary task), which eventually improves the performance of the mortality prediction task (main task).
(2) The developed self-adjusting strategy enables a stable and optimised convergence training process for multi-task learning, which avoids the problem of local optima that may occur in two-stage algorithms.
(3) The proposed method could cope with incomplete data in the real world to build a prediction model, which augments the number of training data. Remarkably, even when the patient missed some inspections, they could still benefit from this prognosis model.

The remainder of this paper is organised as follows. We first review related works in Section 2. Details of the proposed method are provided in Section 3. Section 4 presents and discusses the experimental results. Finally, Section 5 provides concluding remarks.

2. Related works

We first review existing works on two-stage methods for prediction tasks with missing data (Section 2.1). Then, we review related end-to-end studies on data imputation and prediction (Section 2.2).

2.1. Two-stage methods

Real-world data, particularly medical data, often have a high rate of missing values. Two-stage approaches are commonly used to handle missing data, which involves performing missing value imputation before subsequent tasks.

2.1.1. Data imputation

Several studies have highlighted the risks associated with using deletion methods to handle missing data [29,30]. To address this challenge, various methods have been proposed for missing data imputation, including interpolation techniques, ranging from simple approaches, such as zero-value and mean-value imputation, to more complex methods, such as K-nearest neighbours (KNN) [19] interpolation based on distance metrics, MissForest [31] interpolation utilising tree models, and Multiple Imputation by Chained Equations (MICE) [32] based on iterative processes. However, it is important to note that the interpolation-filling process can easily introduce bias. In multi-view learning, the primary objective of incomplete multi-view clustering [33] is to categorise data points from different perspectives

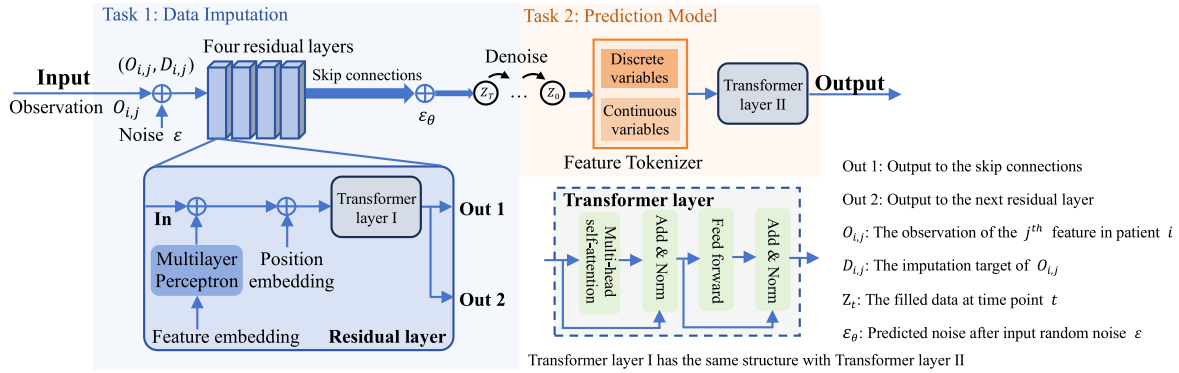


Fig. 2. Overview of the proposed Multi-task Oriented Diffusion Model (MODM). This algorithm incorporates feature and position embedding during the diffusion process for the auxiliary task of data imputation. Furthermore, the filled data are partitioned into discrete and continuous variables using a Feature Tokenizer and Transformer layer for the main prediction task.

into distinct clusters. Meanwhile, higher-order incomplete multiview subclustering [34] places greater emphasis on the higher-order feature relationships. For instance, GCFAGG [35] learns the global structure among samples using an encoder, ensuring that data representations within the same cluster are similar. Yan et al. [36] proposed a technique for obtaining data patterns in a high-dimensional space via feature learning and structure learning. Their approach involved a multi-view learning method designed to mitigate noise and eliminate redundant features in samples by recognising the projection direction of the data. These multi-view learning methods have proven highly effective in handling incomplete data. In 2014, generative models became popular due to the emergence of GANs [9], followed by diffusion models. OpenAI showed that a diffusion model [11,37] can outperform GANs in terms of image synthesis quality [38]. Song et al. proposed score-based generative models [39–41], and in 2021, they proposed a time series imputation method based on a conditional score-based diffusion model (CSDI) [42]. Although simple imputation methods may be sufficient for certain applications, they may not be appropriate for all cases. Meanwhile, machine learning-based approaches are effective, but they do not allow for back-propagation and cannot be updated in conjunction with multiple tasks.

2.1.2. Prediction models

Mortality prediction can be seen as a classification problem, and machine learning-based algorithms, such as random forest, XGboost [43], Catboost [44], and LightGBM [45], are good at handling such discrete data. Support Vector Machine (SVM) are primarily used for classification. Tree-based integrated models have several advantages, such as faster training and ease of interpretation. However, compared with deep learning models, they lack the ability to perform backpropagation. Deep learning methods, such as the Gated Recurrent Neural network (GRU) [24], ResNet [23], and Transformer [46] work well on one-dimensional data. SAINT [47] focuses on discrete data using a modified Transformer architecture. Gorishniy [25] proposed a Feature Tokenizer Transformer (FTT) that incorporates discrete and continuous numerical variables into the embedding before inputting to the Transformer. The goal of TransTab [48] is to convert each sample into a generalisable embedding vector and then apply a stacked converter for feature encoding.

2.2. End-to-end methods

Real-world datasets often contain missing values, which make it challenging to evaluate the performance of models trained on such data. Random missing validation metrics, such as RMSE and MAE, may not accurately reflect the distribution of complete data, leading to difficulties in identifying effective methods for the imputation of missing values. Consequently, researchers often indirectly assess their models' performance by evaluating the quality of predictions made on an incomplete dataset. One potential solution to this problem is to use GAN-based methods to fill in missing values and make end-to-end predictions. Two recent studies [49,50] employed this approach and demonstrated that end-to-end approaches have significant advantages over two-stage networks when using metrics such as AUC and MSE. In another study, RNN-based GANs were used to fill in missing data during the training process.

The above studies demonstrated the potential of using data-driven and end-to-end methods to address prediction problems with missing data in the real world.

3. Proposed method

We aimed to develop an end-to-end network, named a Multi-task Oriented Diffusion Model (MODM), to simultaneously fill in incomplete data and predict the mortality of shock patients in the ICU. The overall structure of our MODM method is shown in Fig. 2. In this section, we first introduce the structure of the end-to-end model (Section 3.1). Then, we explain the proposed method for multi-task oriented learning in Section 3.2, and the self-adjusting training strategy is explained in Section 3.3.

3.1. The overall structure

The MODM method simultaneously fills in missing data and predicts the mortality of shock patients, where the main task of mortality prediction could benefit from the auxiliary task of data imputation. $O_{i,j}$ is the observation of the jth feature of patient i. We first input $O_{i,j}$ along with the imputation target $D_{i,j}$, where the missing values in $D_{i,j}$ are filled with random Gaussian noise ε. Four residual layers are used to alleviate the potential vanishing gradient problem. Each layer contains a feature-embedding layer and diffusion-position-embedding layer for training, followed by a Transformer layer I. The feature-embedding layer is tailored to the number of features in the one-dimensional data. The position embedding layer corresponds to random diffusion time steps, aiming to prevent the diffusion training process from becoming stuck in local optima and enhance the model robustness. In the Transformer layer I, multi-head self-attention directs the model's focus towards internal dependencies between features, facilitating a better understanding of the actual data distribution. The outputs of the Transformer layer I contain two parts: one serves as input for the subsequent residual layer and the other acts as a skip connection. The predicted noise ε_θ is used to fill in the missing data through a reverse diffusion process. For continuous variables, the Feature Tokenizer embeds the learnable parameters, including weights and biases. For discrete variables, biases are embedded as learnable parameters. Ultimately, the Transformer layer II predicts the mortality of shock patients.

3.2. Multi-task oriented learning

Real-world data contain both discrete and continuous variables. To obtain better imputation results, we encoded discrete variables in a one-hot manner and standardised the continuous variables. The observation $O_{i,j}$ indicates the value of the ith patient's jth feature. The imputation targets $D = \{D_{1:N,1:M}\} \in \mathbb{R}^{N\times M}$ are expressed as

$$D_{i,j} = \begin{cases} 0, & O_{i,j} \text{ is nan (when the data is missing)} \\ 1, & O_{i,j} \text{ is not nan} \end{cases} \tag{1}$$

where N is the number of data (patients), and M is the number of features per patient. The intrinsic missing rate σ_0 can be computed as follows:

$$\sigma_0 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} D_{i,j}}{N \times M} \tag{2}$$

3.2.1. Data imputation

Let us consider learning a model distribution $p_\theta(x_0)$ that approximates a data distribution $q(x_0)$. The diffusion model is mainly utilised for the data imputation task (auxiliary task), which ensures that the data belongs to a Gaussian distribution after each addition of noise. Subsequently, during the denoising process, the Gaussian distribution is transformed back to the original distribution via a reversible Markov chain. Therefore, the diffusion model consists of forward and reverse processes. The forward process is defined by the following Markov chain:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{3}$$

where x_t is a variable that is subjected to a Gaussian distribution at time t centred on x_{t-1} with mean $\mu_t(x_t) = \sqrt{\alpha_t}x_{t-1}$ and variance $\gamma_t(x_t) = (1-\alpha_t)\mathbf{I}$. We set a fixed α_t in $D_{i,j}$ when adding random noise as follows:

$$x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\varepsilon, \ \ where\ \varepsilon \sim N(\varepsilon; \mathbf{0}, \mathbf{I}) \tag{4}$$

As demonstrated in [51], one can obtain x_t from x_0 as follows:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon_0, \ \ where\ \bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i \tag{5}$$

The inverse process of a Markov chain in the denoising process is used to obtain the original distribution for data imputation, which is formulated as follows:

$$p_\theta(x_{0:T}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{6}$$

In the diffusion process, predicting the original data x_0 can be regarded as a noise prediction process [51]. The optimisation problem can be formulated as follows:

$$\min_\theta \mathcal{L}(\theta) = \min_\theta \mathrm{E}_{x_0\sim q(x_0),\varepsilon\sim\mathcal{N}(\mathbf{0},\ \mathbf{I}),\ t}\| \ \varepsilon - \varepsilon_\theta(x_t,\ t)\ \|_2^2 \tag{7}$$

where ε is input noise, and $\varepsilon_\theta(x_t,\ t)$ is predicted noise at time t.

Let $O_{i,\ j,\ k}$ and $D_{i,\ j,\ k}$ for $k = 0, 1, \ldots, T$ be the observation of $O_{i,\ j}$ and imputation target of $D_{i,\ j}$ at moment k, respectively. The reverse process of the diffusion model in Eq. (6) in our MODM is obtained as follows:

$$p_\theta(D_{i,\ j,\ 0:T} \mid O_{i,\ j,\ 0}) = p(D_{i,\ j,\ T})\prod_{t=1}^{T} p_\theta(D_{i,\ j,\ t-1} \mid D_{i,\ j,\ t}, O_{i,\ j,\ 0}) \tag{8}$$

In simple terms, we only require a noise prediction network ε_θ that can accurately predict the size of the noise after each addition of random noise, and then we obtain filled data that is very close to the real data distribution. According to Eq. (7), the optimisation objective of MODM is to minimise the following loss function:

$$\min_\theta S(\theta) = \min_\theta \mathrm{E}_{x_0\sim q(x_0),\varepsilon\sim\mathcal{N}(\mathbf{0},\ \mathbf{I}),\ t}\| \ \varepsilon - \varepsilon_\theta(D_{i,\ j,\ t},\ t \mid O_{i,\ j,\ 0})\ \|_2^2 \tag{9}$$

To evaluate the imputation effect, we introduce a random missing rate σ_1, and a binary matrix $B_{i,j}$:

$$\sum_{i=1}^{N}\sum_{j=1}^{M} B_{i,j} = N \times M \times (1-\sigma_1),\ \ \sigma_1 \in (0,1) \tag{10}$$

where different $B_{i,j}$ lead to different observations $O_{i,j}^{\sigma_1}$:

$$O_{i,j}^{\sigma_1} = B_{i,j} \times O_{i,j} \tag{11}$$

We calculate the mean-squared loss function ($\mathcal{L}_{mse1}$) by taking the filled data obtained from the imputation network with the ground truth at a random missing rate σ_1 as follows:

$$\mathcal{L}_{mse1} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(\hat{Z}_{i,j}^{\sigma_1} - Z_{i,j}^{\sigma_1})^2}{N \times M} \tag{12}$$

where N is the number of data (patients), and M is the number of features for each patient. $\hat{Z}_{i,j}^{\sigma_1}$ is the filled data of the ith patient's jth feature at the pre-set missing rate σ_1, and $Z_{i,j}^{\sigma_1}$ is the ground truth at the pre-set missing rate σ_1.

3.2.2. Mortality prediction

The predicted noise ε_θ from the data imputation task is used to fill in the missing data through a reverse diffusion process. Subsequently, we calculate the Euclidean distance between the discrete variables and category labels. For continuous variables, the Feature Tokenizer embeds learnable parameters, including weights and biases. For discrete variables, biases are embedded as learnable parameters. This separation helps prevent gradient vanishing in deep learning networks when dealing with discrete variable inputs. In the Transformer layer II, the multi-head self-attention directs the model's focus towards internal dependencies between features and improves the prediction performance. The mean-squared loss function ($\mathcal{L}_{mse2}$) evaluates the similarity between the predicted values and true labels, while the cross-entropy loss function ($\mathcal{L}_{ce}$) measures the discrepancy between the predicted class probabilities and true labels. The $\mathcal{L}_{mse2}$ and $\mathcal{L}_{ce}$ we used are as follows:

$$\mathcal{L}_{mse2} = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(\hat{Z}_i))^2 \tag{13}$$

$$\mathcal{L}_{ce} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(f(\hat{Z}_i)) + (1-y_i)\log(1-f(\hat{Z}_i))) \tag{14}$$

where N is the number of data (patients), $\hat{Z}_i$ is the filled value of the ith patient, y_i is the label of data (patients), and f is the prediction network.

3.3. Self-adjusting training strategy

We utilise both $\mathcal{L}_{mse1}$ and $\mathcal{L}_{mse2}$ to train this multi-task network. The domain gap between the different data points decreases in the imputation task, which eventually benefitted the mortality prediction task. The MSE loss of the network can be denoted as:

$$\mathcal{L}_{t-mse} = \lambda\mathcal{L}_{mse1} + (1-\lambda)\mathcal{L}_{mse2} \tag{15}$$

where λ is a tuning factor. The training of the diffusion model involves an iterative denoising process. Finding a balance between the convergence speed and other tasks is challenging. Building on the work of [52], we propose a new method for balancing the weight of the multi-task loss function:

$$\mathcal{L}_{total} = \log(\mathcal{L}_{ce}) + \frac{1}{2}\log((S(\theta) + \mathcal{L}_{t-mse})) \tag{16}$$

where $\mathcal{L}_{ce}$ can be calculate by Eq. (14), and $S(\theta)$ is obtained using Eq. (9). This transformation allows larger loss terms to receive smaller update weights, while smaller loss terms receive larger update weights

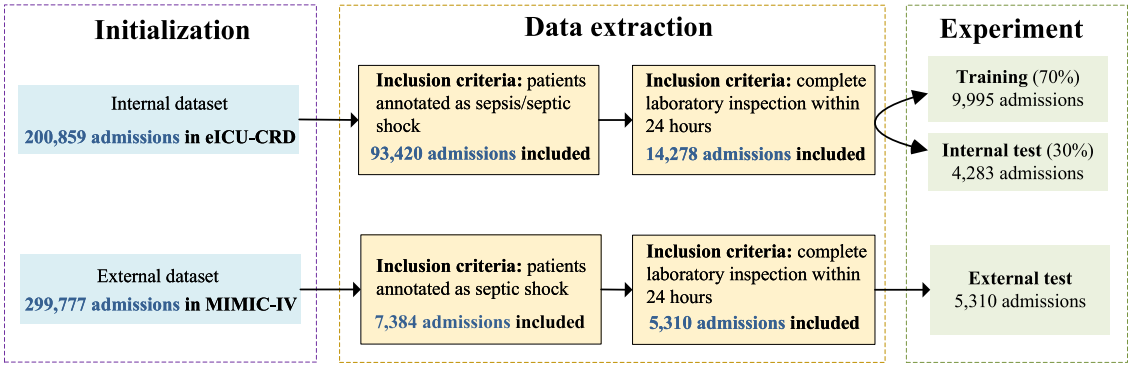


Fig. 3. Schematic illustration of the dataset construction.

when computing the gradient of the overall loss function. In this manner, we automatically balance the update speed of each loss function. By dynamically adjusting the update weights based on the magnitude of the loss, this technique can contribute to a smoother convergence and improved balance between the diffusion model and other tasks in a multitask setting. The pseudo-code is shown below (we used the Adam optimiser with epochs $N = 200$ and learning rate = 0.0001):

Algorithm 1 Multi-task Oriented Diffusion Model.

Input: $O_{i,j}^{\sigma_1}$ and $D_{i,j}^{\sigma_1}$ are the observation data and imputation target with a missing rate σ_1, respectively.
Output: mortality prediction

for epoch = 0, 1, 2, ..., N **do**
  Set the fixed parameters α_t of the forward process of the diffusion model (Eq. (5))
  Add random noise ε to $D_{i,j}^{\sigma_1}$
  Input ($O_{i,j}^{\sigma_1}$, $D_{i,j}^{\sigma_1}$) into the proposed network to predict noise ε_θ (Eq. (9))
  Input ε_θ in the reverse process of the diffusion model to obtain the filled data $\hat{Z}_{i,j}$
  Divide $\hat{Z}_{i,j}$ into continuous and discrete variables using the Feature Tokenizer
  Calculate the loss function of Eq. (16).
  Repeat the above steps until the loss converges.
end for
return prediction

4. Experiments

Because mortality prediction based on shock patient electronic medical records is crucial for treatment decisions in the ICU, we conducted a multi-centre study using the two largest world-known ICU databases, namely, eICU-CRD [16] and MIMIC-IV [17].

In this section, we first describe the dataset construction and implementation details. Subsequently, the evaluation metrics used in this study are provided. Finally, we present the proposed MODM algorithm evaluation from five aspects: comparison among data imputation methods, comparison among prediction models, comparison among end-to-end methods, ablation experiment, and key parameter analysis.

4.1. Dataset

The two largest world-known publicly available ICU databases, eICU-CRD [16] and MIMIC-IV [17], were utilised in this work to conduct a multi-centre study (Fig. 3).

eICU-CRD (version 2.0) [16]: This database was developed by Philips Healthcare and includes 200,859 patients from 208 hospitals in the United States in 2014 and 2015.

Table 1
The features of each shock patient that were used in the experiment.

Feature dimension	Feature names
Demographic characteristics	Age, gender, weight, height, race
Laboratory data (max, min)	Aniongap, albumin, bands, bicarbonate, bilirubin, creatinine, chloride, glucose, hematocrit, hemoglobin, lactate, platelet, potassium, ptt, inr, pt, sodium, bun, wbc

MIMIC-IV (version 2.1) [17]: This database covers 299,777 patients admitted to the intensive care unit or emergency department at the Beth Israel Deaconess Medical Centre (BIDMC) from 2008 to 2019.

Data extraction: From the datasets, we first involved the patients diagnosed with "septic shock". Among them, only those who underwent laboratory inspections within 24 h preceding their ICU stay were finally admitted to the experiment (Fig. 3). For each shock patient in eICU-CRD, 45 features were recorded, including demographic characteristics (age, sex, weight, height, and race) and laboratory data (blood gas, blood cell differential, liver function, renal function, respiratory function, and coagulation function). Correspondingly, each shock patient in MIMIC-IV had 91 features, including demographic characteristics, vital signs (heart rate, temperature, respiratory rate, mean arterial pressure, and central venous pressure), laboratory data, and urine output. Owing to duplication, redundancy, name confusion, and conceptual ambiguity in the data, we performed feature alignment with eICU-CRD and MIMIC-IV by doctors in terms of medical knowledge. Finally, 43 features were selected for the experiment, as listed in Table 1.

It is difficult to obtain complete datasets in the real world, especially for medical data. For patients included in eICU-CRD, the missing rate of all features was up to 14.50%, whereas for those in MIMIC-IV, the missing rate was 10.03%. The distribution of missing features is shown in Fig. 4, which highlights the challenges encountered by doctors in assessing a patient's condition when multiple features are missing. One can observe that the patient with the most missing features has only six features obtained, which further emphasises the difficulties in mortality prediction using real-world data.

Experimental construction: For the internal experiment, 14,278 patients from eICU-CRD were included. Among them, 9,995 patients' data (70%) were used for training, and the remaining 4,283 (30%) were used for internal testing. To further evaluate the generalisation performance of our model, 5,310 patients from MIMIC-IV were used as an external test set.

The proposed algorithm was implemented in Python 3.9 using an Nvidia GTX 4080 GPU and Intel i7-13700KF CPU. In our experiments, we set the hyperparameters of random missing rate σ_1 in Eq. (10) and λ in Eq. (15) to 0.1 and 0.2, respectively.

4.2. Evaluation metrics

For the different tasks in this study, we adopted corresponding metrics to evaluate the performance.

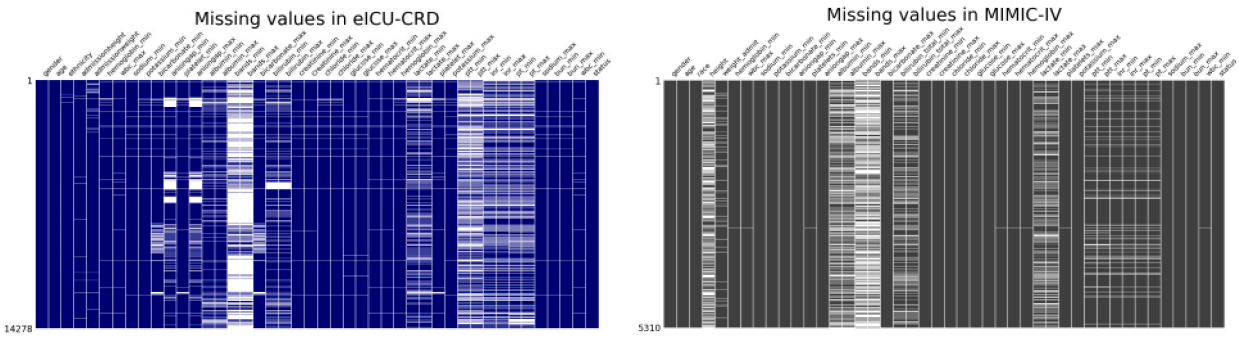


Fig. 4. Distributions of the missing features in eICU-CRD and MIMIC-IV. Each row represents a patient, and each column represents a feature. The black cells indicate the presence of a feature for a particular patient, while the blank cells represent missing features.

Table 2
RMSE and MAE values obtained by applying different data imputation methods at various missing rates (the results shown in **red bold** indicate the best performance, while values with ***violet italic*** are the second best).

Missing rate	RMSE							MAE						
	Zero	Mean	KNN [19]	GAIN [20]	CSDI [42]	ReMasker [22]	MODM	Zero	Mean	KNN [19]	GAIN [20]	CSDI [42]	ReMasker [22]	MODM
10%	0.0078	0.0077	0.0071	0.0067	0.0070	*0.0060*	**0.0056**	0.6300	0.6278	0.5722	0.4881	0.5049	*0.4655*	**0.3842**
30%	0.0044	0.0043	0.0040	0.0040	0.0039	*0.0035*	**0.0034**	0.6195	0.6172	0.5276	0.5032	0.5221	*0.4678*	**0.4358**
50%	0.0034	0.0034	0.0037	0.0032	0.0033	**0.0027**	*0.0030*	0.6177	0.6153	0.6793	0.5322	0.5490	**0.4696**	*0.4975*
70%	0.0029	0.0029	0.0033	0.0028	0.0028	**0.0023**	*0.0027*	0.6208	0.6170	0.6948	0.5631	0.5797	**0.4693**	*0.5435*

4.2.1. Metrics for data imputation

In this study, Root Mean Squared Error (RMSE) and Mean Absolute Error(MAE) were adopted, which are commonly used to evaluate the performance of data imputation methods. Lower RMSE and MAE values indicate more accurate and reliable performance.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(\hat{Z}_{i,j} - Z_{i,j})^2}{N \times M}} \tag{17}$$

$$\text{MAE} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}|\hat{Z}_{i,j} - Z_{i,j}|}{N \times M} \tag{18}$$

where N is the number of patients, and M is the number of the features for each patient. $\hat{Z}_{i,j}$ is the filled value of the ith patient's jth feature, and $Z_{i,j}$ is the ground truth.

4.2.2. Metrics for mortality prediction

The performance of the prediction network was evaluated using the Area Under the Curve (AUC), which is a widely used metric for medical data analysis. AUC is calculated by measuring the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve is created by plotting the True Positive Rate (TPR) on the vertical axis against the False Positive Rate (FPR) on the horizontal axis. A higher AUC indicates better performance.

$$TPR = \frac{TP}{TP + FN} \tag{19}$$

$$FPR = \frac{FP}{FP + TN} \tag{20}$$

where TP represents True Positive, FN represents False Negative, FP represents False Positive, and TN represents True Negative.

4.3. Comparisons and analysis

Next, we tested the proposed MODM algorithm on electronic medical records (explained in Section 4.1) from the real world. First, we compared our MODM method with other two-stage approaches. The SOTA end-to-end algorithms were also compared within internal and external tests. An ablation experiment was conducted to determine the contributions of each designed component. Furthermore, we evaluated the robustness of the proposed method using different key parameter settings.

4.3.1. Comparison among data imputation methods

To assess the performance of the methods in the data imputation stage, we intentionally set the missing rates to 10%, 30%, 50%, and 70% in the training set, following the same experimental settings as in [42]. Six other commonly used algorithms were evaluated: zero-value imputation, mean-value imputation, KNN imputation [19], GAIN imputation [20], CSDI imputation [42], and ReMasker imputation [22]. The results are reported using RMSE and MAE, as shown in Table 2. One can see that our MODM method showed the best imputation results at missing rates of 10% and 30%, and comparably good performance was observed with missing rates of 50% and 70%.

4.3.2. Comparison among prediction models

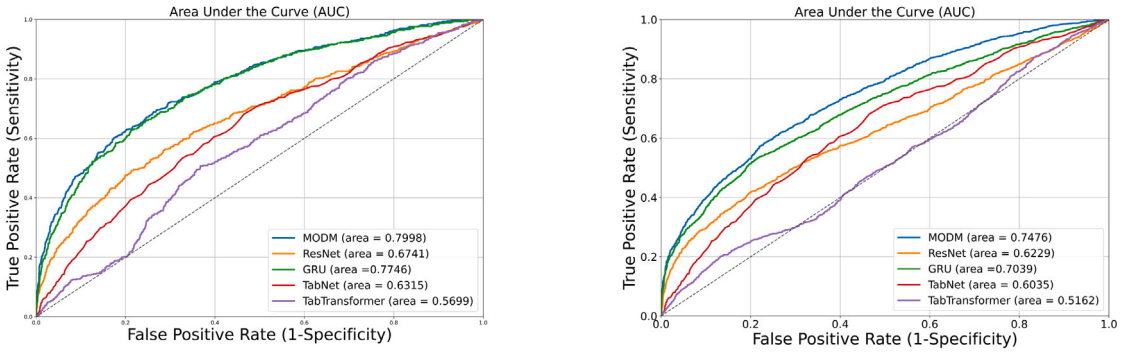
To evaluate the algorithms in the mortality prediction task, we compared our methodology to other SOTA algorithms: Linear Support Vector Machine (SVM) [12], Random Forest (RF) [13], ResNet34 [23], Gated Recurrent Neural network (GRU) [24], Feature Tokenizer Transformer (FTT) [25], TabNet [26], TabTransformer [27], and TabAttention [28]. The results of the internal and external tests are presented in Table 3. The AUC obtained from the external tests demonstrated the generalisation ability of the models. Based on the results, it is evident that our end-to-end approach (MODM) outperformed others on both the internal validation set and external test set. In comparison, the two-stage method involving GAIN padding followed by FTT prediction exhibited a 1% lower performance on the internal validation set, and the two-stage method employing zero padding followed by RF prediction demonstrated a 1% lower performance on the external test set compared to our method.

4.3.3. Comparison among end-to-end methods

We further compared our MODM method to other SOTA revised prediction algorithms in our end-to-end manner: ResNet34 [23], GRU [24], TabNet [26], and TabTransformer [27]. The internal and external AUCs are shown in Fig. 5. From Figs. 5(a) and 5(b), we found that the AUC of MODM outperformed that of the other algorithms for both internal datasets (0.7998) and external datasets (0.7476). The results indicate that our end-to-end approach, utilising the diffusion model in conjunction with Feature Tokenizer, delivers the most favourable outcomes. It achieved an AUC of 0.7998 on the internal validation set and 0.7476 on the external test set.

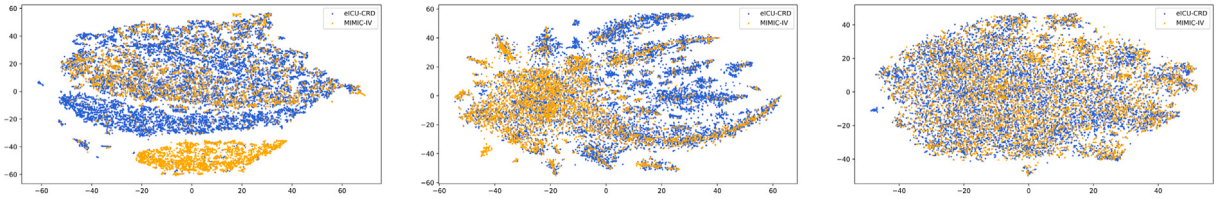
Table 3
AUC of prediction models built with the results from different data imputation methods on both internal and external datasets (the results shown in **red bold** indicate the best performance, while values with ***violet italic*** are the second best).

PM	DI													
	Internal test							External test						
	Zero	Mean	KNN [19]	GAIN [20]	CSDI [42]	ReMasker [22]	MODM	Zero	Mean	KNN [19]	GAIN [20]	CSDI [42]	ReMasker [22]	MODM
SVM [12]	0.7666	0.7665	0.7666	0.7697	0.6248	0.4448	0.5440	0.7150	0.7152	0.7116	0.7184	0.5765	0.4714	0.5788
RF [13]	0.7769	0.7716	0.7756	0.7741	0.6025	0.6128	0.6191	*0.7355*	0.7350	0.7328	0.7345	0.5936	0.5939	0.6141
RestNet34 [23]	0.7464	0.7129	0.6931	0.7007	0.4566	0.5090	0.5342	0.6662	0.6725	0.6412	0.6558	0.4631	0.5279	0.4904
GRU [24]	0.6231	0.6192	0.6222	0.6184	0.5011	0.4970	0.4921	0.5736	0.5668	0.5681	0.5726	0.4936	0.5042	0.5039
FTT [25]	0.7827	0.7818	0.7790	*0.7840*	0.5456	0.5267	0.7157	0.7192	0.7212	0.7188	0.7226	0.5036	0.5078	0.6839
TabNet [26]	0.7502	0.7335	0.7583	0.7529	0.5448	0.5774	0.6562	0.7034	0.6594	0.7157	0.7077	0.5679	0.5460	0.6220
TabTransformer [27]	0.6022	0.5117	0.5753	0.5469	0.4999	0.6736	0.5550	0.6164	0.5344	0.4837	0.5450	0.4984	0.6587	0.5216
TabAttention [28]	0.7594	0.7573	0.7483	0.7653	0.4307	0.3639	0.4429	0.5671	0.5426	0.5815	0.5533	0.4309	0.3725	0.4330
MODM	N/A	N/A	N/A	N/A	N/A	N/A	**0.7998**	N/A	N/A	N/A	N/A	N/A	N/A	**0.7476**



(a) Internal validation (b) External test

Fig. 5. ROC curves of different end-to-end algorithms: (a) validation on the **internal** dataset and (b) testing on the **external** dataset.



(a) Before training (b) During training (c) After training

Fig. 6. Visualisation of data distribution during the training process.

Furthermore, we employed t-SNE [53] dimension reduction techniques to visualise the distributions of two databases, namely eICU-CRD and MIMIC-IV, during the MODM training process in Fig. 6. It can be seen that the domain gap between the two databases decreased during the training process. This is attributed to the auxiliary task, data imputation, which improved the prediction model.

4.3.4. Ablation experiment

To further investigate the designed methodology, we conducted an ablation study. RMSE and MAE were used to evaluate the auxiliary task of data imputation. The performance of the main task, mortality prediction, was reported using the metric of AUC on both the validation and test data. The improvements in the important components, such as Transformer layer I, Feature Tokenizer, and Transformer layer II, are shown in Table 4. One can see that each component contributed to the final performance. A significant improvement in the data imputation task was found with Transformer layer I (demonstrated by the RMSE and MAE), which mainly learns the data distribution. Feature Tokenizer and Transformer layer II are mainly used for mortality prediction with little improvement in the data imputation task.

Additionally, we evaluated the effectiveness of the proposed self-adjusting strategy explained in Section 3.3. In Fig. 7, the green line represents the loss of $S(\theta)$ (Eq. (9)) with the designed self-adjusting training strategy for the data imputation task, while the blue line represents the corresponding loss curve without our self-adjusting strategy.

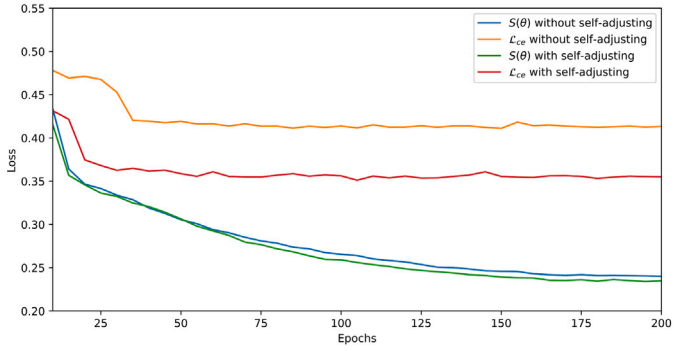
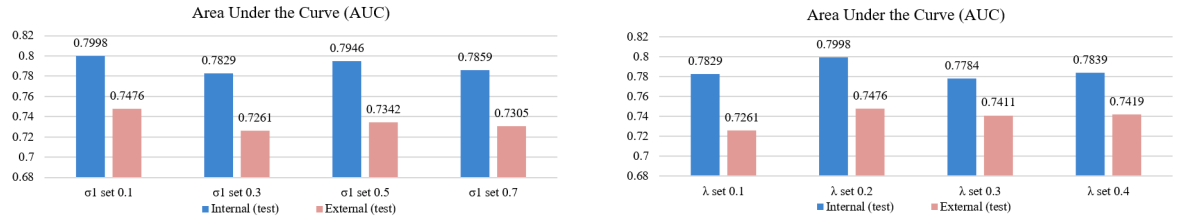


Fig. 7. Loss curves of $S(\theta)$ in Eq. (9) and $\mathcal{L}_{ce}$ in Eq. (14).

Both cases effectively captured the underlying data distribution, while the self-adjusting strategy enabled faster convergence. A more notable difference could be seen in the cross-entropy loss of $\mathcal{L}_{ce}$(Eq. (14)) for the prediction model. With our training approach, $\mathcal{L}_{ce}$ decreased by approximately 20% compared to the case without the self-adjusting strategy. This demonstrates that the self-adjustment strategy balances the convergence rates of multiple tasks.

Table 4
Ablation study with the key components. The best results are shown in **red bold**, while values with ***violet italic*** are the second best.

Methods	Model Components			Metrics			
	Transformer layer I	Feature Tokenizer	Transformer layer II	RMSE	MAE	AUC (valid)	AUC (test)
(a)		✓	✓	0.0087	0.6832	0.7901	0.7305
(b)	✓		✓	0.0057	0.3897	0.7047	0.6733
(c)	✓	✓		*0.0057*	*0.3886*	*0.7414*	*0.6934*
(d)	✓	✓	✓	**0.0056**	**0.3842**	**0.7998**	**0.7476**



(a) Performance with various σ_1 settings

(b) Performance with various λ settings

Fig. 8. AUC of MODM algorithm with different parameter settings.

4.3.5. Key parameter analysis

We changed the values of the hyperparameter random missing rate σ_1 in Eq. (10) and λ in Eq. (15) to evaluate the robustness of the MODM algorithm. From Figs. 8(a) and 8(b), it can be seen that with acceptable changes in σ_1 or λ, the MODM could still obtain considerably good results. The results indicate that the hyperparameter settings for σ_1 and λ have a relatively minor impact on the overall outcomes, with variations between the best and worst results falling within a 2% range.

5. Conclusion

In this paper, we proposed a Multi-task Oriented Diffusion Model (MODM) that simultaneously fills in missing values and predicts the mortality of shock patients using real-world data. Specifically, the model incorporates label information from different tasks to guide the optimal training direction and effectively reduces uncertainty in the diffusion process. In addition, we proposed a self-adjusting training strategy that balances convergence rates among different tasks. The two largest well-known ICU datasets were used in this study, where 14,278 patients from eICU-CRD [16] were included for an internal experiment and 5,310 patients from MIMIC-IV [17] were used for external test.

The experimental results demonstrate the advantages of end-to-end algorithms over two-step methods. Although the ReMasker method achieved comparably good performance in the data imputation stage, it did not translate into superior prediction results. In other words, these two-stage methods cannot harness information from each stage to improve overall model performance. Our findings provide a strong foundation for future end-to-end research.

CRediT authorship contribution statement

Weijie Zhao: Data extraction and experimental design. **Zihang Chen:** Conceptualization. **Puguang Xie:** Conceptualization, Request permissions for eICU-CRD and MIMIC-IV. **Jinyang Liu:** Ablation experiments. **Siyu Hou:** Ablation experiments. **Liang Xu:** Feasibility of the experiment, Provide clinical knowledge. **Yuan Qiu:** Feasibility of the experiment, Provide clinical knowledge. **Dongdong Wu:** Feasibility of the experiment, Provide clinical knowledge. **Jingjing Xiao:** Funding, Experimental conditions, Technical support for the whole experiment. **Kunlun He:** Funding, Experimental conditions, Technical support for the whole experiment.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The eICU-CRD and MIMIC-IV are open public databases, and both have been ethically reviewed.

Data availability

Our data were mainly obtained from eICU-CRD (version 2.0), https://eicu-crd.mit.edu/, and MIMIC-IV (version 2.1), https://mimic.mit.edu/, studying patients who were diagnosed with “septic shock”. Access to the MIMICIV and eICU-CRD databases was approved by the Massachusetts Institute of Technology (Cambridge, MA) and Beth Israel Deaconess Medical Centre (Boston, MA), and consent was obtained for the data collection. These databases are public, all patient data is anonymized, and data extracted from the databases do not require individual informed consent. The author (P.X.) attended a series of courses offered by the National Institutes of Health and was granted access to these databases after passing the required assessment (Record ID: 51524821). The study was conducted following the Declaration of Helsinki and protocols were approved by the Ethics Committee of the Second Hospital Affiliated to the Army Medical University.

Acknowledgements

The present study was supported by National Natural Science Foundation of China (NO. 62076247, NO. 61701506), Independent Research Project of Medical Engineering Laboratory of Chinese PLA General Hospital, China (2022SYSZZKY07), Medical research project of Chongqing Municipal Health Commission, China (2023WSJK011).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.inffus.2023.102207.

References

[1] Jean-Louis Vincent, Daniel De Backer, Circulatory shock, N. Engl. J. Med. 369 (18) (2013) 1726–1734.
[2] Shuihua Wang, M Emre Celebi, Yu-Dong Zhang, Xiang Yu, Siyuan Lu, Xujing Yao, Qinghua Zhou, Martinez-Garcia Miguel, Yingli Tian, Juan M Gorriz, et al., Advances in data preprocessing for biomedical data fusion: An overview of the methods, challenges, and prospects, Inf. Fusion 76 (2021) 376–421.

[3] Zeyu Gao, Anyu Mao, Kefei Wu, Yang Li, Liebin Zhao, Xianli Zhang, Jialun Wu, Lisha Yu, Chao Xing, Tieliang Gong, et al., Childhood leukemia classification via information bottleneck enhanced hierarchical multi-instance learning, IEEE Trans. Med. Imaging (2023).

[4] Dwarikanath Mahapatra, Zongyuan Ge, Mauricio Reyes, Self-supervised generalized zero shot learning for medical image classification using novel interpretable saliency maps, IEEE Trans. Med. Imaging 41 (9) (2022) 2443–2456.

[5] CHMP Ich, E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials, in: Proceedings of the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, 2019.

[6] SWJ Nijman, AM Leeuwenberg, I Beekers, I Verkouter, JJL Jacobs, ML Bots, FW Asselbergs, KGM Moons, TPA Debray, Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review, J. Clin. Epidemiol. 142 (2022) 218–229.

[7] Shuihua Wang, M. Emre Celebi, Yu-Dong Zhang, Xiang Yu, Siyuan Lu, Xujing Yao, Qinghua Zhou, Martínez-García Miguel, Yingli Tian, Juan M Gorriz, Ivan Tyukin, Advances in data preprocessing for biomedical data fusion: An overview of the methods, challenges, and prospects, Inf. Fusion 76 (2021) 376–421.

[8] Diederik P Kingma, Max Welling, Auto-encoding variational bayes, stat 1050 (2014) 1.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.

[10] Ivan Kobyzev, Simon J.D. Prince, Marcus A. Brubaker, Normalizing flows: An introduction and review of current methods, IEEE Trans. Pattern Anal. Mach. Intell. 43 (11) (2020) 3964–3979.

[11] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, Surya Ganguli, Deep unsupervised learning using nonequilibrium thermodynamics, in: ICML, PMLR, 2015, pp. 2256–2265.

[12] Woojae Kim, Ku Sang Kim, Jeong Eon Lee, Dong-Young Noh, Sung-Won Kim, Yong Sik Jung, Man Young Park, Rae Woong Park, Development of novel breast cancer recurrence prediction model using support vector machine, J. Breast Cancer 15 (2) (2012) 230–238.

[13] Steven J. Rigatti, Random forest, J. Insur. Med. 47 (1) (2017) 31–39.

[14] Gu-Wei Ji, Chen-Yu Jiao, Zheng-Gang Xu, Xiang-Cheng Li, Ke Wang, Xue-Hao Wang, Development and validation of a gradient boosting machine to predict prognosis after liver resection for intrahepatic cholangiocarcinoma, BMC Cancer 22 (1) (2022) 258.

[15] Zhiwei Zhang, Jingjing Xiao, Shandong Wu, Fajin Lv, Junwei Gong, Lin Jiang, Renqiang Yu, Tianyou Luo, Deep convolutional radiomic features on diffusion tensor images for classification of glioma grades, J. Digit. Imaging 33 (4) (2020) 826—837.

[16] mit.edu, eICUDataset, 2018, https://eicu-crd.mit.edu/.

[17] mit.edu, MIMICDataset, 2012, https://mimic.mit.edu/.

[18] Edgar Acuna, Caroline Rodriguez, The treatment of missing values and its effect on classifier accuracy, in: Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15–18 July 2004, Springer, 2004, pp. 639–647.

[19] R. Malarvizhi, Antony Selvadoss Thanamani, K-nearest neighbor in missing data imputation, Int. J. Eng. Res. Dev. 5 (1) (2012) 5–7.

[20] Jinsung Yoon, James Jordon, Mihaela Schaar, Gain: Missing data imputation using generative adversarial nets, in: ICML, PMLR, 2018, pp. 5689–5698.

[21] Shuhan Zheng, Nontawat Charoenphakdee, Diffusion models for missing value imputation in tabular data, 2022, arXiv preprint arXiv:2210.17128.

[22] Tianyu Du, Luca Melis, Ting Wang, Remasker: Imputing tabular data with masked autoencoding, 2023, arXiv preprint arXiv:2309.13793.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Identity mappings in deep residual networks, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part IV 14, Springer, 2016, pp. 630–645.

[24] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, Yoshua Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, in: NIPS 2014 Workshop on Deep Learning, December 2014, 2014.

[25] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, Artem Babenko, Revisiting deep learning models for tabular data, NeurIPS 34 (2021) 18932–18943.

[26] Sercan Ö Arik, Tomas Pfister, Tabnet: Attentive interpretable tabular learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 6679–6687.

[27] Xin Huang, Ashish Khetan, Milan Cvitkovic, Zohar Karnin, Tabtransformer: Tabular data modeling using contextual embeddings, 2020, arXiv preprint arXiv:2012.06678.

[28] Michal K Grzeszczyk, Szymon Płotka, Beata Rebizant, Katarzyna Kosińska-Kaczyńska, Michał Lipa, Robert Brawura-Biskupski-Samaha, Przemysław Korzeniowski, Tomasz Trzciński, Arkadiusz Sitek, TabAttention: Learning attention conditionally on tabular data, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 347–357.

[29] Roderick J.A. Little, Donald B. Rubin, Statistical Analysis with Missing Data, Vol. 793, John Wiley & Sons, 2019.

[30] Donald B. Rubin, Multiple Imputation for Nonresponse in Surveys, Vol. 81, John Wiley & Sons, 2004.

[31] Daniel J. Stekhoven, Peter Bühlmann, MissForest—non-parametric missing value imputation for mixed-type data, Bioinformatics 28 (1) (2012) 112–118.

[32] Stef Van Buuren, Karin Groothuis-Oudshoorn, mice: Multivariate imputation by chained equations in R, J. Stat. Softw. 45 (2011) 1–67.

[33] Hang Gao, Yuxing Peng, Songlei Jian, Incomplete multi-view clustering, in: Intelligent Information Processing VIII: 9th IFIP TC 12 International Conference, IIP 2016, Melbourne, VIC, Australia, November 18-21, 2016, Proceedings 9, Springer, 2016, pp. 245–255.

[34] Zhenglai Li, Chang Tang, Xiao Zheng, Xinwang Liu, Wei Zhang, En Zhu, High-order correlation preserved incomplete multi-view subspace clustering, IEEE Trans. Image Process. 31 (2022) 2067–2080.

[35] Weiqing Yan, Yuanyang Zhang, Chenlei Lv, Chang Tang, Guanghui Yue, Liang Liao, Weisi Lin, GCFAgg: Global and cross-view feature aggregation for multi-view clustering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19863–19872.

[36] Weiqing Yan, Meiqi Gu, Jinlai Ren, Guanghui Yue, Zhaowei Liu, Jindong Xu, Weisi Lin, Collaborative structure and feature learning for multi-view clustering, Inf. Fusion 98 (2023) 101832.

[37] Jiaming Song, Chenlin Meng, Stefano Ermon, Denoising diffusion implicit models, in: International Conference on Learning Representations, 2020.

[38] Prafulla Dhariwal, Alexander Nichol, Diffusion models beat gans on image synthesis, NeurIPS 34 (2021) 8780–8794.

[39] Yang Song, Stefano Ermon, Generative modeling by estimating gradients of the data distribution, NeurIPS 32 (2019).

[40] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, Ben Poole, Score-based generative modeling through stochastic differential equations, in: International Conference on Learning Representations, 2020.

[41] Yang Song, Stefano Ermon, Improved techniques for training score-based generative models, NeurIPS 33 (2020) 12438–12448.

[42] Yusuke Tashiro, Jiaming Song, Yang Song, Stefano Ermon, CSDI: Conditional score-based diffusion models for probabilistic time series imputation, NeurIPS 34 (2021) 24804–24816.

[43] Tianqi Chen, Carlos Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[44] Anna Veronika Dorogush, Vasily Ershov, Andrey Gulin, CatBoost: gradient boosting with categorical features support, 2018, arXiv preprint arXiv:1810.11363.

[45] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu, Lightgbm: A highly efficient gradient boosting decision tree, Adv. Neural Inf. Process. Syst. 30 (2017).

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, Illia Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[47] Gowthami Somepalli, Avi Schwarzschild, Micah Goldblum, C Bayan Bruss, Tom Goldstein, SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training, in: NeurIPS 2022 First Table Representation Workshop, 2022.

[48] Zifeng Wang, Jimeng Sun, Transtab: learning transferable tabular transformers across tables, Advances in Neural Information Processing Systems 35 (2022) 2902–2915.

[49] Ying Zhang, Baohang Zhou, Xiangrui Cai, Wenya Guo, Xiaoke Ding, Xiaojie Yuan, Missing value imputation in multivariate time series with end-to-end generative adversarial networks, Inform. Sci. 551 (2021) 67–82.

[50] Yonghong Luo, Ying Zhang, Xiangrui Cai, Xiaojie Yuan, E2gan: End-to-end generative adversarial network for multivariate time series imputation, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, AAAI Press, 2019, pp. 3094–3100.

[51] Calvin Luo, Understanding diffusion models: A unified perspective, 2022, arXiv preprint arXiv:2208.11970.

[52] Sumanth Chennupati, Ganesh Sistu, Senthil Yogamani, Samir A Rawashdeh, Multinet++: Multi-stream feature aggregation and geometric loss strategy for multi-task learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.

[53] Anna C Belkina, Christopher O Ciccolella, Rina Anno, Richard Halpert, Josef Spidlen, Jennifer E Snyder-Cappione, Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets, Nat. Commun. 10 (1) (2019) 5415.