Chinese Journal of Electronics
vol. 33, no. 3, pp. 776–792, May 2024
https://doi.org/10.23919/cje.2023.00.181

RESEARCH ARTICLE

# Extracting Integrated Features of Electronic Medical Records Big Data for Mortality and Phenotype Prediction

Fei LI[1,2], Yiqiang CHEN[1], Yang GU[1], and Yaowei WANG[2]

1. *Institute of Computing Technology, University of Chinese Academy of Sciences, Beijing 100080, China*
2. *Pengcheng Laboratory, Shenzhen 518055, China*

Corresponding author: Yiqiang CHEN, Email: yqchen@ict.ac.cn
Manuscript Received May 10, 2023; Accepted August 7, 2023


**Abstract** — The key to synthesizing the features of electronic medical records (EMR) big data and using them for specific medical purposes, such as mortality and phenotype prediction, is to integrate the individual medical event and the overall multivariate time series feature extraction automatically, as well as to alleviate data imbalance problems. This paper provides a general feature extraction method to reduce manual intervention and automatically process large-scale data. The processing uses two variational auto-encoders (VAEs) to automatically extract individual and global features. It avoids the well-known posterior collapse problem of Transformer VAE through a uniquely designed "proportional and stabilizing" mechanism and forms a unique means to alleviate the data imbalance problem. We conducted experiments using ICU-STAY patients' data from the MIMIC-III database and compared them with the mainstream EMR time series processing methods. The results show that the method extracts visible and comprehensive features, alleviates data imbalance problems and improves the accuracy in specific predicting tasks.

**Keywords** — Variational auto-encoder, Transformer, Electronic medical records, Posterior collapse, Imbalance data.

**Citation** — Fei LI, Yiqiang CHEN, Yang GU, *et al.*, "Extracting Integrated Features of Electronic Medical Records Big Data for Mortality and Phenotype Prediction," *Chinese Journal of Electronics*, vol. 33, no. 3, pp. 776–792, 2024. doi: 10.23919/cje.2023.00.181.

## I. Introduction

The analysis of electronic medical records (EMR) of big data is of great significance for diagnosis and treatment [1]. Over the years, hospitals have built many applications, almost all built on top of the database systems. A comprehensive collection of data related to electronic medical records from the back-end databases of these applications form high-volume, diverse, and time-spanning EMR big data, which can then be used for data analysis based on defined objectives (e.g., mortality and phenotype prediction), and using all relevant data often leads to better results.

However, there are many vital obstacles and challenges to EMR's extensive data analysis. One of the most significant difficulties is effectively extracting comprehensive features objectively from the massive data [2]–[4].

Hospital applications are often built around business processes, and their database systems prioritize serving business management over data analysis. Data related to analytic objectives are often scattered among various process details. Therefore, even when EMR big data has been collected and organized, the information available for defined analytic objectives (e.g., mortality and phenotype prediction) is difficult to access directly and is often hidden in many nested hierarchical database tables with different structures and specifications. At this point, each piece of data will likely be high-value data containing necessary information but more likely to be almost useless, and overall, the vast majority of data has a meager average value. Therefore, it is imperative for EMR big data analytics to automatically condense the valuable information in a single data record around the analysis objectives.

At the same time, for most healthcare analytics, the information in a single piece of healthcare data is insuffi-

Associate Editor: Prof. Yi PAN, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences.

cient to predict a specific medical diagnosis (e.g., mortality and phenotype). We need to combine multivariate time series characteristics of patients over time to give a relatively definitive assessment, sometimes even in the context of multiple hospital admissions.

In addition, EMR data analysis often encounters data imbalance problems, with some categories of samples being huge and others being particularly scarce. The immediate consequence of data imbalance is that it makes the analysis conclusions unfair and, in extreme cases, even makes the conclusions seem absurd in common sense. The problem in EMR big data analysis is severe, requiring additional natural correlations for large data tables. The imbalance in the original data tends to be amplified in the process, making the data imbalance almost unavoidable.

The current EMR analysis methods for addressing these issues could include two ideas:

One idea is first to assume a correlation between certain phenomena and specific disease outcomes based on professional knowledge, then filter the data to do a correlation analysis between phenomena and diseases, and then conclude. In this "small" data analysis scenario, all three of these problems can be avoided. However, the disadvantage is obvious:

It requires professionals to perform the delicate preprocessing work such as manual screening and manual completion of the data in advance. Manual processing is often inefficient and will become costly and unavailable. Additionally, people often have prior mental expectations and conjectures and will often operate according to subjective expectations in these feature selection works. It will be easy to fall into "survivorship bias". The researcher may ignore data subjectively considered irrelevant or weakly correlated (but may affect the analysis results) and then get wrong or unreliable conclusions.

The other idea is first to assume that certain specific intermediate features in a single record are meaningful from a professional perspective, then extract these intermediate features from single records to form a time series, and finally extract the overall features of the time series for prediction.

These methods have many advantages incorporating prior expertise. However, it is still a semi-automated process that requires human involvement in determining which intermediate features in a single record are meaningful for the outcome, which is still based on psychological expectations, and still has the potential to introduce an element of unfairness. In addition, it tends to lose implicit features by emphasizing the explicit features related to the final target orientation, which may lead to less accurate predictions.

In particular, when solving the imbalanced data problem, the first idea usually relies on the original distribution of the observed data, which is affected by the solution of most unbalanced data problems. The second idea involves two stages, which can also pose challenges in exerting controlled influence on the original distribution.

Instead, this paper proposes a generalized dual variational auto-encoder (VAE) feature extraction implementation to solve the related problems and uses mortality and phenotype prediction analysis objectives as examples for illustration. Its general block diagram is shown in Figure 1.

We first generate sparse multidimensional data by concatenating related database tables around patients and apply a specially designed 1D-CNN VAE to extract the dimensionality reduction features of a single medical record (considered as an event). After that, we treat the reduced-dimensional features of patients as multivariate time series and process them by Transformer VAE to extract global fusion features of patients for mortality prediction and phenotype prediction. In general, our proposed method has the following characteristics:

1) We provide an automatic generic method for feature extraction. It minimizes human intervention and enables the automatic processing of data at scale, and is suitable for big data analysis scenarios.

2) We consider a combination of single recording and global time-series feature auto-extraction. Both stages can be automatically optimized according to universal prediction objectives.

3) Based on the inherent advantage of VAE to generate any random data according to the distribution features, we can generate any number of different data on demand, thus forming a unique means to solve the data imbalance problem.

4) By a uniquely designed "proportional and stabilizing" mechanism, we avoid the well-known intractable posterior collapse problem [5] of Transformer VAE.

5) The on-demand generation of an arbitrary amount of data can also improve prediction accuracy.

## II. Background and Related Work

### 1. Data background

This paper uses the MIMIC-III database [6] for related research, but the readers can extend the proposed method to other databases.

The Medical Information Market of Intensive Care (MIMIC-III) is a sizeable center database that contains information about patients admitted to intensive care units in large tertiary care hospitals. Its data include vital signs, drugs, laboratory measurements, doctor's orders, operation codes, diagnostic codes, image reports, hospital stay, and survival data.

MIMIC-III contains data related to 53423 hospital admissions of adult patients (16 years old or older) admitted to the intensive care unit between 2001 and 2012 and 7870 newborns admitted between 2001 and 2008.

The database contains 26 data tables. Besides the dictionary tables, the tables are connected by a patient number (subject_id), one or more admission record num-

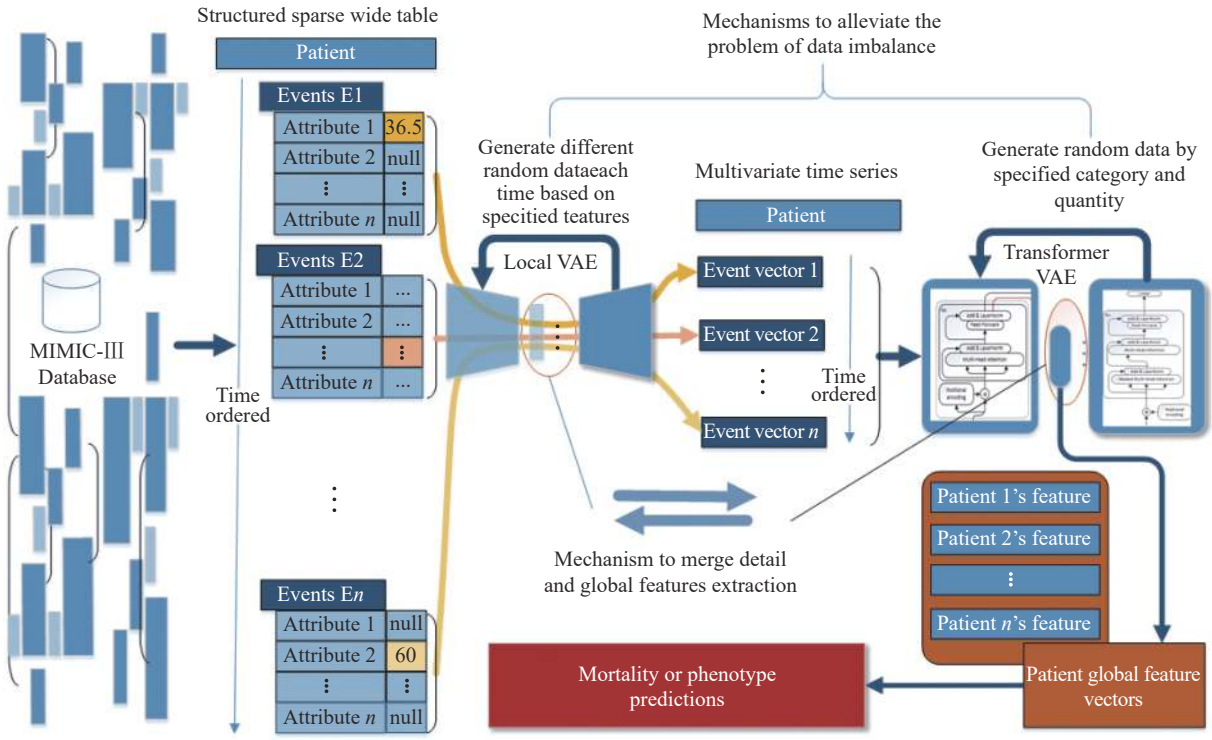


**Figure 1** The general structure of the proposed Dual VAE.

bers (hadm_id), and ICU numbers (icustay_id).

The dictionary table includes item code (d_item), surgery code (d_icd_procedures), and laboratory examination code (d_labs). One patient corresponds to one subject_id, but there may be multiple admissions and multiple hadm_id; one admission may involve multiple ICU entries, i.e., one hadm_ id may correspond to numerous icustay_id. Each ICU stay (indicated by icustay_ id) contains various events after the patient entered the ICU, such as DateTimeEvents, InputEvents (CV, MV), OutputEvents, LabEvents, MicrobiologyEvents, NoteEvents, Prescriptions, ProceduresEvents, ChartEvents, and CPTEvents.

## 2. Related works

1) Feature extraction in current EMR analysis tasks

Our primary purpose is to extract features from EMR big data for specific analysis purposes (prediction of mortality and phenotype). In this direction, there are three usual approaches: extracting only single record features; single record features directly composing time series and extracting time series features; single record extracting features, features composing time series and then extracting time series features for EMR data analysis.

In the current literature [7]–[11], biological, pharmaceutical, and medical researchers prefer to extract single record features for EMR analysis. The approaches usually begin with hypothesizing that there is some correlation between the phenomenon and the disease based on knowledge and principles within the field of expertise, hypothesizing one or more conclusions, and then selecting multiple features in the EMR relevant to the patient for validation or disproof. This analysis focuses on discovering or validating some professionally relevant scientific theories related to biology, pathology, and pharmacology. The amount of data used is usually tiny compared with big data, but each piece of data is rigorously screened, and data quality is ensured. The analytical approach is simple but highly interpretable. When massive extensive data are involved, it involves only simple statistical features, usually setting confidence interval thresholds for the association of individual features of the phenomenon with the disease and validating and disproving the original hypothesis by calculating p-values. Clearly, rather than relying on data, such analyses depend more on the background of the analyst's expertise.

There is also some literature [1], [2], [12]–[18] on automatic feature extraction using machine learning (including deep learning) methods for single records. These methods usually construct single records carefully, then extract certain features of single records by machine learning methods, and finally, use logistic regression or softmax to achieve relevant predictions. Since their feature extraction process does not consider the integration of time and space, part and whole, veracity and diversity of the analysis domain, usually, these methods are used to solve some analysis problems with explicit inference processes and obvious conclusions (including intermediate results). The difficulty of their work usually lies in the manual collation of raw data or the formulation of filtering rules.

Several studies have noted the importance of analyzing EMR data, composing time series from individual

records and extracting time series features for regression or classification. At this point, the excessive number of variables in multivariate time series is highly detrimental to the analysis, so compressing the number of variables becomes critical. In the literature [19], only fields of two tables (LabEvents, ChartEvents) were used to construct time series, thus limiting the number of variables; in the literature [20], [21], only four features of Disease, Age, Position, and Segment were extracted for time series analysis; in the literature [22], time series analysis networks were constructed separately for continuous features, clinical orders, and category data, the number of input variables for each time series network could be limited in this case.

Several papers have used two stages to extract features for EMR data analysis with exciting analytical results. They first customize a set of condensed features in a single record that can be obtained computationally and then use them to form a time series for analysis. The literature [23] uses EHR data through sequences of ICD-9 codes, CPT codes, medication lists, and codes representing patient visits, which are first unsupervised and pretrained through masked language modeling (MLM) and then fine-tuned downstream using another network in conjunction with a predicting depression task. Moreover, the literature [24], [25] first builds a semi-synthetic EHR dataset based on MIMIC-III using custom rules and then uses Transformer to train on the semi-synthetic dataset. While the literature [26] first transforms EHR data into topological map features and then applies multi-headed attention and Transformer on the topological map to accomplish the prediction task. In summary, all these methods use unsupervised methods to extract dimensionality reduction features in the first stage and then use a temporal analysis network in the second stage to complete the specific prediction task. The first stage needs to be optimized according to the prediction objective.

2) Imbalanced data problem and its solutions

Generally, the EMR big data is distributed unevenly in specific analysis tasks, and imbalanced data problem always exists in EMR/EHR analysis [27].

Traditionally, four Strategies for methods to deal with unbalanced category data exist [28]–[32]:

**Under-sampling** When a small number of samples does not affect the model training, the balance between a small number of samples and a large number of samples can be achieved by undersampling a large number of samples.

**Oversampling** When a small number of samples do not support model training, the balance between a small number of samples and a large number of samples can be achieved by oversampling a small number of samples.

**Weight allocation** By introducing a weighted model algorithm, focus on fitting a small number of samples to improve the learning of the characteristics of a small number of samples.

**Generate composite data** Create new composite points from small samples to increase their cardinality.

From the literature in recent years, many solutions in the field of EMR analysis are still a combination or refinement of these four ideas [33]–[36]. Solving problems based on VAE in this paper is still a method of generating composite data, but it has unique advantages.

3) VAE and its posterior collapse problem-solving

We study a global fusion feature extraction method based on the VAE (variational auto-encoder) [37] technique.

As an auto-encoder, VAE first has the function of extracting features by dimensionality reduction, after which it constructs a set of incomplete independent normally distributed random variables based on these features, lets these variables generate features, and then decodes them to obtain an output similar to the input. In this process, we can get the dimensionality reduction features and generate an arbitrary number of synthetic samples that are different from each other based on the features. Since the data is generated based on the category label orientation and overall data distribution characteristics, using VAE-generated data to solve the data imbalance problem is feasible and effective.

However, to utilize VAE in this scenario, we must overcome the well-known posterior collapse problem in VAE [38], which is particularly likely to occur in time-series-related VAE networks. In general, the cause of the posterior collapse in VAE models is the disappearance of KL divergence, which results in the decoder ignoring the posterior distribution generated by the Encoder and only sampling from the noise $N(0, I)$, thus invalidating the VAE network.

The usual idea to solve this problem is to transform KL to find more flexible priors and posteriors for the later variables. For example, setting a weight value for the KL loss so that the weights gradually anneal [38]–[40] is a good approach, but its drawback is that the annealing rate needs to be adjusted for different datasets.

In addition, Kingma *et al.* (2016) [41], Chen *et al.* (2017) [42], Razavi *et al.* (2019) [5], and Zhu *et al.* (2020) [43] proposed the concept of "free bits" so that each dimension of the KL term "retains a little space". In other words, if the KL value of this dimension is too small, we leave it untouched until it increases beyond a threshold value. This approach is easy to follow, but the proper threshold $\epsilon$ must be obtained after several manual attempts.

Kingma *et al.* [41] also proposed the idea of normalizing flow. Namely, sampling the latent variable by simple distribution and then making the latent variable more flexible by multiple reversible distribution transformations. Thus, the latent variable sampling is no longer restricted to Gaussian distribution to avoid vanishing KL.

Techniques applicable to other scenarios can also be used to solve the KL vanishing problem [44]–[56] (for NLP, music, music, video, molecular design, text-image generation, multimedia learning, text generation, material/

drug discovery, social relations, saliency prediction, sentence generation, computer vision, respectively), however, they are too closely related to specific application scenarios and are difficult to migrate to the scenario of this article. Therefore, we propose a novel "proportional and stabilizing" mechanism to solve the posterior collapse problem.

## III. Method

### 1. Global data collecting and organization

The current relational database system's "entity relationship" is relatively straightforward. Therefore, using foreign keys to collect all data related to the target entity in the database is feasible through repeated SQL "join" operations.

Take MIMIC-III as an example. The "subject_id" (which can be regarded as the patient ID) can be associated with "hadm_id" (stand for hospital admission ID or visits ID) and "icustay_id", then through the "icustay_id", the patient can be associated with various events. In various events, there may be various dictionary code fields, such as item code (defined in the "d_item" table), summary code ("d_icd_procedures"), and laboratory examination code ("d_labs"). The exact meaning of these codes can be obtained by looking up the corresponding dictionary table to get the code value pair, and then the value represented by each code can be obtained.

Take the column names (field names) involved in all event tables in the database as a collection, and use the elements of this collection as the column names of the new table to obtain a structured, sparse, multidimensional empty table. In this way, almost all SQL "join" outputs can find the corresponding fields and populate themselves.

This table has the following characteristics:

1) In the database, For events' attributes related to its "subject_id" and its affiliated "hadm_id", "icustay_id", the corresponding column names can always be found in the table. Therefore, the information about each event can be filled in the table. Therefore, the event information can be fully collected in the "subject_id".

2) For different table records, the position and quantity of columns with data may differ. The structure of the table can make it compatible with vibrant diversity.

3) A record in the table represents an event during a patient's stay in an ICU ward. Events can be sorted by time. The time sequence of events is one of the essential global characteristics.

4) Tables are multi-dimensional and sparse, which means that most columns are empty for any record in the table, and the table has colossal dimension reduction space.

In particular, SQL "join" operations tend to produce many empty columns. In extreme cases, the foreign key and its necessary auxiliary field may be insignificant. We do not do special handling (we impute missing data in the simplest way possible, i.e., 0, blank, or false). For missing features, we form a missing vector for fields' "missing" state, with a default setting of 1 and 0 denoting absence. Our subsequent local VAE is an auto-coding dimensionality reduction network that requires the output to match the input as closely as possible. After training, it will automatically filter out low-density information.

The different "hadm_id" obviously represent different "visit" records of the patient. In other papers [20], [21], [23], the data are specifically processed manually to make them different time series. However, when broken down completely, there are deeper sub-series within the time series, such as ICU stay, ward transfer, surgery, and drug administration, all of which can be seen as sub-visits within a visit. We do nothing special in our method except time ordering since the global Transformer VAE designed for multivariate time series will capture long-range dependent.

For multimodal data fusion, we only do simple processing in the pre-processing stage: the category data is processed as one hot vector; all-natural language description fields will be merged and converted into numerical vectors using TensorFlow's default vocabulary processor.

### 2. The design of Dual-VAE

1) The local VAE

The primary purpose of designing the unique local feature extraction VAE is to extract dimension-reduction features from a single ICU-STAY event. Figure 2 shows its necessary auxiliary information and structure.

In this process, we first define the "proportional" mechanism: the proportional mechanism refers to the inclusion of a loss function term in the VAE design so that the relationship between the high-dimensional reconstructed data and the low-dimensional feature hidden variables can be maintained at a roughly proportional distance, which makes the magnitude of change of the reconstructed variables consistent with the magnitude of change of the hidden variables.

We add a loss function item to constrain hidden variables:

$$l_{\text{e-proportional}} = \left| ||\hat{x} - \hat{x}'|| - k||z_0 - z_0^{'}|| \right| \tag{1}$$

where $z_0$ is a dimension reduction feature extracted from the input single medical event record $x$, and $z_0'$ is a randomly generated vector consistent with the $z_0$ structure. $\hat{x}$ and $\hat{x}'$ are the reconstructed samples according to $z_0$ and $z_0'$, respectively. It requires that, for any pair of vectors $z_0$ and $z_0'$, if there is a distance between them, there should also be a distance between the reconstruction vectors $\hat{x}$ and $\hat{x}'$ generated by taking $z_0$ and $z_0'$ as their hidden variables. In other words, the distance between them should be proportional to the distance between $z_0$ and $z_0'$ as far as possible, which is also described in (2).

$$||\hat{x} - \hat{x}'|| \approx k||z_0 - z_0'|| \tag{2}$$

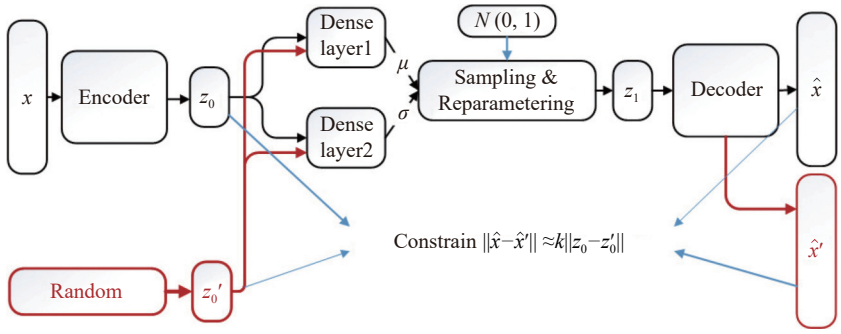


**Figure 2** The structure of our local VAE.

The advantage of this is that it projects the hidden variable and the generation record each into a space similar in Euclidean distance so that two identical generation records have a short distance in both the original high-dimensional space and the hidden variable low-dimensional space. In comparison, the two different generation records are extended in both spaces. On the one hand, this can facilitate the position coding in the next Transformer VAE, and there will be no gradient explosion; On the other hand, it enables the KL calculation to consider other distributions different from the Gaussian distribution, which is convenient for solving the posterior collapse problem.

Another key to the local VAE design is the design of its Encoder and decoder. Here, we implement encoding and decoding based on 1D-CNN as shown in Figure 3.

For encoding, a 1D convolutional neural network (1D-CNN) is used for dimensionality reduction (shown in Figure 3(a)), while for decoding, the corresponding deconvolutional design is used (shown in Figure 3(b)).

In Figure 3(a), the input data is normalized and discretized to 2074 dimensions. It undergoes three convolutions and then global pooling to reduce the dimensionality to 34 dimensions, and we perform a batch normalization and pooling operation for each convolution. We finally use a Dense layer to align the dimensionality reduction data to 28 dimensions.

In Figure 3(b), 28-dimensional hidden variables are expanded to 34 dimensions by a Dense layer, then up-

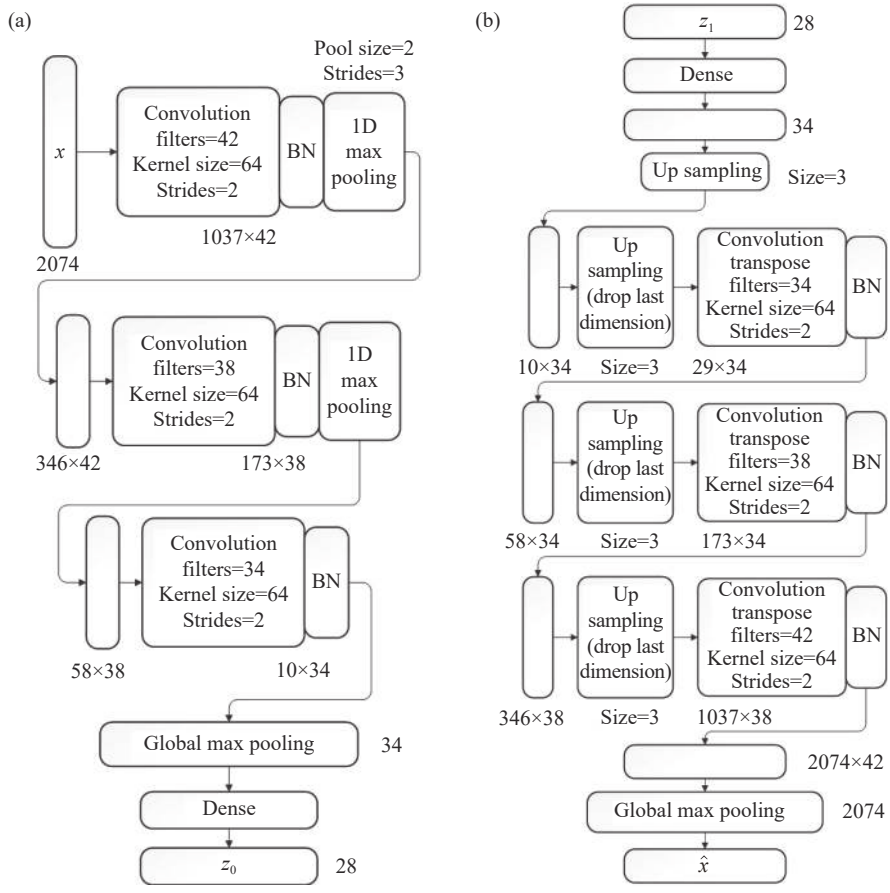


**Figure 3** The encoder and decoder design of the local VAE. (a) Endocer; (b) Decoder.

sampled to 10 × 34. After three 1D deconvolutions, the dimension is expanded to 2074 by global pooling.

In this way, we realize the coding and decoding of the 1D-CNN process, whose encoding and extraction of features are almost the same as decoding and reconstructing data.

2) The Transformer VAE

A specially designed Transformer VAE is used to extract the time series features of one patient. The key to this is to solve the posterior collapse problem of the Transformer VAE.

First, we design the structure of Transformer VAE as Figure 4. It accepts a collection of hidden variable vectors as multivariate times series as its input.

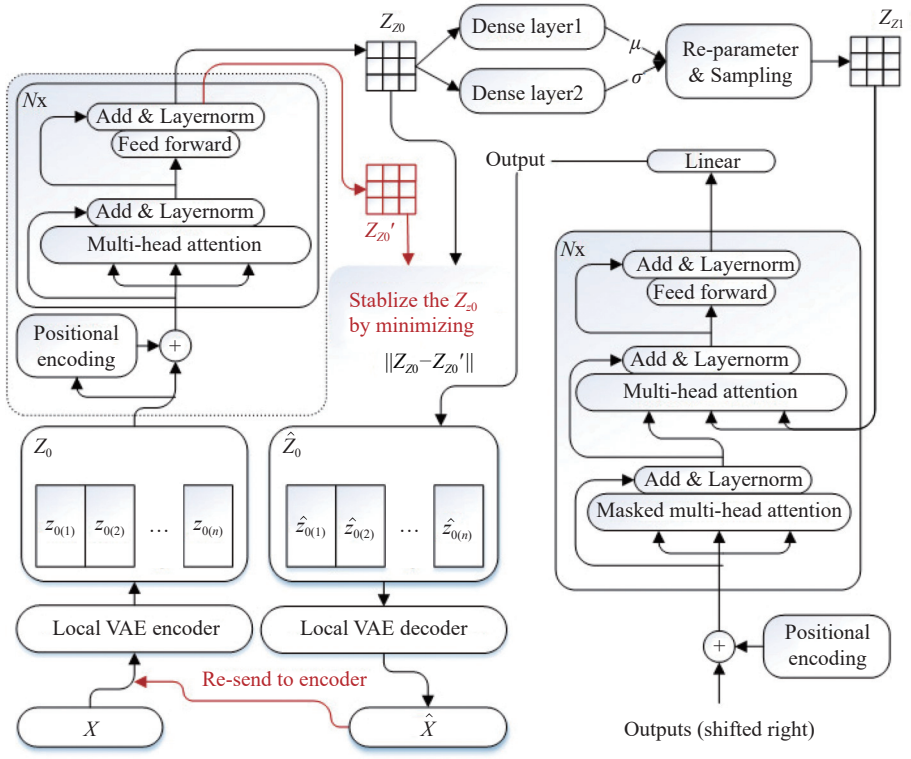


**Figure 4** The diagram of the Transformer VAE.

We can minimize the loss in the local VAE. Therefore, if Transformer VAE generates the required hidden variable $Z_{Z0}$, it is deemed to create a series of $z_0$ and then a series of $\hat{x}$. After training, the Transformer VAE can ensure the creation of proper $z_0$ series approximately, but cannot ensure the creation of approximate proper $\hat{x}$ series without the "proportional" mechanism because there always be a slight error between the $\hat{z}_0$ and $z_0$. Once $\hat{z}_0$ and $z_0$ exists little difference, there would be a total complete new $\hat{x}$ created. The "proportional" mechanism ensures that, if there is a slight deviation of $\hat{z}_0$, the departure of $\hat{x}$ is proportional to it, avoiding the situation that the slight variation of $\hat{z}_0$ causes a sharp expansion of $\hat{x}$ deviation.

For the solved $\hat{Z}_0$ by the Transformer VAE decoder, we will re-send it to the encoders (local VAE encoder and Transformer VAE encoder) to generate $Z'_{Z0}$ and require the minimum distance between $Z'_{Z0}$ and the original $Z_{Z0}$.

This additional process makes the decoder inverted into an "encoder". In contrast, the original encoder is switched to "decoder" this time, giving the original encoder a chance to optimize from the representation when participating in the optimization. Thus, when the original codec extracts consistent features, the gradient descent has a clear and convenient path.

3) Balanced sampling design

This paper uses the feature that VAE can generate different reconstruction data for the same sample and designs a balanced sampling mechanism to solve the imbalanced data problem.

We conduct a balanced sampling with the replacement method for each data category to ensure that every category of sampled data is equal when training.

Suppose that the number of multivariate time series in the dataset is $n$, the number of label categories is $k$, and the number of iterations in the dataset is $c$.

We take $[n/k]$ as the balanced sampling times of each category in the entire dataset. That is to say, no matter whether the number of data records in each category is greater than or less than this number, the data within the category sampled before will be put back for sampling next time until the whole category sampled $[n/k]$ times.

In this way, for small class data, even if each data is sampled multiple times, each sampling generates non-

repeated reconstruction data with a variation. Therefore, we can extract class features as much as possible, and the whole network can pay full attention to small class data.

For large categories of data, the key lies in the participation of each data since each data is precious. We need to let each data have a high probability of participating in the training process.

Suppose there are $s$ records in a large category. In each training iteration, only $[n/k]$ records are sampled into the training process. Therefore, the probability of not being collected for each record in this category is $1-n/(ks)$. After $c$ rounds of iteration, the chance of not being collected is $(1-n/(ks))^c$. Therefore, no matter how uneven the data is, as long as $c$ is large enough, the probability of not being collected for each record will be small enough to guarantee that every record participates in the training process. In extreme cases, for a data set with 25 categories and 1/2 of them is large categories, the probability of each record not being hit has been reduced to less than 50% after only 9 iterations. While the general situation accounts for 1/10 in significant categories, the probability of not being taken after 10 iterations decrease to 0.01%.

## 3. Loss function constraint in training process

1) The local VAE's loss function and its training

The local VAE must implement the function of basic VAE, Which will be realized through the formulas below:

$$\begin{cases} l_{e\_1}(x) = \mathrm{RMSE}(x,\hat{x}) = \sqrt{\dfrac{1}{d}\displaystyle\sum_{k=1}^{d}\left(x_{(k)}-\hat{x}_{(k)}\right)^2} \\ l_{e\_2}(x) = \mathrm{KL}(p(z_0 \mid x)\|q(z_0)) \\ \qquad\quad = \dfrac{1}{2}\displaystyle\sum_{k=1}^{d}\left(\mu^2_{(k)}(x)+\sigma^2_{(k)}(x)-\ln\sigma^2_{(k)}(x)-1\right) \\ l_{e\text{-vae}}(x) = l_{e\_1}(x)+l_{e\_2}(x) \end{cases} \tag{3}$$

The previously mentioned formula 1 uses $l_{e\text{-proportional}}$ to support the prevention of posterior collapse, and it should be included. Therefore, the loss of VAE during initial training can be expressed by the formula below:

$$l_{e\text{-dividual}}(x) = \alpha_e l_{e\text{-vae}}(x) + \beta_e l_{e\text{-proportional}}(x) \tag{4}$$

where $\alpha_e$ and $\beta_e$ are the weights of various losses, which can be adjusted during training.

When the loss of the Transformer VAE training is relatively stable, we add it to the alternative training so that the transformer VAE optimization can be considered in the local VAE training, thus promoting the efficiency improvement of the entire model.

In order to achieve this goal, the individual record features $z_0$ involved in the patient are combined to form a time series $Z_0$, which is then sent to the Transformer VAE, and then the Transformer VAE's loss of $Z_0$ is obtained from the output. The loss will be uniformly distributed to $z_0$. In addition, this process is suitable for a process at the end of the epoch.

Therefore, in alternative training, at the end of every epoch, the weights of the local VAE will be adjusted once according to the equation below:

$$\begin{cases} l_t(x) = \bar{l}_t(Z_0) \\ l_{e\text{-co}}(x) = l_{e\text{-dividual}}(x) + l_t(x) \end{cases} \tag{5}$$

2) The Transformer VAE's training and its loss functions

Similarly, for Transformer VAE, to achieve VAE, we use the formulas below:

$$\begin{cases} l_{t\_1}(Z_0) = RMSE(Z_0,\hat{Z}_0) = \sqrt{\dfrac{1}{D}\displaystyle\sum_{k=1}^{D}\left(Z_{0(k)}-\hat{Z}_{0(k)}\right)^2} \\ l_{t\_2}(Z_0) = KL(p(Zz_0 \mid Z_0)\|q(Z_0)) \\ \qquad\quad = \dfrac{1}{2}\displaystyle\sum_{k=1}^{D}\left(\mu^2_{(k)}(Z_0)+\sigma^2_{(k)}(Z_0)-\ln\sigma^2_{(k)}(Z_0)-1\right) \\ l_{t\text{-vae}}(Z_0) = l_{t\_1}(Z_0)+l_{t\_2}(Z_0) \end{cases} \tag{6}$$

When the $Z_0$ and $\hat{Z}_0$ dimensions do not match, the missing dimensions will be filled with zeroes.

To support the mechanism for preventing posterior collapse, we add the following:

$$l_{t\text{-stablize}}(Z_0) = ||Z_{z0} - Z'_{z0}|| \tag{7}$$

To support the binary classification of predicting mortality, we add the following:

$$l_{\text{task}}(Z'_{z0}, y) = -[y\log\hat{y} + (1-y)\log(1-\hat{y})] \tag{8}$$

where $Z'_{z0}$ is the output of the Transformer VAE labeled in Figure 3, following a sigmoid output layer, a binary probability $\hat{y}$ output can be obtained. When predicting mortality, the $y$ is 1 for mortality and 0 for surviving.

Phenotyping is a combination of binary classifications. Assuming that there are $K$ phenotypes, the network needs $K$ sigmoid outputs $\hat{y}_k$ for the ground truth $y_k$, where each output represents the probability of predicting the presence of a particular phenotype. Thus, the classification loss will be:

$$l_{\text{task}}(Z'_{z0}, Y) = \frac{1}{k}\sum_{k=1}^{K}[y_k\log(\hat{y}_k) - (1-y_k)\log(1-\hat{y}_k)] \tag{9}$$

Therefore, the loss of the Transformer VAE during training can be expressed by

$$l_t(Z_0, y) = \alpha_t l_{t\text{-vae}}(Z_0) + \beta_t l_{t\text{-stablize}}(Z_0) + l_{\text{task}} \tag{10}$$

where $\alpha_t$ and $\beta_t$ are the weights of various losses, which can be adjusted during training, and $l_{\text{task}}$ is chosen as (8) or (9), depending on the specific task.

3) Alternate training and combination of the dual VAE

First, we can pre-train the local VAE to encode a single medical event $x$ into a hidden variable $z_0$. These variables $z_0$ are collected under the patient's name to form time series data $Z_0$ for the Transformer VAE. This time we only use (1) as the optimization loss function without considering the Transformer VAE loss.

Then, the Transformer VAE can be trained with these time series data ($Z_0$) as input. Refer to (10) as its loss function.

When the Transformer VAE trains for several rounds and gets stable, it can train the local VAE alternately. The Transformer VAE losses can be added to the training using the loss function of (5). After alternately training the local VAE and the Transformer VAE several times, the entire network can achieve optimal status.

## IV. Experiments

We conduct our experiments with the following hardware and software environments:

Hardware: CPU XEON 2698v4 (20 cores and 40 threads), memory 64 GB, SSD 1 TB, GPU Nvidia Tesla V100.

Software: operating system Ubuntu 18.04.3 LTS, Nvidia driver version 440.33.01, CUDA v10.2, cuDNN v7.6.5, Keras GPU 2.3.1, Tensorflow GPU 1.15.0, Sci-kit Learn 0.19.0.

### 1. Unsupervised feature extraction

We expect to judge the natural feature extraction effect of the method in this paper by observations. Therefore, we extract features by the Dual VAE without specifying any analysis objectives and observing the visualization of the features using PCA and t-SNE.

We labeled the individual samples (time series data with patients as the integrated object) to observe the features. That is, we color-code the samples with various phenotype combinations.

In the MIMIC-III data, the phenotypes are 25, and these phenotypes are attached to each sample in the form of multiple labels, i.e., the same sample can be attributed to multiple phenotypes. 25 phenotypes can be combined into $(2^{25}-1)$ categories, but the most common ones are only a few hundred. Therefore, we extract the most common 467 combinations to color the samples with 467 colors.

Figure 5 is the visualization of the effects of the proposed Dual VAE.

From Figure 5, we can see that:

The sample points in Figure 5(a) are not easily classified. Although these samples are normalized and discretized, most are concentrated in a small area, and a minimal number of them have a massive variance, so the maximum PCA scale is stretched to hundreds for all samples with unidimensional values around 1. The situation is significantly improved for the sample points in Figure 5(b), where the sample points are more evenly distributed, and the scale is reduced to a maximum of 1.25.

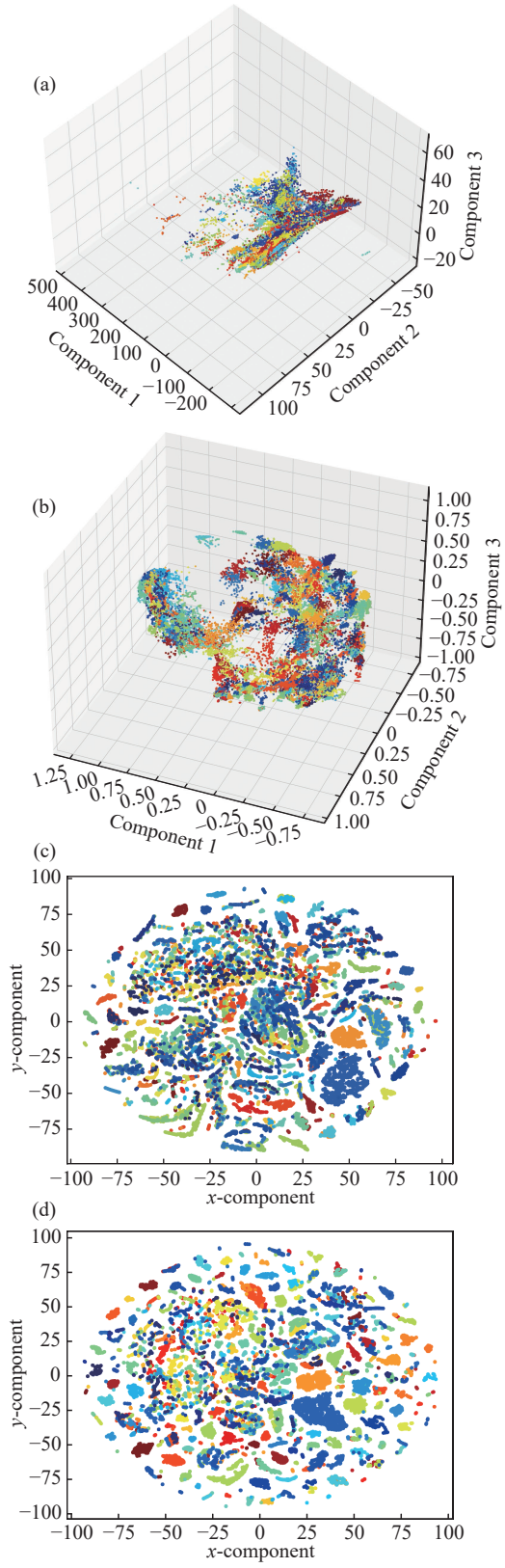


**Figure 5** The overall visualization (dimensionless) of the effects of the proposed Dual VAE. (a)–(b) The 3-components PCA of (a) the 467 combinations of the original samples, which we have already normalized preprocessed, and (b) the 467 combination features, which the proposed Dual VAE extracts. (c)–(d) The 2-dimension t-SNE of (c) the original samples and (d) the features we extracted by the proposed Dual VAE method.

Figures 5(c) and (d) show the divisibility of features more intuitively. In Figure 5(c), although a few categories are divisible, most categories have mixed and entangled samples, which are difficult to distinguish. In Figure 5(d), the samples clearly show a tendency to cluster in similar categories and move away from each other in different categories, and the separability is significantly improved.

## 2. Feature extraction for predicting mortality task

We can extract global fusion features and optimize them for specific binary classification tasks (mortality and phenotype prediction).

When the global features are extracted completely, we randomly select 400 samples each time and use them to observe the distribution of the samples. We do this because most of the overall samples are not mortal; if they are displayed thoroughly, the mortality samples will be obscured.

The illustration is shown in Figure 6.

In contrast, Figures 6(a2), (b2), (c2) and (d2) show that the sample has been separated clearly after the Dual VAE treatment. Even visual observation can predict efficiently. The usefulness of the method is prominent in this case.

We studied the results using mortality as a positive case and non-mortal as a negative case. Most patients eventually recover after leaving the ICU. Mortal cases are rare compared to recovery cases. In the MIMIC-III database, only 11.5% of the patients were mortal cases [6]. If the test data were sampled naturally proportionally and tested by the 10-fold cross-validation method, the experimental results are shown in Table 1.

However, we are more concerned with positive cases because of the cost of incorrect predictions. When the model predicts a non-mortal patient as mortality, we lose

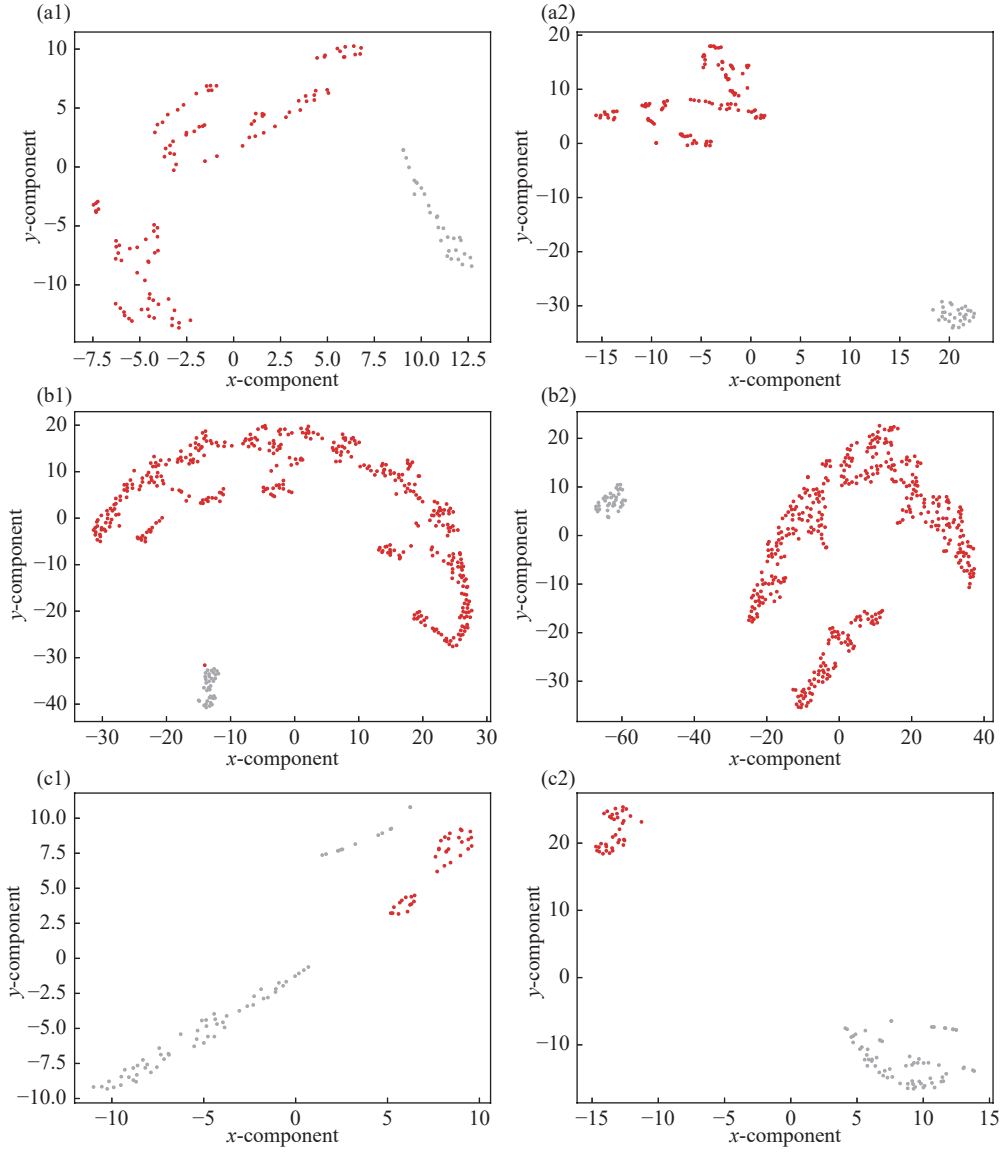


**Figure 6** (to be continued)

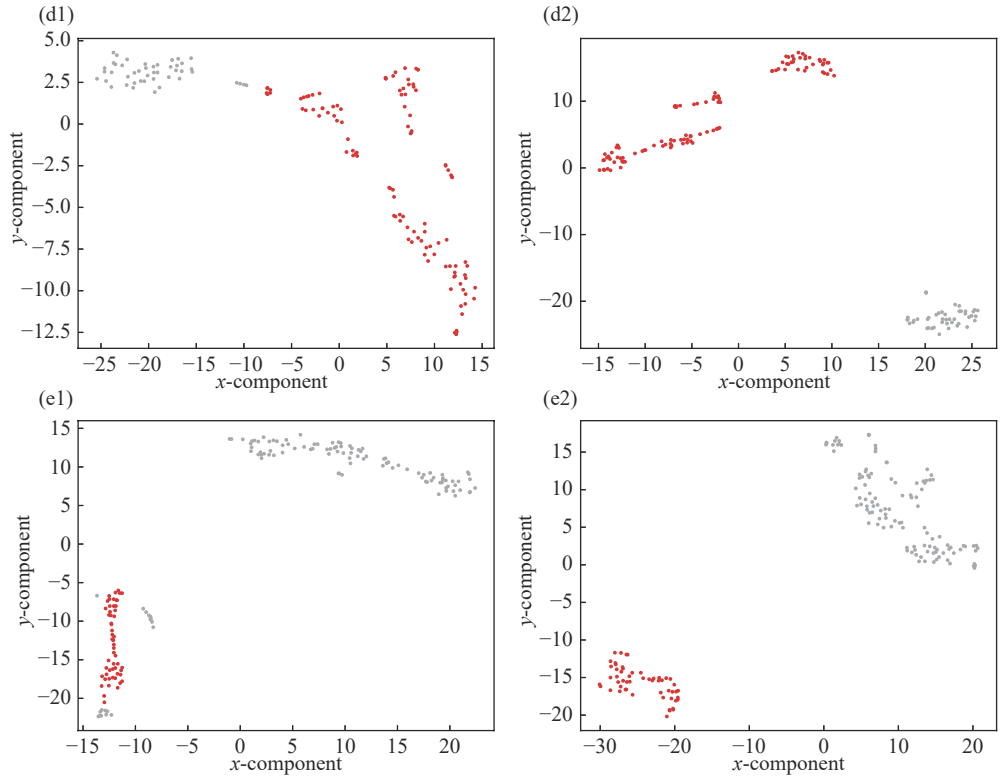


**Figure 6** The left side (a1), (b1), (c1), (d1) and (e1) show the *original* data distribution (dimensionless) of Sample sets 1, 2, 3, 4, 5, respectively, note that the distribution is different each time because a small number of samples are randomly selected. The right side (a2), (b2), (c2), (d2) and (e2) show the distribution of the *extracted features* from the left original data. The features on the right are more favorable for discrimination. (continued)

some medical resources at most, whereas predicting a mortal patient as non-mortal is likely to put the patient in mortal danger. In addition, when a patient reaches a point where their mortality needs to be predicted, the patient or physician must have felt the presence of a death threat, in which case the mortality rate must be higher than the statistical cases. Therefore, it is necessary to artificially draw positive and negative samples in equal proportions to test the model. The results of the equal proportions test for the same model are shown in Table 2. By comparing Tables 1 and 2 we find that:

1) Our model performs better in the equal proportion test set than the natural proportion test set. In a balanced dataset, our model will work well and has a high precision due to the raw data proportion having a large proportion of non-mortality. The AUROC remained virtually unchanged, unaffected by the proportion of samples tested. Thus the artificial sampling does not affect the actual accuracy.

2) There are no significant fluctuations in each score from "In 24 hours" to "In 48 hours" and "Category", which means that the model can extract global features that can capture long-range dependencies.

3) In Table 1, the precision is exceptionally low, leading to an exact low F1 and AUPRC. Comparing Table 2 shows that this is an unfair result due to a severe imbalance in the data. The results are significantly different simply because of the different proportions of positive and negative cases in the test set. Suppose we invert the positive and negative cases, with non-mortality as positive and mortality as negative (which means that we emphasize the precisions of non-mortality predictions more). In that case, we get Table 3.

**Table 1** The mortality prediction results (natural proportions)

| Case | Acc | Pre | Rec | F1 | ROC | PR |
|---|---|---|---|---|---|---|
| In 24 h | 0.881 | 0.309 | 0.871 | 0.456 | 0.876 | 0.590 |
| In 48 h | 0.882 | 0.412 | 0.872 | 0.560 | 0.877 | 0.642 |
| Category | 0.871 | 0.467 | 0.867 | 0.607 | 0.869 | 0.667 |

**Table 2** The mortality prediction results (equal proportions)

| Case | Acc | Pre | Rec | F1 | ROC | PRC |
|---|---|---|---|---|---|---|
| In 24 h | 0.875 | 0.880 | 0.869 | 0.874 | 0.875 | 0.874 |
| In 48 h | 0.876 | 0.881 | 0.870 | 0.875 | 0.876 | 0.875 |
| Category | 0.870 | 0.871 | 0.869 | 0.870 | 0.870 | 0.870 |

**Table 3** The mortality prediction results (Invert the positive and negative)

| Case | Acc | Pre | Rec | F1 | ROC | PR |
|---|---|---|---|---|---|---|
| In 24h | 0.880 | 0.991 | 0.881 | 0.933 | 0.876 | 0.936 |
| In 48h | 0.877 | 0.986 | 0.877 | 0.929 | 0.875 | 0.932 |
| Category | 0.881 | 0.981 | 0.883 | 0.929 | 0.875 | 0.932 |

The scores in Table 3 look very pleasing and create

the illusion that the model is much more capable than Table 1, while these are just different test results for the same model.

## 3. Feature extraction for predicting phenotypes

We can extract global fusion features according to the purpose of a specific multi-label classification task and optimize the global features to favor phenotypic classification.

First, we observe the PCA and t-SNE visualization images of the extracted features, as shown in Figure 7.

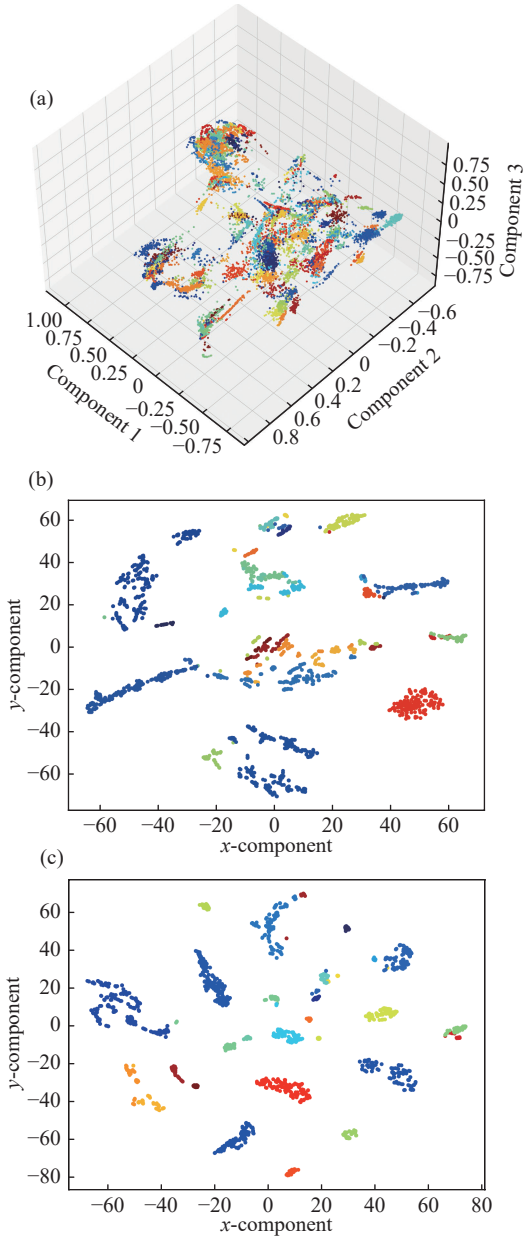


**Figure 7** The visualization (dimensionless) of phenotypic objective feature extraction. (a) The 3-component PCA illustration of phenotypies; (b) The unsupervised feature extraction of t-SNE from 25 random combinations; (c) The Dual VAE processed features of t-SNE from 25 random com-binations.

From Figure 7(a), it can be seen that the goal-oriented PCA makes the orientation of the data more apparent compared to the unsupervised feature extraction without imposing analysis goals (refer to Figure 5(b)). Moreover, comparing Figure 7(b) and (c), it can be seen that although this is a multi-label classification, the analysis goal orientation makes the sample distribution more consistent with the multi-class classification characteristics. We display the multi-label classification results in Table 4 as below.

**Table 4** Phenotypies prediction task results

| Phenotype | ROC | PR |
|---|---|---|
| Acute and unspecified renal failure | 0.844 | 0.729 |
| Acute cerebrovascular disease | 0.946 | 0.815 |
| Acute myocardial infarction | 0.812 | 0.696 |
| Cardiac dysrhythmias | 0.708 | 0.622 |
| Chronic kidney disease | 0.801 | 0.695 |
| Chronic obstructive pulmonary disease | 0.713 | 0.622 |
| Complications of surgical/medical care | 0.758 | 0.654 |
| Conduction disorders | 0.760 | 0.667 |
| Congestive heart failure; Nonhypertensive | 0.795 | 0.690 |
| Coronary atherosclerosis and related | 0.828 | 0.724 |
| Diabetes mellitus with complications | 0.915 | 0.793 |
| Diabetes mellitus without complication | 0.830 | 0.712 |
| Disorders of lipid metabolism | 0.751 | 0.659 |
| Essential hypertension | 0.703 | 0.618 |
| Fluid and electrolyte disorders | 0.771 | 0.662 |
| Gastrointestinal hemorrhage | 0.771 | 0.683 |
| Hypertension with complications | 0.772 | 0.671 |
| Other liver diseases | 0.802 | 0.697 |
| Other lower respiratory disease | 0.729 | 0.627 |
| Other upper respiratory disease | 0.808 | 0.705 |
| Pleurisy; Pneumothorax; Pulmonary collapse | 0.727 | 0.645 |
| Pneumonia | 0.842 | 0.724 |
| Respiratory failure; Insufficiency; Arrest | 0.949 | 0.822 |
| Septicemia (except in labor) | 0.888 | 0.767 |
| Shock | 0.929 | 0.804 |
| All acute diseases (macro-averaged) | 0.829 | 0.718 |
| All mixed (macro-averaged) | 0.796 | 0.694 |
| All chronic diseases (macro-averaged) | 0.771 | 0.671 |
| All diseases (macro-averaged) | 0.806 | 0.700 |

Table 4 shows that compared to the mortality prediction task, the method performs relatively worse in the phenotype prediction tasks but is still available for consideration. This downside is mainly because the imbalance between positive and negative cases in phenotype prediction is more severe. However, the model still shows a relatively good side: due to balanced sampling by phenotypes (with the same number of samples for each phenotype and an equal proportion for positive and negative classes), the AUROC and AUPRC of each phenotype tend to be consistent.

## 4. Ablation experiment

1) Ablation of the local VAE

Instead of extracting single record features, we let the sparse and high-dimensional data go directly to Transformer VAE for mortality and phenotypes prediction.

In this case, too many variables would make it difficult to process the time series data in Transformer, so here we use the method in the literature[19] to process the raw data, i.e., we only use the data from the ChartEvents and LabEvents tables and normalize and discretize the data before entering the Transformer. The sparse high-dimensional data $X$ is directly substituted into $Z_0$, and (6), (7), (8) are used as the loss functions for training. We observed about a 0.23 reduction in AUROC. In addition, due to the lack of the "proportional and stabilizing" mechanism, there is an approximately 1/3 chance that Transformer VAE will experience a posterior collapse randomly.

2) Ablation of posterior collapse prevention mechanisms

Our posterior collapse prevention mechanism consists of two parts: the use of the "proportional" mechanism in the local VAE, i.e., formula (1); the other is the re-feeding of the reconstructed samples into the Encoder in the Transformer VAE and the use of formula (6). When these two components are missing, posterior collapse is inevitable. We experiment with mortality prediction (the phenotyping experiment is similar and is omitted here).

We observed the training accuracy curve in Transformer VAE training, as shown in Figure 8.

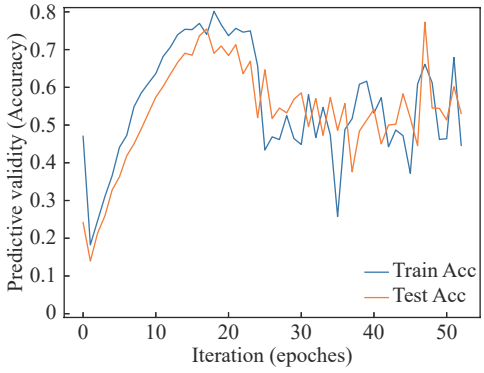


**Figure 8** The train accuracy curve of posterior collapse.

Figure 8 shows a brief upward trend in accuracy during the initial training because the local VAE has been fully trained. However, with the gradual adjustment of the network weights, accuracy violently oscillates, drops sharply, and finally oscillates around 0.5. The network's hidden layer no longer learns from the data but from $N(0,1)$.

3) Ablation of imbalanced data problem alleviation mechanisms

The approach in this paper focuses on alleviating the data imbalance problem by generating high-quality synthetic data on demand. Therefore, we compare the mortality data under the condition of an equal number of data bars (10000 patients) at 1 to 1, 1 to 2, ... and 1 to 18 to compare their ROC-AUC (note that the natural ratio of the original data is 0.085:1, which can be treated as if no data imbalance mitigation mechanism is used). The results are shown in Figure 9.

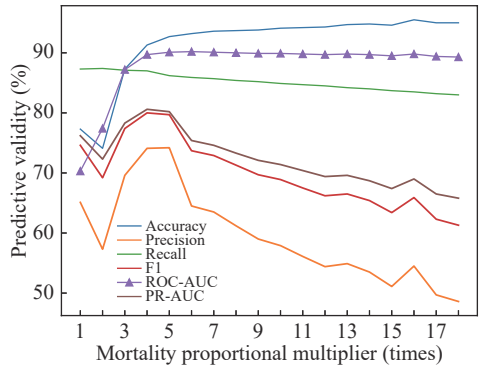


**Figure 9** The results of different mortality proportional multipliers in training set.

As seen from Figure 9, the ROC-AUC of this model is lower than the natural ratio when the mortality samples to non-mortality samples are lower than 1 to 4. Paying particular attention to the mortality category is not for free.

However, when the ratio of mortality samples to non-mortality samples reaches 1:6, the ROC-AUC of the model in this paper reaches its peak, and its value is higher than the natural ratio.

Therefore, it can be concluded that:

i) The sample rationing affects the final model effect.

ii) The natural ratio of mortality and non-mortality samples could be worse for constructing the classification model. However, the sample balance could also differ from the equal number of samples in both categories.

iii) The method in this paper can alleviate the sample imbalance problem in mortality prediction.

This ablation experiment construction method can also be utilized to explore the training sample ratio of the optimal model for binary classification.

## 5. Comparison experiment

1) Comparing with benchmarks

We give the mortality accuracy results compared with the benchmarks [19], as shown in Table 5 (natural proportion in the test set) and Table 6 (natural proportion).

In order to be consistent in caliber with the compared method, only events from two tables (ChartEvents and LabEvents) are used in the experiment (the method being compared does not lend itself to too many tables), which makes the results of our method here slightly inconsistent with the results the previous separate experi-

**Table 5** The mortality prediction results with benchmarks (natural proportion)

| Method | Acc | Pre | Recall | F1 | ROC | PR |
|---|---|---|---|---|---|---|
| Dual VAE | 0.879 | 0.250 | 0.900 | 0.391 | 0.889 | 0.575 |
| Logistics regression | 0.846 | 0.200 | 0.850 | 0.323 | 0.848 | 0.525 |
| Channel-wise LSTM | 0.862 | 0.220 | 0.860 | 0.350 | 0.861 | 0.540 |
| Multi-task chanel-wise LSTM | 0.869 | 0.232 | 0.880 | 0.367 | 0.874 | 0.556 |

**Table 6** The mortality prediction results compared with benchmarks (equal proportion)

| Method | Acc | Pre | Recall | F1 | ROC | PR |
|---|---|---|---|---|---|---|
| Dual VAE | 0.889 | 0.881 | 0.900 | 0.890 | 0.889 | 0.890 |
| Logistics regression | 0.848 | 0.847 | 0.850 | 0.848 | 0.848 | 0.848 |
| Channel-wise LSTM | 0.861 | 0.862 | 0.860 | 0.861 | 0.861 | 0.861 |
| Multi-task chanel-wise LSTM | 0.874 | 0.870 | 0.880 | 0.875 | 0.874 | 0.875 |

ments. Using only two tables limits the method's performance, and the true capability of our method is not fully realized here.

As can be seen from the two tables:

i) The main reason for pulling down the Precision is the extreme data imbalance.

ii) The method in this paper can alleviate the data imbalance problem.

iii) The method in this paper improves prediction accuracy (AUC-ROC) while alleviating the data imbalance problem.

2) Comparing with the latest methods

The method in this paper is compared with the current methods related to EHR and EMR time series analysis. Since the prediction objectives of each method are different, the methods of constructing training and test sets are also different. For this reason, this paper takes their basic ideas and modifies them appropriately according to the mortality prediction task objectives. Meanwhile, in the training set and test set, the ratio of mortality to non-mortality was adjusted to 1:6 according to the random sampling, which is likely to be the optimal ratio for achieving the best model in the MIMIC-III database based on the results of the ablation experiments of alleviating the data imbalance problem in this paper.

The results of the experiments are shown in Table 7.

From the table, the experimental results of the methods in this paper are relatively good when compared with the latest methods.

**Table 7** The mortality prediction results compared with the latest methods

| Method | Acc | Pre | Recall | F1 | ROC | PR |
|---|---|---|---|---|---|---|
| **Dual-VAE** (ours) | **0.920** | **0.667** | 0.877 | **0.758** | **0.902** | **0.772** |
| BEHRT [21] | 0.914 | 0.648 | 0.873 | 0.744 | 0.897 | 0.761 |
| HI-BEHRT [20] | 0.917 | 0.657 | 0.874 | 0.750 | 0.899 | 0.766 |
| T-BEHRT [25] | 0.905 | 0.618 | 0.872 | 0.723 | 0.891 | 0.745 |
| HiTANet [57] | 0.820 | 0.435 | 0.870 | 0.580 | 0.841 | 0.653 |
| BRLTM [23] | 0.847 | 0.479 | 0.862 | 0.616 | 0.853 | 0.671 |
| MUFASA [22] | 0.844 | 0.476 | **0.901** | 0.623 | 0.868 | 0.689 |
| SETOR [26] | 0.888 | 0.574 | 0.833 | 0.680 | 0.865 | 0.704 |

## V. Discussion

In this paper, we construct a Dual VAE, which integrates two feature extraction processes (the single medical event and the integrating time series feature extraction). By using these means, we can finally extract global fusion features of the whole database for medical electronic medical record analysis.

Extracting individual medical event features makes the implied variables proportional to the reconstructed variables. It stabilizes the reconstructed variables to recode the values when extracting patient-related time series features, making it possible to avoid the posterior collapse phenomenon in complex VAE design when extracting patient global fusion features.

Meanwhile, the advantage of generating different implied features for the same sample each time by sampling without replacement and using VAE alleviates the common data imbalance problem in medical electronic records analysis.

The paper demonstrates the feature extraction effect using feature visualization and other means.

## VI. Conclusion

This paper describes the method of establishing Dual VAE, which can extract global fusion features from multi-dimensional sparse and time-related large wide tables. This method can avoid a posteriori collapse which is difficult to avoid in traditional VAE methods for complex time series data, and alleviate the problem of data imbalance.

Experiments show that this method can achieve better results than the benchmark methods in binary and multi-label classification tasks. Therefore, it can be used as a reference for the analysis of EMR, which has a reasonable prospect today.

## References


[1] B. K. Beaulieu-Jones, P. Orzechowski, and J. H. Moore, "Mapping patient trajectories using longitudinal extraction and deep learning in the MIMIC-III critical care database," in *Proceedings of the Pacific Symposium*, Kohala Coast, HI, USA, pp. 123–132, 2018.

[2] A. Budrionis, M. Miara, P. Miara, *et al.*, "Benchmarking PySyft federated learning framework on MIMIC-III dataset," *IEEE Access*, vol. 9, pp. 116869–116878, 2021.

[3] J. F. Chen, L. L. Sun, C. H. Guo, *et al.*, "A fusion framework to extract typical treatment patterns from electronic medical records," *Artificial Intelligence in Medicine*, vol. 103, article no. 101782, 2020.

[4] Y. Cheng, F. Wang, P. Zhang, *et al.*, "Risk prediction with electronic health records: A deep learning approach," in *Proceedings of the 2016 SIAM International Conference on Data Mining*, Miami, FL, USA, pp. 432–440, 2016.

[5] A. Razavi, A. van den Oord, B. Poole, *et al.*, "Preventing posterior collapse with delta-VAEs," in *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, LA, USA, 2019.

[6] A. E. W. Johnson, T. J. Pollard, L. Shen, *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, article no. 160035, 2016.

[7] S. K. Bashar, M. B. Hossain, E. Ding, *et al.*, "Atrial fibrillation detection during sepsis: Study on MIMIC III ICU data," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 11, pp. 3124–3135, 2020.

[8] H. Chen, Z. Zhu, C. Y. Zhao, *et al.*, "Central venous pressure measurement is associated with improved outcomes in septic patients: An analysis of the MIMIC-III database," *Critical Care*, vol. 24, no. 1, article no. 433, 2020.

[9] B. H. Cheng, D. W. Li, Y. Q. Gong, *et al.*, "Serum anion gap predicts all-cause mortality in critically ill patients with acute kidney injury: Analysis of the MIMIC-III database," *Disease Markers*, vol. 2020, article no. 6501272, 2020.

[10] Z. Dai, S. R. Liu, J. F. Wu, *et al.*, "Analysis of adult disease characteristics and mortality on MIMIC-III," *PLoS One*, vol. 15, no. 4, article no. e0232176, 2020.

[11] M. L. Feng, J. I. McSparron, D. T. Kien, *et al.*, "Transthoracic echocardiography and mortality in sepsis: Analysis of the MIMIC-III database," *Intensive Care Medicine*, vol. 44, no. 6, pp. 884–892, 2018.

[12] E. Y. Ding, D. Albuquerque, M. Winter, *et al.*, "Novel method of atrial fibrillation case identification and burden estimation using the MIMIC-III electronic health data set," *Journal of Intensive Care Medicine*, vol. 34, no. 10, pp. 851–857, 2019.

[13] M. Böck, J. Malle, D. Pasterk, *et al.*, "Superhuman performance on sepsis MIMIC-III data by distributional reinforcement learning," *PLoS One*, vol. 17, no. 11, article no. e0275358, 2022.

[14] N. Ding, C. R. Guo, C. L. Li, *et al.*, "An artificial neural networks model for early predicting in-hospital mortality in acute pancreatitis in MIMIC-III," *BioMed Research International*, vol. 2021, article no. 6638919, 2021.

[15] T. Gentimis, A. J. Alnaser, A. Durante, *et al.*, "Predicting hospital length of stay using neural networks on MIMIC III data," in *Proceedings of the 15th International Conference on Dependable, Autonomic and Secure Computing, 15th International Conference on Pervasive Intelligence and Computing, 3rd International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress*, Orlando, FL, USA, pp. 1194–1201, 2017.

[16] N. Z. Hou, M. Z. Li, L. He, *et al.*, "Predicting 30-days mortality for MIMIC-III patients with sepsis-3: A machine learning approach using XGboost," *Journal of Translational Medicine*, vol. 18, no. 1, article no. 462, 2020.

[17] C. J. McWilliams, D. J. Lawson, R. Santos-Rodriguez, *et al.*, "Towards a decision support tool for intensive care discharge: Machine learning algorithm development using electronic healthcare data from MIMIC-III and Bristol, UK," *BMJ Open*, vol. 9, no. 3, article no. e025925, 2019.

[18] S. Nuthakki, S. Neela, J. W. Gichoya, *et al.*, "Natural language processing of MIMIC-III clinical notes for identifying diagnosis and procedures with neural networks," *arXiv preprint*, arXiv: 1912.12397, 2019.

[19] H. Harutyunyan, H. Khachatrian, D. C. Kale, *et al.*, "Multitask learning and benchmarking with clinical time series data," *Scientific Data*, vol. 6, no. 1, article no. 96, 2019.

[20] Y. K. Li, S. Rao, J. R. A. Solares, *et al.*, "BEHRT: Transformer for electronic health records," *Scientific Reports*, vol. 10, no. 1, article no. 7155, 2020.

[21] Y. K. Li, M. Mamouei, G. Salimi-Khorshidi, *et al.*, "Hi-BEHRT: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 2, pp. 1106–1117, 2023.

[22] Z. Xu, D. R. So, and A. M. Dai, "MUFASA: Multimodal fusion architecture search for electronic health records," in *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, Online, pp. 10532–10540, 2021.

[23] Y. W. Meng, W. Speier, M. K. Ong, *et al.*, "Bidirectional representation learning from transformers using multimodal electronic health record data to predict Depression," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 8, pp. 3121–3129, 2021.

[24] P. Chatha, Y. X. Wang, Z. K. Wu, *et al.*, "Dynamic survival transformers for causal inference with electronic health records," *arXiv preprint*, arXiv: 2210.15417, 2022.

[25] S. Rao, M. Mamouei, G. Salimi-Khorshidi, *et al.*, "Targeted-BEHRT: Deep learning for observational causal inference on longitudinal electronic health records," *IEEE Transactions on Neural Networks and Learning Systems*, In Press, 2022.

[26] X. P. Peng, G. D. Long, T. Shen, *et al.*, "Sequential diagnosis prediction with transformer and ontological representation," in *Proceedings of the 2021 IEEE International Conference on Data Mining (ICDM)*, Auckland, New Zealand, pp. 489–498, 2021.

[27] J. L. Wu, J. Roy, and W. F. Stewart, "Prediction modeling using EHR data: Challenges, strategies, and a comparison of machine learning approaches," *Medical Care*, vol. 48, no. 6, pp. S106–S113, 2010.

[28] H. B. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[29] H. X. Guo, Y. J. Li, J. Shang, *et al.*, "Learning from class-imbalanced data: Review of methods and applications," *Expert*

*Systems with Applications*, vol. 73, pp. 220–239, 2017.
[30] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.
[31] Y. M. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 4, pp. 687–719, 2009.
[32] T. C. Zhang, J. L. Chen, F. D. Li, *et al.*, "Intelligent fault diagnosis of machines with small & imbalanced data: A state-of-the-art review and possible extensions," *ISA Transactions*, vol. 119, pp. 152–171, 2022.
[33] H. D. Ma, Z. C. Dong, M. C. Chen, *et al.*, "A gradient boosting tree model for multi-department venous thromboembolism risk assessment with imbalanced data," *Journal of Biomedical Informatics*, vol. 134, article no. 104210, 2022.
[34] Y. Wang, Y. K. Wei, H. Yang, *et al.*, "Utilizing imbalanced electronic health records to predict acute kidney injury by ensemble learning and time series model," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, article no. 238, 2020.
[35] H. K. Chang, C. T. Wu, J. H. Liu, *et al.*, "Early detecting in-hospital cardiac arrest based on machine learning on imbalanced data," in *Proceedings of the 2019 IEEE International Conference on Healthcare Informatics (ICHI)*, Xi'an, China, pp. 1–10, 2019.
[36] K. Fujiwara, Y. K. Huang, K. Hori, *et al.*, "Over- and under-sampling approach for extremely imbalanced and small minority data problem in health record analysis," *Frontiers in Public Health*, vol. 8, article no. 178, 2020.
[37] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.
[38] S. R. Bowman, L. Vilnis, O. Vinyals, *et al.*, "Generating sentences from a continuous space," in *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, Berlin, Germany, pp. 10–21, 2016.
[39] H. Fu, C. Y. Li, X. D. Liu, *et al.*, "Cyclical annealing schedule: A simple approach to mitigating KL vanishing," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, MN, USA, pp. 240–250, 2019.
[40] H. J. Shao, S. C. Yao, D. C. Sun, *et al.*, "ControlVAE: Controllable variational autoencoder," in *Proceedings of the 37th International Conference on Machine Learning*, Online, pp. 8655–8664, 2020.
[41] D. P. Kingma, T. Salimans, R. Jozefowicz, *et al.*, "Improved variational inference with inverse autoregressive flow," in *Proceedings of the 30th Conference on Neural Information Processing Systems*, Barcelona, Spain, pp. 4743-4751, 2016.
[42] X. Chen, D. P. Kingma, T. Salimans, *et al.*, "Variational lossy autoencoder," in *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France, pp. 1-17, 2017.
[43] Q. L. Zhu, W. Bi, X. J. Liu, *et al.*, "A batch normalized inference network keeps the KL vanishing away," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 2636–2649, 2020.
[44] A. D. McCarthy, X. Li, J. T. Gu, *et al.*, "Addressing posterior collapse with mutual information for improved variational neural machine translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 8512–8525, 2020.
[45] H. Y. Wu and M. Tavakol, "MuseBar: Alleviating posterior collapse in recurrent VAEs toward music generation," in *Proceedings of the 20th International Symposium on Intelligent Data Analysis*, Rennes, France, pp. 365–377, 2022.
[46] S. L. Wu and Y. H. Yang, "MuseMorphose: Full-song and fine-grained piano music style transfer with one transformer VAE," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1953–1967, 2023.
[47] Y. Z. Zhou, C. Luo, X. Y. Sun, *et al.*, "VAE^2: Preventing posterior collapse of variational video predictions in the wild," *arXiv preprint*, arXiv: 2101.12050, 2021.
[48] O. Dollar, N. Joshi, D. A. C. Beck, *et al.*, "Attention-based generative models for *de novo* molecular design," *Chemical Science*, vol. 12, no. 24, pp. 8362–8372, 2021.
[49] M. Ding, Z. Y. Yang, W. Y. Hong, *et al.*, "CogView: Mastering text-to-image generation via transformers," in *Proceedings of the 35th Conference on Neural Information Processing Systems*, Online, pp. 19822–19835, 2021.
[50] J. T. Chien, "Deep Bayesian multimedia learning," in *Proceedings of the 28th ACM International Conference on Multimedia*, Seattle, WA, USA, pp. 4791–4793, 2020.
[51] J. Y. Hu, X. Y. Yi, W. H. Li, *et al.*, "Fuse it more deeply! A variational transformer with layer-wise latent variable inference for text generation," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, WA, USA, pp. 697–716, 2022.
[52] M. Kondo, "Developing a generative model utilizing self-attention networks: Application to materials/drug discovery," *Molecular Informatics*, vol. 40, no. 10, article no. 2100102, 2021.
[53] C. Tang, W. Zhan, and M. Tomizuka, "Exploring social posterior collapse in variational autoencoder for interaction modeling," in *Proceedings of the 35th Conference on Neural Information Processing Systems*, Virtual Event, pp. 8481–8494, 2021.
[54] J. Zhang, J. W. Xie, N. Barnes, *et al.*, "Learning generative vision transformer with energy-based latent space for saliency prediction," in *Proceedings of the 35th Conference on Neural Information Processing Systems*, Virtual Event, pp. 15448–15463, 2021.
[55] M. Ş. Bilici and M. F. Amasyali, "Transformers as neural augmentors: Class conditional sentence generation via variational Bayes," *arXiv preprint*, arXiv: 2205.09391, 2022.
[56] D. M. Arroyo, J. Postels, and F. Tombari, "Variational transformer networks for layout generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 13637–13647, 2021.
[57] J. Y. Luo, M. C. Ye, C. Xiao, *et al.*, "HiTANet: Hierarchical time-aware attention networks for risk prediction on electronic health records," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Virtual Event, pp. 647–656, 2020.

**Fei LI** received the M.S. degree from the Lingnan College of Sun Yat-sen University, Guangzhou, China, in 2009. In 2013, she was a Visiting Scholar with the Department of Electrical Information, MIIT, Beijing, China, for 19 months. After Jan. 2018, she worked as a senior engineer in the field of electronic information technology and also as a Ph.D. candidate in the AI and database field. Now she is a doctoral student jointly trained by the Institute of Computing Technology, University of Chinese Academy of Sciences and Pengcheng Laboratory, Shenzhen, China. Her research interests include machine learning, big data and database.
(Email: lifei21@mails.ucas.ac.cn)

**Yiqiang CHEN** is currently a Professor and the Director of the Pervasive Computing Research Center, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. In 2004, he was a Visiting Scholar researcher with the Department of Computer Science, Hong Kong University of Science and Technology, Hong Kong, China. His research interests include artificial intelligence, pervasive computing, and human–computer interaction. He received the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, in 2003. He serves as the AE of the *IEEE Transactions on Emerging Topics in Computational Intelligence* and *IEEE Access*. He is a Senior Member of the IEEE.
(Email: yqchen@ict.ac.cn)

**Yang GU** was born in 1988. She is Ph.D., Research Associate of the Institute of Computing Technology, Chinese Academy of Sciences. Her main research interests include generative adversarial networks and machine learning.
(Email: guyang@ict.ac.cn)

**Yaowei WANG** received the Ph.D. degree in computer science from the University of Chinese Academy of Sciences in 2005. He worked with the Department of Electronics Engineering, Beijing Institute of Technology, from 2005 to 2019. Currently, he is a Professor at the Peng Cheng Laboratory in Shenzhen, China. He servers the Chair of the IEEE Digital Retina Systems Working Group and a Member of IEEE, CIE, CCF, CSIG. He was the recipient of the second prize of the National Technology Invention in 2017 and the first prize of the CIE Technology Invention in 2015. He has co-authored more than 120 technical articles in international journals and conferences, including IEEE TIP, CVPR, ICCV, etc. His research interests include machine learning, multimedia content analysis and understanding. He promoted digital retina technology, and made efforts to establish system standards for digital retina. He trained a vision model named "Pengcheng · Dasheng" with 1 billion parameters, achieving an over 10% performance gain in the detection and recognition task in more than 20 application scenarios. He led the development of the first digital retina verification system, which has been applied to the urban traffic management field of over 30 large and medium-sized cities in China.
(Email: wangyw@pcl.ac.cn)