Medical & Biological Engineering & Computing (2023) 61:1857–1873
https://doi.org/10.1007/s11517-023-02816-z

ORIGINAL ARTICLE

Predicting heart failure in-hospital mortality by integrating longitudinal and category data in electronic health records

Meikun Ma[1,2,3] · Xiaoyan Hao[1] · Jumin Zhao[1,2,4] · Shijie Luo[1] · Yi Liu[2,3,5] · Dengao Li[2,3,5]

Received: 22 July 2022 / Accepted: 2 March 2023 / Published online: 24 March 2023

Abstract

Heart failure is a life-threatening syndrome that is diagnosed in 3.6 million people worldwide each year. We propose a deep fusion learning model (DFL-IMP) that uses time series and category data from electronic health records to predict in-hospital mortality in patients with heart failure. We considered 41 time series features (platelets, white blood cells, urea nitrogen, etc.) and 17 category features (gender, insurance, marital status, etc.) as predictors, all of which were available within the time of the patient's last hospitalization, and a total of 7696 patients participated in the observational study. Our model was evaluated against different time windows. The best performance was achieved with an AUC of 0.914 when the observation window was 5 days and the prediction window was 30 days. Outperformed other baseline models including LR (0.708), RF (0.717), SVM (0.675), LSTM (0.757), GRU (0.759), GRU-U (0.766) and MTSSP (0.770). This tool allows us to predict the expected pathway of heart failure patients and intervene early in the treatment process, which has significant implications for improving the life expectancy of heart failure patients.

Keywords Deep learning · Heart failure · Fatal outcome · Electronic health records · Feature fusion

Abbreviations

EHR	Electronic health record
HF	Heart failure
LR	Logistic regression
RF	Random forest
SVM	Support vector machine
LSTM	Long short-term memory
GRU	Gate recurrent unit

1 Introduction

Heart failure (HF) is a life-threatening syndrome diagnosed in 3.6 million people worldwide each year with 35% of patients dying within the first year and the rest within 5 years [1]. HF is increasingly common in China, with a standardized prevalence of HF of 1.38% in patients older than 35 years of age, about 50% than in the survey dated in 2000 [2]. A favorable prognosis of HF using deep learning techniques can delay disease progression and improve patients' quality of life and life expectancy. Accurately predicting in-hospital mortality in HF can assist physicians in diagnosing, which plays an essential role in clinical decision-making [3].

Electronic health record (EHR) contains healthcare information about patients, including diagnoses, procedures, medications, laboratory measurements, and imaging data [4], typically used to develop clinical decision support systems [5]. For example, several studies have constructed EHR risk models for adverse event prediction in HF [6–8]. In these analyses, traditional machine learning algorithms such as logistic regression (LR), random forest (RF), support vector machine (SVM), decision tree (DT), etc. are applied to EHR data to identify early HF or predict patient outcomes. However, these models use only limited

✉ Dengao Li
lidengao@tyut.edu.cn

[1] College of Information and Computer, Taiyuan University of Technology, Taiyuan 030024, China

[2] Key Laboratory of Big Data Fusion Analysis and Application of Shanxi Province, Taiyuan 030024, China

[3] Technology Research Center of Spatial Information Network Engineering of Shanxi, Taiyuan 030024, China

[4] Intelligent Perception Engineering Technology Center of Shanxi, Taiyuan 030024, China

[5] College of Data Science, Taiyuan University of Technology, Taiyuan 030024, China

information based on traditional statistical methods, which has the potential limitation of information loss.

In recent years, deep learning has been used in bioinformatics and healthcare with great success [13, 14] and is used for risk prediction in different clinical situations [15–17]. However, most studies rarely consider longitudinal time series information regarding inpatient treatment trajectories. Recurrent neural network (RNN) has achieved better performance by capturing temporal patterns present in EHR longitudinal series data [18]. However, using only the time series features of the EHR as a single input model, without considering categorical features such as demographics or without additional processing of categorical features, results in insufficient features and thus biases the decision direction of the final model, limiting the accuracy of the model. Feature fusion techniques extract complementary and more complete information by fusing data from multiple modalities. The completeness of the data allows for better execution of machine learning models, thus improving the accuracy of decisions [19].

In this study, we propose a deep fusion learning model, DFL-IMP, for HF mortality prediction. The model refines time series features and category features in structured EHR data and fuses the refined features for analysis, fully exploiting different forms of data information under a single model. Specifically, in our DFL-IMP, we propose a novel GRU cell, stochastic-decay gate recurrent unit, called GRU-S, which introduced a stochastic decay factor to capture the information on time series features in EHR. The category features are input to the fully connected layer for feature dimensionality reduction. The reduced-dimension features are fused with the output of the last hidden layer of GRU-S. The overall features are then fed to the fully connected layer and further fed to the classifier to predict the in-hospital mortality of HF. In addition, to address the problem of many missing values commonly found in EHR serial data, in our model, we invoked a variational recurrent neural network (VRNN) to estimate the missing values of the serial data.

The remainder of this paper is organized as follows. We review the related work in Section 2. Section 3 introduces the data, formulates the problem, and presents our proposed approach in detail. The experimental setup and results using a real clinical dataset are presented in Section 4. We discuss in Section 5 and conclude our work and highlight future research directions in Section 6.

2 Related work

2.1 Machine learning in EHR applications

Early prognostic work of HF mainly relies on traditional machine learning modeling techniques, such as LR, RF, SVM, DT, etc. A large number of studies have shown that machine learning plays a crucial role in the prognostic study of HF based on EHR data. Konig et al. [9] used four machine learning algorithms (i.e., LR, RF, XGBoost, NNet) to predict in-patient mortality to explore whether managing routine data alone could improve future care for HF. The problem is that it ignores information about patients' heart imaging, laboratory results, medication, and treatment-related data, making it impossible to add more objective measures of disease severity. Adler et al. [10] trained an enhanced DT algorithm to capture the correlation between patient characteristics and mortality. To capture the higher-order correlation between input variables, it excluded missing data, which may introduce additional bias. Angraal et al. [11] used five machine learning methods to predict the risk of death and hospitalization in HF with ejection fraction retention, but it only used the baseline characteristics of patients without follow-up data, and the model explored was not enough to be applied to a wider population. Davide et al. [12] applied 10 machine learning classifiers to predict the survival of patients, and the results showed that serum creatinine and ejection fraction were sufficient to predict the survival of patients, but the dataset used was only 299 pieces, and the data scale was small. In general, the above studies only use traditional statistical methods, such as LR model multivariate analysis, which may have the limitation of loss of information.

2.2 Deep learning in EHR applications

In recent years, an increasing number of scholars have applied deep learning techniques to medical research. One of the advantages of deep learning is that features and relationships can be learned automatically from given data without the need for feature engineering. Wang et al. [20] proposed a deep learning system based on feature rearrangement to predict heart failure mortality. This framework dealt with the problem of unbalanced datasets and achieved better feature representation. However, a limitation of their study is that only the aggregated features of events within a single observation window are used, ignoring the temporal relationship between events within the observation window.

Recently, RNN models have been used effectively in many complex machine learning tasks, such as image captioning and language translation. The ability of RNNs to model high-dimensional non-linear long-term dependencies between elements has attracted the attention of researchers in healthcare, with a series of studies using RNN models to capture the temporal patterns present in EHR longitudinal series data for disease progression and risk prediction. Jun et al. [21] introduced a stochastic gradient variational Bayesian (SGVB) approach to RNN sequence models to capture the underlying sequence structure and generate missing values for multivariate time series data inference. Missing values expressed as variance were used as fidelity using uncertainty and a new uncertainty-gated stochastic sequence model was proposed for clinical time series prediction. Men et al. [22] used a long and short-term memory network with a time-aware and attention-based mechanism to classify multi-labeled diseases based on patients' clinical attendance records, but with a single dataset and without using demographic data. Priyanga et al. [23] used a multilayer bidirectional LSTM algorithm for feature selection, the LCBWO algorithm for structural improvement and fast convergence, and the LSTM algorithm for predicting heart disease within 5 years, achieving high accuracy. Yoon et al. et al. [24] proposed the M-RNN, which uses a bidirectional RNN to estimate missing values. The M-RNN in which the estimated values are considered constant and cannot be updated sufficiently. It uses a bidirectional RNN-based consistency loss to prevent the propagation of errors. McGilvray et al. [25] integrated a deep learning model using time series and densely connected networks developed based on standard EHR data to assist clinicians to identify HF drug treatment non-response and predict death in a timely and accurate manner. Li et al. [38] proposed an MTSSP model for predicting survival, which interpolates missing values by combining mask and time interval information to obtain a global view using a bidirectional RNN architecture and a local view using a one-dimensional dilated convolution, combining a missing value complementation approach with a time series classification prediction approach. Shickel et al. [39] proposed a dynamic approach to label discrete and continuous patient data and proposed a transformer classifier that uses a joint embedding space to integrate different temporal patient measures. Six mortality and readmission outcomes were also predicted simultaneously. However, these studies focused on other clinical conditions and did not focus on the area of heart failure.

We further explored studies in the HF domain, where Chu et al. [26] used an adversarial learning scheme to distinguish generated feature vectors from true feature vectors, using the prediction of HF feature vectors as an adjunct to endpoint prediction, but its data used only part of the treatment trajectory, missing the value-rich outpatient information with follow-up data. Radhachandran et al. [27] developed a gradient-enhanced decision tree to predict 7-day mortality in AHF patients by applying continuous information from the first 8 h of the patient's hospitalization.

It can be concluded from the above studies that previous studies in the field of heart failure are often limited by the dataset, where a single dataset or a small or unbalanced dataset may lead to a decrease in model accuracy. Most of the models using EHR time series have a single input, ignoring categorical features such as demographics and no additional features have been applied to them. Our research focuses on exploring the integration of time series data with categorical data for heart failure patients to predict their in-hospital mortality.

2.3 Feature fusion analysis of EHR

Feature fusion techniques improve decision accuracy by fusing data from multiple modalities to extract complementary and more comprehensive information for better execution of machine learning models [28]. Most studies [31–33] have shown that models using fusion perform better compared to the performance of a single model. Feature fusion plays a crucial role in medical decision making. Zhi et al. [29] proposed a multimodal fusion model based on a multilayer perceptron and a two-dimensional CNN to ingest EHR data and CT images for pulmonary embolism diagnosis. In addition, multidimensional scaling (MDS) algorithms were used to reduce the dimensionality of EHR data. Zheng et al. [30] used a novel longitudinal data fusion approach to model disease progression for chronic disease care. A temporal regularization term was designed to maintain the temporal inheritance of data at different time points, and data from both the source level and feature level were analyzed based on a sparse regularized regression approach. Li et al. [40] proposed an enhanced BEHRT model, Hi-BEHRT, for risk prediction. It allows integration of long EHR series from different modalities, addresses the shortcomings of common transformers in handling long series data and avoids the loss of important historical information in risk prediction, and has achieved superior results in four investigated (heart failure, diabetes, CKD and stroke) risk prediction tasks. Liu et al. [41] proposed a new multimodal PLM for jointly modeling unstructured and structured data in electronic medical records that learns cross-modal interactions while maintaining unimodal representation capabilities.

In summary, the above studies focus on data fusion in multiple modalities, ignoring the different forms of data types in a single modality. Instead, we argue that a granular analysis of different types of data in a single modality can help to mine more adequate information with fewer data. In addition, data from other modalities of medical clinics are not readily available.

3 Methods

3.1 Data preparation

We used the Medical Information Mart for Intensive Care (MIMIC-III[1]) dataset, an extensive medical center database

[1] Available at https://mimic.physionet.org/

with EHR data related to 53,423 adult patient admissions to intensive care units at the Beth Israel Deaconess Medical Center between 2001 and 2012. It includes vital signs, medications, laboratory measurements, observations, and notes recorded by nursing staff, fluid balance, procedure codes, diagnosis codes, imaging reports, length of stay, survival data, and more [36]. Criteria for the incident onset of HF were adopted from Gurwitz et al. [37], which relied on qualifying International Classification of Diseases, Ninth Revision (ICD-9) codes. Our target sample was HF patients with ICD-9 code 4280.

We selected 41 laboratory longitudinal measured time series features and 17 category features. Time series features are mainly laboratory measurements. These data share the same property that they change over time, and the distribution is nonuniform in time. For instance, a blood test is a discrete event that happens sometime during the admission. The category features are the patient's gender, insurance, etc. Usually, it will not change during hospital admission. In all our observation windows, 41 time series features and 17 category features are included, and the number of features did not decrease due to the decrease in observation windows. Table 1 shows the information on several of the critical patient features we selected. Differences between the positive and negative classes were assessed for significance using a two-proportion *t*-test.

For the time series features in our data, we consider different observation and prediction windows to evaluate the performance of our model, each patient was considered for multiple observations, the observation window length lasted for days, choose one observation per day, corresponding to a time window. If the patient had multiple observations in a day, we used the average value to represent the observations for that day, and the unobserved values we set to Null. As shown in Fig. 1, the observation windows were selected as 5, 7, and 10 days, and the prediction windows were selected as 10, 20, and 30 days. The reason for choosing the window this way is that the data drives us to do so; an observation window that is too short would not take full advantage of the information in the time series data, and too long an observation window would exclude too many positive samples resulting in unbalanced data. As our goal was to predict patient death, for patients with multiple hospitalizations, we selected data from the patient's last hospitalization. If the patient's last hospitalization was less than the length of our observation window (5 days/7 days/10 days), this hospitalization data was excluded and the patient's most recent hospitalization data was then selected. If all of the patient's hospital admissions are less than our observation window, the patient is excluded. For category features, such as gender, and co-morbidities, that did not change during the hospital admission.

Overall, we extracted a total of 13,112 EHR data from the MIMIC-III database for patients with HF. After excluding non-compliant patients and after data pre-processing, 7696/6296/4472 samples remained, with sample sizes varying by observation window, as shown in Table 2, with sample sizes varying by observation window. Patients were included in the positive class only if the death date occurred within the selected prediction windows for mortality prediction. Patients with mortality dates later than the observation window were included in the negative class, as they were still alive within the prediction window. Surviving patients with no documented mortality were also included in the negative class.

Since the data values of the time series are continuous, to prevent the influence of outliers on the study results, we remove the outliers with a Winsorize process. As each variable has other evaluation indicators, there are different units of magnitude, and they are in different magnitudes. *Z*-score normalization is performed for all datasets so that each feature is in the same order of magnitude for a comprehensive comparison. For the category features, we use one-hot coding to transform the category variables.

3.2 Problem formulation

We define a patient sample with D time series features and Q category features (time-invariant features). Given a time series feature X, which is observed at T time points, we denote it as $X = \left[x_1, \cdots, x_d, \cdots, x_D\right] = \begin{bmatrix} x_1^1 & \cdots & x_D^1 \\ \vdots & & \vdots \\ x_1^T & \cdots & x_D^T \end{bmatrix}$, where x_d^t is the observed value of the feature d at the moment t. Given a category feature C, which is denoted as $C = \left[c_1, \cdots, c_q, \cdots, c_Q\right]$, c_q denotes the observed value of the feature q.

We introduce the masking vector M that marks the observed and missing values in the time series, denoted as $M = \left(m_1, \ldots, m_d, \ldots, m_D\right)$, if x_d^t is observed, then $m_d^t = 1$; otherwise, $m_d^t = 0$. Based on the mask vector, we define a new time series containing missing values $\widetilde{X} = \left(\widetilde{x}_1, \ldots, \widetilde{x}_d, \ldots, \widetilde{x}_D\right)$, as follows:

$$\widetilde{x}_d^t = \begin{cases} x_d^t, m_d^t = 1 \\ *, m_d^t = 0 \end{cases} \tag{1}$$

where $*$ indicates the missing values, and the initial $*$ is set to 0. In addition, we define the time interval $\Delta = \left(\delta_1, \ldots, \delta_d, \ldots, \delta_D\right)$ as the difference between the time stamp of the last observed value and the current timestamp with the following equation:

$$\delta_d^t = \begin{cases} s_d^t - s_d^{t-1} + \delta_d^{t-1}, \ if \ t > 1, m_d^{t-1} = 0 \\ s_d^t - s_d^{t-1}, \ if \ t > 1, m_d^{t-1} = 1 \\ 1, \ if \ t = 1 \end{cases} \tag{2}$$

Table 1 Baseline features of experimental HF dataset

Feature type	Feature	No. of participants ($n=6296$)	Death ($n=941$)	Survival ($n=5355$)	*P*-value
Time series features	Platelet, mean (SD)	236.12 (129.57)	232.45 (140.23)	236.76 (127.61)	0.346
	Creatinine, mean (SD)	1.74 (1.68)	1.98 (1.75)	1.70 (1.66)	<0.001
	Urea nitrogen, mean (SD)	34.02 (25.15)	41.39 (29.12)	32.73 (24.16)	<0.001
	White blood cell, mean (SD)	11.31 (12.82)	13.27 (25.84)	10.97 (8.67)	<0.001
	Hematocrit, mean (SD)	32.96 (7.79)	32.68 (6.65)	33.01 (7.97)	0.23
	Lymphocyte, mean (SD)	8.88 (10.76)	9.05 (10.97)	8.85 (10.72)	0.615
	Neutrophil, mean (SD)	51.87 (38.52)	59.58 (36.40)	50.52 (38.72)	<0.001
	Calcium, mean (SD)	5.58 (4.18)	6.21 (3.86)	5.47 (4.22)	<0.001
	Potassium, mean (SD)	4.13 (1.21)	4.32 (0.95)	4.10 (1.25)	<0.001
	Chlorine, mean (SD)	97.45 (22.11)	99.56 (16.10)	97.08 (22.98)	0.002
	Sodium, mean (SD)	129.77 (33.11)	134.60 (21.57)	128.92 (34.68)	<0.001
	Magnesium, mean (SD)	1.39 (0.99)	1.51 (0.95)	1.37 (1.00)	<0.001
	Albumin, mean (SD)	1.27 (1.67)	1.31 (1.56)	1.26 (1.69)	0.387
Category features	Insurance (%)				<0.001
	Missing	22 (0.3)	2 (0.2)	20 (0.4)	
	Government	88 (1.4)	10 (1.1)	78 (1.5)	
	Medicaid	378 (6.0)	36 (3.8)	342 (6.4)	
	Medicare	4505 (71.6)	734 (78.0)	3771 (70.4)	
	Private	1278 (20.3)	155 (16.5)	1123 (21.0)	
	Self-pay	25 (0.4)	21 (0.4)	21 (0.4)	
	Gender (%)				0.503
	Missing	22 (0.3)	2 (0.2)	20 (0.4)	
	Female	2896 (46.0)	421 (44.7)	2475 (46.2)	
	Male	3378 (53.7)	518 (55.0)	2860 (53.4)	
	Medical comorbidities (%)				
	Myocardial infarction (suffered)	355 (5.6)	52 (5.5)	303 (5.7)	0.923
	Diabetes (suffered)	2353 (37.4)	330 (35.1)	2023 (37.8)	0.199
	Acute renal failure (suffered)	2439 (38.7)	529 (56.2)	1910 (35.7)	<0.001
	Respiratory support (supported)	1742 (27.7)	469 (49.8)	1273 (23.8)	<0.001

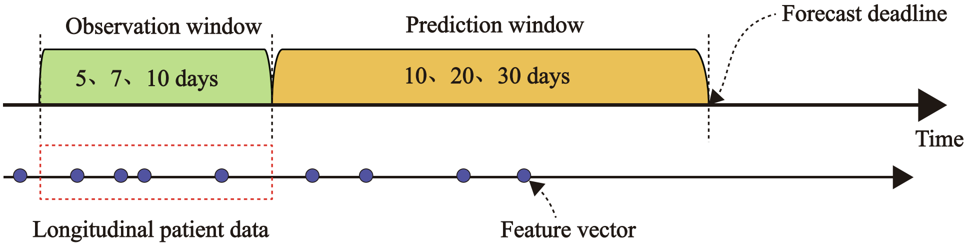


Fig. 1 Framework for data extraction for predictive modeling tasks. Relation of prediction window, observation window

where s_d^t represents the time stamp t observed by the feature d, assuming that the timestamp of the first observation is 0 (i.e.$s_1 = 0$).

For N samples, given the dataset $V = \left\{ \left(\widetilde{X}^{(n)}, M^{(n)}, \Delta^{(n)} \right) \circ C^{(n)} \right\}_{n=1}^{N}$, $\circ$ denotes the matrix juxtaposition. Y is the predictive label for whether the patient died in the hospital (1 for death and 0 for survival), denoted as:$Y_n = \left[y_1, \cdots, y_n \right]$.

3.3 Model description

The proposed DFL-IMP consists of three main components: (1) VRNN missing value imputation, (2) GRU-S, and (3) mortality prediction of feature fusion. The idea of the bidirectional recurrent neural network is applied in our model, specifically, the inputs of DFL-IMP are both the serial forward direction $X = \left\{ x_1, x_2, \ldots, x_D \right\}$ and backward direction $X' = \left\{ x_1', x_2', \ldots, x_D' \right\}$. The final result is the average of the

Table 2 Total number of samples for different observation and prediction windows

Observation window (days)	Prediction window (days)	Missing rate (%)	Total	Positive	Negative
5	10	57.93	7696	772	6924
	20			1045	6651
	30			1190	6505
7	10	58.82	6296	589	5707
	20			821	5475
	30			941	5355
10	10	59.11	4472	420	4052
	20			621	3851
	30			701	3771

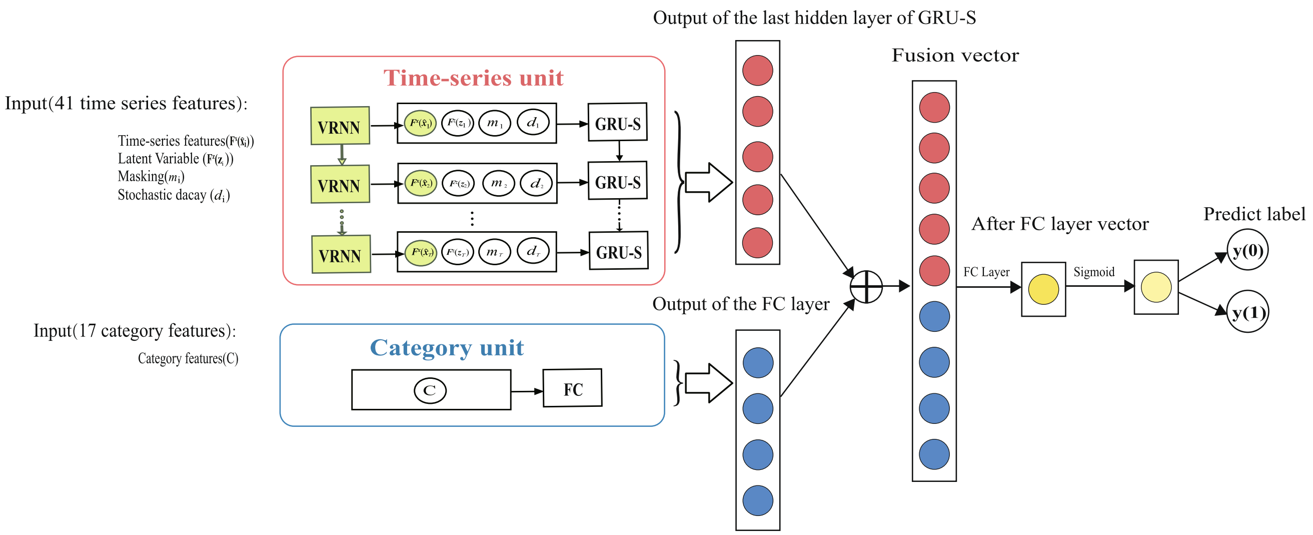


Fig. 2 Whole architecture of the DFL-IMP

forward DFL-IMP and backward DFL-IMP calculations. The overall architecture of the DFL-IMP is shown in Fig. 2.

3.3.1 VRNN missing value imputation

We adopt VRNN [34] to fill in the missing values in the time series. Figure 3 shows the flowchart of the model, which consists of three main steps: (1) prior, (2) inference, and (3) generation. VRNN cyclically updates the hidden layer state as follows:

$$h_t = f_\theta\left(\mathcal{H}^x\left(\hat{x}_t\right), \mathcal{H}^z\left(z_t\right), h_{t-1}\right) \tag{3}$$

The derivation of the formula is detailed in Appendix A.

3.3.2 GRU-S

Our proposed model GRU-S is based on the GRU-U [21] made improvement; the model architecture is shown in Fig. 4. We propose two different decay factors, the stochastic decay factor and the time decay factor, which mainly address the problem of input variables disappearing over time due to the long-term absence of EHR time series data.

Firstly, we investigated the effectiveness of the stochastic decay factor propagation in the GRU hidden state h_t. Since the imputation value are stochastic, we capture stochastically in VRNN with a stochastic estimate of $\sigma_{x,t}$, as:

$$s_t = \left(1 - m_t\right) \odot \sigma_{x,t} \tag{4}$$

We introduced a training decay rate in the model using a negative exponential rectifier to make the stochastic decay factor $d_t \in (0, 1]$ monotonically decreasing, as follows:

$$d_t = \exp\left\{-\max\left(0, W_d s_t + b_d\right)\right\} \tag{5}$$

where s_t is a stochastic factor and W_d is a diagonal matrix to ensure that the decay factors of each variable are independent of each other. The stochastic decay factor d_t is fed into our network by updating the state of $\boldsymbol{h}_{t-1}$:

$$h'_{t-1} = h_{t-1} \odot d_t \tag{6}$$

Secondly, we consider a time decay factor γ_t to decay the input features. The calculation is similar to the stochastic decay factor as follows:

$$\gamma_t = \exp\left\{-\max\left(0, W_\gamma \delta_t + b_\gamma\right)\right\} \tag{7}$$

W_γ, b_γ is the trainable model parameter, and δ_t is the time interval. Based on the time decay factor γ_t, $\widetilde{x}_t$ and the estimated mean values $\mu_{x,t}$ from VRNN weighted calculation to obtain the time decay estimate c_t as follows:

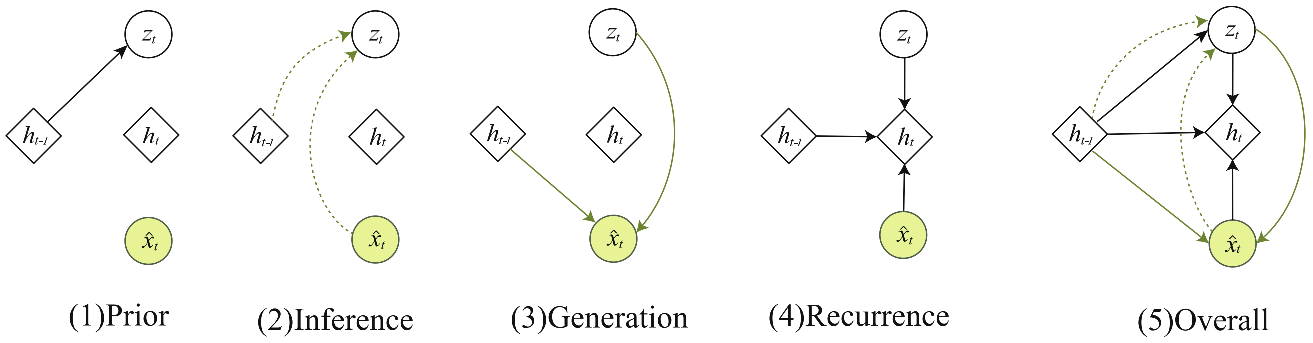
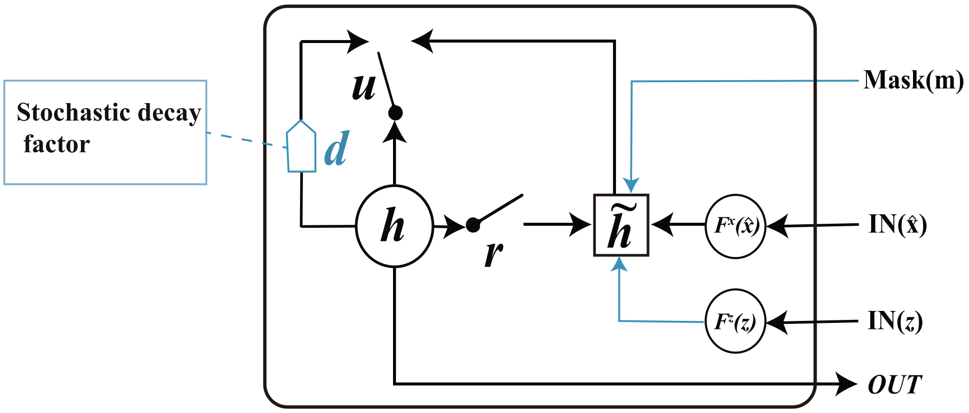


Fig. 3 Graphical illustrations of each operation of VRNN: (1) computation of conditional prior; (2) inference step; (3) generation step; (4) recurring update of hidden states of RNN; (5) overall structure of VRNN

Fig. 4 Graphical illustrations of the proposed GRU-S



$$c_t = \widetilde{x}_t \odot \mathcal{H}^{decay}(\gamma_t, m_t) + \mu_{x,t} \odot \left(1 - \mathcal{H}^{decay}(\gamma_t, m_t)\right) \tag{8}$$

where m_t is the mask vector, and $\mathcal{H}^{decay}$ is a fully connected layer. We finally express the imputation estimates in the following equation:

$$\hat{x}_t = m_t\widetilde{x}_t + \left(1 - m_t\right)\mathcal{H}^{pool}(\mathcal{H}^{cor}(\mu_{x,t})\circ c_t) \tag{9}$$

where $\mathcal{H}^{cor}$ is a feature regression layer to calculate the relationship between features, $\mathcal{H}^{pool}$ is a neural network layer. From this, we can obtain the hidden layer state h_t as follows:

$$h_t = f_\theta\left(\mathcal{H}^x\left(\hat{x}_t\right), \mathcal{H}^z\left(z_t\right), h'_{t-1}\right) \tag{10}$$

In GRU-S, the information is controlled by resetting gate r and updating gate u with the following equations:

$$r_t = \sigma\left(W_r\left[\mathcal{H}^x\left(\hat{x}_t\right)\circ\mathcal{H}^z\left(z_t\right)\right] + U_r h'_{t-1} + V_r m_t + b_r\right) \tag{11}$$

$$u_t = \sigma\left(W_u\left[\mathcal{H}^x\left(\hat{x}_t\right)\circ\mathcal{H}^z\left(z_t\right)\right] + U_u h'_{t-1} + V_u m_t + b_u\right) \tag{12}$$

$$\widetilde{h}_t = \tanh\left(W_h\left[\mathcal{H}^x\left(\hat{x}_t\right)\circ\mathcal{H}^z\left(z_t\right)\right] + U_h\left[r_t \odot h'_{t-1}\right] + V_h m_t + b_h\right) \tag{13}$$

$$h_t = \left(1 - u_t\right) \odot h'_{t-1} + u_t \odot \widetilde{h}_t \tag{14}$$

where σ is a nonlinear activation function. It is worth noting that x_t and h_{t-1} in the original GRU formula have been replaced with $\mathcal{H}^x\left(\hat{x}_t\right)\circ\mathcal{H}^z\left(z_t\right)$ and h'_{t-1}. In addition, we input the masking vector m_t additionally to the model.

3.3.3 Mortality prediction of feature fusion

Feature fusion is the process of combining data and knowledge from different sources to maximize useful information content. It improves the reliability or discriminant capability and offers the opportunity to minimize the data retained [35]. There are three main types of integration strategies, namely early fusion, joint fusion, and late fusion [19]. In this work, joint fusion was adopted. For time series features, we take the output of the features from the last hidden layer of GRU-S, and for category, we use a fully connected layer (FC) to extract features. Specifically, a new feature vector is composed which consists of time series features extracted using GRU-S, and category features extracted by FC, as shown in Fig. 2. We directly concatenate the features extracted from the two networks.

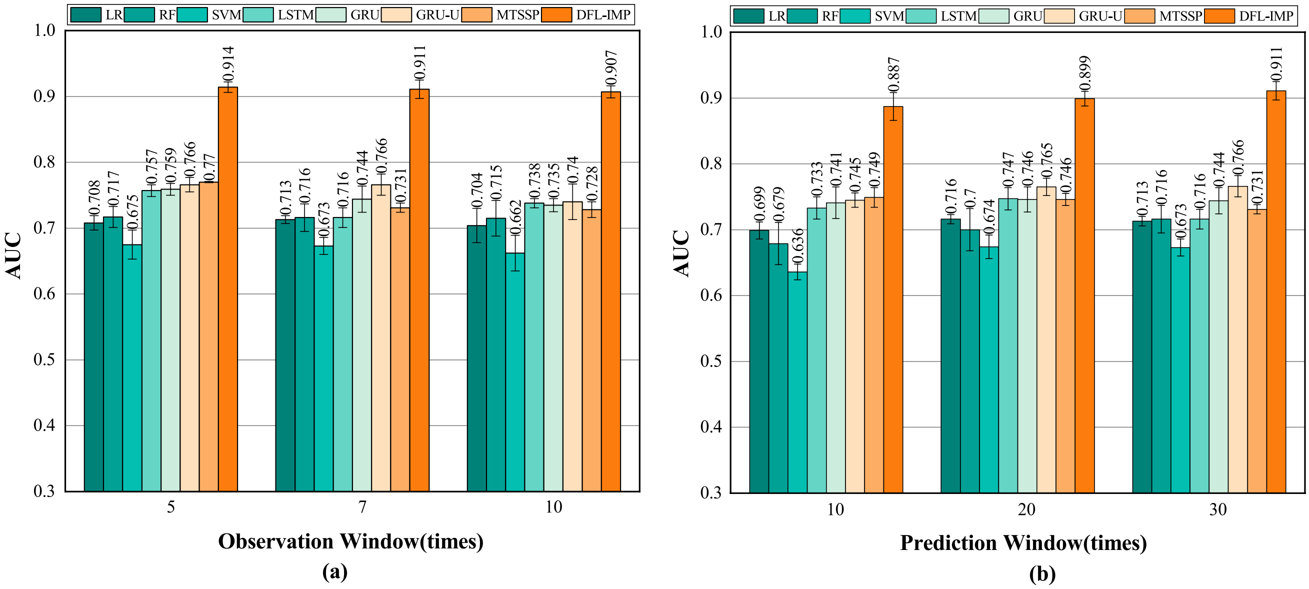


Fig. 5 (**a**) AUC values of models on mortality prediction with the changing observation window, where the prediction window is set to be 30 days. (**b**) AUC values of models on mortality prediction with the changing prediction window, where observation window is set to be 7 days

Firstly, we extract category features as follows:

$$C_{fc} = \tanh(\mathcal{H}(C_{one-hot})) \tag{15}$$

where $C_{one-hot}$ is the category features after one-hot encoding, which is because our category features are mainly composed of patient demographic features (gender, insurance, etc.) and patient co-morbidities, and these feature values are discrete and disordered, and using one-hot encoding to process the data will make the calculation of the distance between features more reasonable. $\mathcal{H}$ is a fully connected layer that extracts and reduces the dimensionality of the features and alleviates the sparsity and high dimensionality of the one-hot coding input vector. tanh is a nonlinear activation function.

Secondly, since the last GRU-S hidden state contains temporal information encoded across all time steps, we use the last GRU-S hidden state h_t to fuse with the category features as follows:

$$V_t = h_t \oplus C_{fc} \tag{16}$$

where $\oplus$ is a matrix concatenation, the fused features V_t contain richer data information, and V_t is fed into a fully connected layer, followed by a sigmoid activation function, which is used to perform our binary classification task. This is shown below:

$$p\left(y = 1|V_t\right) = \sigma\left(W_o V_t\right) \tag{17}$$

where W_o is a classifier parameter.

3.4 Training and testing

We use a joint learning strategy throughout defining the loss function of the model as a composite function with four components: (1) VRNN loss $\mathcal{L}_{VRNN}$; (2) consistency loss $\mathcal{L}_{cons}$; (3) masked imputation value loss $\mathcal{L}_{imp}$; and (4) classification loss $\mathcal{L}_{bce}$.

(1) We refer to the loss calculation function of VAE to define the VRNN loss $\mathcal{L}_{VRNN}$, the main objective is to integrate the reconstruction error and Kullback–Leibler (KL) divergence of N samples over the time series, as follows:

$$\mathcal{L}_{VRNN} = \sum_{n=1}^{N}\sum_{t=1}^{T}(-KL(q(z_t|\widetilde{x}^{(n)}_{1:t}, z_{1:t-1}) \parallel p(z_t|\widetilde{x}^{(n)}_{1:t}, z_{1:t-1})) + \log p(\widetilde{x}^{(n)}_t|_{1:t}, \widetilde{x}^{(n)}_{1:t-1})) \tag{18}$$

(2) Consistency loss is defined as the difference between the estimate of VRNN in the forward DFL-IMP $\widehat{x}_t$ and the estimate of VRNN in the backward DFL-IMP $\widehat{x}'_t$ by time variation, calculated from the mean absolute error (MAE).

$$\mathcal{L}_{cons} = \frac{1}{N}\sum_{n=1}^{N}\left|\widehat{X}'^{(n)} - \widehat{X}^{(n)}\right| \tag{19}$$

(3) Masked imputation value loss $\mathcal{L}_{imp}$, we calculate the masked MAE between the original sample X as the ground truth and the input sample $\widehat{X}$, M_{imp} is the initial masking matrix.

$$\mathcal{L}_{imp} = \frac{1}{N}\sum_{n=1}^{N}\left|X^{(n)} \odot M_{imp}^{(n)} - \widehat{X}^{(n)} \odot M_{imp}^{(n)}\right| \tag{20}$$

(4) Classification loss $\mathcal{L}_{bce}$, we calculate from the binary cross entropy loss.

$$\mathcal{L}_{bce} = \frac{1}{N}\sum_{n=1}^{N}(-[y_n \log\sigma(p_n) + (1 - y_n)\log(1 - \sigma(p_n))] \tag{21}$$

y_n is the label, p_n is the predicted probability and σ is the sigmoid activation function.

Therefore, all losses are defined by integrating the forward and reverse losses and the composite loss is defined as $\mathcal{L} = \omega_1\left(\mathcal{L}_{VRNN} + \mathcal{L}'_{VRNN}\right) + \omega_2\mathcal{L}_{cons} + \omega_3\left(\mathcal{L}_{imp} + \mathcal{L}'_{imp}\right) + \omega_4\left(\mathcal{L}_{bce} + \mathcal{L}'_{bce}\right)$,where ω_1, ω_2, ω_3, and ω_4 are the hyperparameters that control the loss ratio. We optimize all parameters of the model in an end-to-end manner through this composite loss.

3.5 Experimental setup

In our mortality prediction task, missing value imputed and outcome prediction are performed simultaneously during training. We trained our model using the RAdam optimizer with an initial learning rate of 0.001, epoch set to 80, learning rate decay set to every 10 epochs, decay 0.01 times, and batch size set to 64. The composite loss parameter ω_1, ω_2, ω_3, and ω_4 is set to $1e^{-5}$, 1, $1e^{-2}$ and 1. We set masking scenarios is 5% of the observations that were additionally masked for dataset. We selected the final optimal model based on the performance of the validation set.

4 Results

Our baseline model consists of three main components: machine learning models (LR, RF, SVM), deep learning models (LSTM, GRU), and published approaches from the literature (GRU-U [21], MTSSP [38]). LR is a linear classifier with a linear combination of features as independent variables, RF is an algorithm that integrates multiple decision trees, and SVM is a binary classification model defined on the feature space with the largest interval. The above model is difficult to process sequence data, the model is prone to underfitting problems, and the experimental results are not very ideal. The time series model LSTM learns long-term dependencies through three gate mechanisms, while the GRU greatly improves the training speed by discarding forgetting gates, but these two models are less effective when there is a large amount of missing data.

Our proposed model, DFL-IMP, captures missing time series through the attenuation mechanism of GRU-S and introduces the VRNN model to interpolate missing values, which solves the problem of missing time series values in medical data of heart failure well. In addition, we fused category features and extracted features using FC to make full use of different forms of data information in patients with HF.

For all models, we used a five-fold cross-validation strategy, with 60% of the data trained, 20% used for validation, and the performance of the trained models evaluated on the remaining 20% of the data (the final test set), which was not used during training. The experiment was repeated five times and the final performance was the mean and standard deviation of these five repetitions. AUC (area under the ROC curve), accuracy, precision, F1 score, and specificity were used as evaluation metrics.

The algorithms were implemented in Python 3.7. LSTM and GRU were trained using Pytorch 1.10.0 for model training and implemented using the Torch.nn.Module library. LR, RF, and SVM were implemented using Python Scikit-Learn 0.24.2.

4.1 Performance by DFL-IMP

We evaluated the validity of the proposed model DFL-IMP, based on the observation window shown in Fig. 1.

First, we investigated the effect on the AUC values obtained by the HF mortality prediction model by changing the observation and prediction windows. We set one of the windows by default and conducted experiments by changing the value of the other window, as shown in Fig. 5. The results show that the performance of our model decreases as the length of the observation window increases (Fig. 5(a)) and increases as the prediction window increases (Fig. 5(b)).

With an observation window of 5 days and a prediction window of 30 days, our model achieves an AUC of 0.914. Machine learning models have AUCs around 0.7, specifically: LR (0.708), RF (0.717), and SVM (0.675). Our model improved by 20.6% over LR and 19.7% over RF. The AUC values for the deep learning models were approximately 0.75 for LSTM (0.757) and GRU (0.759), GRU-U (0.766), and MTSSP (0.770). Our model improved by 15.7% over LSTM, 14.8% over GRU-U, and 14.4% over MTSSP. Overall, our model results are around 0.9 for all evaluation windows, significantly better than other benchmark models. Secondly, we investigated the model's performance in terms of accuracy, precision, F1 score, and specificity under specific prediction windows, as shown in Table 3. When the observation window is 5 days and the prediction window is 30 days, the accuracy (0.928), precision (0.867), and F1 score (0.734) of our model all rank first, although the specificity (0.982) of our model is not as good as SVM (0.999), but in general, our model metrics outperform other models.

In addition, we explored the ROC curves for all the models evaluated over a 7-day observation window and a 30-day prediction window, as shown in Fig. 6, and the results show that our models achieved excellent results.

From the results, our model achieves such excellent results mainly due to the refinement analysis and fusion processing of time series data with category data, as well as the inclusion of a stochastic attenuation factor in GRU-S to capture the missing data. Specifically, firstly, the missing time series data are captured by the stochastic attenuation in GRU-S to reduce the non-fidelity of the information passed downstream. We also introduce VRNN to interpolate the sequence data to reduce the information loss of the missing time series, capturing the trend information well. Secondly, the patient demographic information is captured by extracting category features through FC. Our proposed model provides significant improvements in all metrics, fully explores patients' survival health information, and more accurately predicts patient survival outcomes. This has important implications for researchers and provides a reference for future research work.

4.2 Performance by GRU-S

To analyze the effectiveness of our proposed GRU-S in handling time series data, we analyzed the performance of HF mortality prediction based on time series features. We evaluated GRU-S according to the window shown in Fig. 1, and the results are shown in Table 4. The results show that the deep learning model outperforms the traditional machine learning model in handling time series data, while our proposed GRU-S outperforms MTSSP and GRU-U in most cases. When the observation window is 5 days and the prediction window is 30 days, the AUC value of GRU-S reaches 0.775, while the AUC value of MTSSP is only 0.770, GRU-U (0.766), GRU (0.759), and LSTM (0.757). The GRU-S model worked best and outperformed the other benchmark models in most evaluation windows.

Table 3 The prediction performance of all the models

Prediction window (days)	Observation window (days)	Models	Accuracy	Precision	F1 score	Specificity
30	5	LR	0.845 ± 0.005	0.506 ± 0.114	0.138 ± 0.041	0.988 ± 0.001
		RF	0.847 ± 0.002	0.586 ± 0.111	0.093 ± 0.019	0.998 ± 0.000
		SVM	0.845 ± 0.000	0.483 ±0.409	0.009 ± 0.009	**0.999 ± 0.009**
		LSTM	0.844 ± 0.010	0.509 ± 0.073	0.254 ± 0.033	0.967 ± 0.013
		GRU	0.848 ± 0.008	0.563 ± 0.093	0.238 ± 0.032	0.976 ± 0.011
		GRU-U	0.840 ± 0.008	0.475 ± 0.049	0.333 ± 0.034	0.947 ± 0.008
		MTSSP	0.841 ± 0.001	0.454 ± 0.003	0.202 ± 0.002	0.971 ± 1.244
		DFL-IMP	**0.928 ± 0.003**	**0.867 ± 0.021**	**0.734 ± 0.018**	0.982 ± 0.003
	7	LR	0.849 ± 0.002	0.478 ± 0.056	0.126 ± 0.022	0.988 ± 0.001
		RF	0.852 ± 0.002	0.626 ± 0.169	0.091 ± 0.016	0.998 ± 0.000
		SVM	0.850 ± 0.000	0.200 ± 0.266	0.006 ± 0.008	**0.999 ± 0.000**
		LSTM	0.848 ± 0.006	0.480 ± 0.051	0.254 ± 0.029	0.967 ± 0.004
		GRU	0.854 ± 0.002	0.557 ± 0.028	0.241 ± 0.014	0.978 ± 0.002
		GRU-U	0.845 ± 0.010	0.473 ± 0.056	0.364 ± 0.037	0.941 ± 0.009
		MTSSP	0.845 ± 0.003	0.434 ± 0.031	0.197 ± 0.016	0.970 ± 0.001
		DFL-IMP	**0.928 ± 0.005**	**0.857 ± 0.035**	**0.726 ± 0.024**	0.981 ± 0.004
	10	LR	0.843 ± 0.003	0.557 ± 0.16	0.128 ± 0.025	0.989 ± 0.002
		RF	0.843 ± 0.002	0.595 ± 0.211	0.008 ± 0.011	0.998 ± 0.006
		SVM	0.843 ± 0.000	0.333 ± 0.421	0.066 ± 0.011	**0.999 ± 0.000**
		LSTM	0.841 ± 0.004	0.467 ± 0.031	0.227 ± 0.027	0.968 ± 0.003
		GRU	0.841 ± 0.005	0.517 ± 0.063	0.216 ± 0.038	0.975 ± 0.006
		GRU-U	0.820 ± 0.015	0.432 ± 0.040	0.423 ± 0.040	0.895 ± 0.025
		MTSSP	0.841 ± 0.006	0.471 ± 0.028	0.273 ± 0.022	0.960 ± 0.002
		DFL-IMP	**0.922 ± 0.004**	**0.863 ± 0.017**	**0.709 ± 0.023**	0.982 ± 0.003

The highest performance of all models is shown in bold

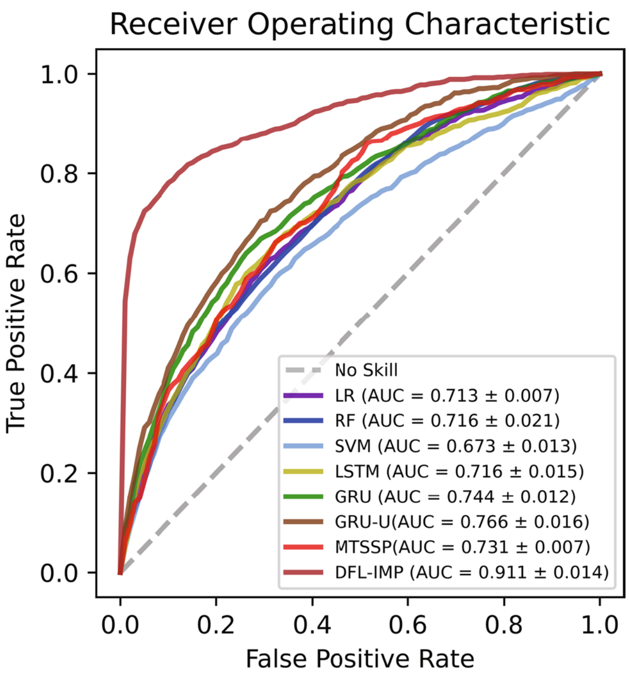


Fig. 6 ROC curves for the evaluated all models (observation window of 7 days and prediction window of 30 days)

It is shown that the update and reset gates of the GRU-S control the inflow of more useful information and the discarding of unnecessary information, respectively. For time series with missing values in the EHR, the impact of low-fidelity input data on the prediction task is reduced by adding a stochastic decay factor to the data input and hidden layers, effectively propagating randomness within the GRU unit gates at each timestamp, thus allowing the estimated data to be combined with this randomness in a non-linear manner.

4.3 Performance by feature fusion strategy

First, we explore whether data integration can improve the accuracy of the prediction models. For the machine learning models LR, RF and SVM, we concat mean of time series features and category features, fed into the model to predict patient mortality. For LSTM, GRU, GRU-U and MTSSP, we applied our fusion ideas to these baseline models by fusing the RNN final hidden layer state with the category features. The experimental results under all evaluation windows are shown in Table 5. Our model works best with an observation window of 5 days and a prediction window of 30 days, with an AUC value of 0.914, compared to LR (0.898), RF (0.900), SVM (0.881), LSTM (0.909), GRU (0.910), GRU-U (0.913), and MTSSP (0.792). Notably, with an observation window of 5 days and a prediction window of 20 days, GRU-U achieves an AUC of 0.902 after fusing features, which is higher than our model (0.900); with an observation window of 10 days and a prediction window of 20 days, GRU-U achieves an AUC of 0.891 after fusing features, which is higher than our model (0.890). However, in most cases, our model performed better than the other benchmark models.

Figure 7 compares the model performance before and after data integration, with the same window settings as in Fig. 5 by default. With an observation window of 5 days and a prediction window of 30 days, the AUC of LR can reach 0.898 after data integration, which is 19% higher than before integration, the AUC of LSTM can reach 0.909, which is 15.2% higher than before integration, and the AUC of GRU-U can reach 0.913, which is 14.7% higher than before integration, and MTSSP can achieve an AUC of 0.792, 2.2% higher than before integration. Our model achieves an AUC of 0.914, 13.9% higher than before integration. The experimental results show that data integration of time series

Table 4 The achieved AUC values of the GRU-S and benchmarks in different windows

Prediction window (days)	Observation window (days)	LR	RF	SVM	LSTM	GRU	GRU-U	MTSSP	GRU-S
10	5	0.700±0.017	0.697±0.018	0.650±0.013	0.747±0.018	0.752±0.015	0.748±0.012	0.746±0.001	**0.768±0.013**
	7	0.699±0.013	0.679±0.032	0.636±0.012	0.733±0.017	0.741±0.024	0.745±0.011	0.749±0.015	**0.767±0.015**
	10	0.686±0.036	0.689±0.004	0.637±0.048	0.721±0.023	0.722±0.029	0.732±0.022	0.698±0.012	**0.747±0.024**
20	5	0.702±0.011	0.709±0.012	0.665±0.015	0.748±0.013	0.750±0.015	0.756±0.016	**0.772±0.001**	0.771±0.012
	7	0.716±0.007	0.700±0.032	0.674±0.018	0.747±0.017	0.746±0.019	0.765±0.013	0.746±0.009	**0.777±0.014**
	10	0.704±0.036	0.708±0.021	0.654±0.025	0.735±0.019	0.739±0.019	0.747±0.021	0.739±0.009	**0.757±0.027**
30	5	0.708±0.011	0.717±0.016	0.675±0.022	0.757±0.009	0.759±0.009	0.766±0.011	0.770±0.001	**0.775±0.006**
	7	0.713±0.007	0.716±0.021	0.673±0.013	0.716±0.015	0.744±0.020	0.766±0.016	0.731±0.007	**0.775±0.012**
	10	0.704±0.026	0.715±0.027	0.662±0.027	0.738±0.007	0.735±0.010	0.740±0.027	0.728±0.012	**0.760±0.016**

The highest performance of all models is shown in bold

Table 5 The achieved AUC values using the fusion strategy under different windows

Prediction window (days)	Observation window (days)	LR + Fusion	RF + Fusion	SVM + Fusion	LSTM + Fusion	GRU + Fusion	GRU_U + Fusion	MTSSP + Fusion	DFL-IMP
10	5	0.862 ± 0.008	0.884 ± 0.017	0.845 ± 0.025	0.887 ± 0.014	0.891 ± 0.017	0.893 ± 0.017	0.866 ± 0.001	**0.894 ± 0.015**
	7	0.859 ± 0.010	0.878 ± 0.024	0.834 ± 0.029	0.874 ± 0.022	0.874 ± 0.022	0.886 ± 0.018	0.837 ± 0.001	**0.887 ± 0.021**
	10	0.826 ± 0.027	0.858 ± 0.033	0.823 ± 0.036	0.848 ± 0.033	0.851 ± 0.029	0.859 ± 0.028	0.856 ± 0.003	**0.865 ± 0.024**
20	5	0.883 ± 0.015	0.894 ± 0.017	0.864 ± 0.021	0.897 ± 0.014	0.896 ± 0.016	**0.902 ± 0.012**	0.866 ± 0.006	0.900 ± 0.022
	7	0.885 ± 0.018	0.888 ± 0.028	0.861 ± 0.027	0.885 ± 0.012	0.883 ± 0.015	0.896 ± 0.018	0.884 ± 0.006	**0.899 ± 0.011**
	10	0.869 ± 0.017	0.882 ± 0.015	0.855 ± 0.018	0.877 ± 0.006	0.874 ± 0.014	**0.891 ± 0.009**	0.869 ± 0.008	0.890 ± 0.004
30	5	0.898 ± 0.012	0.900 ± 0.010	0.881 ± 0.012	0.909 ± 0.008	0.910 ± 0.010	0.913 ± 0.009	0.792 ± 0.013	**0.914 ± 0.008**
	7	0.894 ± 0.014	0.902 ± 0.016	0.870 ± 0.018	0.898 ± 0.010	0.897 ± 0.010	0.908 ± 0.013	0.852 ± 0.008	**0.911 ± 0.014**
	10	0.884 ± 0.026	0.891 ± 0.028	0.864 ± 0.012	0.895 ± 0.027	0.888 ± 0.014	0.906 ± 0.010	0.789 ± 0.007	**0.907 ± 0.009**

The highest performance of all models is shown in bold

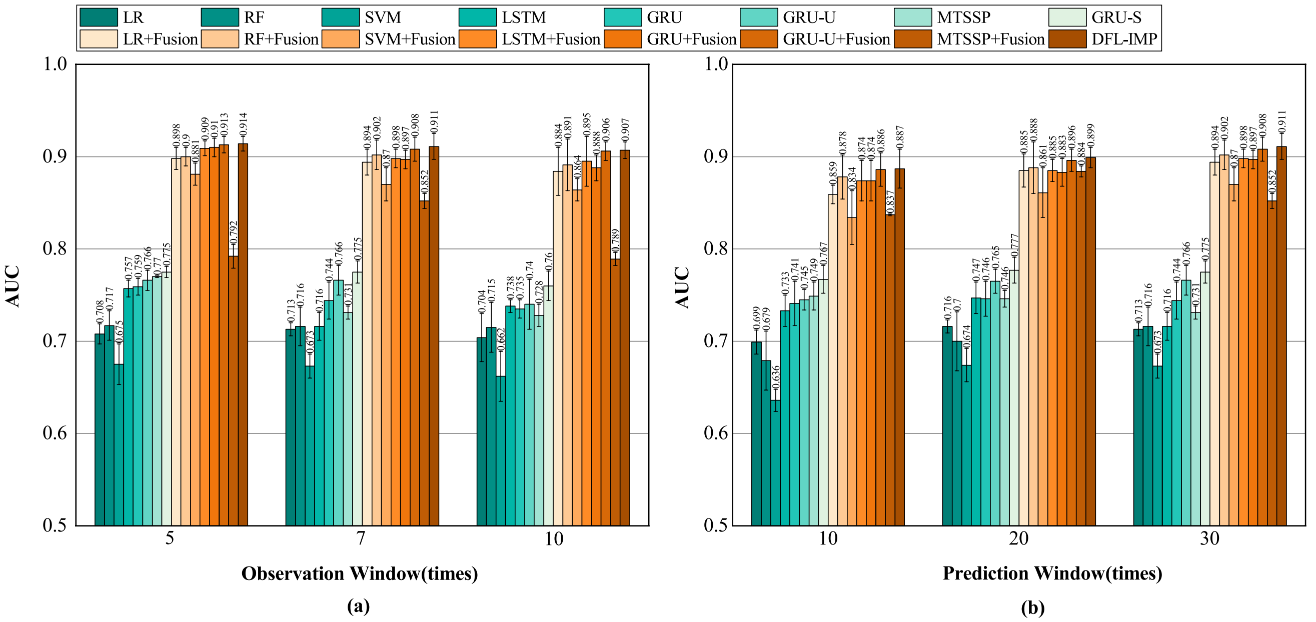


Fig. 7 The plots before and after fusion. (**a**) show the AUC values achieved by the models with different observation windows when the prediction window is set to 30 days, and (**b**) shows the AUC values achieved by the models with different prediction windows when the observation window is set to 7 days

features with category features is effective in improving the accuracy of the models compared to prediction models using a single time series feature or a single aggregated feature.

In addition, we explored the effectiveness of fine-grained analysis of the data on the sequence models, as shown in Fig. 8. "Model + Fusion" denotes our fusion strategy, which explicitly treats time series data separately from category data using different models. "Model + Integration" means that the category data are concatenated to the time series data based on the length of the time series data and input to a separate model. The results show that our model achieves an AUC of 0.914 under the applied fusion strategy, while under data integration the AUC is 0.911. GRU-U achieves an AUC of 0.913 under the applied fusion strategy, while under data integration, the AUC is 0.896. LSTM achieves an AUC of 0.909 under the applied fusion strategy, while under data integration the AUC is 0.881. The results show that the model fusion strategy is particularly effective in improving outcomes. We believe that refined analysis of healthcare data can help improve prediction accuracy.

The study shows that feature fusion in heart failure mortality prediction research questions, introducing multiple forms of patient electronic health data helps to fully exploit patient information and obtain a closer approximation to the patient's true physical condition. In addition, with the joint fusion strategy we used, the model better achieved interactions between data in different formats,

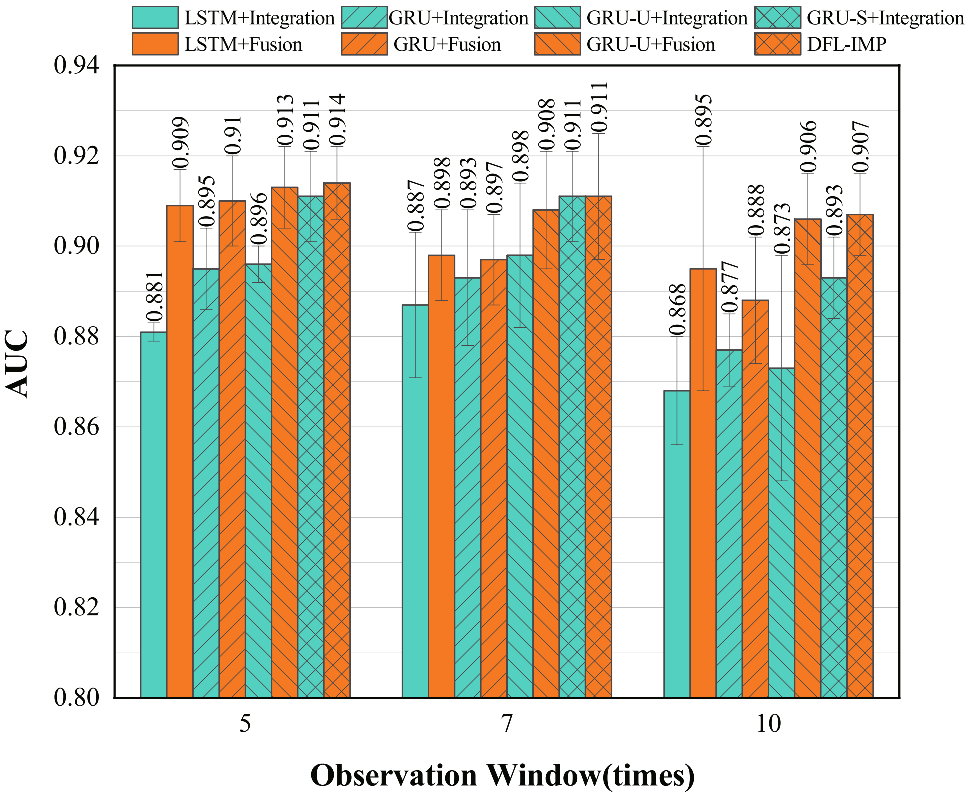


Fig. 8 Comparative plot of fine-grained analysis, AUC values achieved by the model with different observation windows when the prediction window is set to 30 days

improving the accuracy of the model for heart failure prediction tasks.

4.4 Ablation studies

In this study, we performed two sets of ablation experiments under two different observation windows to validate the results obtained by our method.

Firstly, we conducted an ablation study of the proposed GRU-S and fusion strategy, and the results are presented in Table 6. Our proposed DFL-IMP model exhibited degraded performance by ablating the GRU-S and fusion strategies. With an observation window of 5 days, the model performance was 0.775 when GRU-S was applied without the fusion strategy, and conversely, the model performance reached 0.910. It can be concluded that the gain in classification performance by applying the GRU-S model was not as significant as the fusion strategy. The model performance is best when both are applied and worst when neither is applied. It can be seen that the DFL-IMP model is not a single module that improves the algorithm performance, but the best result is produced by the combination.

Secondly, we used a convolutional neural network (CNN) instead of the FC used in the category features and the results of the study are shown in Table 7. With an observation window of 5 days, the AUC value reached 0.914 when using FC, whereas when using CNN, the AUC value was only 0.887. This indicates that FC performs better than CNN. We speculate that this is because FC mainly maps vectors, while CNN is mainly used for feature extraction and is good at extracting Euclidean structured data. For the category features after preprocessing (one-hot) in the experiment, convolution cannot extract feature information effectively, while FC can learn effective mapping transformations.

Table 6 DFL-IMP submodule ablation study results

Prediction windows	Observation windows	GRU-S	Fusion	AUC
30	5	×	×	0.759 ± 0.009
		×	√	0.910 ± 0.010
		√	×	0.775 ± 0.006
		√	√	**0.914 ± 0.008**
	7	×	×	0.750 ± 0.020
		×	√	0.897 ± 0.010
		√	×	0.775 ± 0.012
		√	√	**0.911 ± 0.014**

The highest performance of all models is shown in bold

Table 7 Category data model ablation study results

Prediction windows	Observation windows	FC	CNN	AUC
30	5	√	×	**0.914 ± 0.008**
		×	√	0.887 ± 0.005
	7	√	×	**0.911 ± 0.014**
		×	√	0.873 ± 0.010

The highest performance of all models is shown in bold

5 Discussion

In this study, it was first hypothesized that a refined analysis of time series and category data in the electronic health record would help to improve the accuracy of in-hospital mortality prediction models for patients with HF. The prediction performance results for all models are shown in Fig. 5 and Table 3. The proposed model, DFL-IMP, is considerably more effective after the introduction of category data and significantly outperforms other baseline models (LR, RF, SVM, LSTM, GRU, GRU-U, MTSSP). This leads to the conclusion that mining the hidden information contained in each of the different forms of data and introducing auxiliary information (i.e., demographic information) is crucial to the HF mortality prediction model.

Secondly, we believe that for a large number of missing values in medical time series data, the model adopts additional processing mechanisms to help extract information from the data and improve prediction accuracy. As can be seen from Table 4, our proposed GRU-S outperforms other sequence models (LSTM, GRU, GRU-U, MTSSP), 0.9% increase compared to GRU-U, achieving good results. This suggests that propagating uncertainty within the GRU unit gate reduces the loss of information in the time series due to missing values, allowing the estimated data to be combined with randomness in a non-linear manner. This helps to reduce the impact of low-fidelity input data on mortality prediction results. That is, in the presence of missing values, key information in the sequence data can still be captured.

For the first time, we use a feature fusion strategy to construct a deep learning model to predict in-hospital mortality in patients with heart failure. In our evaluation results, we confirmed that our model results represent the most advanced predictive performance that can be achieved by applying a deep learning model that uses feature fusion strategies to discover complex relationships in EHR data. For small differences in window length, the model effect does not differ much. For the same prediction window, the prediction effect decreases with the increase of the observation window. Under the same observation window, the prediction effect of the model increases with the increase of the prediction window, as shown in Fig. 5(b), and Fig. 7(b). This is due to our dataset; the longer the observation window (the shorter the prediction window), the more positive samples we exclude, the more unbalanced the data and the decreasing model effect. This suggests that our model prefers a more balanced dataset and works relatively better on a more balanced dataset. We expect that model performance will benefit from using a more balanced sample of cases.

Limitations of this study focus on three aspects. Firstly, the length of our time series feature observation window is short and insufficient, and the performance of the model is limited by the equilibrium state of the dataset. Secondly, our fusion is a simple fusion without further exploration of other fusion strategies. Third, we mined different forms of data in a single mode without combining multiple modes, such as chest radiographs, medical prescriptions, and medications, to obtain adequate prognostic information on heart failure.

6 Conclusion and future work

This paper aims to apply deep learning methods to in-hospital mortality prediction in HF using classification methods. We propose a new model DFL-IMP to refine the analysis of time series features and category features in EHR data and perform feature fusion modeling. In this, GRU-S is proposed to capture time series features, FC to capture category features and VRNN to estimate missing values of time series features. The experimental results show that our proposed model DFL-IMP dramatically improves the accuracy of in-hospital mortality decisions in HF, which helps physicians to make timely interventions in high-risk patients, delay the progression of HF disease, and improve the quality of life and life expectancy of patients.

In future work, we will combine data from clinical hospitals, select samples with fewer missing values, and choose observation windows of appropriate length. We will also introduce a priori knowledge of clinical medicine and use statistical analysis to find the impact of specific features on the results to enhance the interpretability of deep learning. Further, we will explore whether other advanced fusion strategies can improve model performance, as well as consider multimodal data fusion studies, e.g., physician's orders versus chest films, to produce more clinically meaningful models.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1007/s11517-023-02816-z.

Acknowledgements The paper supported by National Major Scientific Research Instrument Development Project (62027819): High-speed Real-time Analyzer for Laser Chip's Optical Catastrophic Damage Process; The General Object of National Natural Science Foundation(62076177): Research on Risk Assessment Model for Heart Failure Incorporating Multi-modal Big Data; Shanxi Province key technology and generic technology R\&D project (2020XXX007): Energy

Internet Integrated Intelligent Data Management and Decision Support Platform; Key research and development program of Shanxi Province NO.202102020101006.

References

1. Chen J, Aronowitz P (2022) Congestive heart failure[J]. Medical Clinics 106(3):447–458. https://doi.org/10.1016/j.mcna.2021.12.002
2. Wang H, Chai K, Du M et al (2021) Prevalence and incidence of heart failure among urban patients in China: a national population-based analysis[J]. Circ Heart Fail 14(10):e008406. https://doi.org/10.1161/CIRCHEARTFAILURE.121.008406
3. Błaziak M, Urban S, Wietrzyk W et al (2022) An artificial intelligence approach to guiding the management of heart failure patients using predictive models: a systematic review[J]. Biomedicines 10(9):2188. https://doi.org/10.3390/biomedicines10092188
4. Kao DP (2022) Electronic health records and heart failure[J]. Heart Fail Clin 18(2):201–211. https://doi.org/10.1016/j.hfc.2021.12.004
5. Reimer AP, Dai W, Smith B et al (2021) Subcategorizing EHR diagnosis codes to improve clinical application of machine learning models[J]. Int J Med Inform 156:104588. https://doi.org/10.1016/j.ijmedinf.2021.104588
6. Lv H, Yang X, Wang B et al (2021) Machine learning-driven models to predict prognostic outcomes in patients hospitalized with heart failure using electronic health records: retrospective study[J]. J Med Internet Res 23(4):e24996. https://doi.org/10.2196/24996
7. Javeed A, Khan S U, Ali L et al (2022) Machine Learning-Based Automated Diagnostic Systems Developed for Heart Failure Prediction Using Different Types of Data Modalities: A Systematic Review and Future Directions[J]. Computational and Mathematical Methods in Medicine 2022:9288452. https://doi.org/10.1155/2022/9288452
8. Benke KK (2019) Data Analytics and Machine Learning for Disease Identification in Electronic Health Records[J]. JAMA Ophthalmol 137(5):497–498. https://doi.org/10.1001/jamaophthalmol.2018.7055
9. König S, Pellissier V, Hohenstein S et al (2021) Machine learning algorithms for claims data-based prediction of in -hospital mortality in patients with heart failure[J]. ESC Heart Fail 8(4):3026–3036. https://doi.org/10.1002/ehf2.13398
10. Adler ED, Voors AA, Klein L et al (2020) Improving risk prediction in heart failure using machine learning[J]. Eur J Heart Fail 22(1):139–147. https://doi.org/10.1002/ejhf.1628
11. Angraal S, Mortazavi BJ, Gupta A et al (2020) Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction[J]. JACC Heart Fail 8(1):12–21. https://doi.org/10.1016/j.jchf.2019.06.013
12. Davide C, Giuseppe J (2020) Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone[J]. BMC Medical Informatics and Decision Making 20(1):16. https://doi.org/10.1186/s12911-020-1023-5
13. Xie F, Yuan H, Ning Y et al (2022) Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies[J]. J Biomed Inform 126:103980. https://doi.org/10.1016/j.jbi.2021.103980
14. Penso M, Solbiati S, Moccia S et al (2022) Decision Support Systems in HF based on Deep Learning Technologies[J]. Current Heart Failure Reports 19(2):38–51. https://doi.org/10.1007/s11897-022-00540-7
15. Haq IU, Chhatwal K, Sanaka K et al (2022) Artificial intelligence in cardiovascular medicine: current insights and future prospects[J]. Vasc Health Risk Manag 18:517. https://doi.org/10.2147/VHRM.S279337
16. Doppalapudi S, Qiu R G, Badr Y (2021) Lung cancer survival period prediction and understanding: Deep learning approaches[J]. International Journal of Medical Informatics 148:104371
17. Gupta VK, Gupta A, Kumar D et al (2021) Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model[J]. Big Data Min Anal 4(2):116–123. https://doi.org/10.26599/BDMA.2020.902001
18. Tong R, Lei L, Zhou Y et al (2019) Representation learning for clinical time series prediction tasks in electronic health records[J]. BMC Med Inform Decis Mak 19(Suppl 8):259. https://doi.org/10.1186/s12911-019-0985-7
19. Amal S, Safarnejad L, Omiye JA et al (2022) Use of Multi-Modal Data and Machine Learning to Improve Cardiovascular Disease Care[J]. Frontiers in Cardiovascular Medicine 9:840262. https://doi.org/10.3389/fcvm.2022.840262
20. Wang Z, Zhu Y, Li D et al (2020) Feature rearrangement based deep learning system for predicting heart failure mortality[J]. Comput Methods Prog Biomed 191:105383. https://doi.org/10.1016/j.cmpb.2020.105383
21. Jun E, Mulyadi A W, Choi J, et al (2020). Uncertainty-Gated Stochastic Sequential Model for EHR Mortality Prediction[J]. IEEE Transactions on Neural Networks and Learning Systems 32(9):4052–4062. https://doi.org/10.1109/TNNLS.2020.3016670
22. Men L, Ilk N, Tang X et al (2021) Multi-disease prediction using LSTM recurrent neural networks[J]. Expert Syst Appl 177:114905. https://doi.org/10.1016/j.eswa.2021.114905
23. Priyanga P, Pattankar VV, Sridevi S (2021) A hybrid recurrent neural network-logistic chaos-based whale optimization framework for heart disease prediction with electronic health records[J]. Comput Intell 37(1):315–343. https://doi.org/10.1111/coin.12405
24. Yoon J, Zame W R, van der Schaar M (2018) Estimating Missing Data in Temporal Data Streams Using Multi-directional Recurrent Neural Networks[J]. IEEE Transactions on Biomedical Engineering 66(5):1477–1490. https://doi.org/10.1109/TBME.2018.2874712
25. McGilvray MMO, Heaton J, Guo A et al (2022) Electronic health record-based deep learning prediction of death or severe decompensation in heart failure patients[J]. Heart Fail 10(9):637–647. https://doi.org/10.1016/j.jchf.2022.05.010
26. Chu J, Dong W, Huang Z (2020) Endpoint prediction of heart failure using electronic health records[J]. J Biomed Inform 109:103518. https://doi.org/10.1016/j.jbi.2020.103518
27. Radhachandran A, Garikipati A, Zelin N S, et al (2021) Prediction of short-term mortality in acute heart failure patients using minimal electronic health record data[J]. BioData Mining 14(1):23. https://doi.org/10.1186/s13040-021-00255-w
28. Zheng G, Han G, Soomro NQ (2019) An inception module CNN classifiers fusion method on pulmonary nodule diagnosis by signs[J]. Tsinghua Sci Technol 25(3):368–383. https://doi.org/10.26599/TST.2019.9010010
29. Zhi Z, Elbadawi M, Daneshmend A, et al (2022) Multimodal Diagnosis for Pulmonary Embolism from EHR Data and CT Images[C]. 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, pp 2053–2057. https://doi.org/10.1109/EMBC48229.2022.9871041
30. Zheng Yi, Xiangpei Hu (2020) Healthcare predictive analytics for disease progression: a longitudinal data fusion approach. J Intell Inf Syst 55(2):351–369. https://doi.org/10.1007/s10844-020-00606-9
31. Niu K, Lu Y, Peng X et al (2022) Fusion of sequential visits and medical ontology for mortality prediction[J]. J Biomed Inform 127:104012. https://doi.org/10.1016/j.jbi.2022.104012

32. Lin M, Wang S, Ding Y, et al (2021) An empirical study of using radiology reports and images to improve ICU-mortality prediction[C]. 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI). IEEE, Victoria, BC, Canada, pp 497–498. https://doi.org/10.1109/ICHI52183.2021.00088
33. Hess LM, Han Y, Zhu YE et al (2021) Characteristics and outcomes of patients with RET-fusion positive non-small lung cancer in real-world practice in the United States[J]. BMC Cancer 21(1):1–12
34. Chung J , Kastner K , Dinh L et al (2015) A recurrent latent variable model for sequential data[J]. Advances in neural information processing systems 28
35. Ross A, Jain A (2003) Information fusion in biometrics[J]. Pattern Recogn Lett 24(13):2115–2125. https://doi.org/10.1016/S0167-8655(03)00079-5
36. Alistair LS, Johnson EW, Pollard TJ (2016) Data descriptor: MIMIC-III a freely accessible critical care database[J]. Thromb Haemost 76(2):258–262. https://doi.org/10.1038/sdata.2016.35
37. Gurwitz JH, Magid DJ, Smith DH et al (2013) Contemporary prevalence and correlates of incident heart failure with preserved ejection fraction[J]. Am J Med 126(5):393–400. https://doi.org/10.1016/j.amjmed.2012.10.022
38. Li B, Shi Y, Cheng L, et al (2022) MTSSP: Missing value imputation in multivariate time series for survival prediction[C]. 2022 International Joint Conference on Neural Networks (IJCNN). Padua, Italy, pp 1–8. https://doi.org/10.1109/IJCNN55064.2022.9892806.
39. Shickel B, Silva B, Ozrazgat-Baslanti T et al (2022) Multi-dimensional patient acuity estimation with longitudinal EHR tokenization and flexible transformer networks[J]. Front Digital Health 4:1029191. https://doi.org/10.3389/fdgth.2022.1029191
40. Li Y, Mamouei M, Salimi-Khorshidi G et al (2022) Hi-BEHRT: Hierarchical Transformer-Based Model for Accurate Prediction of Clinical Events Using Multimodal Longitudinal Electronic Health Records[J]. IEEE J Biomed Health Inform27(2):1106–1117. https://doi.org/10.1109/JBHI.2022.3224727
41. Liu S, Wang X, Hou Y et al (2022) Multimodal Data Matters: Language Model Pre-Training Over Structured and Unstructured Electronic Health Records[J]. IEEE Journal of Biomedical and Health Informatics 27(1):504–514. https://doi.org/10.1109/JBHI.2022.3217810

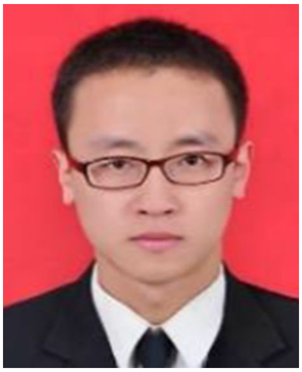
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Meikun Ma was born in 1997. She is currently pursuing an M.S. degree at the Taiyuan University of Technology. Her research interests include heart failure outcome prediction and deep learning.

Xiaoyan Hao is an associate professor in the School of Information and Computer Science at the Taiyuan University of Technology. Her main research interests include computational linguistics.

Jumin Zhao is a professor at the College of Information and Computer at the Taiyuan University of Technology. Her main research includes intelligent perception and internet of things technology.

Shijie Luo was born in 1998. He is currently pursuing an M.S. degree at the Taiyuan University of Technology. His research interests include complex networks, representation learning, and deep learning.



Yi Liu obtained an M.S. degree from Ningxia University in 2018. He is currently pursuing a doctoral degree at the Taiyuan University of Technology. His research interest is medical datasets.

Dengao Li is a professor at the College of Data Science at the Taiyuan University of Technology. He is the chairman of the Taiyuan Branch of the International Computer Society (ACM).