HGV4Risk: Hierarchical Global View-guided Sequence Representation Learning for Risk Prediction

YOURU LI and ZHENFENG ZHU, Institute of Information Science, Beijing Jiaotong University and Beijing Key Laboratory of Advanced Information Science and Network Technology, China

XIAOBO GUO, Institute of Information Science, Beijing Jiaotong University and MYBank, Ant Group, China

SHAOSHUAI LI, MYBank, Ant Group, China

YUCHEN YANG, Department of Biology, Johns Hopkins University, USA

YAO ZHAO, Institute of Information Science, Beijing Jiaotong University and Beijing Key Laboratory of Advanced Information Science and Network Technology, China

Risk prediction, usually achieved by learning representations from patient’s physiological sequence or user’s behavioral sequence data, and has been widely applied in healthcare and finance. Despite that, some recent time-aware deep learning methods have led to superior performances in such sequence representation learning tasks, such improvement is limited due to a lack of guidance from hierarchical global view. To address this issue, we propose a novel end-to-end Hierarchical Global View-guided (HGV) sequence representation learning framework. Specifically, the Global Graph Embedding (GGE) module is proposed to learn sequential clip-aware representations from temporal correlation graph (TCG) at instance level. Furthermore, following the way of key-query attention, the harmonic $\beta$-attention ($\beta$-Attn) is also developed for making a global tradeoff between time-aware decay and observation significance at channel level adaptively. Moreover, the hierarchical representations at both instance level and channel level can be coordinated by the heterogeneous information aggregation under the guidance of global view. Experimental results on both healthcare risk prediction benchmark and SMEs credit overdue risk prediction task from the real-world industrial scenario in MYBank, Ant Group, have illustrated that the proposed model can achieve competitive prediction performance compared with other known baselines. The code has been released public available at: https://github.com/LIYouru0228/HGV.

This work was supported in part by Science and Technology Innovation 2030 – New Generation Artificial Intelligence Major Project under Grant 2018AAA0102101, Beijing Natural Science Foundation, China under Grant No. 7222313, National Natural Science Foundation of China under Grant No. 61976018 and No. U1936212, and National High Level Hospital Clinical Research Funding under grant No. 2022-PUMCH-C-041 and also in part by Ant Group RI Program. We also would like to thank the anonymous reviewers and editors for their insightful comments and valuable suggestions.

Authors’ addresses: Y. Li, Z. Zhu, (corresponding author) and Y. Zhao, Institute of Information Science, Beijing Jiaotong University and Beijing Key Laboratory of Advanced Information Science and Network Technology, No. 3, Shangyuancun, Haidian, Beijing 100044, China; emails: {liyouru, zhfzhu, yzhao}@bjtu.edu.cn; X. Guo, Institute of Information Science, Beijing Jiaotong University and MYBank, Ant Group, No. 1 East Third Ring Middle Road, Chaoyang, Beijing 100026, China; email: xb_guo@bjtu.edu.cn; S. Li, MYBank, Ant Group, No. 1 East Third Ring Middle Road, Chaoyang, Beijing 100026, China; email: lishaoshuai.lss@antgroup.com; Y. Yang, Department of Biology, Johns Hopkins University, Baltimore, MD 21218; email: yuchen.yang@jhu.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM Transactions on Knowledge Discovery from Data, Vol. 18, No. 1, Article 1. Publication date: August 2023.
1 INTRODUCTION

Aiming at predicting the occurrence probability of the event of interest, whether the patient will develop a disease [36] or whether the user will engage in non-conforming behavior, the risk prediction modeling is commonly used in the world of medicine [9, 11, 23] to help guide clinical decision-making but are also used in other fields such as finance [24, 44] and E-commerce [47, 50] to avoid economic losses. To solve these tasks can rely on representation learning techniques [25]. Specifically, the user/patient sequence patterns usually can be captured by learning trends of physiological sequence [27] or sequential behavior [38, 39] from historical data. Recent studies show that some deep learning techniques, particularly Recurrent Neural Networks (RNNs) and their variants [4, 16], have achieved impressive performance in such a sequence representation learning task [14].

Beyond plain RNNs, some studies have started to utilize attention mechanism to characterize the importance of the individualization in some general tasks such as computer vision [29] and natural language processing [1]. Inspired by this, others have employed attention-based models [33] for structured data representation learning. Such typical time-aware models focus on learning weights for each time interval and capturing the long-term temporal dependencies at instance level. However, they fail to explore more granular information encoded in the original channel-level signal. To solve such problem, models in another branch such as RETAIN [6] and RetainEX [21] are proposed by exploring both temporal relationships and variable significance with a two-level neural attention framework.

In this context, however, the irregularity of time intervals in historical sequence has not been addressed. To this end, some works [2] have been proposed to introduce decay modules for jointly learning time-decay and contextual dependencies during the representation learning with irregular time series data, achieving better prediction performance. Although the aforementioned studies have addressed both long-term dependencies and time irregularity, the effect is still far from satisfactory. Self-attention architecture has been introduced in models such as SAnD [34] and HiTANet [25] to further capture the sequential dependencies in time series data, thus achieving better prediction performance. To explore the relationship between dynamic information and static information, ConCare [28] has further improved the performance by modeling both channel-wise sequential dependencies and feature-dependencies while utilizing static information.

However, most of the above-mentioned temporal modeling work relies on progressive manners, which typically captures the sequential dependencies merely from the time series data. From this perspective, although dynamic and static information is available, some existing studies are yet to be done in capturing inherent patterns other than sequential dependencies in time series data over long time spans. This will to some extent help to reveal the temporal rhythmic variation of the observed status in time series, which can be characterized by the temporal correlation graph ....
(TCG) induced from dynamic time series data, as shown in Figure 1. For these heterogeneous data, a representation learning framework should be established for jointly modeling both sequential dependencies in time series data and the global status correlation in TCG data, meanwhile adopting static information.

In general, the major difference between our work and the above models is that we provide a new perspective for risk prediction modeling with the help of hierarchical representations learned from heterogeneous data. Given the limitations of existing methods, a workable model needs to meet three main challenges: (1) how to learn the temporal correlation among status in sequence data at instance level; (2) how to capture the specific patterns against long-term decay from channel-level irregular time series data in an explicit and end-to-end manner; (3) how to aggregate the channel-wise representations and instance-wise representations together with static information by a union hierarchical heterogeneous representation learning framework. Therefore, to address these issues, this article makes the following contributions:

— For the purpose of obtaining sequential clip-aware representation from TCG data at instance level, a global graph embedding method is proposed.
— A novel key-query attention, i.e., the harmonic $\beta$-attention, is proposed to learn a global tradeoff between time-aware decay and observation significance for irregular time series data at channel level adaptively.
— To coordinate the hierarchical representation learned from both instance-level and channel-level data, a heterogeneous information aggregation strategy is introduced for modeling dependencies among the multi granularity information.
— We advance the representation learning framework Hierarchical Global View-guided (HGV) by testing it on two real-world risk prediction tasks across both healthcare and financial domains, and show its effectiveness.

The rest of the article is organized as follows. Section 2 reviews some related work. In Section 3, the problem definition and symbol notation is presented. Section 4 describes the proposed approach. Section 5 presents the experimental results and Section 6 concludes.
2 RELATED WORK

In recent years, risk prediction has been successfully applied in many real-world tasks, especially in healthcare and financial areas [3, 43, 49]. Indiscriminately, as a sequence representation learning problem, here, we only take the healthcare risk prediction as an example, and summarize the following typical modeling paradigms:

**Time-aware Models.** As a typical time-aware model, RETAIN [6] claims that existing deep learning methods often have to face the challenge of tradeoff between interpretability and performance. It developed two RNNs in the reversed time and then generated the context vector with an attention-based representation learning module. Although it can improve interpretability to some extent, its performance is limited. Moreover, as a patient subtyping model, T-LSTM [2] was proposed to address the challenge that the traditional LSTMs suffer from suboptimal performance when handling data with time irregularities. It learns a subspace decomposition of the cell memory, which enables time decay to discount the memory content according to the elapsed time. Although they have achieved better performance, it lacks considering the impact of time-aware decay [28].

**Attention-based Models.** Prior efforts usually leverage attention or self-attention architecture in risk prediction models. For example, as a framework composed of an attention-based RNN and a conditional deep generative model, MCA-RNN [22] was proposed for capturing the heterogeneity of EHRs by considering essential context into the sequence modeling. To effectively handle long sequences in a time-series modeling task, SAnD [34] has employed a masked self-attention mechanism and introduced positional encoding and dense interpolation strategies for incorporating temporal order. Furthermore, for the better exploration of the personal characteristics during the sequences and the improvement of the time-decay assumption for covering all conditions, ConCare [28] was proposed by combining a new time-aware attention and multi-head self-attention with a cross-head decorrelation loss. Moreover, various attention structures were also proposed to explore other general sequence representation learning-based tasks, such as skeleton-based action recognition [41, 42]. However, as most of the temporal modeling methods have done, they are also ineffective at capturing inherent patterns such as temporal correlations among status in time series data.

**Knowledge-enhanced Models.** Due to the sparsity and low quality of data, some knowledge-enhanced models have been raised recently. The models such as GRAM [5], KAME [26], MMORE [35], and HAP [48] were proposed to improve the prediction performance on healthcare-related tasks by incorporating the medical ontology such as ICD codes. Moreover, to better model sequences of ICD codes, HiTANet [25] assumes the non-stationary disease progression and proposes a hierarchical time-aware attention network to predict diagnosis codes by employing a time-aware Transformer at visit level and a time-aware key-query attention mechanism among timestamps. However, such knowledge encoded in medical ontology is not always available for all issues [28]. Furthermore, to fully extract the correlation between similar patients inside the dataset, GRASP [46] clusters patients who have similar conditions and results, and then improves the performance by leveraging knowledge extracted from these similar patients. However, this risk prediction model leads to insufficient improvement from noisy knowledge due to the inevitable gap between the extracted knowledge and the patients themselves.

**Graph-based Models.** Meanwhile, to better model the dependencies between temporally-order user behavioral sequences, some graph-based models were raised, especially in those existing session-based recommendation techniques. For example, [40] combined self-supervised learning with co-training for informative session-based data augmentation in session-based recommendation. [18] proposed a multi-task learning framework to jointly learn item transition dynamics in automatic and hierarchical manner. Moreover, [30] introduced a collaborative graph learning approach for session-based recommendation. However, although these works
can capture high-order behavior-wise dependencies from click sequences by using graph neural networks, they still pay insufficient attention to learning temporal correlation between multiple behaviors from hierarchical views.

3 NOTATIONS AND FRAMEWORK

3.1 Notations

To facilitate the elaboration of the risk prediction task to be dealt with, some notations used throughout the article are given first in Table 1.

Let $U = \{u_1, u_2, \ldots, u_{|U|}\}$ be the set of users/patients with the observed dynamic status information $S = \{S^i \in \mathbb{R}^{N_d \times T} \}_{i=1, \ldots, |U|}$ and the corresponding ground truth label $y = \{y_i\} \in \{0, 1\}^{|U|}$, where $N_d$ and $T$ represent the number of channels and the time step for feature observation of time series, respectively, and $|U|$ denotes the volume of $U$. In addition, it is also assumed that a $N_b$-dimensional basic features (also called static information) $F_b = \{f^i_b \in \mathbb{R}^{N_b}\}_{i=1, \ldots, |U|}$ is also available for each user/patient together with the dynamic information $S$, e.g., the static information including age and weight, and so on, in healthcare risk prediction. Particularly, for the dynamic status information $S^i = [s^i_{n,t}]_{n,t} \in \mathbb{R}^{N_d \times T}$ of the $i$th user/patient, we use the row vector $S^i_n = [s^i_{n,1}, s^i_{n,2}, \ldots, s^i_{n,T}] \in \mathbb{R}^T$ of $S_i$ denote the observed $T$ historical sequence of statuses from the $n$-th channel, and likewise, the column vector $S^i_t = [s^i_{1,t}, s^i_{2,t}, \ldots, s^i_{N_d,t}] \in \mathbb{R}^{N_d}$ of $S^i$ denote the observed historical statuses from $N_d$ channels at the $t$-th time step. With the given historical statuses $S^i$, a global status correlation graph $g_i$ can be built by using $\{S^i_t\}_{t=1, \ldots, T}$ as the nodes. With the above notations, we formally define the global view and channel view on characterizing a patient/user as follows.

**Definition 1 (Global View).** The global view at instance level is defined as a fused observation of both the global status correlation graph $g_i$ and static information $f^i_b$ from the $i$th instance.

**Definition 2 (Channel View).** The channel view is defined as a globally coordinated observation of each status $s^i_{n,1}, s^i_{n,2}, \ldots, s^i_{n,T}$ in the historical sequence of statuses $S^i_n$, from the $n$th channel.

As opposed to the representations of the users/patients from global view, the channel view provides a more granular characterization of them.

---

**Table 1. Notations and Description**

| Notation | Description |
|----------|-------------|
| $U = \{u_i\}$ | The set of $|U|$ users/patients |
| $y_i, \bar{y}_i$ | The ground truth and predicted labels for $u_i$ |
| $T$ | No. of time steps for making observation |
| $N_d$ | No. of channels for making observation |
| $S^i$ | The dynamic status information for $u_i$ |
| $S^i_{t, n} \in \mathbb{R}^{N_d}$ | The observed statuses at the $t$th time step from the $n$-th channel |
| $f^i_b \in \mathbb{R}^{N_b}$ | $N_b$-dimensional basic features for $u_i$ |
| $g_i \in \mathbb{R}^{T \times T}$ | Temporal clip global correlation graph using $\{S^i_t\}_{t=1, \ldots, T}$ as nodes |
| $E_d(\cdot), E_b(\cdot)$ and $E_g(\cdot)$ | Embeddings for $S^i_n, f^i_b$, and $g_i$, respectively |

---

1Here, the risk prediction is formulated as a binary classification problem, and without loss of generality, it is trivial to extend our model to multilevel risk prediction.
3.2 Overall Framework

An overall illustration of the proposed hierarchical global view-guided sequence representation learning for risk prediction is given in Figure 2. Specifically, the proposed framework mainly consists of three modules:

1. **Global Graph Embedding (GGE).** It aims at learning global clip-aware representations via convolution neural network on a TCG at the sequential clip level;

2. **Harmonic $\beta$-attention.** Inspired by the F-score, the proposed harmonic $\beta$-attention attempts to make a global tradeoff between time-aware decay and observation significance;

3. **Heterogeneous Information Aggregation.** To aggregate heterogeneous information, the instance-wise representations and channel-wise representations are weighted and formed a unified representation with multi-granularity information, which makes the hierarchical guidance on two global views well coordinated.

4 METHODOLOGY

4.1 Global Graph Embedding

Distinguished from the sequential dependencies in time series data, the correlations between temporal status should also be considered from a global perspective, thus enhancing the acquisition of more valuable historical statuses for the risk prediction in the future. For this purpose, a GGE approach is proposed.
4.1.1 Temporal Correlation Graph. Empirically, in medical diagnosis, the physical signs of the human body at different times can be similar and interrelated to each other. For example, a person’s blood glucose in general rises gradually after a meal and returns to its premeal status in about two hours [10]. Therefore, as a usual fact, the blood glucose status \( S_{n,t} \) at each layer \( t \) denotes the concatenated vector of the output of the last layer of convolution neural network. \( S_{n,1} \) and \( S_{n,t} \) for patient \( u_i \) at 12 p.m. and 6 p.m. (time for lunch or dinner) can be similar, even though these two moments are not closely adjacent.

To fully explore the correlations among nonadjacent status and mine heterogeneous correlation beyond homogeneous sequences, a graph of status is constructed to represent the correlative similarity in different clips under the global view. Specifically, for each user/patient \( u_i \) with historical status set \( \{S_{i,t}\}_{t=1}^{T} \), we define the TCG as \( G(V_i, E_i, g_i) \), where \( V_i = \{S_{i,t}\}_{t=1}^{T} \) denotes the nodes of the graph, \( E_i = V_i \times V_i \) is the set of edges, and \( g_i \in \mathbb{R}^{T \times T} \) represents the graph adjacency matrix with its element \( g_i[t_1, t_2] \in [0, 1] \) being the normalized cosine similarity between nodes \( S_{i,t_1} \) and \( S_{i,t_2} \).

4.1.2 Global Graph Embedding. Different from the conventional unordered graph, the TCG \( g_i \) is constructed in a temporal order as shown in Figure 2, it means that the convolution neural network can be directly applied to \( g_i \) using a clip-aware sliding convolution kernel, thus obtaining a global embedding of TCG \( g_i \). One thing worth pointing out is that the traditional used GNNs [20], such as GCN and GAT, are only feasible to graph node representation rather than the graph representation itself, i.e., global graph representation. Specifically, let \( W_i^l \) and \( b_i^l \) be the convolution kernel parameters of layer \( l \) in 2-D convolutional networks, the output \( g_i^{(l)} \) at each layer \( l \in \{1, \ldots, L\} \) is given as follows:

\[
g_i^{(l)} = f \left( g_i^{(l-1)} \star W_i^l + b_i^l \right),
\]

where \( \star \) is the convolutional operation, and \( f(\cdot) \) denotes a non-linear activation function (ReLU [13]) is used in our case. With the sliding of convolution kernel, the clip-aware patterns encoded in temporal status correlation can be extracted effectively. Following the convolution neural network is a fully connected layer to is to obtain the final GGE \( E_g(g_i) \) and we have

\[
E_g(g_i) = f \left( W_i^{FC} \cdot g_i^{(L)} + b_i^{FC} \right),
\]

where \( W_i^{FC} \) and \( b_i^{FC} \) are parameters of the fully connected layer, and \( g_i^{(L)} \) denotes the concatenated vector of the output of the last layer of convolution neural network.

4.2 Harmonic \( \beta \)-attention

4.2.1 Temporal Modeling at Channel Level. To capture the temporal dependencies within an individualized sequential signal at channel level, we set the Long Short-Term Memory networks (LSTMs) [17] as the backbone. Specifically, one of time series channels \( S_{n} = (S_{n,1}, S_{n,2}, \ldots, S_{n,T}) \in S \) is fed into an LSTM network and the output for feature \( S_{n,t} \) at time \( t \) can be obtained by

\[
h_{n,1}^i, h_{n,2}^i, \ldots, h_{n,T}^i = LSTM_\Theta \left( S_{n,1}^i, S_{n,2}^i, \ldots, S_{n,T}^i \right),
\]

where \( h_{n,t}^i \in \mathbb{R}^{d_i} \) is the hidden representation for \( S_{n,t}^i \) and \( \Theta \) is the parameter space need to be learned for each LSTM network.

4.2.2 \( \beta \)-Attn: An Adaptive Key-query Attention. To make a global tradeoff between time-aware decay and observation significance in trends learning for sequence signal from both healthcare and financial domains, an adaptive key-query attention is proposed in this section. To be specific, existing studies [2, 28] have shown the effectiveness of considering time-aware decay in long sequences and capturing the irregularity among different visit records. However, some significant
variable values in specific visit record should be given more weights than those less significant values in nearer ones. For example, patients with severe diabetes in ICU may experience sudden blood glucose spikes and gradually recover as the result of therapeutic interventions. Obviously, such abnormal values deserve more attention [31], which calls for a global tradeoff between time-aware decay and observation significance. The time-aware decay $d^i_\omega$ and the observation significance $o^i_{n,t}$ are defined as

$$d^i_\omega = 1 - \frac{\Delta t}{\max(\Delta t)},$$

$$o^i_{n,t} = \frac{\sigma(S^i_{n,t})}{\sigma(\max(|S^i_n|))},$$

where $\Delta t$ is the time interval from time $t$ to the latest observation time $T$, $\max(\Delta t) = T$, and $\sigma(\cdot)$ the non-linear mapping function $sigmoid$. Meanwhile, the F-score for model performance evaluation is a harmonic means of precision and recall, namely, the reciprocal of the average of the reciprocal of precision and the reciprocal of recall [32]. Inspired by this, we learn the $\beta$-Attn, a key-query attention, to make a global tradeoff between time-aware decay and observation significance adaptively. Specifically, the tradeoff between $d^i_\omega$ and $o^i_{n,t}$ is measured as

$$\beta^i_{n,t} = \frac{1}{\frac{1}{\beta + 1} \cdot \frac{1}{d^i_\omega} + \frac{\beta}{p + 1} \cdot \frac{1}{o^i_{n,t}}},$$

where $\beta$ is the trainable tradeoff parameter. Furthermore, partly following the manner of [28], we define attention weights for $\beta$-Attn as

$$\theta^i_{n,t} = \tanh \left( \frac{(W^q_n \cdot h^i_{n,T})^T \cdot W_n^k \cdot h^i_{n,t}}{y^i_n \cdot \log(c + (1 - \sigma((W^q_n \cdot h^i_{n,T})^T \cdot W_n^k \cdot h^i_{n,t}))) \cdot \beta^i_{n,t} \cdot T} \right),$$

where $y^i_n$ is also a parameter that needs to be learned, $c$ represents a constant, and $W^q_n \in \mathbb{R}^{d_q \times d_i}$, $W_n^k \in \mathbb{R}^{d_k \times d_i}$ are projection matrices to map the query and key vectors in key-query attention, respectively. Finally, the normalized attention weights can be obtained by

$$\alpha^i_n = [\alpha^i_{n,1}, \alpha^i_{n,2}, \ldots, \alpha^i_{n,T}] \approx \text{softmax} \left( \theta^i_{n,1}, \theta^i_{n,2}, \ldots, \theta^i_{n,T} \right).$$

Based on Equation (8), the weighted channel-wise representation $E_d(S_n^i) \in \mathbb{R}^{d_i}$ for the channel signal $S_n^i$ can be calculated by

$$E_d \left( S_n^i \right) = h^i_n \cdot (\alpha^i_n)^T,$$

where $h^i_n = [h^i_{n,1}, h^i_{n,2}, \ldots, h^i_{n,T}] \in \mathbb{R}^{d_i \times T}$.

### 4.3 Heterogeneous Information Aggregation

Let $E_b(f^i_b)$ be the embedding for the static information $f^i_b$ of the $i$th instance, and then, both the $E_b(f^i_b) \in \mathbb{R}^{d_b}$ and the GGE $E_g(g^i) \in \mathbb{R}^{d_g}$ are concatenated to obtain a fused instance level representation $G_i \in \mathbb{R}^{d_i}$ via a linear Fusenet (a one-layer MLP). Furthermore, by stacking the instance level representation $G_i$ and the channel-level representations $E^i_d = [E_d(S_n^i)]_{n=1,\ldots,N_d} \in \mathbb{R}^{d_i \times N_d}$, we have hierarchical representations $E_i = [E^i_d, G^i] \in \mathbb{R}^{d_i \times (N_d+1)}$ for the $i$th instance.

To capture the interdependencies among these hierarchical representations learned from both instance level and channel level, a strategy of multi-head attention [37] is adopted. To be specific,
given the hierarchical stacked representations $E_i$ as the input, let $\text{head}_i^{(h)}$ be the embedding through the $h$th attention head given by

$$\text{head}_i^{(h)}(E_i) = \text{softmax}\left(\frac{Q_i^{(h)} \cdot (K_i^{(h)})^T}{\sqrt{d_i}}\right) \cdot V_i^{(h)},$$  \hspace{1cm} (10)

where $\{Q_i^{(h)}, K_i^{(h)}, V_i^{(h)}\} \in \mathbb{R}^{(N_d+1) \times d_i}$ are the query, key, and value matrices, respectively, for the $h$th head, and $\{W_q^{(h)}, W_k^{(h)}, W_v^{(h)}\} \in \mathbb{R}^{d_i \times d_i}$ are the corresponding projection matrices. To further integrate the multi-head embeddings $\text{head}_i^{(h)}(E_i), h = 1, \ldots, N_H$, we have:

$$H^i = \text{MultiHead}(E_i)$$

$$= (W^H)^T \cdot \text{Concat}\left(\text{head}_1^{(1)}(E_i), \ldots, \text{head}_{N_H}^{(N_H)}(E_i)\right),$$  \hspace{1cm} (12)

where $H^i \in \mathbb{R}^{d_i \times (N_d+1)}$ and $W^H \in \mathbb{R}^{(d_i \times N_d+1) \times d_i}$ is a linear projection matrix.

Clearly, as a hierarchical description of the instance $i$, the first $N_d$ columns of $H^i$ reflect indeed the characterization of instance $i$ at the channel level, whereas the last one $H_{N_d+1}^i$ reflects the global view on it at the instance level. Intuitively, in a real risk prediction task, the channel level characterizations of dynamic statuses should be consistent as far as possible with the global view with embedded static information and TCG. It means we can obtain a unified representation with multi-granularity information via the guidance of global view. Specifically, we have the unified representation $H_{rep}^i \in \mathbb{R}^{d_i}$ as follows:

$$H_{rep}^i = \sum_{n=1}^{N_d} \underbrace{\mu_n \cdot H_n^i}_{\text{channel-level}} + \underbrace{\mu_{N_d+1} \cdot H_{N_d+1}^i}_{\text{instance-level}},$$  \hspace{1cm} (13)

where $\{\mu_m = \text{softmax}([{(H_m^i)^T \cdot H_{N_d+1}^i}]_{m=1,\ldots,N_d+1})\}$ denote the global view guided weights on representations at both channel level and instance level. In fact, it is not hard to see, the more relevant the channel level characterization to the global guidance at the instance level, the larger the weight will be confirmed. Finally, the risk probability $\hat{y}_i$ can be obtained via an MLP predictor:

$$\hat{y}_i = \text{MLP}\left(H_{rep}^i\right),$$  \hspace{1cm} (14)

The pipeline of the heterogeneous information aggregation is given in Figure 3.

### 4.4 Model Optimization

#### 4.4.1 Model Regularization

With the practical experience [7, 28], the DeCov loss is widely introduced for obtaining non-redundant representations by minimizing the cross-covariance of hidden activation [8]. Specifically, to capture the coupling among multi-head attention layers, the covariances between all pairs of activation $i$ and $j$ from matrix $C$ can be calculated as

$$C_{i,j} = \frac{1}{B} \sum_b \left( H_b^i - \frac{1}{B} \sum_b H_b^i \right) \left( H_b^j - \frac{1}{B} \sum_b H_b^j \right)^T,$$  \hspace{1cm} (15)

where $B$ is the batch size, $H_b^i$ the $n$th refined representation for the $b$th case in the batch (for convenience, $n$ is not distinguished here). To capture similar dependencies among different heads, we need to minimize covariance by penalizing the norm of $C$. However, the diagonal of $C$ should
ALGORITHM 1: Hierarchical Global Views-guided Sequence Representation Learning: \( \tilde{y}_i = \text{HGV}(U, S, F_b, y, T) \)

Require:

\( U \): User/patient set, \( S \): Dynamic information tensor, \( F_b \): Static information matrix, \( y \): Historical risk probability vector, \( T \): Time steps

Ensure:

\( \tilde{y}_i \): Risk probability for user/patient \( u_i \)

1: while \( u_i \in U \) do
2:     for \( S^i \in S \) do
3:         \( g_i \leftarrow \text{TemporalCorrelationGraph}(S^i, T) \)
4:         \( E_d(g_i) \leftarrow \text{GlobalGraphEmbedin}(g_i) \)
5:     end for
6:     \( E_b(f^i) \leftarrow \text{EmbeddingNet}(f^i) \)
7:     \( G^i \leftarrow \text{FuseNet}(\text{Concat}(E_d(g_i), E_b(f^i))) \)
8:     for \( n \in \{1, 2, \ldots, N_d\} \) do
9:         \( h^i_n \leftarrow \text{LSTMs}(S^i_n, T) \) //Equation(3)
10:        \( a^i_n \leftarrow \text{Softmax}(\beta \cdot \text{Attn}(S^i_n, T)) \) //Equation(7-8)
11:        \( E_d(S^i_n) \leftarrow h^i_n \cdot (a^i_n)^T \) //Equation(9)
12:     end for
13:     \( H^i \leftarrow \text{MutilHead}([E^i_d, G^i]) \) //Equation(12)
14: \( \tilde{y}_i \leftarrow F(H^i) \) //Equation(13-14)
15: end while

not be minimized as well as the norm due to the fact that it measures dynamic range of activation, which has nothing to do with our purpose. Thus, the regularization loss can be defined as

\[
L_{\text{DeCov}} = \frac{1}{2} \left( \|C\|_F^2 - \|\text{diag}(C)\|_2^2 \right),
\]

where \( \| \cdot \|_F \) is the Frobenius norm and the \( \text{diag}(\cdot) \) is the operator of extracting a vector from the main diagonal of a matrix. Moreover, to avoid over-fitting, we have also deployed some strategies such as residual connection and dropout operation [15].

4.4.2 Risk Prediction. Generally, the risk prediction task is defined as a binary classification problem, where an observed risky endpoint status is assigned a target value 1, otherwise 0. Specifically, we use the cross entropy as the loss functions for classification,

\[
L_C = - \left( \sum_{y_i \in R^+} \log \tilde{y}_i + \sum_{y_i \in R^-} \log (1 - \tilde{y}_i) \right),
\]

where \( R^+ \) and \( R^- \) are the positive and negative class sets, respectively.
Table 2. Detailed Statistics of MIMIC-III and Ant Group-MYBank Datasets

| Data Sources     | MIMIC-III | Ant Group-MYBank |
|------------------|-----------|------------------|
| #patients/users  | 18,094    | 7,947            |
| #time steps T    | 48        | 14               |
| #Num. of static info. N_b | 7        | 9                |
| #Num. of dyn. info. N_d  | 17      | 26               |
| #sparsity        | 0.6138    | 0.9189           |
| #train samples   | 14,681    | 5,564            |
| #valid samples   | 3,222     | 795              |
| #test samples    | 3,236     | 1,588            |

(sparsity is the proportion of positive samples, indicating the imbalanced and skewed level of the dataset).

in which $R^+$ and $R^-$ are the records with positive and negative target values. Finally, to both capture the coupling among multi-head attention layers and improve the performance of risk prediction, the task can be solved by minimizing the following hybrid loss

$$
L = L_C + \lambda_d \cdot L_{DeCov},
$$

where $\lambda_d$ is a tradeoff parameter.

In summary, the implementation details about the proposed HGV are outlined in Algorithm 1.

5 EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we aim at answering the following research questions:

— RQ1: How to set the risk prediction on both healthcare and financial tasks?
— RQ2: How to evaluate the performance of the HGV?
— RQ3: How does the HGV perform and why?

5.1 Data Description and Settings (RQ1)

5.1.1 Risk Prediction on Healthcare Data. Our healthcare risk prediction task is based on MIMIC-III [19], a large publicly available benchmarking dataset. In this task, the mortality risk needs to predict is a primary outcome of interest in acute care, which is the key to improving outcomes for those at-risk patients. Specifically, following the preprocessing pipeline established by the [14], we split the MIMIC-III dataset into train, val and test set, and the detailed statistics has been summarized in Table 2.

— MIMIC-III Dataset: Medical Information Mart for Intensive Care is a large, single-center database comprising information relating to patients admitted to critical care units at a large

2https://mimic.physionet.org/.
tertiary care hospital. Data includes vital signs, medications, laboratory measurements, observations, and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and more. Noted that one patient may have more than one record.

5.1.2 Risk Prediction on Financial Data. The detailed statistics for the real-world industrial dataset are also given in Table 2. Noted, as a SMEs credit overdue risk prediction task, it is more imbalanced and skewed than MIMIC-III, which brings more challenges.

— **Ant Group-MYBank Dataset:** It includes rich personal profiles and loaning behavior data of SMEs owners of MYBank such as the age, gender, education and loan amount, the current balance, the duration time to the latest loan, the number of loans and so on. We randomly collect about 484,828 traffic logs across one month (e.g., from Dec. 1, 2021 to Jan. 1, 2022) from an online SMEs loan scenario in MYBank, Ant Group. We gather the repayment feedback of each SME owner as the target risk status. Note that all above data are definitely authorized by the SME users since they hope to apply for a loan in our bank, and they should provide their lending history and personal profiles.

5.2 Experimental Settings (RQ2)

5.2.1 Evaluation Metrics. We use **Area Under the Receiver Operating Characteristic (AUROC)** curve, **Area Under the Precision-Recall Curve (AUPRC)**, and the Minimum of Precision and Sensitivity Min(Se, P+) to evaluate the performance of the proposed model on both healthcare risk prediction task and financial credit overdue risk prediction task. In fact, it is a widely used evaluation metric group in this kind of binary risk prediction problem with such imbalanced and skewed data.

5.2.2 Baselines. We compare our proposed HGV with both traditional [12, 45], time-aware [2, 6], attention-based [4, 22, 28, 34, 37], and knowledge-enhanced [46] baselines. It is worth noting that, in order to ensure the fairness of the comparison, we quote the reported results or reproduced results for each baseline from their original literature or official open-source implementations and all the hyper-parameter settings in reproducing experiments are fine-tuned by grid-searching strategy.

— **LR** [45]: It is a classic logistic regression model. We use a more elaborate version of the hand-engineered features given in benchmark pipelines [14].

— **GBDT** [12]: It is also a classic tree-based ensemble learning method. The hand-engineered features used are the same as the LR baseline.

— **Attn-GRU** [4]: It is an attention-based model, where we add the plain attention to a multi-channel GRU.

— **RETAI** [6]: It is a well-known attention-based model, which can explore both temporal relationships and variable significance by using two-level neural attention model.

— **T-LSTM** [2]: It is a time-aware model, it can handle data with time irregularities. In this article, it is modified into a supervised learning model.

— **MCA-RNN** [22]: It is an attention-based model, which utilizes an attention-based RNN and a conditional deep generative model for capturing the heterogeneity in time series data.

— **Transformer** [37]: It is a well-known baseline self-attention-based model. A flatten layer and FFNs in the final step are used to make the risk prediction.
— **SAnD** [34]: It is a self-attention-based model, which applies a masked, self-attention mechanism, and uses positional encoding and dense interpolation strategies for incorporating temporal order.

— **ConCare** [28]: It is one of the state-of-the-art models in risk prediction task, which combine a new time-aware attention and multi-head self-attention networks with a cross-head decorrelation loss.

— **GRASP** [46]: It is a knowledge-enhanced predictor, which extracts a k-nearest neighbor graph clustering from similar samples, and then enhances the performance by introducing cluster centers embedding learned by GCNs [20]. Noted, we use the reported SOTA backbone, GRASP+ConCare, as the baseline.

5.2.3 Parameter Settings. There are some training parameters involved in HGV, i.e., learning rate $lr$ and batch size $B$. In particular, for the batch size $B$ and learning rate $lr$, we set $B = 256, lr = 0.001$ for MIMIC-III and $B = 128, lr = 0.001$ for Ant Group-MYBank dataset, respectively. In addition, there are also some other hyperparameters in the backbone modules, LSTMs and CNNs, i.e., the hidden size $d_1$, layer number $L$, channel number $\lambda$, and kernel size $C_k$ of the layer $l$, stride $C_s$ for the convolution operation in the CNN. For these hyperparameters, we set $L_{\text{LSTM}} = 1$, $L_{\text{CNN}} = 2$, $\lambda_{l=1} = 64$, $\lambda_{l=2} = 128$, $C_k=3$, and $C_s=1$. Moreover, for the hyperparameters in multi-head attention networks and $\beta$-Attn, i.e., the number of head $N_H$, hidden size for two attention layers $d_1, d_2$. Taking both the efficiency and performance into account, the settings for these hyperparameters are: $d_1 = 64$, $d_2 = 32$, and $N_H = 4$. Note that, in order to guarantee the optimal parameters in experiments, we conduct grid searches and set the optimal hyperparameters for both our model and other competitors. As for the computational complexity in our proposed framework, it is mainly embodied in the construction process of the TCG, compared with that in existing typical end-to-end sequence representation networks. In fact, the complexity of this processing module is linearly correlated with the size of the dataset, and it only needs to be pre-computed once when the implementation runs for the first time. More implementation details for the HGV can be referred to the open source code, which has been available at the GitHub repository.4

5.3 Experimental Results and Analysis (RQ3)

5.3.1 Performance Comparison. Table 3 and Figure 4 have shown the experimental results of the HGV and other baselines in two real-world risk prediction tasks. Overall, when compared with all the other methods in performance testing, our proposed HGV consistently achieves the best performance in both tasks.

Specifically, as we can see, although with good interpretability, both classic machine learning methods and plain RNN-based ones [12, 45] perform poorly. Meanwhile, we also find our model outperforms the time-aware methods [2, 6], which shows that it is not sufficient to only consider the effect of time-aware decay. Furthermore, compared to the HGV, the attention-based baselines [4, 22, 28, 34, 37] have also shown insufficient performance due to a lack of capturing clip-aware patterns encoded in temporal status correlation. Faced with the challenge of inevitable noise, the performance of the knowledge-enhanced predictor GRASP [46] is still unsatisfactory.

Furthermore, to evaluate the performance of the HGV on a more imbalanced and skewed risk prediction task (sparsity is 0.6138 on MIMIC-III but 0.9189 on Ant Group-MYBank), we conduct the financial credit overdue risk prediction experiment on a real-world industrial scenario from MYBank, Ant Group. After analyzing the results on the public benchmark dataset, we select the

4https://github.com/LiYouru0228/HGV.
In addition, we also conduct the ablation studies on the benchmark task to demonstrate the effectiveness of heterogeneous information aggregation. To be more specific, we can find:

To demonstrate the effectiveness of considering time-decay during the representation learning with irregular time series data, we replace the \( \beta \)-Attn with a plain attention module.

To demonstrate the effectiveness of making a global tradeoff between time-aware decay and observation significance in sequence representation learning, we replace the \( \beta \)-Attn with a time-decay attention module.

We also remove the GGE to demonstrate the usefulness of mining heterogeneous correlation beyond homogeneous sequences by constructing the TCG.

We also remove the DeCov loss constrained multi-head attention networks to demonstrate the effectiveness of heterogeneous information aggregation.

The results given in Figure 7 have shown that all variants of HGV perform worse than the original HGV, proving its effectiveness in each module. To be more specific, we can find:

Table 3. Performance Comparison on Public Benchmark Healthcare Risk Prediction Task (The Best is in red, the Second is in blue, the Third is in green and the Improvements are in Brackets)

| Model                   | MIMIC-III Dataset (Bootstrapping = 1,000) |
|-------------------------|------------------------------------------|
|                         | AUROC        | AUPRC        | min(Se, P+)           |
| LR [45]                 | 0.8485 ± 0.010 (−2.33%) | 0.4758 ± 0.028 (−6.28%) | 0.4643 ± 0.022 (−5.66%) |
| GBDT [12]               | 0.8468 ± 0.011 (−2.50%) | 0.5032 ± 0.027 (−3.54%) | 0.4916 ± 0.022 (−2.93%) |
| Attn-GRU [4]            | 0.8628 ± 0.011 (−0.90%) | 0.4989 ± 0.022 (−3.97%) | 0.5026 ± 0.028 (−1.83%) |
| RETAIN [6]              | 0.8313 ± 0.014 (−4.05%) | 0.4790 ± 0.020 (−5.96%) | 0.4721 ± 0.022 (−4.88%) |
| T-LSTM [2]              | 0.8628 ± 0.011 (−0.90%) | 0.4989 ± 0.022 (−3.97%) | 0.5026 ± 0.028 (−1.83%) |
| MCA-RNN [22]           | 0.8587 ± 0.013 (−1.31%) | 0.5003 ± 0.028 (−3.83%) | 0.4932 ± 0.024 (−2.77%) |
| Transformer [37]       | 0.8535 ± 0.014 (−1.83%) | 0.4917 ± 0.022 (−4.69%) | 0.5000 ± 0.019 (−2.09%) |
| SAnd [34]              | 0.8382 ± 0.007 (−3.36%) | 0.4545 ± 0.018 (−8.41%) | 0.4885 ± 0.017 (−3.24%) |
| ConCare [28]           | 0.8659 ± 0.009 (−0.59%) | 0.5238 ± 0.027 (−1.48%) | 0.5077 ± 0.022 (−1.32%) |
| GRASP [46]             | 0.8635 ± 0.009 (−0.83%) | 0.5246 ± 0.028 (−1.40%) | 0.5068 ± 0.028 (−1.41%) |
| HGV                     | 0.8718 ± 0.010 | 0.5386 ± 0.028 | 0.5209 ± 0.023 |

Overall Performance of Healthcare Risk Prediction in Benchmark Task.

The experimental results in Figure 4 show the AUROC, AUPRC, and min(Se, P+) on test set. The proposed model, HGV, outperforms the baselines, and has an average improvement rate of 1.55%, 0.19% on AUROC, 1.88%, 0.92% on AUPRC and 4.20%, 2.96% on min(Se, P+), respectively, compared with ConCare and GRASP. Therefore, we can find that the HGV can still outperform other baselines even in the risk prediction task with more sparsity challenges.

5.3.2 Parameter Sensitivity Analysis. To show how the main hyper-parameters involved in HGV affect the model performance, we check the sensitivity of some hyper-parameters, the embedding sizes \( d_1, d_2 \) and the number of head \( N_H \). Figure 5 show the performance under different hyper-parameter combinations of the proposed HGV model for MIMIC-III dataset. We can see that with a consideration of both efficiency and performance, a relatively smaller number of head (not too small) and larger embedding size (not too large) for the hyper-parameter combinations settings leads to the best result. Meanwhile, to evaluate the impact of coefficient \( \lambda_d \), we vary it in \{1e0, 1e1, 1e2, 1e3, 1e4\}. According to the results in Figure 6, the overall performance remains relatively stable and achieves the best performance when \( \lambda_d = 1e3 \).

5.3.3 Ablation Studies. In addition, we also conduct the ablation studies on the benchmark dataset with the setting listed in Table 4, as follows:

1. **HGV (w/o \( \beta \)-Attn)\(_\alpha\)**: To demonstrate the effectiveness of considering time-decay during the representation learning with irregular time series data, we replace the \( \beta \)-Attn with a plain attention module.
2. **HGV (w/o \( \beta \)-Attn)\(_\beta\)**: To demonstrate the effectiveness of making a global tradeoff between time-aware decay and observation significance in sequence representation learning, we replace the \( \beta \)-Attn with a time-decay attention module.
3. **HGV (w/o GGE)**: We also remove the GGE to demonstrate the usefulness of mining heterogeneous correlation beyond homogeneous sequences by constructing the TCG.
4. **HGV (w/o HIA)**: We also remove the DeCov loss constrained multi-head attention networks to demonstrate the effectiveness of heterogeneous information aggregation.
When we did not consider time-decay during the representation learning with irregular time series data, the performance reduced significantly. This clearly shows the necessity of addressing the irregularity of time intervals in historical sequence.

Moreover, we also find that the performance still has room for improvement when we further consider making a global tradeoff between time-aware decay and observation significance. Specifically, in addition to the overall improvement, it will show more explicit outcomes at the case level in the next section.
From the overall performance results, we can see that the GGE is also an important contributor to the proposed HGV framework by effectively modeling the correlations between static and dynamic information with the guidance of global view.

In addition, the removal of the heterogeneous information aggregation module has implied that aggregating the channel-wise representations and instance-wise representations together in a union hierarchical representation learning framework can also significantly affect the model performance.
### Table 4. Detailed Settings for Ablation Studies

| Variants                  | \(\beta\text{-Attn}_a\) | \(\beta\text{-Attn}_b\) | GGE   | HIA   |
|---------------------------|---------------------------|---------------------------|-------|-------|
| HGV (w/o \(\beta\text{-Attn}_a\)) | \(\checkmark\)           |                           | \(\checkmark\) | \(\checkmark\) |
| HGV (w/o \(\beta\text{-Attn}_b\)) |                           | \(\checkmark\)           | \(\checkmark\) | \(\checkmark\) |
| HGV (w/o GGE)             | \(\checkmark\)           | \(\checkmark\)           | \(\checkmark\) |       |
| HGV (w/o HIA)             | \(\checkmark\)           | \(\checkmark\)           |       | \(\times\) |

5.3.4 **Case Study.** To intuitively demonstrate how the main modules, GGE, \(\beta\text{-Attn}\), and heterogeneous information aggregation work in the proposed HGV framework, we give the intuitive case analysis on a randomly selected samples from the MIMIC-III dataset in Figures 8–9.

Specifically, from the left sub-figure in Figure 8, it is easy to see that there are some local bright blocks in the TCG, which shows the clip-aware correlation among different statuses, and figuratively indicates the necessity of extracting such information from time series data. For the second sub-figure, we can also find that the \(\beta\text{-Attn}\) have learned how to make a global tradeoff between time-aware decay and observation significance for weighting the time series, which is significantly distinct from the smooth attention distribution learned by only considering time-aware decay.

Furthermore, in Figure 9, we can find that the heterogeneous information aggregation module can assign different attention weights to each patient for aggregating the channel-wise representations with the hierarchical guidance on two global views.
6 CONCLUSION AND FUTURE WORK

This article proposed a novel end-to-end HGV sequence representation learning framework to predict risk in both healthcare and finance. Specifically, to joint learn hierarchical representations from heterogeneous data, the GGE has achieved to reveal the temporal rhythmic variation of the observed status and the $\beta$-Attn has learned a global tradeoff between time-aware decay and observation significance. In addition, we have conducted experiments on two real-world risk prediction tasks and evaluated the performance of the HGV.

For future work, we will further explore incorporating more explicit prior information into such risk prediction modeling tasks, especially considering the introduction of knowledge graphs into an end-to-end deep learning framework to further improve model interpretability.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015.

[2] Inci M. Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K. Jain, and Jiayu Zhou. 2017. Patient subtyping via time-aware LSTM networks. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017. ACM, 65–74. DOI: https://doi.org/10.1145/3097983.3097997

[3] Chaochao Chen, Jun Zhou, Li Wang, Xibin Wu, Wenjing Fang, Jin Tan, Lei Wang, Alex X. Liu, Hao Wang, and Cheng Hong. 2021. When homomorphic encryption marries secret sharing: Secure large-scale sparse logistic regression and applications in risk control. In Proceedings of the KDD’21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2021. ACM, 2652–2662. DOI: https://doi.org/10.1145/3447548.3467210

[4] Kyunghyun Cho, Bart van Merrienboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine
translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2014. ACL, 1724–1734. DOI: https://doi.org/10.3115/v1/d14-1179

[5] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. 2017. GRAM: Graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017. ACM, 787–795. DOI: https://doi.org/10.1145/3097903.3098126

[6] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter F. Stewart. 2016. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Proceedings of the Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems* 2016. 3504–3512. Retrieved from https://proceedings.neurips.cc/paper/2016/hash/231141b34c82aa95e48810a9d1b33a79-Abstract.html.

[7] Xu Chu, Yang Lin, Yashar Wang, Leye Wang, Jiangtao Wang, and Jingyue Gao. 2019. MLRDA: A multi-task semi-supervised learning framework for drug-drug interaction prediction. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019*. ijcai.org, 4518–4524. DOI: https://doi.org/10.24963/ijcai.2019/628

[8] Michael Cogswell, Faruk Ahmed, Ross B. Girshick, Larry Zitnick, and Dhruv Batra. 2016. Resolving ambiguity in deep networks by decorrelating representations. In *Proceedings of the 4th International Conference on Learning Representations, ICLR 2016*.

[9] Antonio Coronato and Alfredo Cuzzocrea. 2022. An innovative risk assessment methodology for medical information systems. *IEEE Transactions on Knowledge and Data Engineering* 34, 7 (2022), 3095–3110. DOI: https://doi.org/10.1109/TKDE.2020.3023535

[10] Mark E. Daly, Catherine Vale, Mark Walker, Alison Littlefield, K. George Alberti, and John C. Mathers. 1998. Acute effects on insulin sensitivity and diurnal metabolic profiles of a high-sucrose compared with a high-starch diet. *The American Journal of Clinical Nutrition* 67, 6 (1998), 1186–1196.

[11] Nino Fijacko, Ruth Masterson Creber, Lucija Gosak, Primoz Kocbek, Leona Cilar, Peter Creber, and Gregor Stiglic. 2021. A review of mortality risk prediction models in smartphone applications. *Journal of Medical Systems* 45, 12 (2021), 107. DOI: https://doi.org/10.1007/s10916-021-01776-x

[12] Jerome H. Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 29, 5 (2001), 1189–1232.

[13] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, AISTATS 2011*. JMLR.org, 315–323. Retrieved from http://proceedings.mlr.press/v15/glorot11a/glorot11a.pdf.

[14] Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific Data* 6, 1 (2019), 96. DOI: https://doi.org/10.1038/s41597-019-0103-9

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*. IEEE Computer Society, 770–778. DOI: https://doi.org/10.1109/CVPR.2016.90

[16] Sepp Hochreiter and Jürgen Schmidhuber. 1996. LSTM can solve hard long time lag problems. In *Proceedings of the Advances in Neural Information Processing Systems 9, NIPS, 1996*. MIT, 473–479. Retrieved from http://papers.nips.cc/paper/1215-lstm-can-solve-hard-long-time-lag-problems.

[17] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780. DOI: https://doi.org/10.1162/neco.1997.9.8.1735

[18] Chao Huang, Jiahui Chen, Lianghao Xia, Yong Xu, Peng Dai, Yanqing Chen, and Jimmy Xiangji Huang. 2021. Graph-enhanced multi-task learning of multi-level transition dynamics for session-based recommendations. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence, AAAI 2021*. AAAI, 4123–4130. Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/16534.

[19] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, H. Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care data base. *Scientific Data* 3, 1 (2016), 1–9.

[20] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations, ICLR 2017*. OpenReview.net. Retrieved from https://openreview.net/forum?id=SJU4ayY1gl.

[21] Bum Chul Kwon, Min-Je Choi, Joanne Taery Kim, Edward Choi, Young Bin Kim, Soonwook Kwon, Jimeng Sun, and Jaegul Choo. 2019. RetainVis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 299–309. DOI: https://doi.org/10.1109/TVCG.2018.2865028

[22] Wonsung Lee, Sungrae Park, Weonyoung Joo, and Il-Chul Moon. 2018. Diagnosis prediction via medical context attention networks using deep generative modeling. In *Proceedings of the IEEE International Conference on Data Mining, ICDM 2018*. IEEE Computer Society, 1104–1109. DOI: https://doi.org/10.1109/ICDM.2018.00143

ACM Transactions on Knowledge Discovery from Data, Vol. 18, No. 1, Article 1. Publication date: August 2023.
[23] Wu Lee, Yuliang Shi, Hongfeng Sun, Lin Cheng, Kun Zhang, Xinzun Wang, and Zhiyong Chen. 2022. MSIPA: Multiscale interval pattern-aware network for ICU transfer prediction. ACM Transactions on Knowledge Discovery from Data 16, 1 (2022), 17:1–17:17. DOI: https://doi.org/10.1145/3458284

[24] Jianzheng Li, Linyi Yang, Barry Smyth, and Ruihai Dong. 2020. MAEC: A multimodal aligned earnings conference call dataset for financial risk prediction. In Proceedings of the CIKM’20: The 29th ACM International Conference on Information and Knowledge Management, 2020. ACM, 3063–3070. DOI: https://doi.org/10.1145/340531.3412879

[25] Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. 2020. HitANet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In Proceedings of the KDD ’20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2020. ACM, 647–656. DOI: https://doi.org/10.1145/3394486.3403107

[26] Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. 2018. KAME: Knowledge-based attention model for diagnosis prediction in healthcare. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018. ACM, 743–752. DOI: https://doi.org/10.1145/3269206.3271701

[27] Liantao Ma, Junyi Gao, Yasha Wang, Chaohue Zhang, Jiangtao Wang, Wenjie Ruan, Wen Tang, Xin Gao, and Xinyu Ma. 2020. AdaCare: Explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration. In Proceedings of the 34th AAAI Conference on Artificial Intelligence, AAAI 2020. AAAI, 825–832. Retrieved from https://aaaai.org/ojs/index.php/AAAI/article/view/5427.

[28] Liantao Ma, Junyi Gao, Yasha Wang, Chaohue Zhang, Jiangtao Wang, Wenjie Ruan, Wen Tang, Xinyu Ma, Xin Gao, and Junyi Gao. 2020. ConCare: Personalized clinical feature embedding via capturing the healthcare context. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020. AAAI, 833–840. Retrieved from https://aaaai.org/ojs/index.php/AAAI/article/view/5428.

[29] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent models of visual attention. In Proceedings of the Annual Conference on Neural Information Processing Systems 2014. 2204–2212. Retrieved from https://proceedings.neurips.cc/paper/2014/hash/09c6c3783b4a70054da74f2538ed47c6-Abstract.html.

[30] Zhiqiang Pan, Fei Cai, Wanyu Chen, Changhao Chen, and Honghui Chen. 2022. Collaborative graph learning for session-based recommendation. ACM Transactions on Information Systems 40, 4 (2022), 72:1–72:26. DOI: https://doi.org/10.1145/34900479

[31] Byung Sam Park, Ji Sung Yoon, Jun Sung Moon, Kyu Chang Won, and Hyoung Woo Lee. 2013. Predicting mortality of critically ill patients by blood glucose levels. Diabetes and Metabolism Journal 37, 5 (2013), 385–390.

[32] David Powers. 2011. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies 2, 1 (2011), 37–63.

[33] Yao Qin, Dongqin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison W. Cottrell. 2017. A dual-stage attention-based recurrent neural network for time series prediction. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI 2017. ijcai.org, 2627–2633. DOI: https://doi.org/10.24963/ijcai.2017/366

[34] Huan Song, Deepta Rajan, Jayaraman J. Thiagarajan, and Andreas Spanias. 2018. Attend and diagnose: Clinical time series analysis using attention models. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence, (AAAI-18). AAAI, 4091–4098. Retrieved from https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16325.

[35] Liuhong Song, Chin Wang Cheong, Kejing Yin, William K. Cheung, Benjamin C. M. Fung, and Jonathan Poon. 2019. Medical concept embedding with multiple ontological representations. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019. ijcai.org, 4613–4619. DOI: https://doi.org/10.24963/ijcai.2019/641

[36] Sindhu Tipirneni and Chandan K. Reddy. 2022. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. ACM Transactions on Knowledge Discovery from Data 16, 6 (2022), 105:1–105:17. DOI: https://doi.org/10.1145/3516367

[37] Ashish Vaswani, Noam Shazeer, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 2017. 5998–6008. Retrieved from https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[38] Chenyang Wang, Weizhi Ma, Min Zhang, Chong Chen, Yiqun Liu, and Shaoping Ma. 2021. Toward dynamic user intention: Temporal evolutionary effects of item relations in sequential recommendation. ACM Transactions on Information Systems 39, 2 (2021), 16:1–16:33. DOI: https://doi.org/10.1145/3432244

[39] Daixin Wang, Zhiqiang Zhang, Jun Zhou, Peng Cui, Jingli Fang, Quanhui Jia, Yanming Fang, and Yuan Qi. 2021. Temporal-aware graph neural network for credit risk prediction. In Proceedings of the 2021 SIAM International Conference on Data Mining, SDM 2021. SIAM, 702–710. DOI: https://doi.org/10.1137/1.9781611976700.79

[40] Xin Xia, Hongzhi Yin, Junliang Yu, Yingxia Shao, and Lizhen Cui. 2021. Self-supervised graph co-training for session-based recommendation. In Proceedings of the CIKM’21: The 30th ACM International Conference on Information and Knowledge Management, 2021. ACM, 2180–2190. DOI: https://doi.org/10.1145/3459637.3482388

[41] Binqian Xu and Xiangbo Shu. 2023. Pyramid self-attention polymerization learning for semi-supervised skeleton-based action recognition. arXiv:2302.02327. Retrieved from https://arxiv.org/abs/2302.02327.

ACM Transactions on Knowledge Discovery from Data, Vol. 18, No. 1, Article 1. Publication date: August 2023.
[42] Binqian Xu and Xiangbo Shu. 2023. Spatiotemporal decouple-and-squeeze contrastive learning for semisupervised skeleton-based action recognition. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[43] Shuo Yang, Zhiqiang Zhang, Jun Zhou, Yang Wang, Wang Sun, Xingyu Zhong, Yanming Fang, Quan Yu, and Yuan Qi. 2020. Financial risk analysis for SMEs with graph-based supply chain mining. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence, IJCAI 2020*. ijcai.org, 4661–4667. DOI: https://doi.org/10.24963/ijcai.2020/643

[44] Zhen Ye, Yu Qin, and Wei Xu. 2020. Financial risk prediction with multi-round Q&A attention network. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence, IJCAI 2020*. ijcai.org, 4576–4582. DOI: https://doi.org/10.24963/ijcai.2020/631

[45] Hsiang-Fu Yu, Fang-Lan Huang, and Chih-Jen Lin. 2011. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning* 85, 1-2 (2011), 41–75. DOI: https://doi.org/10.1007/s10994-010-5221-8

[46] Chaohe Zhang, Xin Gao, Liantao Ma, Yasha Wang, Jiangtao Wang, and Wen Tang. 2021. GRASP: Generic framework for health status representation learning based on incorporating knowledge from similar patients. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence, AAAI 2021*. AAAI, 715–723. Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/16152.

[47] Ge Zhang, Zhao Li, Jiaming Huang, Jia Wu, Chuan Zhou, Jian Yang, and Jianliang Gao. 2022. eFraudCom: An e-commerce fraud detection system via competitive graph neural networks. *ACM Transactions on Information Systems* 40, 3 (2022), 47:1–47:29. DOI: https://doi.org/10.1145/3474379

[48] Muhan Zhang, Christopher R. King, Michael Avidan, and Yixin Chen. 2020. Hierarchical attention propagation for healthcare representation learning. In *Proceedings of the KDD’20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2020. ACM, 249–256. DOI: https://doi.org/10.1145/3394486.3403067

[49] Wen Zhang, Shaoshan Yan, Jian Li, Xin Tian, and Taketoshi Yoshida. 2022. Credit risk prediction of SMEs in supply chain finance by fusing demographic and behavioral data. *Transportation Research Part E: Logistics and Transportation Review* 158 (2022), 102611.

[50] Yadong Zhu, Xiliang Wang, Qing Li, Tianjun Yao, and Shangsong Liang. 2022. BotSpot++: A hierarchical deep ensemble model for bots install fraud detection in mobile advertising. *ACM Transactions on Information Systems* 40, 3 (2022), 50:1–50:28. DOI: https://doi.org/10.1145/3476107

Received 10 December 2022; revised 15 March 2023; accepted 19 June 2023