BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Med-MGF: multi-level graph-based framework for handling medical data imbalance and representation

Tuong Minh Nguyen[1*], Kim Leng Poh[1], Shu-Ling Chong[2,3] and Jan Hau Lee[3,4]

Abstract

Background Modeling patient data, particularly electronic health records (EHR), is one of the major focuses of machine learning studies in healthcare, as these records provide clinicians with valuable information that can potentially assist them in disease diagnosis and decision-making.

Methods In this study, we present a multi-level graph-based framework called MedMGF, which models both patient medical profiles extracted from EHR data and their relationship network of health profiles in a single architecture. The medical profiles consist of several layers of data embedding derived from interval records obtained during hospitalization, and the patient-patient network is created by measuring the similarities between these profiles. We also propose a modification to the Focal Loss (FL) function to improve classification performance in imbalanced datasets without the need to imputate the data. MedMGF's performance was evaluated against several Graphical Convolutional Network (GCN) baseline models implemented with Binary Cross Entropy (BCE), FL, class balancing parameter α, and Synthetic Minority Oversampling Technique (SMOTE).

Results Our proposed framework achieved high classification performance (AUC: 0.8098, ACC: 0.7503, SEN: 0.8750, SPE: 0.7445, NPV: 0.9923, PPV: 0.1367) on an extreme imbalanced pediatric sepsis dataset (n=3,014, imbalance ratio of 0.047). It yielded a classification improvement of 3.81% for AUC, 15% for SEN compared to the baseline GCN+αFL (AUC: 0.7717, ACC: 0.8144, SEN: 0.7250, SPE: 0.8185, PPV: 0.1559, NPV: 0.9847), and an improvement of 5.88% in AUC and 22.5% compared to GCN+FL+SMOTE (AUC: 0.7510, ACC: 0.8431, SEN: 0.6500, SPE: 0.8520, PPV: 0.1688, NPV: 0.9814). It also showed a classification improvement of 3.86% for AUC, 15% for SEN compared to the baseline GCN+α BCE (AUC: 0.7712, ACC: 0.8133, SEN: 0.7250, SPE: 0.8173, PPV: 0.1551, NPV: 0.9847), and an improvement of 14.33% in AUC and 27.5% in comparison to GCN+BCE+SMOTE (AUC: 0.6665, ACC: 0.7271, SEN: 0.6000, SPE: 0.7329, PPV: 0.0941, NPV: 0.9754).

Conclusion When compared to all baseline models, MedMGF achieved the highest SEN and AUC results, demonstrating the potential for several healthcare applications.

Keywords Pediatric sepsis, Patient network, Graphical models, Message passing, Machine learning

*Correspondence:
Tuong Minh Nguyen
minh.t.nguyen@u.nus.edu
Full list of author information is available at the end of the article

Introduction

Making an accurate medical diagnosis for a patient requires consideration of several aspects of clinical information and evidence. This includes reviewing the patient's medical history, performing physical examinations, ordering tests, interpreting test results, and consulting with other professionals if necessary. The data collected during this process are mainly stored as tabular Electronic health records (EHR) (e.g., vital signs, laboratory results), high-frequency physiologic waveforms (e.g., electrocardiogram), imaging (e.g., radiograph), or other forms of medical data. Using these data, clinicians are able to monitor the patient's disease progression and make informed treatment decisions. As it contains a large volume of rich clinical information, EHR can potentially be used to support clinical research as well [1, 2]. The use of EHR as a data source for Machine learning (ML) studies has increased significantly over the past few years, and modeling EHR data has been one of the major focuses of ML applications in the healthcare sector [3, 4].

The concept of Patient similarity network (PSN) is an emerging research field within the context of precision medicine [5, 6]. The diagnosis made using this network is based on the premise that if patients' medical data are similar in several aspects, then their clinical progress should be similar as well. It is hypothesized that a common disease trajectory resulting in a specific outcome may establish a similarity between patients, thereby making the insight gained using PSN more reliable and robust [7]. Recent advances in ML techniques have led to the development of a variety methods to construct PSN. The International classification of diseases (ICD) was often utilized to establish connections between patients [8, 9]. In some instances, medical inputs are converted into feature vectors and the distance between these vectors will determine the degree of similarity between them [7, 10]. Studies usually treat the medical inputs as a flat structure or embed them within several layers of neural networks without preserving their structure or interpretation. The latter often requires a separate training process to create the medical embedding before they are introduced into the PSN for further training, which could result in an increase in training costs.

In this work, we propose Medical Multilevel Graph-based Framework (MedMGF), a framework that is capable of modeling medical data, as well as representing the patient's individual medical profile and their similarity to other patients within a single architecture. Depending on data availability, the medical profile can be constructed from EHR, physiologic waveforms, imaging data, or a combination thereof. In this study, we demonstrate the feasibility of the framework using EHR data. In contrast to most studies which treat EHR as a flat structure, we preserve its natural hierarchical structure and provide an intuitive way to describe it by incorporating interval data from multiple hospitalizations. A multi-level embedding process allows the medical inputs to pass directly through the PSN, where embedding and PSN are optimized through a single training procedure. We also propose a modification of the Focal Loss (FL, [11]) function to improve classification performance on imbalanced datasets without having to imputate the data, thus reducing the amount of preprocessing needed. In general, MedMGF encapsulates the following characteristics: (1) generality and modality, (2) multi-purpose, (3) intuitive interpretation, and (4) minimal data requirements.

In this study, our objective is to present the framework architecture and feasibility of MedMGF on an imbalanced pediatric sepsis EHR datasets and evaluate its classification performance against several Graph Convolutional Network (GCN, [12]) baselines implemented with Binary Cross Entropy (BCE, [13]), FL, class balancing parameter α, and Synthetic Minority Over-sampling Technique (SMOTE, [14]).

Related works

Electronic health record modeling

The use of ML in modelling EHR has become more prevalent as EHR contains rich clinical information that can potentially assist clinicians in making diagnosis and treatment decisions. Although most studies model the EHR in a flat manner [15, 16], exploring its structural aspects may reveal new possibilities for enhancing the model. In particular, Choi et al. developed Multi-layer Representation Learning for Medical concepts (Med2Vec, [17]), and continue to explore this approach with Graph-based Attention Model (GRAM, [18]) and Multilevel Medical Embedding (MiME, [19]). By leveraging the parent-child relationship on the knowledge-based directed graph, GRAM can learn the representation of medical concepts (e.g., ICD codes) with attention mechanisms and predict the next hospital visit's diagnosis code. Based on GRAM, Li et al. developed Multimodal Diagnosis Prediction model (MDP, [20]) that allows clinical data to be integrated into the framework. Although clinical data from EHR can be weighed dynamically to highlight the most important features, the data is still processed in a flat manner. With MiME, Choi et al. constructed a hierarchical structure of EHR data based on the relationship between symptoms and treatments, where the hospital visit consists of a number of symptoms, each corresponding to a number of specific treatments. This influential interaction is encapsulated in the patient's data embedding representation, which is used for prediction purposes. The precision-recall area under the curve (PR-AUC) for heart failure prediction showed a 15% improvement compared to the baseline model. As most EHR datasets lack the connection between symptoms and treatments, MiME may require extensive data preprocessing, and EHR data may need to undergo a rigorous pre-processing procedure before being mapped to MiME. In addition, the current MiME structure may not capture all aspects of EHR data, other than the relationship between symptoms and treatments. In light of these two drawbacks, we propose MedMGF, a framework for modeling EHR data that can capture all aspects of EHR efficiently and effectively with minimal data preprocessing required.

Patient similarity network

There are several approaches to constructing a PSN using ICD codes [8, 9]. One of the approaches is to create a bipartite graph to connect patients to their corresponding ICDs in a similar manner to what Lu and Uddin did in 2021. This bipartite graph is then converted into a weighted PSN, in which the weight of the edge is determined by the number of mutual ICD codes between the patients [9]. In this approach, the number of mutual ICD codes used to connect patients is highly dependent upon cohort and ICD code selection. In Rouge et al.'s study, an inverse document frequency measured vector of 674 ICD10 codes was constructed for each patient. A cosine similarity between these vectors was calculated for all possible pairs of patients. The PSN was then constructed using a pre-defined threshold on the calculated distances [8]. As the number of patients increases, it becomes more difficult to process the large ICD matrix computationally. In other cases, the medical input is mapped into feature vectors, and distance metrics (e.g. Euclidean, Cosine, Manhattan) are applied to determine the degree of similarity [8, 10]. In the work of Navaz et al., two similarity matrices were calculated separately for static data (e.g. age) and dynamic data (e.g. vital signs). These matrixes were then fused together to construct the PSN [7].

Focal loss function

FL function was first introduced by Lin et al. in 2018 [11]. On the training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_i^N$ independently drawn from an i.i.d probability distribution, the FL for a binary classification problem is defined as follows:

$$\mathcal{L}_{FL}(p_t) = -(1-p_t)^{\gamma} log(p_t) \tag{1}$$

$$p_t = \begin{cases} p, & \text{if } y = 1 \\ 1-p, & \text{otherwise} \end{cases} \tag{2}$$

where p is the predicted probability and $\gamma \geq 0$ is a user-defined hyperparameter to control the rate at which easy samples are down-weighted. It can be observed that FL reduces to the Cross Entropy (CE) when $\gamma = 0$. FL introduces a modulating factor $(1-p_t)^{\gamma}$ to the CE to dynamically adjust the loss based on the difficulty of each sample. This factor is higher for misclassified samples and lower for well-classified samples. Thus, FL reduces the impact of the dominant class by focusing on difficult samples. Researchers typically perform cross-validation to find the optimal value of gamma [11, 21]. In a strategic policy proposed by Mukhoti et al., a higher value of γ would be allocated to predicted probabilities, which is less than a pre-calculated threshold and a lower value of γ for probabilities greater than the threshold [22]. The results of their work showed that the dynamic value of γ could improve FL calibration. In another work, Ghosh et al. proposed to dynamically adjusted γ based on its value from the previous steps [23]. Either way, the classification performance is strongly influenced by and dependent on the value of γ. Considering this dependence, we propose a modification that allows us to dynamically adjust the modulating factor in a similar manner without relying on the hyperparameter γ.

Methods

The framework consists of three main components: the patient's medical profile, which represents the health data extracted from the EHR data, the patient-patient network, which represents the similarity among the patients, based on their profiles, and the modification of FL function. An individual's medical profile is constructed based on a hierarchical representation that embeds several layers of information derived from interval data collected during hospitalizations and medical modules. In this study, we present the medical representation for EHR data. The overall framework is illustrated in Fig. 1. The notation for patient's medical profile and patient-patient network are listed in Tables 1 and 2.

Patient's medical profile

Suppose that an individual's medical profile for a specific disease contains health data from a sequence of hospitalizations $\left(\mathcal{V}^{(1)}, \mathcal{V}^{(2)}, ..., \mathcal{V}^{(i)}, ..., \mathcal{V}^{(T)}\right)$, whereat each hospitalization $\mathcal{V}^{(i)}$, a sequence of medical data $\left(\mathcal{D}^{(1)}, \mathcal{D}^{(2)}, ..., \mathcal{D}^{(j)}, ..., \mathcal{D}^{(t)}\right)$ is entered at time intervals $(\Delta_1, \Delta_2, ..., \Delta_j, ..., \Delta_t)$, with Δ_j being the time interval between $\mathcal{D}^{(j)}$ and $\mathcal{D}^{(j-1)}$. Medical data $\mathcal{D}^{(j)}$ collected at interval j-th includes medical module from EHR data $\mathcal{S}_E^{(j)}$, imaging data $\mathcal{S}_I^{(j)}$, signal data $\mathcal{S}_S^{(j)}$, or a combination thereof, then $\mathcal{D}^{(j)} = \oplus\left(\mathcal{S}_E^{(j)}, \mathcal{S}_I^{(j)}, \mathcal{S}_S^{(j)}, ...\right)$, where $\oplus(.)$ represents the CONCAT data aggregation function. Let $\mathbf{d}^{(j)}$ be the vector representation of $\mathcal{D}^{(j)}$ at j-th interval, $\mathbf{v}^{(i)}$ be a vector representation of i-th

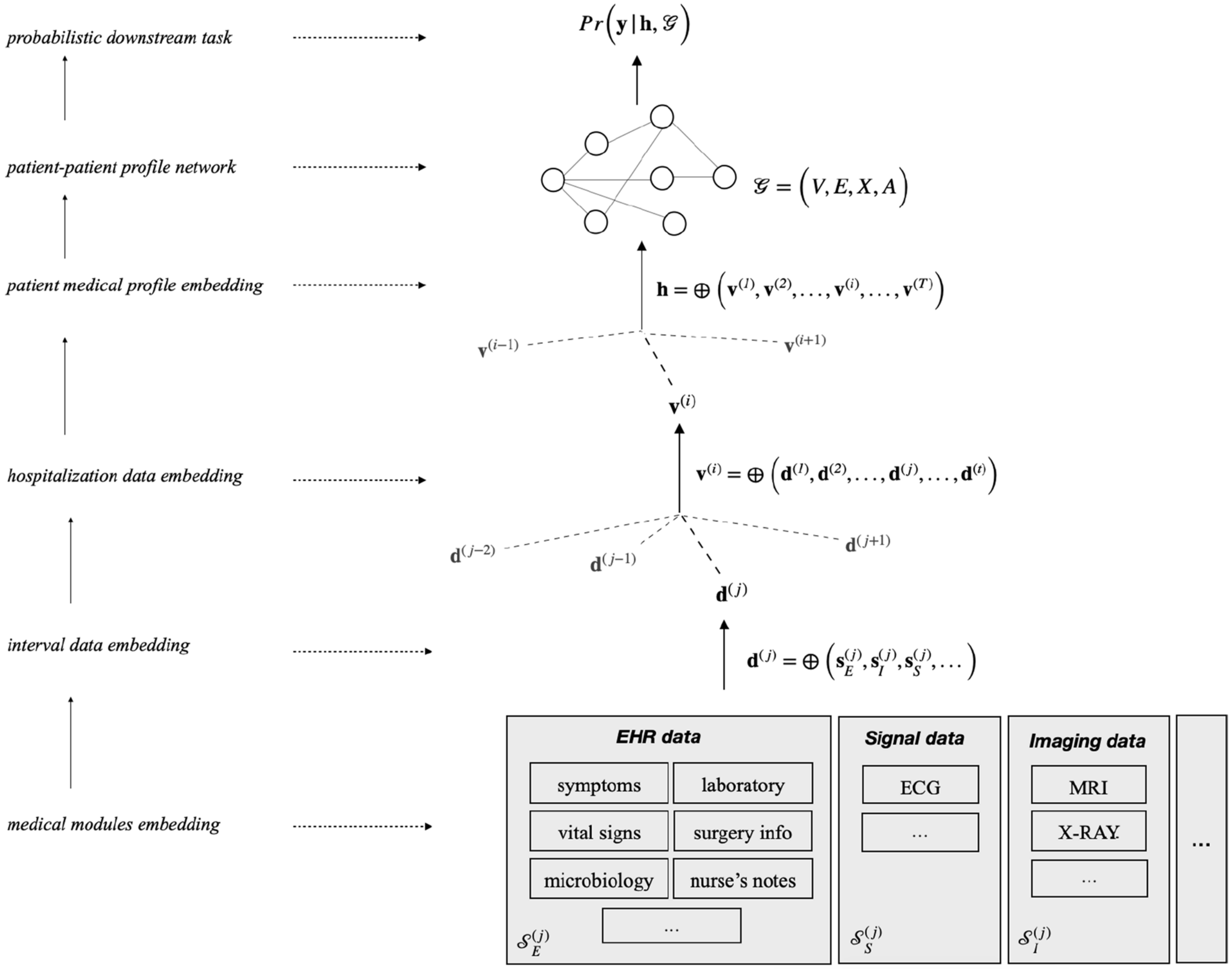

probabilistic downstream task
$Pr(\mathbf{y} \mid \mathbf{h}, \mathcal{G})$
patient-patient profile network
$\mathcal{G} = (V, E, X, A)$
patient medical profile embedding
$\mathbf{h} = \oplus(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(i)}, \ldots, \mathbf{v}^{(T)})$
$\mathbf{v}^{(i-1)}$
$\mathbf{v}^{(i+1)}$
$\mathbf{v}^{(i)}$
hospitalization data embedding
$\mathbf{v}^{(i)} = \oplus(\mathbf{d}^{(1)}, \mathbf{d}^{(2)}, \ldots, \mathbf{d}^{(j)}, \ldots, \mathbf{d}^{(t)})$
$\mathbf{d}^{(j-2)}$
$\mathbf{d}^{(j-1)}$
$\mathbf{d}^{(j+1)}$
$\mathbf{d}^{(j)}$
interval data embedding
$\mathbf{d}^{(j)} = \oplus(\mathbf{s}_E^{(j)}, \mathbf{s}_I^{(j)}, \mathbf{s}_S^{(j)}, \ldots)$
medical modules embedding
EHR data
symptoms
laboratory
vital signs
surgery info
microbiology
nurse's notes
...
$\mathcal{S}_E^{(j)}$
Signal data
ECG
...
$\mathcal{S}_S^{(j)}$
Imaging data
MRI
X-RAY
...
$\mathcal{S}_I^{(j)}$
...

Fig. 1 The MedMGF framework consists of several layers of medical data embedding

Table 1 Notation for patient's medical profile

Notation	Definition
$\mathbf{y}$	A vector represent patient's outcome
$\mathbf{h}$	A vector represent patient's medical profile
$\mathcal{D}^{(j)}$	A patient's medical data collected at *j*-th interval
$\mathcal{V}^{(i)}$	A patient's *i*-th hospitalization
$\mathbf{d}^{(j)} \in \mathbb{R}^z$	A vector representation of $\mathcal{D}^{(j)}$
$\mathbf{v}^{(i)} \in \mathbb{R}^{t\times z}$	A vector representation of $\mathcal{V}^{(i)}$
Δ_j	The time interval between $\mathcal{D}^{(j)}$ and $\mathcal{D}^{(j-1)}$
$\oplus(.)$	An aggregation CONCAT function
$\mathcal{S}_E^{(j)}$	EHR dataset at interval *j*-th
$\mathcal{S}_I^{(j)}$	Signal dataset at interval *j*-th
$\mathcal{S}_S^{(j)}$	Imaging data at interval *j*-th
$\mathbf{s}_E^{(j)}$	A vector represent EHR data at interval *t*-th
$\mathbf{s}_I^{(j)}$	A vector represent Signal data at interval *j*-th
$\mathbf{s}_S^{(j)}$	A vector represent Imaging data at interval *j*-th

hospitalization $\mathcal{V}^{(i)}$, and $\mathbf{s}_E^{(j)}, \mathbf{s}_I^{(j)}, \mathbf{s}_S^{(j)}$ be the vector representation of $\mathcal{S}_E^{(j)}, \mathcal{S}_I^{(j)}, \mathcal{S}_S^{(j)}$, then $\mathbf{d}^{(j)} = \oplus\left(\mathbf{s}_E^{(j)}, \mathbf{s}_I^{(j)}, \mathbf{s}_S^{(j)}, ...\right)$ and $\mathbf{v}^{(i)} = \oplus\left(\mathbf{d}^{(1)}, \mathbf{d}^{(2)}, ..., \mathbf{d}^{(j)}, ..., \mathbf{d}^{(t)}\right) \in \mathbb{R}^{t\times z}$, where z represents the number of the medical modules. We define $\mathbf{h}$ to be the vector presentation of a patient's medical profile, then $\mathbf{h} = \oplus\left(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, ..., \mathbf{v}^{(i)}, ..., \mathbf{v}^{(T)}\right)$.

The interval sequence $(\Delta_1, \Delta_2 ..., \Delta_t)$ represents the irregular periodicity of the hospital data, where Δ_i can vary to match the requirement of the desired analysis. For this study, we fix $\Delta_1 = \Delta_2 = ... = \Delta_t = \Delta$ so that the medical data will be extracted at a fixed interval Δ. Different variables are collected at different intervals, resulting in three possible scenarios: no value is recorded, one value is recorded, or multiple values are recorded. We extract the variables value of an interval as follows: if no data are available for interval *j*-th, the value from the previous interval will be carried forward. If more than one

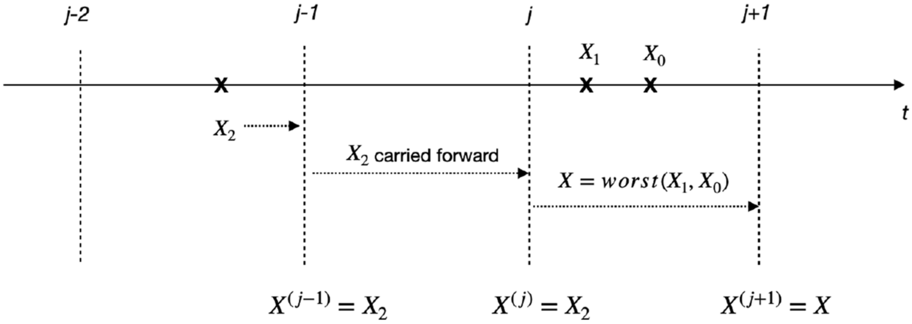


Fig. 2 Data extraction rule at an interval: when no value is available during the interval, the value from previous interval is carried forward. A worst value is selected if more than one value is available in the interval

Table 2 Notation for patient-patient medical profile network

Notation	Definition
$\mathcal{G}$	A graphical network for patients' medical profile
V	A set of nodes, each node represent a patient
E	A set of edges representing the similarity between patients' medical profile
$X \in \mathbb{R}^{N\times T}$	A node feature matrix representing the patients' medical profiles
$A \in \mathbb{R}^{N\times N}$	An adjacency matrix of $\mathcal{G}$
$\mathbf{x}_i$	A vector represent patient's medical profiles. $\mathbf{x}_i = \mathbf{h}_i$
ξ	A predefined similarity threshold to determine the edges of the network
$d(u, v)$	Euclidean distance between node u and v

value is recorded in the interval, the worst value will be taken (Fig. 2).

Patient-patient medical profile network

The patient's medical profile network is defined as a graphical network $\mathcal{G} = (V, E, X, A)$ with $|V| = N$ nodes and $|E|$ edges, where nodes represent patients and the edge weights represent the degree of similarity between them. The node feature matrix $X = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_n) \in \mathbb{R}^{N\times T}$ contains the feature vector of all nodes. A single row $\mathbf{x}_i$ from the node matrix X is a representation of a patient's medical profile from T hospitalizations that has been described in "Patient's medical profile" section. Hence, $\mathbf{x}_i = \mathbf{h}_i = \left\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, ..., \mathbf{v}^{(i)}, ..., \mathbf{v}^{(T)}\right\}$. In order to determine a the similarity between patients, we measure the similarity between their medical profiles. Since the medical profile is represented as a data vector, we can measure the similarity between patient's medical profile by calculating the Euclidean distance between them. Let $u, v \in V$ be the two nodes representing patient u and v on $\mathcal{G}$, the similarity distance $d(u, v)$ is defined as follows:

$$d(u, v) = d(\mathbf{h}_u, \mathbf{h}_v) = ||\mathbf{h}_u - \mathbf{h}_v||_2 \tag{3}$$

Using Eq. 3, an Euclidean distance matrix can be constructed for $\mathcal{G}$. This distance matrix allows us to construct a patient-patient medical profile network $\mathcal{G}$. If we assume that no two patient's profiles are absolutely identical, then $\mathcal{G}$ will be a complete network. Patients with similar profiles will stay close to each other, forming several clusters in the network representation. As connections between very different profiles may produce noise data for classification, we define a similarity threshold ξ to control the number of connections on $\mathcal{G}$.

The connection between nodes u, v is represented by $(u, v) \in E$, and $(u, v) = \mathbb{1}\{d(u, v) \leq \xi : u, v \in V\}$. The adjacency matrix is then expressed as $A \in \mathbb{R}^{N\times N}$, $A_{uv} = \mathbb{1}\{(u, v) = 1 : (u, v) \in E, u, v \in V\}$. The construction of patient's medical profile network consists of the following steps:

1. Calculate the Euclidean similarity matrix using the node feature matrix X and Eq. 3.
2. Setting a threshold ξ for the similarity matrix .

3. Using the thresholded similarity matrix, construct the adjacency matrix A and network $[G]$.

Tree-structure representation of EHR

EHRs are often formatted similarly to relational databases, where variables are categorized by their interpretation into tables, such as demographic information, vital signs, and laboratory results. By leveraging this relationship, the EHR can be easily represented as a tree structure, where a table i is mapped with an object denoted as $\mathbf{o}_i^{(t)}$ and variable j recorded under the table is denoted as $\mathbf{o}_{ij}^{(t)}$ (Fig. 3). In this section, the t-th interval will be dropped to simplify the notation. i and j will be used as the notation for the nodes representing the corresponding objects. The tree-based representation of EHR data is defined as a $\mathcal{T} = (\mathcal{P}, \mathcal{C}, \mathcal{A})$, where $\mathcal{P}$ is a set of parent nodes and $\mathcal{C}$ is a set of child nodes. Let $i \in \mathcal{P}$ and $j \in \mathcal{C}$ then the connection between parent and child is represented by $(i,j) \in \mathcal{A}$. The adjacency matrix is expressed as $\mathcal{A} \in \mathbb{R}^{|\mathcal{P}| \times |\mathcal{C}|}$, $\mathcal{A}_{ij} = 1$ if there is a connection between them or $\mathcal{A}_{ij} = 0$ otherwise. An empty root node $R_{\mathcal{T}}$ is added to $\mathcal{P}$ to receive the final data embedding and its connection to the existing parent nodes are added to $\mathcal{A}$. The data embedding in the tree structure is carried from child node to parent node recursively from the bottom to the root node. The notation summary is listed in Table 3. The data embedding at any parent node is as follows:

$$\mathbf{o}_i = \sigma\left(\sum_{j \in C(i) \cup \{i\}} \frac{1}{\sqrt{|\mathcal{C}(i)|}.\sqrt{|\mathcal{C}(j)|}}.(\mathbf{W}_i.\mathbf{o}_{ij})\right) \tag{4}$$

$$\mathbf{s}_E = \oplus(\{\mathbf{o}_i : i \in \mathcal{C}(R_{\mathcal{T}}) \cup \{R_{\mathcal{T}}\}\}) \tag{5}$$

In Eq. 4, the data of child nodes $\mathbf{o}_{ij}$ are transformed by multiplying with weight matrix $\mathbf{W}_i \in \mathbb{R}^j$, which are then summed together to obtain the embedding of the object group $\mathbf{o}_i$. At the root node, the data is aggregated with a CONCAT function. Hence, the data embedding vector at the root node $\mathbf{s}_E \in \mathbb{R}^{|\mathcal{C}(\mathcal{R}_{\mathcal{T}})|}$ will have the dimension of the number of its child nodes (Eq. 5) .

Proposed modification of loss function

Given the training data $D = \{\mathbf{x}_i, y_i\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the feature vector of d dimensions and $y_i \in \{0, 1\}$ is the label of the sample i-th. $\mathbf{x}_i$ is extracted from EHR data and is used to construct the patient's medical profile and patient-patient network as described in the previous sections. Let p_t be the predicted probability of a patient at node i in the positive class, α_t is a balancing parameter for imbalanced class, and γ is a user-defined hyperparameter,

Table 3 Notation for tree-structure representation of EHR

Notation	Definition
$\mathcal{S}_E^{(t)}$	EHR dataset at interval *t*-th
$\mathbf{s}_E^{(t)}$	A vector represents EHR data at interval *t*-th
$\mathcal{T}$	A tree structure represent EHR data
$\mathcal{P}$	A set of parent nodes, contains all $\mathbf{o}_i^{(t)}$
$\mathcal{C}$	A set of child nodes, contains all $\mathbf{o}_{ij}^{(t)}$
$\mathcal{A} \in \mathbb{R}^{\lvert\mathcal{P}\rvert \times \lvert\mathcal{C}\rvert}$	An adjacency matrix of $\mathcal{T}$
$R_{\mathcal{T}} \in \mathcal{P}$	An empty root node to receive the EHR data embedding
$\sigma(.)$	A non-linear activation function, default to sigmoid
$\oplus(.)$	An aggregation CONCAT function

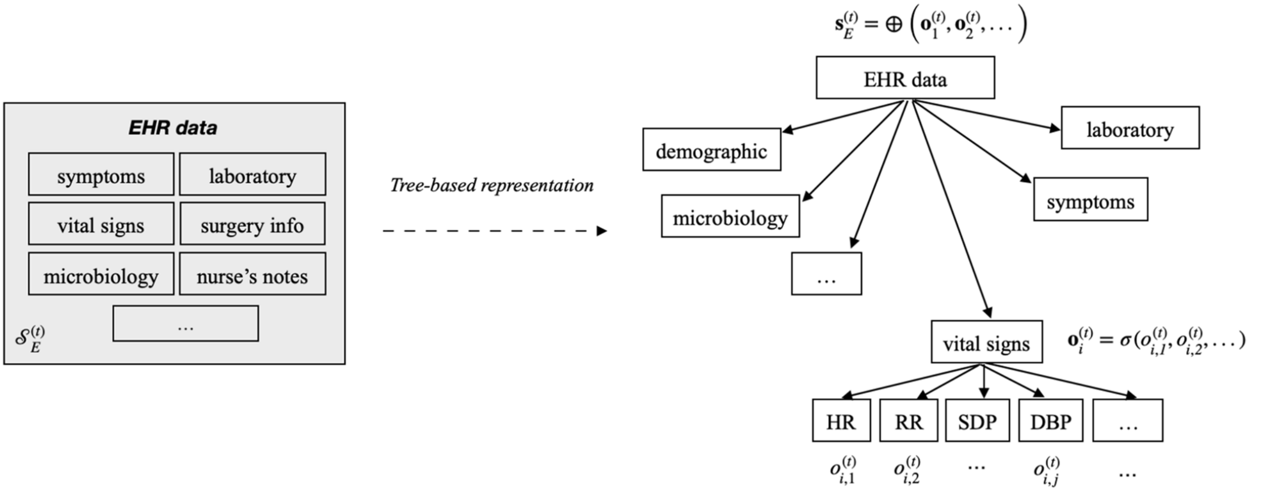
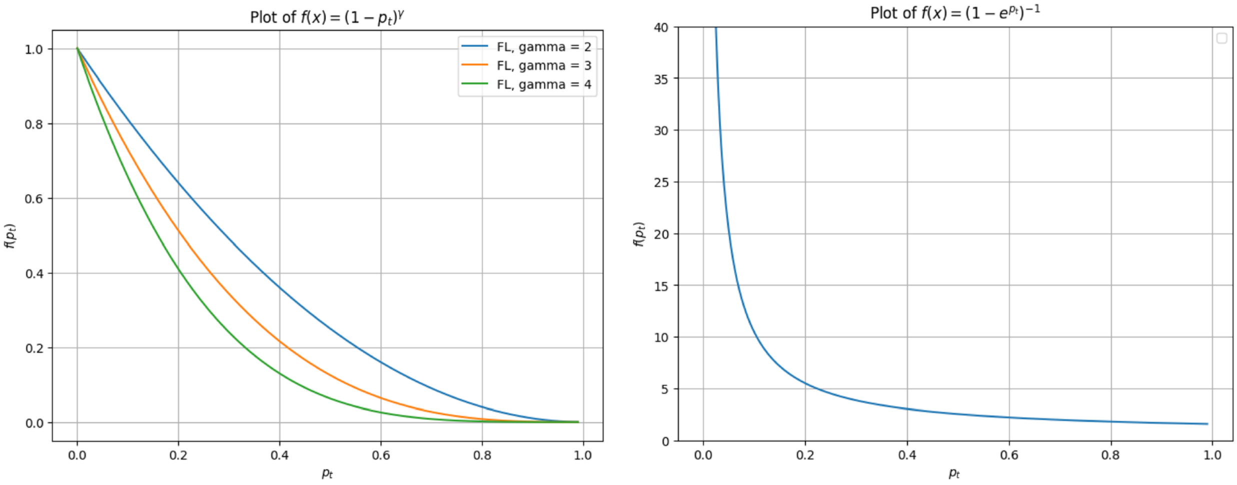


Fig. 3 Tree-structure representation of the EHR data



(a) Original modulating factor (b) Modified modulating factor

Fig. 4 The visualization of the sample weight assigned to sample in the original FL and the proposed modification of FL function

we propose a modification of the FL function for binary classification as follows:

$$\mathcal{L}_{eFL} = -\alpha_t(1 - e^{p_t})^{-1}log(p_t) \tag{6}$$

We propose to use $(1 - e^{p_t})^{-1}$ instead of the original factor $(1 - p_t)^\gamma$ to control the sample weight. Figure 4 shows the weight distribution the modulating term assign to different predicted probability. The proposed modulating factor imposes a more severe penalty for a predicted probability that is further away from the actual probability as compared to the original modulating factor. In this way, it strongly draws the attention of the loss function during the learning process to the wrongly predicted sample, emphasizing the punishment for predicted probabilities that are close to zero. The advantage of this approach over the original FL is that the sample weight can be dynamically adjusted without being dependent on γ, thereby eliminating the need to tune a hyperparameter. A large penalty assigned to a sample that is greatly mispredicted is the driving force behind an improved classification (Fig. 4).

Multi-level data embedding & model learning

The data is embedded in a bottom-up manner, folding several layers of the information: medical modules embedding, interval data embedding, and hospitalization embedding. A patient's medical profile is encoded through the following embedding sequence:

$$\mathbf{d}^{(j)} = \oplus\left(\mathbf{s}_E^{(j)}, \mathbf{s}_S^{(j)}, \mathbf{s}_I^{(j)}, ...\right) \tag{7}$$

$$\mathbf{v}^{(i)} = \oplus\left(\left\{\mathbf{d}^{(j)} : 0 < j \le t\right\}\right) \tag{8}$$

$$\mathbf{h} = \oplus\left(\left\{\mathbf{v}^{(i)} : 0 < i \le T\right\}\right) \tag{9}$$

In Eqs. 7, 8, and 9, $\oplus(.)$ represents a CONCAT aggregation function. The embedding $\mathbf{h}$ is then used to construct the patient-patient network $\mathcal{G}$ described in "Patient-patient medical profile network" section. Let n be a node on $\mathcal{G}$ and $\mathcal{N}(n)$ be the neighbors of n on the network, then the final embedding of node n on $\mathcal{G}$ is encoded as follows:

$$\mathbf{h}_n = \sigma\left(\sum_{u\in\mathcal{N}(n)\cup\{n\}} \frac{1}{\sqrt{|\mathcal{N}(n)|}\cdot\sqrt{|\mathcal{N}(u)|}} \cdot (\mathbf{W}\cdot\mathbf{h}_u)\right) \tag{10}$$

where $\mathbf{W}$ is a trainable weight matrix to transform the embedding of the neighbor nodes, σ is a *softmax* activation function. The learning loss is measured by the proposed loss function as described in "Proposed modification of loss function" section. The framework is trained and validated in a transductive manner. The training algorithm is shown in Algorithm 1.

Algorithm 1 Psuedo code of the framework training

```
Require: $D = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$, $Epochs$: number of epochs, $\xi$: similarity threshold, $\alpha$: balancing parameter, $\eta$: learning rate

  $x\_train, y\_train \leftarrow mask\_train(D)$                ▷ Split train and validation data
  $x\_val, y\_val \leftarrow mask\_val(D)$
  $\mathcal{T} \leftarrow initiate\_EHR\_Tree(x\_train)$              ▷ Initiate tree structure
  $\mathcal{G} \leftarrow initiate(x\_train, \xi)$                    ▷ Initiate patient network
  $w_{\mathcal{T}}, w_{\mathcal{G}} \leftarrow initate()$                          ▷ Initiate weight matrices
  $\mathcal{M} \leftarrow initiate(\mathcal{T}, \mathcal{G}, w_{\mathcal{T}}, w_{\mathcal{G}})$                 ▷ Initiate model
  for $epoch = 1, 2, ...Epochs$ do
      $h\_profile \leftarrow forward(\mathcal{T}, x\_train, y\_train, w_{\mathcal{T}})$   ▷ Embed data along the tree structure
      $\mathbf{M}_h \leftarrow distance(h\_profile, \text{“}euclidean\text{”})$      ▷ Compute similarity distance matrix
      $\mathcal{G} \leftarrow construct\_Network(\mathbf{M}_h, \xi)$           ▷ Construct patient network
      $h \leftarrow forward(\mathcal{G}, h\_profile, w_{\mathcal{G}})$          ▷ Embed data along patient network
      $\mathcal{L}_{eFL} \leftarrow compute()$                      ▷ Compute learning loss
      $w_{\mathcal{T}} \leftarrow w_{\mathcal{T}} - \eta \nabla \mathcal{L}_{eFL}$                 ▷ Update weight matrices
      $w_{\mathcal{G}} \leftarrow w_{\mathcal{G}} - \eta \nabla \mathcal{L}_{eFL}$
      $\mathcal{M} \leftarrow update(\mathcal{T}, \mathcal{G}, w_{\mathcal{T}}, w_{\mathcal{G}})$              ▷ Update model
  end for
```

Dataset and data processing

This study was conducted using the public dataset Pediatric intensive care dataset (PICD), version 1.1.0, which is available in the PhysioNet repository [24]. The dataset consists of patients aged 0-18 years admitted to the Intensive care units (ICUs) at the Children's Hospital of Zhejiang University School of Medicine, Zhejiang, China, from 2010-2019. Our previous work, published in 2023, described the method of selecting cohort samples and extracting data [25]. We follow the same procedure for collecting data and defining sepsis in this study. However, in the current study, only continuous variables were used, and raw demographic, vital sign, and laboratory data were used instead of category-coded data. This study was approved by the National University of Singapore's Institutional Review Board (NUS-IRB-2024-396).

Evaluation metrics

The evaluation task is to predict the sepsis outcome of the patients in the test set. As it is a binary classification task, we used Accuracy (ACC), Sensitivity (SEN), Specificity (SPE), Negative predictive value (NPV), Positive predictive value (PPV), and Area under the receiver operating characteristic curve (AUC) to evaluate the model performance.

$$ACC = \frac{TP+TN}{TP+N} \quad \begin{matrix} SEN = \frac{TP}{TP+FN} & PPV = \frac{TP}{TP+FP} \\ SPE = \frac{TN}{TN+FP} & NPV = \frac{TN}{TN+FN} \end{matrix} \tag{11}$$

AUC was measured by comparing the true positive rate against the false positive rate. A high AUC indicates the model ability to distinguish the classes in binary classification. The rest of the metrics are derived from the confusion matrix (Table 4). SEN, SPE is the proportion of TP among all positives and TN among all negative while PPV, NPV measures the TP and TN among predicted positives and predicted negatives.

Study design

The data was masked in 70% for training and 30% for testing. We trained the MedMGF on training data and reported the model performance on the masked testing

Table 4 Performance confusion matrix

	Positive (P)	Negative (N)
Prediction Positive	True positive (TP)	False positve (FP)
Prediction Negative	False negative (FN)	True negative (TN)

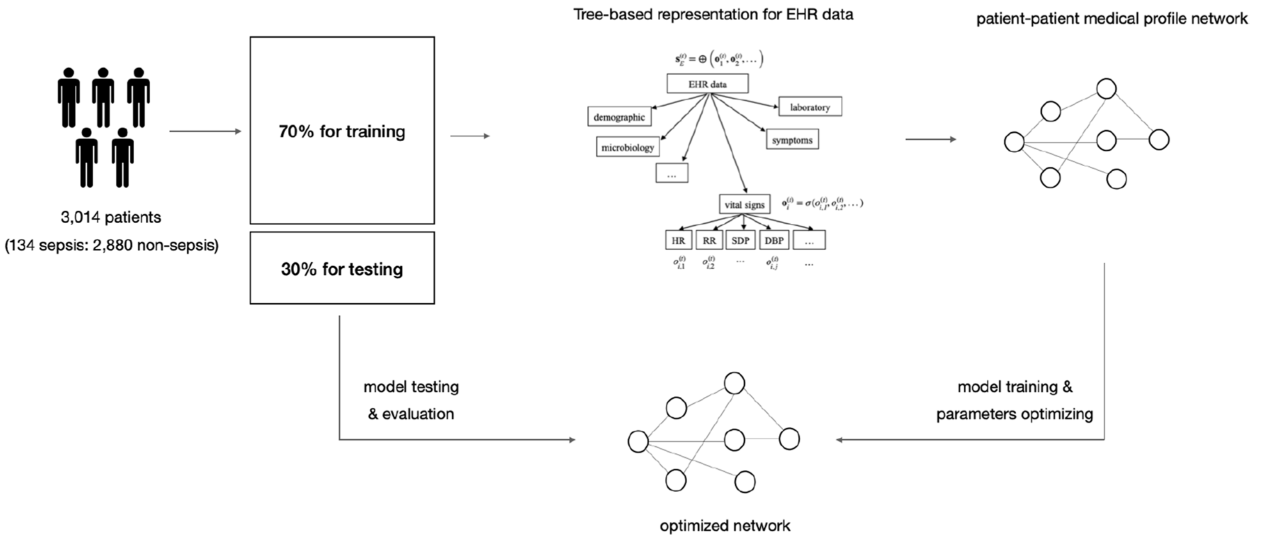

3,014 patients
(134 sepsis: 2,880 non-sepsis)
70% for training
30% for testing
Tree-based representation for EHR data
$\mathbf{s}_E^{(t)} = \oplus\left(\mathbf{o}_1^{(t)}, \mathbf{o}_2^{(t)}, \dots\right)$
EHR data
demographic
microbiology
...
laboratory
symptoms
vital signs
$\mathbf{o}_i^{(t)} = \sigma(o_{i,1}^{(t)}, o_{i,2}^{(t)}, \dots)$
HR
RR
SDP
DBP
...
$o_{i,1}^{(t)}$
$o_{i,2}^{(t)}$
...
$o_{i,j}^{(t)}$
...
patient-patient medical profile network
model testing
& evaluation
model training &
parameters optimizing
optimized network

Fig. 5 The training and validation workflow of MedMGF

data (Fig. 5). The evaluation aimed: (1) to validate the overall performance of the framework compared to the baseline models, (2) to compare its effectiveness against the oversampling method, and (3) to verify that the proposed loss function is comparable to existing loss functions. In the first evaluation, we used three sets of baseline models, including Logistic Regression (LR), GCN implemented with BCE, FL, and balancing parameter α. In the second evaluation, we used GCN+BCE and GCN+FL+SMOTE as the baseline models. As SMOTE is the most common oversampling technique for imbalance data, it was selected for this study. In the third evaluations, we implemented our proposed framework using BCE, FL, with balancing parameter α as the baseline models. Finally, we used t-distributed stochastic neighbor embedding (t-SNE) plot to visualize the data embedding produced by the MedMGF+eFL and the best two baseline models (GCN+αFL and GCN+αBCE) to demonstrate the learning process. The performance of all models was summarized in Table 5. A summary of the proposed MedMGF and the previous studies (MiME, GRAM, and MDP) was also provided in Table 6 to highlight the differences of our approach.

Models were fine-tuned to perform optimally. γ was selected in the manner that would optimize the model performance and the balancing parameter was set at the imbalance ratio of the dataset $\alpha_+ = 0.047$. All models except LR models were trained with Adam optimizer, 10,000 maximum epochs, a learning rate of 0.01. The training was implemented with an early-stopping

Table 5 Performance results on public dataset PICD

Models + Methods	γ	AUC	ACC	SEN	SPE	PPV	NPV
LR (Baseline)	-	0.5110	0.9546	0.0256	0.9965	0.25	0.9578
LR + SMOTE (Baseline)	-	0.7740	0.8252	0.7179	0.8300	0.1600	0.9849
GCN+BCE (Baseline)	-	0.5000	0.9558	0.0000	1.0000	-	0.9558
GCN+αBCE (Baseline)	-	0.7712	0.8133	0.7250	0.8173	0.1551	0.9847
GCN+SMOTE+BCE (Baseline)	-	0.6665	0.7271	0.6000	0.7329	0.0941	0.9754
GCN+FL (Baseline)	4.0	0.5000	0.9558	0.0000	1.0000	-	0.9558
GCN+αFL (Baseline)	4.0	0.7717	0.8144	0.7250	0.8185	0.1559	0.9847
GCN+SMOTE+FL (Baseline)	4.0	0.7510	0.8431	0.6500	0.8520	0.1688	0.9814
MedMGF+αBCE	-	0.7975	0.7724	0.8250	0.7699	0.1422	0.9896
MedMGF+αFL	5.0	0.7998	0.7768	0.8250	0.7746	0.1447	0.9897
MedMGF+ e FL	**-**	**0.8098**	**0.7503**	**0.8750**	**0.7445**	**0.1367**	**0.9923**

Abbreviation: *ACC* Accurary, *AUC* Area under the receiver operating characteristic curve, *BCE* Binary Cross-Entropy, *FC* Focal Loss, *LR* Logistic regression, *MedMGF* Multi-level graph-based framework for medical knowledge representation, *NPV* Negative predicted value, *PPV* Positive predicted value, *SEN* Sensitivity, *eFL* Modified focal loss, *SMOTE* Synthetic Minority Oversampling Technique, *SPE* Specificity mechanism, such that the training would be stopped when the validation loss did not decrease after 10 epochs, otherwise the results would be reported at the conclusion of the training process. BCE with logit loss was set up with a mean reduction. The data split in 70% for training and 30% for testing using the *sklearn* library. The SMOTE oversampling algorithm was implemented using the *imblearn* library. The t-SNE plot was produced by *sklearn* library. The framework was implemented in Spyder IDE (MIT, version 5.5.0, Python version 3.9.14).

Statistical methods

We calculated medians [interquartile ranges (IQRs)] and absolute counts (percentage) for continuous and categorical variables, respectively. Differences between the sepsis and non-sepsis cohort were assessed with Mann-Whitney U on continuous and Pearson's Chi-squared tests on categorical variables. All statistical analyses were performed using Microsoft Excel (version 16.55, Microsoft, USA) with a statistical significance taken as $p < 0.05$.

Results

Demographic and baseline clinical characteristics of patients

The cohort contains 3,014 admissions with a median age of 1.13 (0.15-4.30) years old and 1,698 (56.3%) males. The number of sepsis-positive cases is 134 (4.4%), which results in an imbalance ratio of 0.047 between classes. A total of three demographic variables (age, length of stay in the intensive care unit, length of stay in the hospital), five vital signs (temperature, heart rate, respiratory rate, diastolic and systolic blood pressure), and 15 laboratory variables are included in the study (Appendix A). An overview of cohort demographics and clinical outcomes can be referred to [25].

Model performance comparison against baseline model

On PICD (imbalance ratio of 0.047), LR produced predictions overwhelmingly in favor of the dominant class, resulting in a low SEN (0.0256), high ACC (0.9546), SPE (0.9965), and NPV (0.9578). With LR+SMOTE, the classification improved significantly in AUC and SEN (AUC: 0.7740, SEN: 0.7179). Comparing to these baseline models, MedMGF+eFL showed higher classification performance for AUC, SEN, and NPV (AUC: 0.8098, ACC: 0.7503, SEN: 0.8750, SPE: 0.7445, PPV: 0.1367, NPV: 0.9923). Specifically, MedMGF+eFL obtained an increase of 29.88% in AUC, and 84.94% in SEN when compared to LR.

For both GCN+BCE and GCN+FL, we observed that there was no effective learning for the minority class. However, integrating with balancing parameter, α, improved the results. αFL (AUC: 0.7717, ACC: 0.8144, SEN:0.7250, SPE: 0.8185, PPV: 0.1559, NPV: 0.9847) gave a slightly higher performance than αBCE (AUC: 0.7712, ACC: 0.8133, SEN: 0.7250, SPE: 0.8173, PPV: 0.1551, NPV: 0.9847), though the difference was not considered significant. Compared to the αFL, the proposed MedMGF+eFL framework demonstrated a 3.81% increase in AUC and a 15% increase in SEN.

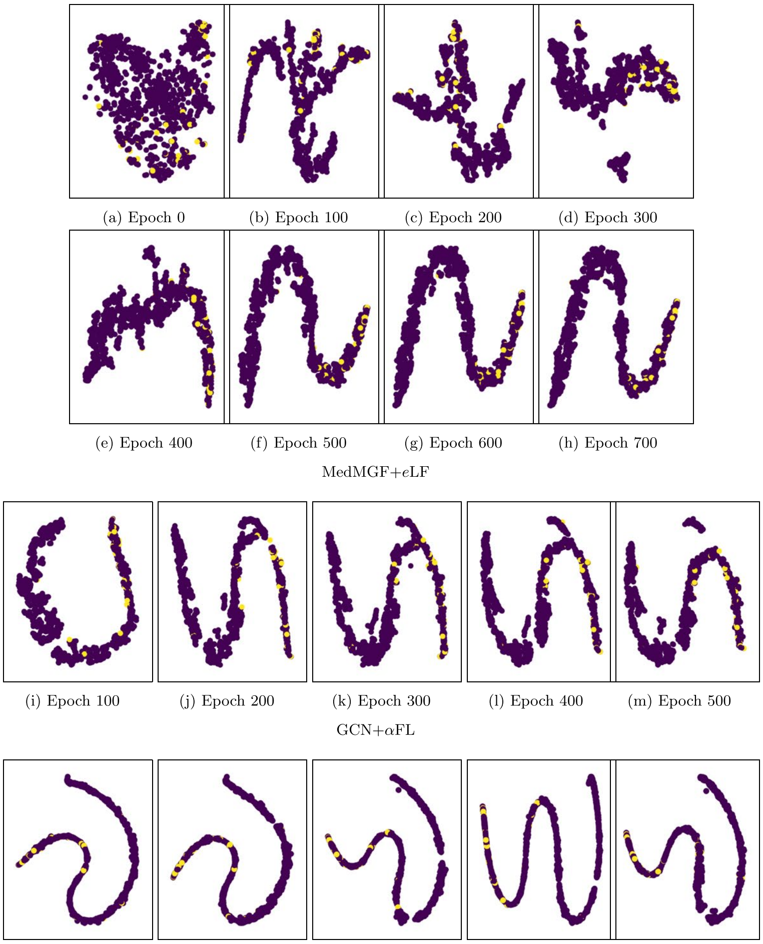
Model performance comparison with different loss functions & SMOTE

Using BCE and FL alone does not lead to effective learning during training due to the extreme imbalance ratio of the dataset. Performance improvements were only achieved with the inclusion of SMOTE. The GCN+SMOTE+FL model (AUC: 0.7510, ACC: 0.8431, SEN: 0.6500, SPE: 0.8520, PPV: 0.1688, NPV: 0.9814) yielded better results compared to GCN+SMOTE+BCE (AUC: 0.6665, ACC: 0.7271, SEN: 0.6000, SPE: 0.7329, PPV: 0.0941, NPV: 0.9754).When compared with GCN+SMOTE+FL, the MedMGF+eFL model showed a 5.88% increase in AUC and a 22.5% increase in SEN, although there was a decrease of 8.05% in SPE and 3.21% in PPV. Additionally, MedMGF+eFL achieved a 3.58% increase in AUC and a 4.98% increase in SEN when compared to LR+SMOTE.

Model performance comparison for the proposed loss function

We observed that MedMGF achieved high SEN (αBCE: 0.8750, αFL: 0.8250, eFL: 0.8750), high AUC (αBCE: 0.7975, αFL: 0.7998, eFL: 0.8098), and high NPV (αBCE: 0.9896, αFL: 0.9897, eFL: 0.9923) when compared to all baseline models. The best SEN (0.8750), AUC (0.8098), and NPV (0.9923) were achieved with the proposed loss function eFL. However, MedMGF+eFL experienced a decrease in PPV (0.1367) and SPE (0.7445) compared to the other two models.

Figure 6 presents the final patient embeddings within the patient network, generated by the proposed MedMGF+eFL framework, alongside GCN+αFL and GCN+αBCE. In this visualization, yellow dots represent the positive class, while purple dots represent the negative class. For MedMGF+eFL, we observed that the yellow dots initially intermingle with the purple dots, making it challenging to establish a clear boundary between them. However, as training progresses, the yellow dots gradually cluster together, and by epoch 700, most of them have concentrated at one end, facilitating easier classification. In contrast, the other two baseline models quickly separated the dots, but the separation process slowed down starting from epoch 300 for GCN+αFL and from epoch 400 for GCN+αBCE. Learning in these models ceased around epoch 500, whereas it


(a) Epoch 0 (b) Epoch 100 (c) Epoch 200 (d) Epoch 300

(e) Epoch 400 (f) Epoch 500 (g) Epoch 600 (h) Epoch 700

MedMGF+eLF

(i) Epoch 100 (j) Epoch 200 (k) Epoch 300 (l) Epoch 400 (m) Epoch 500

GCN+αFL

(n) Epoch 100 (o) Epoch 200 (p) Epoch 300 (q) Epoch 400 (r) Epoch 500

GCN+αBCE

Fig. 6 The data embedding transformation during training with MedMGF+*e*LF, GCN+*α*FL, and GCN+*α*BCE. Yellow dots represent positive samples and purple dots represent negative samples

continued with MedMGF+*e*FL, leading to a higher SEN for MedMGF+*e*FL.

Discussion

In this study, we propose a novel multi-level graph-based framework designed to represent clinical knowledge that can be utilized for several downstream applications. It consists of three components: a tree structure that captures an individual patient's medical information, a patient-patient network, and an modified loss function specifically for imbalanced datasets. The integration of patient medical profiles and patient networks within a unified architecture facilitates multiple types of analyses, including patient stratification and cohort discovery. Our results demonstrated the framework's effectiveness, achieving improved classification performance on a highly imbalanced pediatric sepsis dataset (imbalance ratio of 0.047) compared to baseline models. Furthermore, the proposed loss function has shown improvements in classification performance over BCE and FL. In the following section, we will discuss the framework's properties, its clinical implications, as well as its limitations and potential direction for future research.

Framework approach. Our approach focuses on preserving the EHR's inherent structure and interpretability by leveraging its existing groupings. By utilizing this structure and organizing it in a tree-like manner, we effectively reduce the dimensionality of the data input while maintaining a minimum level of data embedding interpretation. This dimensionality reduction leads to faster training times and a less complex learning process. The approach has also been designed to facilitate the integration of domain experts' knowledge, allowing them to construct the structure intuitively, thereby enhancing interpretability. Compared to the creation of graphical models like Bayesian networks [26] by domain experts, constructing a tree structure is simpler and more cost-effective. Through this graph-based architecture, we can visually represent both the patient's medical profile and their relationships with other patients. This architecture incorporates several layers of information, including interval data and hospitalization records. Essentially, it encompasses the entire hospitalization of the patient and the data for each visit in a compact, easily visualized format. Depending on the context, this can be presented either as an individual medical profile or as a cluster of similar patients. Furthermore, by integrating patient medical profiles and patient networks into a unified architecture and training process, we achieve a reduction in training costs.

Framework properties. MedMGF has the following key properties: (1) generality and modality, (2) multipurpose functionality, (3) intuitive interpretation, and (4) minimal data processing requirements.

Firstly, the framework is designed to seamlessly integrate with various types of medical data by embedding and extending the number of modules to accommodate additional data sources. This modular approach allows for the effortless incorporation of new information, enabling the framework to be easily modified and updated in response to evolving medical data. With its flexible module structure, the MedMGF framework efficiently utilizes available information, enhancing its adaptability and scalability.

Secondly, the framework demonstrates potential for a wide range of tasks, such as disease diagnosis, cohort discovery, and risk prediction. For instance, the similarity between patient profiles can be leveraged to predict another patient's risk of rehospitalization or their likely response to treatment. Clinicians can also utilize the framework to identify individuals at risk for certain diseases or adverse reactions by comparing medical profiles. Additionally, MedMGF can serve as a bedside monitoring system, tracking patients' conditions and the progression of their diseases. In some scenarios, the framework could be adapted to alert clinicians when a patient's medical profile closely resembles that of a specific disease or when certain characteristics are present, enhancing early detection and intervention.

A third characteristic of the framework is its ease of interpretation, which enables clinicians to easily understand concepts related to the structure of the EHR, patient profiles, and the patient network. By presenting the data in a clear and concise manner, the framework can assist clinicians in making informed decisions and gaining valuable insights from framework visualizations. This intuitive interpretability enhances the framework's effectiveness and usability in various medical contexts, ultimately contributing to improved patient care and outcomes.

Last but not least, it requires minimal processing of EHR data since it does not require oversampling techniques to improve learning, or additional processing to map data to the multilevel graph-based structure. However, it still requires basic processing tasks such as handling missing data, removing outliers, and selecting variables for the EHR tree-based structure.

Handling data imbalance. Class imbalances in medical data are common and can significantly impair classification performance [27, 28].Due to these imbalances, ML models may struggle to accurately differentiate between classes, often leading to biased predictions that favor the dominant class. Various techniques can address this issue at both the data and algorithmic levels. Data-level approaches include oversampling and undersampling, while algorithmic-level approaches involve heuristics that prioritize minority classes [29].

Oversampling techniques, such as SMOTE, have shown effectiveness in improving ML model performance by generating synthetic samples during training. This, however, may introduce unwanted additional noise and bias to the training process. On the other hand, undersampling, which reduces the number of samples in the dominant class, is not beneficial when dealing with extremely imbalanced or small medical datasets. In this study, we address the imbalance problem at the algorithmic level by modifying the focal loss function. By assigning a modulating term to samples, the loss function can concentrate more on hard-to-classify samples during training. The modulating term proposed in our study creates a flexible sampling weight that adapts based on the framework's learning at each training round, eliminating the need to rely on a hyperparameter.

Framework explainability. It is essential for clinicians to understand how machine learning models make decisions to apply these models to their practice. For this reason, models should be able to explain how data is used, identify the factors influencing decisions, and clarify how those decisions are reached. Given this need, it is not surprising that Explainable AI (XAI) has seen rapid growth in recent years [30–32]. XAI plays a critical role in bridging the gap between proof-of-concept studies and applied ML in medicine [33]. By leveraging XAI, potential biases or errors in the model can be identified, offering insights into the reasons behind specific decisions. Moreover, it can be used to tune parameters or correct errors in the model. XAI techniques commonly used in ML-based studies include Shapley Additive Explanations (SHAP, [34]) and Local Interpretable Model-Agnostic Explanations (LIME, [30]). Currently, our framework does not use XAI, but it can easily be adapted to do so. For example, it is possible to identify different nodes' attention weights with SHAP or LIME or with Graph Attention Networks (GAT, [35]) integrated into the framework. An alternative is to integrating a GAT-like approach to the hierarchical embeddings to enhance its explainability during the model learning.

Framework complexity. The framework consists of four operations: (1) the tree-structure representation and embedding for EHR data $\mathcal{O}_{\mathcal{T}}$, (2) the multilevel data embedding for patient's medical profile $\mathcal{O}_{\mathcal{P}}$, (3) the construction of patient-patient medical profile network $\mathcal{O}_{\mathcal{G}}$, and (4) inference for downstream tasks on the medical profile network $\mathcal{O}_{\mathcal{I}}$. Hence, the time complexity of the overall framework will be the sum of these operations:

$$\mathcal{O}_{MedMGF} = \mathcal{O}_{\mathcal{T}} + \mathcal{O}_{\mathcal{P}} + \mathcal{O}_{\mathcal{G}} + \mathcal{O}_{\mathcal{I}} \tag{12}$$

The patient's medical profile network is constructed based on a Euclidean distance matrix $\in \mathbb{R}^{N\times N}$ with N number of patients. Hence, the complexity is estimated to be $\mathcal{O}(N^2)$. The core operation in (1), (2), and (4) is based on the message passing mechanism. This mechanism includes the feature transformation, neighborhood aggregation, and updating via activation function in both forward and backward pass for one layer. In the forward pass, the feature transformation is a multiplication between node feature matrix $X \in \mathbb{R}^{N\times T}$ and transformation weight matrix $W \in \mathbb{R}^{T\times T}$, hence, $\mathcal{O}(NT^2)$. Neighbor aggregation is a multiplication between matrices of size $N \times N$ and $N \times T$, yielding $\mathcal{O}(N^2T)$. Finally, the cost for using activation function is a $\mathcal{O}(N)$. In practice, we could use a sparse operator, therefore the cost of the neighbor aggregation can be reduce to $\mathcal{O}(|E|T)$. Hence, the total cost of the forward pass is $\mathcal{O}(NT^2) + \mathcal{O}(|E|T) + \mathcal{O}(N)$. In the backward pass, the cost of performing the backprobagation for X and W is $\mathcal{O}(NT^2) + \mathcal{O}(|E|T)$.

In the tree-structure representation $\mathcal{T}$ for EHR data, there are $(|\mathcal{P} - 1|)$ message passing and one aggregation operation at the root node. Additionally, the multilevel data embedding for interval and hospitalizations consist of three aggregation functions. Hence, the time complexity of the framework from the four mentioned operations is:

$$\mathcal{O}_{MMGF} = \mathcal{O}_{\mathcal{T}}((|\mathcal{P}| - 1)(|\mathcal{C}| + |\mathcal{A}|)) + \mathcal{O}_{\mathcal{P}}(N) + \mathcal{O}_{\mathcal{G}}(N^2) + \mathcal{O}_{\mathcal{I}}(NT^2 + |E|T + N) \tag{13}$$

As we used embedding to reduce the dimension of the feature vectors, T is rather small compared to the original dimension of the feature vectors. Therefore, overall complexity of MedMGF is reduced to $\mathcal{O}(N^2)$. The complexity depends on the number of sample in the dataset. As the number of sample grows, it will be taxing to construct the patient network.

Comparison with previous studies. Table 6 provides a summary of our proposed framework, MedMGF, in comparison with MiME, GRAM, and MDP, all of which employ a multi-level embedding approach to medical data representation. While MiME and GRAM may be limited to diagnosis and treatment codes, MedMGF encompasses more aspects of EHR data and can be extended to incorporate imaging, signals, and other data types into the representation. Although MDP integrated clinical data into GRAM, the data is still handled in a flat manner. While existing works exploit the hierarchy between medical codes, MedMGF exploits the hierarchy between EHR data itself. MiME and GRAM are capable of representing complex and general medical concepts beyond just data alone.

All methods are capable of handling both small and large medical datasets. With MedMGF, the complexity increases significantly as the number of patients increases. None of the methods have integrated XAI, with model interpretation primarily derived from the framework architecture. GRAM and MDP are notable for their use of attention mechanisms, which allow for better model interpretation and feature importance determination. In this regard, MedMGF relies on the intuitive tree structure of EHR data as well as the integration of a network of patient similarity to enhance the interpretation of the model. As of now, MedMGF does not have a mechanism for determining the importance of features.

In comparison with existing methods, MedMGF has the advantage of handling imbalanced data and does not require additional data processing. Existing methods do not address imbalanced data directly and may require additional steps to process medical codes when applied to other EHR datasets.

Clinical impact. The MedMGF framework demonstrates significant improvements in AUC and SEN when compared to the baseline models on an extremely imbalanced dataset. The improvement in these metrics suggests that MedMGF may improve diagnostic precision and accuracy in real-world medical settings, such as sepsis diagnosis, where the sepsis population is much smaller than the healthy population. Furthermore, a false negative in a sepsis diagnosis can have a more detrimental effect on the patient's well-being than a false positive. It is therefore more desirable to achieve a high level of performance in SEN to reduce the number of false negatives. It cannot be ignored that false positives can result in a greater incidence of antibiotic resistance. However, the well-being of the patient as well as his or her mortality

Table 6 Comparison between the proposal MedMGF and previous studies

Criteria	MedMGF	MiME	GRAM	MDP
Architecture				
Multilevel embedding	Hospitalization data + interval medical data	Hospital visit data	Hospital visit data	Hospital visit data
EHR representation	Demographic + vital signs + laboratory + symptoms + etc.	Diagnosis codes (e.g., cough, fever) + treatment codes (e.g., Acetaminophen)	Medical codes + medical concepts	Diagnosis codes + clinical data
Approaches	EHR tree-based structure	Medical code tree-based structure	Medical code knowledge-graph	GRAM's knowledge-graph
	Patient similarity network			Clinical data vector
Efficiency				
Complexity	$\mathcal{O}(N^2)$	Not mentioned	Not mentioned	Not mentioned
Parameters	Embedding weightsNetwork similarity threshold	Embedding weights	Embedding weightsAttention weights	Embedding weightsAttention weights
Robustness & Scalability				
Small Dataset	Yes	Yes	Yes	Yes
Large Dataset	Yes, complexity will increase	Yes, not clear on the complexity	Yes, not clear on the complexity	Yes, not clear on the complexity
Data types	EHR, imaging, signal data, etc.	Diagnosis, procedure, medication codes	Diagnosis, procedure, medication codes, medical concepts	EHR, diagnosis codes
Modularity	Yes	No	No	No
Interpretability & Explainability				
XAI Incorporation	No	No	No	No
Feature Importance	No	No	No	Yes, via attention mechanism
Model Interpretation	Yes, via framework representation & patient network	Yes, via framework representation	Yes, via framework representation & attention mechanism	Yes, via framework representation & attention mechanism
Data Processing				
Missing Data	Yes	No	No	Yes
Outlier Data	No	No	No	No
Imbalanced Data	Yes	Yes	Yes	Not validated
Additional processing	No	Yes, preprocess medical codes	Yes, preprocess medical codes	Yes, preprocess medical codes

The comparison between the proposed MedMGF, MiME [19], GRAM [18], and MDP [20]. Each method demonstrates distinct strengths and approaches in various aspects. It is essential to note that the criteria listed in this table are not intended to be comprehensive. *Abbreviations*: *EHR* Electronic health records, *XAI* Explainable AI

status should usually take precedence. Our MedMGF framework is therefore advantageous, since it can deliver a high SEN on imbalanced data. By effectively addressing the challenges posed by imbalanced datasets, MedMGF can potentially open up new possibilities for more accurate and reliable clinical applications. In terms of development, deployment, and application, MedMGF can be tailored to meet a variety of needs in the hospital, including disease diagnosis, bedside monitoring, and research assistance. With its versatility, it can be used for a variety of purposes, thereby eliminating the need for multiple systems and frameworks, resulting in cost savings.

Study limitation & future works. There are, however, a number of limitations to our study. First, the limited number of datasets used for evaluation may raise concerns about MedMGF's generalizability, which is an important aspect of ML to ensure that the model can perform well on unseen data. Without a diverse range of datasets, the model may fail to accurately predict outcomes in real-world scenarios, leading to unreliable results and limited practical applications. Second, for demonstration purposes, we used only a portion of the data collected within 24 hours of ICU admission. To validate the generality of the framework when modeling patient medical profiles with several hospitalizations and intervals, more data points should be included. The number of features in the dataset is relatively small to be able to validate the complexity of the framework, as discussed in the section on complexity analysis. As our experiment only work on EHR data, more research should be conducted to validate on other data types (e.g., imaging, waveforms).

Furthermore, a comparison of our work with previous studies would provide additional evidence and validation of the MedMGF's efficiency. However, previous studies such as MiME and Med2Vec require data on the relationship between symptoms and treatments, which is not available in our dataset. The implementation of MiME or GRAM with our current data is therefore challenging. We did not include the performance results of MedMGF with MiME, GRAM, and MDP in our comparison for the following reasons. Each framework uses different performance metrics to measure the performance: MiME uses PR-AUC for predicting the onset of Heart Failure (HF), GRAM uses Accuracy@K (where K represents the top diagnosis code guesses of the next hospital visit) to count the number of correct-predicted codes and AUC for predicting the onset of HF, MDP measures Accuracy@K for the top k diagnosis code guesses of the next hospital visit, our MedMGF uses AUC, ACC, SEN, SPE, NPV, PPV for sepsis prediction. Considering the differences in nature of the tasks defined in the various experiments and the metrics used, it is challenging for us to compare the results among studies. In addition, we were not able to reproduce MiME or GRAM as our dataset lacks the relationship between treatment and diagnosis codes.

In addition, the inferred characteristics of the framework are inferred from its design. Currently, demonstrating modality characteristics is challenging due to our dataset lacking imaging or waveform data. Therefore, more research is needed to confirm the feasibility of the framework for a variety of other analysis purposes as well as to confirm its multipurpose characteristic. Finally, we have not yet incorporated an XAI mechanism into the framework. To address these limitations, future research can consider collecting data over a longer period in order to conduct evaluations that are more comprehensive and diverse. Additionally, incorporating data from multiple healthcare institutions or collaborating with other researchers could enhance the framework's generalizability and validate its effectiveness across different settings. It is also beneficial to integrate an XAI mechanism into the framework in order to enhance its interpretability. XAI is an emerging area in applied ML within healthcare, and it has the potential to significantly enhance model interpretation and promote the practical ML application in clinical settings. Literature reviews and model development related to XAI and the contribution of various types of medical data to clinical prediction models, could be valuable areas for further research.

Conclusion

Our study proposes MedMGF, a framework that integrates medical profile representation and patient-patient profile network within a single architecture. It utilizes the hierarchical structure of EHR data to represent the patients' medical data and the graphical structure of patient-patient networks to perform supervised tasks. Additionally, the proposed modification to the focal loss resulted in improved classification performance on imbalance datasets compared to the baseline models. Generally, the framework encapsulates both generality and modality that can easily be adapted to a variety of analyses and applications. Furthermore, it can be further extended by incorporating XAI to enhance its interpretation and transparency in future research.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02649-2.

Supplementary Material 1.

Acknowledgements

This study received no funding.

Authors' contributions

The concept and architecture were designed by TMN under the supervision of PKL, JHL and SLC. PKL provided the technical guidance. JHL and SLC provided clinical interpretation of the results. All authors contributed to manuscript preparation, writing, critical revisions, and have read and approved the final manuscript.

Funding

This study received no funding.

Availability of data and materials

The datasets analysed during the current study are available in the PhysioNet repository, https://physionet.org/content/picdb/1.1.0/.

Code availability

The underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author. Link: https://github.com/zoetmn/med-mgf.

Declarations

Ethics approval and consent to participate

The Institutional Review Board of National University of Singapore approved this study (IRB: NUS-IRB-2024-396).

Consent for publication

This paper did not include any individual data, rights and permissions. Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Industrial Engineering and Management, National University of Singapore, Singapore 117576, Singapore. [2]Children's Emergency, KK Women's and Children's Hospital, Singapore 229899, Singapore. [3]SingHealth-Duke NUS Paediatrics Academic Clinical Programme, Duke-NUS Medical School,

Singapore 169857, Singapore. [4]Children's Intensive Care Unit, KK Women's and Children's Hospital, Singapore 229899, Singapore.

Received: 24 May 2024 Accepted: 23 August 2024
Published online: 02 September 2024

References

1. Wang W, Ferrari D, Haddon-Hill G, Curcin V. Electronic Health Records as Source of Research Data. In: Machine Learning for Brain Disorders, vol. 197. Springer US; 2023. pp. 331–354. https://doi.org/10.1007/978-1-0716-3195-9_11. https://link.springer.com/10.1007/978-1-0716-3195-9_11.
2. Kim MK, Rouphael C, McMichael J, Welch N, Dasarathy S. Challenges in and Opportunities for Electronic Health Record-Based Data Analysis and Interpretation. Gut Liver. 2024;18. https://doi.org/10.5009/gnl230272.
3. Habehh H, Gohel S. Machine learning in healthcare. Curr Genomics. 2021;22:291–300. https://doi.org/10.2174/1389202922666210705124359. https://www.eurekaselect.com/194468/article
4. Amirahmadi A, Ohlsson M, Etminani K. Deep learning prediction models based on EHR trajectories: a systematic review. J Biomed Inform. 2023;144:104430. https://doi.org/10.1016/j.jbi.2023.104430. https://linkinghub.elsevier.com/retrieve/pii/S153204642300151X
5. Pai S, Bader GD. Patient Similarity Networks for Precision Medicine. J Mol Biol. 2018;430:2924–38. https://doi.org/10.1016/j.jmb.2018.05.037. https://linkinghub.elsevier.com/retrieve/pii/S0022283618305321
6. Parimbelli E, Marini S, Sacchi L, Bellazzi R. Patient similarity for precision medicine: A systematic review. J Biomed Inform. 2018;83:87–96. https://doi.org/10.1016/j.jbi.2018.06.001. https://linkinghub.elsevier.com/retrieve/pii/S1532046418301072
7. Navaz AN, T El-Kassabi H, Serhani MA, Oulhaj A, Khalil K. A Novel Patient Similarity Network (PSN) Framework Based on Multi-Model Deep Learning for Precision Medicine. J Personalized Med. 2022;12:768. https://doi.org/10.3390/jpm12050768.
8. Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS Comput Biol. 2011;7. https://doi.org/10.1371/journal.pcbi.1002141.
9. Lu H, Uddin S. A weighted patient network-based framework for predicting chronic diseases using graph neural networks. Sci Rep. 2021;11. https://doi.org/10.1038/s41598-021-01964-2.
10. Panahiazar M, Taslimitehrani V, Pereira NL, Pathak J. Using EHRs for Heart Failure Therapy Recommendation Using Multidimensional Patient Similarity Analytics. Stud Health Technol Inform. 2015;210:369–73.
11. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection. 2018. arXiv:1708.02002.
12. Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. 2017. arXiv:1609.02907. Accessed 24 June 2024.
13. Shannon CE. A Mathematical Theory of Communication. Bell Syst Tech J. 1948;27:379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x. https://ieeexplore.ieee.org/document/6773024
14. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002;16:321–57. https://doi.org/10.1613/jair.953. https://www.jair.org/index.php/jair/article/view/10302
15. Mukherjee P, Humbert-Droz M, Chen JH, Gevaert O. SCOPE: predicting future diagnoses in office visits using electronic health records. Sci Rep. 2023;13:11005. https://doi.org/10.1038/s41598-023-38257-9. https://www.nature.com/articles/s41598-023-38257-9
16. Grout R, Gupta R, Bryant R, Elmahgoub MA, Li Y, Irfanullah K, et al. Predicting disease onset from electronic health records for population health management: a scalable and explainable Deep Learning approach. Front Artif Intell. 2024;6:1287541. https://doi.org/10.3389/frai.2023.1287541. https://www.frontiersin.org/articles/10.3389/frai.2023.1287541/full
17. Choi E, Bahadori MT, Searles E, Coffey C, Sun J. Multi-layer Representation Learning for Medical Concepts. 2016. arXiv:1602.05568.
18. Choi E, Bahadori MT, Song L, Stewart WF, Sun J. GRAM: Graph-based Attention Model for Healthcare Representation Learning. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2017. pp. 787–795. https://doi.org/10.1145/3097983.3098126. https://dl.acm.org/doi/10.1145/3097983.3098126.
19. Choi E, Xiao C, Stewart WF, Sun J. MiME: Multilevel Medical Embedding of Electronic Health Records for Predictive Healthcare. 2018. arXiv:1810.09593.
20. Li R, Ma F, Gao J. Integrating Multimodal Electronic Health Records for Diagnosis Prediction. AMIA Annual Symposium proceedings, vol. 2021. AMIA Symposium; 2021. pp. 726–735.
21. Charoenphakdee N, Vongkulbhisal J, Chairatanakul N, Sugiyama M. On Focal Loss for Class-Posterior Probability Estimation: A Theoretical Perspective. 2020. arXiv:2011.09172.
22. Mukhoti J, Kulharia V, Sanyal A, Golodetz S, Torr PHS, Dokania PK. Calibrating Deep Neural Networks using Focal Loss. 2020. arXiv:2002.09437.
23. Ghosh A, Schaaf T, Gormley MR. AdaFocal: Calibration-aware Adaptive Focal Loss. 2023. arXiv:2211.11838.
24. Zeng X, Yu G, Lu Y, Tan L, Wu X, Shi S, et al. PIC, a paediatric-specific intensive care database. Sci Data. 2020;7:14. https://doi.org/10.1038/s41597-020-0355-4. http://www.nature.com/articles/s41597-020-0355-4
25. Nguyen TM, Poh KL, Chong SL, Lee JH. Effective diagnosis of sepsis in critically ill children using probabilistic graphical model. Transl Pediatr. 2023;12:538–51. https://doi.org/10.21037/tp-22-510.
26. Andersen SK. Probabilistic reasoning in intelligent systems: Networks of plausible inference. Artif Intell. 1991;48:117–24. https://doi.org/10.1016/0004-3702(91)90084-W. https://linkinghub.elsevier.com/retrieve/pii/000437029190084W
27. Thabtah F, Hammoud S, Kamalov F, Gonsalves A. Data imbalance in classification: Experimental evaluation. Inf Sci. 2020;513:429–41. https://doi.org/10.1016/j.ins.2019.11.004. https://linkinghub.elsevier.com/retrieve/pii/S0020025519310497
28. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data. 2019;6:27. https://doi.org/10.1186/s40537-019-0192-5. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0192-5
29. Rezvani S, Wang X. A broad review on class imbalance learning techniques. Appl Soft Comput. 2023;143:110415. https://doi.org/10.1016/j.asoc.2023.110415. https://linkinghub.elsevier.com/retrieve/pii/S1568494623004337
30. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM; 2016. pp. 1135–1144. https://doi.org/10.1145/2939672.2939778. https://dl.acm.org/doi/10.1145/2939672.2939778.
31. Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, et al. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Inf Fusion. 2023;99:101805. https://doi.org/10.1016/j.inffus.2023.101805. https://linkinghub.elsevier.com/retrieve/pii/S1566253523001148
32. Saeed W, Omlin C. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowl-Based Syst. 2023;263:110273. https://doi.org/10.1016/j.knosys.2023.110273. https://linkinghub.elsevier.com/retrieve/pii/S0950705123000230
33. S Band S, Yarahmadi A, Hsu CC, Biyari M, Sookhak M, Ameri R, et al. Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. Inform Med Unlocked. 2023;40:101286. https://doi.org/10.1016/j.imu.2023.101286. https://linkinghub.elsevier.com/retrieve/pii/S2352914823001302.
34. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions. 2017. arXiv:1705.07874.
35. Veličković P, Cucurull G, Casanova A, Romero A, Lió P, Bengio Y. Graph Attention Networks. 2018. arXiv:1710.10903.

Publisher's Note

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