2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) | 979-8-3315-8618-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/EMBC58623.2025.11254005

naviDCN: Navigator-Guided Multi-Modal Deep Clustering for Sepsis Phenotyping in Early ICU Admission

Pi-Ju Tsai[1], Kuan-Fu Chen[2], Charkkri Limbud[3], and Yi-Ju Tseng[4]

***Abstract*— Sepsis is a life-threatening and heterogeneous disease characterized by dysregulated host responses to infection. Although recent studies applied unsupervised algorithms to uncover sepsis phenotypes, the clustering process lacks the ability to incorporate clinical knowledge, potentially resulting in phenotypes with limited interpretability. We propose a novel clustering framework, naviDCN, which integrates a navigator component to align clusters with clinical significance. The naviDCN architecture comprises multi-modal encoders, a deep clustering network (DCN) with reconstruction tasks, and a navigator module. We first encode electronic health records into representative embeddings by introducing an attention mechanism on multi-modal information. The framework then iteratively optimizes network weights of reconstruction and navigator modules, and updates cluster centroids of the clustering module. The navigator incorporates clinical knowledge into the embedding through backpropagation, thereby guiding clustering toward clinically meaningful outcomes. We discover four sepsis phenotypes with unique clinical characteristics, SOFA trajectories, and mortality patterns. Notably, while both α and δ phenotypes show severe conditions in the early stage, naviDCN effectively differentiates between patients likely to show clinical improvement (α) and those at risk of deterioration (δ). Furthermore, the navigator effectively enhances phenotype interpretability without compromising objective clustering performance. This study offers insights into understanding the heterogeneity of sepsis phenotypes.**

***Clinical Relevance*—This study integrates a navigator module into the clustering framework to identify phenotypes with distinct short-term organ dysfunction trajectories and long-term survival status, thus improving the interpretability of sepsis phenotypes. Unraveling the latent information in demographics, laboratory test results, and vital signs with deep learning, our framework identifies characteristic phenotypes and opens new avenues for exploring targeted treatment for sepsis patients across phenotypes.**

I. INTRODUCTION

Sepsis is a life-threatening condition that occurs when the immune system has an extreme response to an infection, leading to organ dysfunction [1]. As one of the primary causes of mortality and morbidity in intensive care unit (ICU) patients, sepsis accounts for approximately 20% of deaths worldwide [2]. A better understanding of sepsis enables more effective patient cohort stratification, thereby enabling the development of personalized therapeutic strategy [3].

The disease progression in sepsis patients varies according to host factors, infection etiology, and the involvement of multiple organ dysfunctions. Due to the significant heterogeneity of sepsis, physicians often integrate diverse patient data from multiple sources to formulate appropriate treatment strategies. Since identifying patterns within complex and dynamic clinical records is challenging, previous studies have employed unsupervised learning algorithms to explore distinct sepsis phenotypes. A retrospective study extracted the most abnormal values from multiple laboratory measurements and applied consensus k-means clustering to identify four distinct phenotypes, laying the foundation for sepsis phenotyping [4]. Recent studies have further focused on incorporating temporal information into clustering. One study analyzed organ dysfunction trajectories using univariate time series [5], and another study applied weighted k-means clustering on multivariate time series, identifying four phenotypes characterized by different organ dysfunction progression or organ failure profiles [6].

However, these studies haven't fully utilized comprehensive information. For instance, the frequency of laboratory tests may carry implicit details regarding patient severity. When applying clustering, features are treated as independent variables, thereby neglecting potential interactions that could reflect the underlying complexity of sepsis. Moreover, current approaches lack mechanisms for incorporating factors that indicate clinical significance, such as important clinical outcomes, during autoencoder's generation of embeddings for clustering [7]. By guiding the autoencoder to capture clinically relevant patterns, the resulting embeddings and phenotypes could demonstrate enhanced interpretability.

This study aims to cluster sepsis patients into distinct phenotypes using multi-modal information collected during early ICU admission. We propose the concept of "navigator" module, which guides the embedding process to incorporate important domain knowledge, such as discharged status or length of stay (LOS). With target-driven clustering, we seek to uncover clinically meaningful sepsis phenotypes that align with indicators for clinical significance and provide valuable insights to support future research in precision medicine.

II. METHODS

Figure 1 illustrates the workflow, including multi-modal encoders, deep clustering network (DCN) with reconstruction tasks [8], and the navigator module. We gather electronic

[1]Pi-Ju Tsai is with the Institute of Data Science and Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan. ruubi0807@gmail.com

[2]Kuan-Fu Chen is with the College of Intelligent Computing and Medical Statistics Research Center, Chang Gung University, Taoyuan, Taiwan, and the Department of Emergency Medicine, Chang Gung Memorial Hospital, Keelung, Taiwan.

[3]Charkkri Limbud is with the EECS International Graduate Program, National Yang Ming Chiao Tung University, Hsinchu, Taiwan.

[4]Yi-Ju Tseng is with the Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan and Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA. yjtseng@nycu.edu.tw

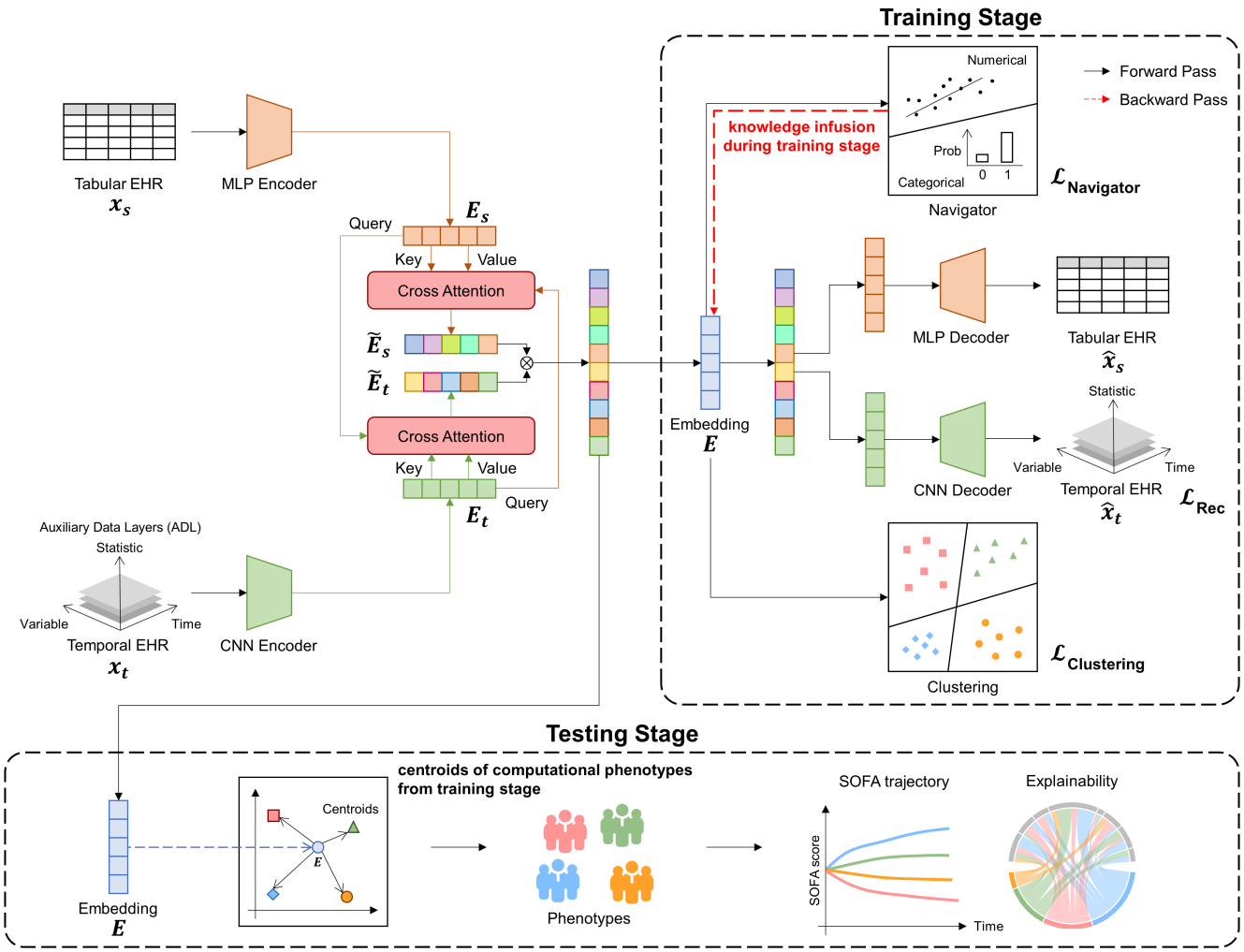

Fig. 1: The framework of naviDCN. Through the MLP and the CNN encoders, we turn tabular and temporal EHR into embeddings E_s and E_t. Using the cross-attention mechanism, we explore interactions between these two embeddings and output the embedding E for the downstream tasks. For the training stage, the model performs backpropagation according to $\mathcal{L}_{\text{rec}}$, $\mathcal{L}_{\text{clustering}}$, and $\mathcal{L}_{\text{navigator}}$. We can derive the embedding E with infused knowledge of the navigator and the computable phenotypes during the training stage. For the testing stage, we only input tabular and temporal EHR, without the navigator, to get the representative embedding. At last, we compute the distance between the embedding and each phenotype in the embedding space, then assign the patient to the nearest phenotype.

health records (EHR) from the first six hours after ICU admission and establish both static tabular EHR and temporal EHR representations. Through the cross-attention mechanism, we explore interactions between static and temporal EHR modalities, enabling the embedding to represent patient information comprehensively. After processing the input to the three downstream tasks of reconstruction, clustering, and navigator, we employ an alternating stochastic optimization algorithm, iteratively optimizing network weights of reconstruction tasks and navigator module, and updating cluster centroids of the clustering module. While optimizing the network weights of the navigator module, the navigator module infuses clinical knowledge into the embedding during backpropagation, guiding the clustering process and offering a possible direction in the embedding space.

A. *Multi-Modal Embedding*

Tabular EHR For static information, such as demographics and comorbidities, we directly employ raw variables from the EHR. Missing values for qualitative and quantitative variables are imputed with mode and mean, respectively. For time-varying information, such as laboratory results and vital signs in EHR, we implement feature aggregation. If there are multiple records, we choose the most abnormal values as the representative point. Missing values are imputed using MICE [9]. The tabular EHR $x_s \in \mathbb{R}^{n_s}$, where n_s is the number of tabular variables, is then sent to the MLP encoder, yielding embedding $E_s \in \mathbb{R}^d$ where d is the embedding dimension.

Temporal EHR Some variables, such as laboratory test results and vital signs, can change over time, with valuable insights embedded in the direction and magnitude of these

changes. To capture the clinical trajectory, we implement the sliding window approach, processing variables with non-overlapping hourly windows. If multiple values are recorded within a single window, we select the value of which timestamp is closest to the next window. Missing values are filled using last observation carried forward (LOCF) or overall mean if LOCF is not available. Additionally, the occurrence and frequency of measurements could reflect external information, such as the urgency of the situation. Thus, we also cover these statistics with the auxiliary data layers [10], resulting in the temporal EHR resembling an image-like, 3D form $x_t \in \mathbb{R}^{n_c \times n_w \times n_v}$ where n_c is the number of channels, n_w is the number of hourly windows, and n_v is the number of temporal variables. The channels encompass original values, missing indicators, which indicate if the variable is missing or not in the window, and the number of measurements of the variable in the window. The temporal EHR is then sent to the CNN encoder, yielding embedding $E_t \in \mathbb{R}^d$ where d is the embedding dimension.

Multi-Head Cross-Attention Mechanism After encoding the tabular and temporal EHRs into embeddings, naively applying fully connected layers is insufficient to describe the complex body systems completely. Therefore, we introduce the multi-head cross-attention mechanism, enabling comprehensive interaction between modalities.

$$\text{MultiHead}(Q, K, V) = [\text{Head}_1, ..., \text{Head}_h]W^O, \quad (1)$$

where $\text{Head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, h is the number of attention heads, $W^O \in \mathbb{R}^{d \times d}$, W_i^Q, W_i^K, $W_i^V \in \mathbb{R}^{d \times \frac{d}{h}}$ are weight matrices of query, key, value of i-th head cross-attention module.

For the tabular EHR, E_s is used as the key and value, while E_t serves as the query to produce the multi-head output $\tilde{E}_s$. To obtain $\tilde{E}_t$, we take E_t as the key and value, while E_s serves as the query for temporal EHR. This allows both EHRs to refine their representation by focusing on relevant information to each other.

$$\tilde{E}_s = \text{MultiHead}(E_t W_t^Q, E_s W_s^K, E_s W_s^V) \quad (2)$$

$$\tilde{E}_t = \text{MultiHead}(E_s W_s^Q, E_t W_t^K, E_t W_t^V) \quad (3)$$

After refining the static embedding $\tilde{E}_s$ and temporal embedding $\tilde{E}_t$ through the attention mechanism, we concatenate the two embeddings and pass them through the linear layer, compressing them into embedding $E \in \mathbb{R}^d$.

$$E = \text{Linear}(\tilde{E}_s \otimes \tilde{E}_t) \quad (4)$$

The resulting embedding captures not only the basic information from the two EHRs but also the multiple correlations between them.

B. Deep Clustering Network (DCN)

DCN uses an alternating stochastic optimization strategy, incorporating both reconstruction loss $\mathcal{L}_{\text{rec}}$ and clustering loss $\mathcal{L}_{\text{clustering}}$ to prevent the embedding from relying solely on reconstruction loss, which makes it less friendly for clustering because the embedding is not designed for clustering, or solely on clustering loss, which causes trivial solutions where all patient embeddings are alike [8].

$$\mathcal{L}_{\text{rec}} = \sum_{i=1}^{N} (\parallel x_s^i - \hat{x}_s^i \parallel_2^2 + \parallel x_t^i - \hat{x}_t^i \parallel_2^2), \quad (5)$$

where N is the number of patients, $\hat{x}_s$ is the reconstructed x_s, and $\hat{x}_t$ is the reconstructed x_t.

$$\begin{aligned} \mathcal{L}_{\text{clustering}} &= \sum_{i=1}^{N} \parallel E_i - M s_i \parallel_2^2 \\ \text{s.t. } & s_{j,i} \in \{0, 1\}, \mathbf{1}^T s_i = 1 \ \forall i, j, \end{aligned} \quad (6)$$

where $M \in \mathbb{R}^{k \times d}$ is the matrix including clustering centroids, k is the number of clusters, and s_i is the one-hot assignment vector.

With DCN, the embedding can retain the ability to represent the original information while forming clusters that are evenly distributed around their centroids in the feature space.

C. Knowledge Infusion with Navigator

When applying clustering approaches, the results are often unforeseen as these approaches rely on predefined features, leaving the discovery of underlying patterns to passive observation. These approaches lack the ability to allow features to learn external knowledge during clustering. However, we often aim for phenotypes that not only disclose unknown outcomes but also exhibit distinct distributions in key targets, such as important clinical outcome indicators. To address this, we introduce the concept of navigator. Through iteratively optimizing network weights of the reconstruction task and navigator module, and updating cluster centroids of the clustering module, the navigator enables the embedding to progressively learn key knowledge through backpropagation while simultaneously refining the composition of phenotypes. When the embedding contains more important knowledge, the clustering results can uncover more relevant heterogeneity among phenotypes. Here, we present examples of the navigator's functionality by the categorical and numerical navigators.

Categorical Navigator We pass the embedding through the linear layer, outputting probabilities for each category.

$$p = \text{softmax}(W^T E + b), \quad (7)$$

where $W \in \mathbb{R}^{d \times c}$ and $b \in \mathbb{R}^c$ are learnable parameters, and c is the number of categories. These probabilities are then compared with ground truth using the loss function:

$$\begin{aligned} \mathcal{L}_{\text{navigator}} &= -\frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{c} w_i (1 - p_{t,i}^j)^2 \log(p_{t,i}^j), \\ p_{t,i}^j &= \begin{cases} p_i^j & \text{if } y^j = i \\ 1 - p_i^j & \text{otherwise,} \end{cases} \end{aligned} \quad (8)$$

where w_i is the weight hyperparameter and p_i is the probability of category i.

Numerical Navigator For the numerical target, we pass the embedding through the linear layer to get the estimation

of the target, which is compared with the ground truth using the loss function:

$$\mathcal{L}_{\text{navigator}} = \frac{1}{n}\sum_{i=1}^{N}(y_i - \hat{y_i})^2, \tag{9}$$

where y_i is the ground truth and $\hat{y_i}$ is the estimation.

Ultimately, the overall loss function of our model is as follows:

$$\mathcal{L} = \lambda_1 * \mathcal{L}_{\text{rec}} + \lambda_2 * \mathcal{L}_{\text{clustering}} + \lambda_3 * \mathcal{L}_{\text{navigator}}, \tag{10}$$

where λ_1, λ_2, and λ_3 are hyperparameters to balance three components. Then, we leverage $\mathcal{L}$ as the optimization objective to adjust the network weights during backpropagation. To update cluster assignments and centroids, we follow the same procedure as in DCN [8]. Through the navigator, clinical knowledge is imparted to the embedding during backpropagation, guiding clustering in an external knowledge-driven direction.

III. RESULTS

A. Dataset

Our dataset is derived from MIMIC-IV v2.2 [11], with sepsis patients identified by the Sepsis-3 criteria. We exclude patients with ICU stays shorter than 24 hours and those whose sepsis onset occurs more than 24 hours before or after ICU admission. Multiple hospitalizations from the same patient are included as independent records if the gap between two hospitalizations is greater than 14 days, resulting in 19,834 hospitalizations. We include 65 variables [4], including demographics, comorbidities, laboratory test results, and vital signs. Comorbidities are determined based on diagnostic codes from prior hospitalizations. Other variables are collected within the first six hours of ICU stay.

B. Implementation Details

We randomly split the dataset into the training set with 80% of them and the testing set with the remaining 20%. During the training stage, we incorporate the navigator to guide the embedding in acquiring clinical knowledge and form the cluster centroids. During the testing stage, without the navigator, the input includes only records of the patient's first six hours after ICU admission to encode the embedding. Patients are then assigned to the nearest phenotype based on the cluster centroids obtained from the training stage. In subsequent experiments, except for the characteristics of the four computable phenotypes, we present results in the testing set to assess the performance of naviDCN in reality where the navigator is not available in the early stage.

For the model settings, the attention module has 4 heads and the embedding dimension is 32. We train our framework 50 epochs with a batch size of 256. The optimizer is Adam with learning rate of 0.01 and weight decay of 5e-7. We set $\lambda_1 = 1$, $\lambda_2 =$ 1e-4, and $\lambda_3 = 1$. We cluster sepsis patients into four phenotypes, following the previous studies [4]–[6]. The discharged status is the main navigator, and ICU LOS is set as an alternative navigator for further evaluation.

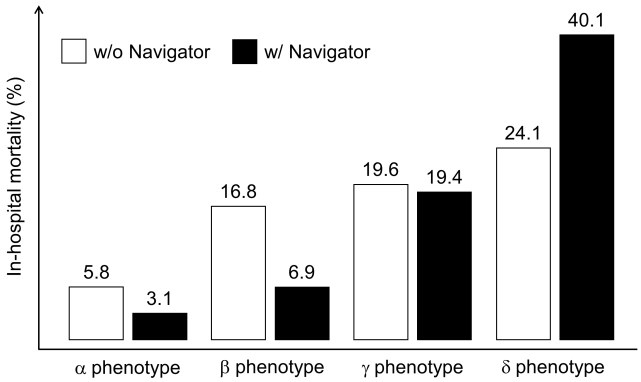

Fig. 2: The in-hospital mortality of computable phenotypes of sepsis in the testing set with and without the discharged status navigator.

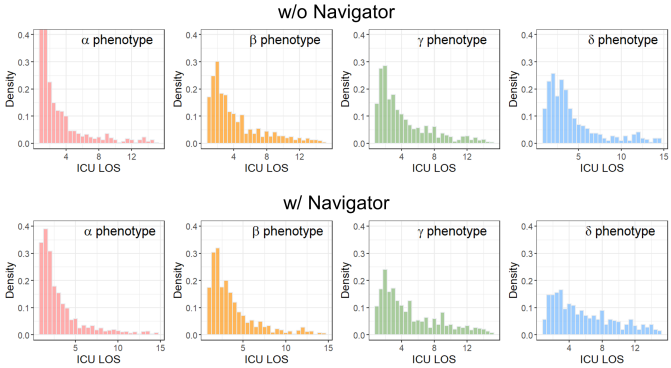

Fig. 3: The ICU LOS of computable phenotypes of sepsis in the testing set with and without the ICU LOS navigator.

C. Sepsis Phenotypes

Patient Characteristics Across Phenotypes The characteristics of the 4 computable phenotypes in the full dataset appear in **Table 1**. The clustering results present four phenotypes α, β, γ, and δ ranging from 13.7% to 34.5% in the study population.

After clustering, the four phenotypes show different outcomes. The α phenotype has almost no in-hospital deaths and it also has the shortest ICU and hospital stays and the lowest 365-day outpatient mortality among the four phenotypes. The δ phenotype has the highest in-hospital mortality, which is 46%, and 75% of patients in this phenotype died at discharge or 365 days after discharge.

Navigator To test naviDCN, we evaluate the in-hospital mortality rates in the testing set across phenotypes with and without the inclusion of the discharged status navigator in the training stage (Figure 2). Before applying the navigator, the mortality rates range from 5.8% to 24.1% in the testing set. After including discharged status as the navigator in the training stage, the rates in the testing set range from 3.1% to 40.1%, with the most notable difference observed in the δ phenotype, which rose from 24.1% to 40.1%. Furthermore, the ICU LOS distributions for β, γ, and δ phenotypes are highly similar without the ICU LOS as a navigator in the

TABLE I: Characteristics of the 4 computable phenotypes of sepsis.

Characteristic	Total (n = 19834)	α (n = 2715)	β (n = 6849)	γ (n = 6696)	δ (n = 3574)
Age, mean (SD)	64.7 (15.8)	61.2 (14.6)	60.3 (16.3)	67.8 (14.5)	70.0 (14.5)
Sex, No. (%)					
Male	11763 (59.3)	1835 (67.6)	4192 (61.2)	3568 (53.3)	2168 (60.7)
Female	8071 (40.7)	880 (32.4)	2657 (38.8)	3128 (46.7)	1406 (39.3)
Elixhauser comorbidities, median (IQR)	0 [0, 5]	0 [0, 0]	0 [0, 4]	2 [0. 6]	4 [0, 7]
SOFA score, mean (SD)	7 [4, 10]	8 [5, 9]	6 [3, 8]	6 [4, 9]	10 [6, 13]
Inflammation					
Temperature, mean (SD)	37.0 (0.9)	36.8 (0.7)	37.1 (0.8)	37.1 (0.8)	36.8 (1.0)
Bands, median (IQR)	3.0 [1.4, 5.6]	2.4 [1.2, 4.4]	2.8 [1.4, 5.2]	3.0 [1.6, 5.8]	3.8 [1.8, 7.0]
CRP, median (IQR)	81 [47, 121]	46 [28, 74]	72 [43, 112]	94 [59, 131]	104 [72, 137]
ESR, median (IQR)	49.0 [36.0, 63.6]	47.6 [35.4, 59.9]	47.8 [35.0, 62.6]	49.6 [36.0, 64.4]	51.6 [38.2, 66.4]
WBC, median (IQR)	12.8 [9.2, 17.3]	13.0 [9.7, 16.8]	12.2 [8.8, 16.4]	12.8 [9.2, 17.2]	14.1 [9.7, 19.9]
Neutrophil, median (IQR)	81.3 [75.0, 86.4]	78.2 [73.0, 82.9]	81.0 [74.9, 85.9]	82.6 [76.4, 87.6]	82.3 [74.9, 87.6]
Metamyelocytes, median (IQR)	0.4 [0.2, 1.0]	0.4 [0.2, 0.8]	0.4 [0.2, 1.0]	0.4 [0.2, 1.0]	0.6 [0.2, 1.4]
Myelocytes, median (IQR)	0.2 [0.0, 0.4]	0.2 [0.0, 0.4]	0.2 [0.0, 0.4]	0.2 [0.0, 0.4]	0.2 [0.0, 0.6]
Promyelocytes, median (IQR)	1.2 [1.0, 1.6]	1.2 [1.0, 1.6]	1.2 [1.0, 1.6]	1.2 [1.0, 1.6]	1.2 [1.0, 1.6]
Lymphocytes, median (IQR)	11.0 [7.0, 16.4]	15.6 [11.3, 20.4]	11.8 [7.9, 17.0]	9.7 [6.0, 14.2]	9.0 [5.2, 14.2]
Pulmonary					
SaO2, median (IQR)	84.2 [73.8, 93.0]	85.0 [76.0, 94.6]	85.8 [75.6, 93.4]	83.8 [73.0, 92.2]	81.4 [68.0, 91.2]
PO2, mean (SD)	101 (75)	143 (80)	109 (76)	89 (68)	78 (63)
FiO2, mean (SD)	76.5 (21.9)	88.5 (19.3)	75.7 (21.7)	72.4 (21.1)	76.5 (22.6)
Respiratory rate, mean (SD)	25.2 (7.6)	21.2 (5.4)	24.0 (6.2)	26.4 (6.9)	28.3 (10.2)
Cardiovascular					
Bicarbonate, mean (SD)	21.8 (5.3)	23.1 (3.4)	22.3 (4.8)	21.7 (5.6)	19.9 (6.3)
Heart rate, mean (SD)	98 (21)	89 (15)	95 (19)	100 (21)	106 (23)
Lactate, median (IQR)	2.3 [1.6, 3.5]	2.3 [1.8, 2.9]	2.1 [1.5, 3.0]	2.3 [1.6, 3.5]	3.3 [2.1, 6.1]
SBP, median (IQR)	94 [85, 107]	94 [86, 104]	97 [87, 110]	94 [84, 107]	90 [80, 100]
Troponin, median (IQR)	0.2 [0.1, 0.4]	0.2 [0.1, 0.5]	0.2 [0.1, 0.4]	0.2 [0.1, 0.4]	0.2 [0.1, 0.5]
Renal					
BUN, median (IQR)	24 [16, 40]	16 [12, 20]	20 [14, 31]	30 [19, 47]	37 [23, 58]
Creatinine, median (IQR)	1.2 [0.8, 2.0]	0.9 [0.7, 1.1]	1.1 [0.8, 1.6]	1.4 [0.9, 2.3]	1.7 [1.1, 2.7]
Hepatic					
AST, median (IQR)	66 [36, 143]	74 [44, 138]	62 [35, 127]	62 [33, 138]	80 [38, 215]
ALT, median (IQR)	42 [24, 92]	48 [29, 91]	41 [24, 85]	40 [22, 89]	46 [24, 124]
Bilirubin, median (IQR)	0.9 [0.5, 1.7]	1.0 [0.6, 1.6]	0.9 [0.5, 1.6]	0.9 [0.5, 1.8]	1.0 [0.5, 2.2]
Hematologic					
Hemoglobin, mean (SD)	10.1 (2.4)	9.3 (2.2)	10.2 (2.4)	10.2 (2.3)	10.0 (2.4)
INR, median (IQR)	1.4 [1.2, 1.7]	1.4 [1.2, 1.5]	1.3 [1.2, 1.5]	1.4 [1.2, 1.8]	1.5 [1.2, 2.2]
Platelets, median (IQR)	174 [120, 241]	143 [112, 187]	177 [124, 243]	188 [127, 258]	175 [110, 251]
Neurologic					
GCS, mean (SD)	9.2 (4.9)	6.0 (4.7)	9.8 (4.8)	10.4 (4.4)	8.2 (4.8)
Others					
Albumin, mean (SD)	3.2 (0.5)	3.3 (0.5)	3.3 (0.5)	3.2 (0.5)	3.0 (0.6)
Chloride, mean (SD)	105 (7)	109 (5)	106 (7)	104 (8)	104 (9)
Glucose, median (IQR)	154 [121, 200]	163 [138, 191]	150 [119, 192]	149 [118, 200]	163 [121, 236]
Sodium, mean (SD)	139 (6)	139 (3)	139 (5)	139 (7)	139 (7)
Outcomes					
In-hospital mortality, No. (%)	3159 (15.9)	16 (0.6)	285 (4.2)	1204 (18.0)	1654 (46.3)
365-day outpatient mortality, No. (%)	5355 (27.0)	307 (11.3)	1670 (24.4)	2339 (34.9)	1039 (29.1)
Length of stay in the ICU, median (IQR)	3.0 [1.8, 5.9]	1.9 [1.3, 3.1]	2.8 [1.7, 5.2]	3.6 [2.1, 7.0]	4.1 [2.3, 8.2]
Length of stay in the hospital, median (IQR)	8.3 [5.2, 14.4]	6.2 [4.9, 9.2]	8.1 [5.3, 13.8]	9.8 [5.9, 16.5]	8.9 [4.4, 16.0]

training stage. After using the navigator, the four phenotypes show clearly distinct ICU LOS distributions in the testing set (Figure 3).

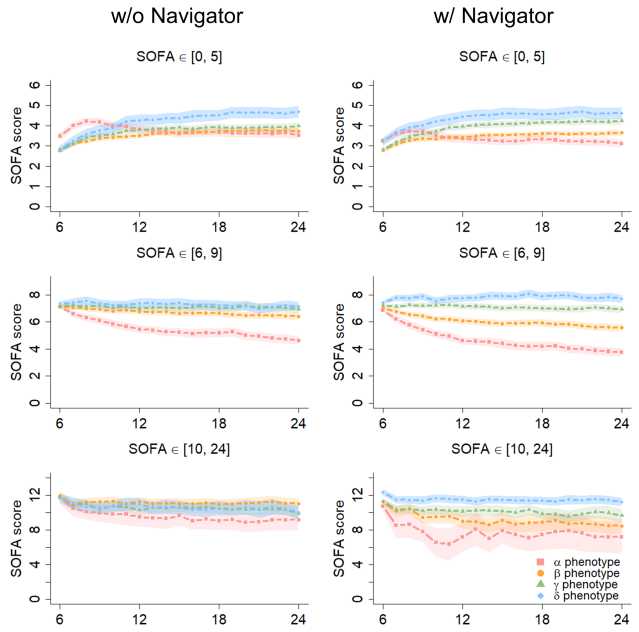

Fig. 4: The SOFA trajectories by phenotypes in the next 18 hours after clustering in the testing set. The patients are stratified by the SOFA score at the 6th hour.

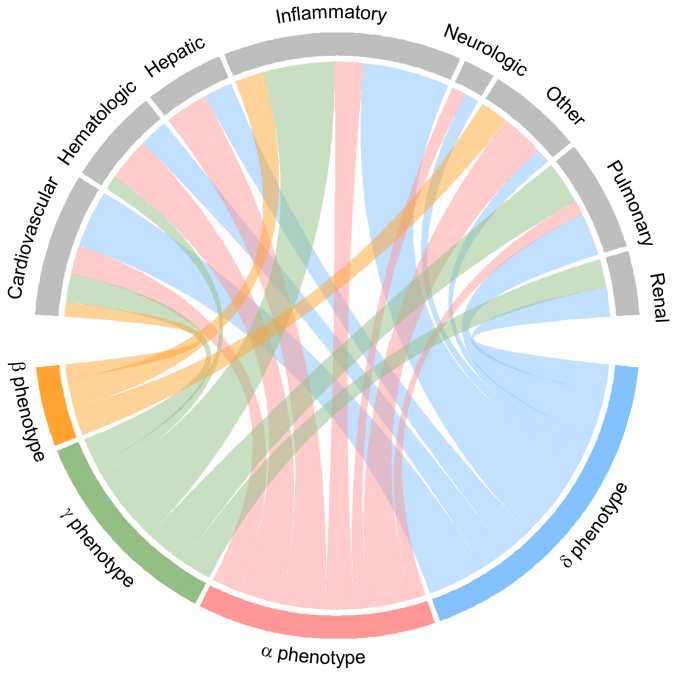

Fig. 5: Chord diagram showing abnormal clinical variables by the computable phenotype of sepsis. The ribbon connects from a phenotype to an organ system if the group median is more abnormal than the overall median for the testing set.

SOFA Trajectories We then further explore the short-term changes in organ dysfunctions across four phenotypes in the testing set. Patients are first stratified by SOFA score at the sixth hour after ICU admission to control the beginning severity. We observe the average SOFA score trajectories for each phenotype over the following 18 hours (the 7th to the 24th hours after the ICU admission). Figure 4 shows SOFA trajectories with and without the discharged status navigator. Compared to the results without the discharged status navigator, the navigator-assisted results have significantly different trajectory patterns across the four phenotypes in the testing set. The δ phenotype exhibits rising or consistently high SOFA scores over the next 18 hours. On the other hand, the α phenotype demonstrates a fast decline, returning to lower SOFA scores. The dynamic SOFA trajectories provide valuable insight into the short-term changes in the organ dysfunctions across different phenotypes, helping to characterize each phenotype's disease progression.

Explainability We group the 32 laboratory tests and vital signs into eight organ systems, which are inflammatory, pulmonary, cardiovascular, renal, hepatic, hematologic, neurologic, and other systems, following the previous research [4]. Figure 5 highlights the systems where each phenotype exhibits the abnormalities in the testing set. The α and δ phenotypes are both abnormal across nearly all systems. However, their clinical outcomes are completely different. Only 3.1% of patients with the α phenotype were discharged by death in the testing set (Figure 2), with SOFA trajectories showing declines over the following 18 hours (Figure 4). Conversely, the δ phenotype had the highest mortality (40.1%), with SOFA scores remaining elevated over time. This finding remains consistent across the entire dataset.

Internal Metrics We assess the clustering performance with three common internal metrics, including the Silhouette Index (SI), Calinski-Harabasz Index (CHI), and Davies-Bouldin Index (DBI). Higher SI and CHI values, alongside a lower DBI, indicate better clustering quality. Since CHI is affected by the dataset size, the value of CHI is divided by the number of patients [6]. The SI, CHI, and DBI are 0.434, 1.489, and 1.008 for the proposed naviDCN in the testing set, while the three internal metrics are 0.475, 2.203, and 0.803 for the original DCN. The two-sample t-test shows no statistically significant differences between DCN and our framework on the three internal metrics. The p-values are 0.424, 0.262, and 0.086, respectively.

IV. DISCUSSION

Integrating the navigator with clustering is a key component of the proposed naviDCN, improving the alignment of computable phenotypes with important clinical outcomes. Without introducing the navigator, the SOFA trajectories of the four phenotypes show overlap, and the differences in mortality and ICU LOS distributions are unclear. With the navigator, we can easily see the differences in mortality we mainly pilot, while SOFA trajectories, which represent the unknown heterogeneity, also show distinct directions. Moreover, this enhancement doesn't compromise clustering internal metrics, maintaining the consistency of phenotypes.

The naviDCN addresses three critical issues in prior studies. First, considering only measurement values may neglect crucial statistical information, such as the frequency of measurements, which could imply the severity status. To address this, we construct an image-like EHR representation to preserve such information [10]. Next, clustering algorithms treat variables as independent entities, neglecting their interactions and relationships. We apply an attention mechanism to generate embeddings that capture correlations of the raw variables. Most importantly, existing clustering approaches lack mechanisms for incorporating external knowledge with autoencoder, potentially leading to suboptimal results. Some studies introduce the prediction task as the domain knowledge into the clustering framework by incorporating their loss [7]. However, our navigator module allows the embedding to infuse key knowledge, thereby ensuring phenotypes exhibit distinct distributions across key knowledge while simultaneously revealing unknown outcomes. The current selection of navigators is based on their clinical significance. Future studies may choose different navigators depending on their specific objectives.

The SOFA score offers insights into organ dysfunction and its progression over time. Studies have shown a strong association between maximum SOFA scores and mortality, with even minor fluctuations reflecting significant shifts in patient prognosis [12]. Moreover, identifying phenotypes through organ dysfunction trajectories has implications for the predictive enrichment of clinical trials. These phenotypes may reflect underlying differences in pathophysiology, making it possible to tailor treatment strategies by identifying patients who may benefit from specific therapeutic interventions or require more intensive monitoring [5]. Our framework effectively distinguishes patients with divergent SOFA trajectories. Within each stratification of SOFA scores, the four phenotypes begin with similar SOFA scores. However, their trajectories diverge over the subsequent 18 hours, resulting in distinct outcomes. These findings underscore the capacity of our approach to inform therapeutic interventions, thereby enabling more effective treatment strategies.

Our framework has limitations. First, the current encoder architecture cannot process complete patient records directly. We aggregate features in the whole period or within the hourly window, which may result in missing valuable information. Other sequence-based networks, such as Transformer [13], can be further chosen to retain the records more completely by regarding irregular temporal EHR as sequences. Nevertheless, since most data points in MIMIC-IV are recorded at hourly intervals, the temporal aggregation within hourly windows is unlikely to impact the results of this study significantly. Moreover, our framework only clusters once at the sixth hour after ICU admission, which lacks the ability to dynamically update clustering results in response to real-time changes in patient conditions. Future work can explore transforming naviDCN into a dynamic monitoring system that adjusts phenotype assignments as patient status evolves, thereby enhancing its applicability. Finally, the lack of external validation restricts the generalizability of our findings. Other datasets with different information, such as biomarkers or pathogens, or from different regions, can help ensure the robustness of our framework.

V. CONCLUSION

We propose **naviDCN**, a deep clustering framework integrating the navigator, infuses target knowledge into the clustering process. The navigator not only allows for clinically meaningful adjustments in clustering but also enhances the interpretability of phenotypes. This approach identifies four sepsis phenotypes in early ICU admission, each with distinct SOFA trajectories and mortality outcomes. After uncovering the heterogeneity of sepsis, further research can explore the relationship between phenotypes and treatments to form more precise treatment strategies for different phenotypes.

ACKNOWLEDGMENT

This study was supported by grants from the National Science and Technology Council, Taiwan (NSTC 111-2628-E-A49-026-MY3), the Higher Education Sprout Project of the National Yang Ming Chiao Tung University and MOE, Taiwan (CGMH-NYCU-113-CORPG2P0071), and Chang Gung Memorial Hospital (CORPG2P0071).

REFERENCES

[1] M. Singer, C. S. Deutschman, C. W. Seymour *et al.*, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *Jama*, vol. 315, no. 8, pp. 801–810, 2016.

[2] K. E. Rudd, S. C. Johnson, K. M. Agesa *et al.*, "Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study," *The Lancet*, vol. 395, no. 10219, pp. 200–211, 2020.

[3] E. J. Giamarellos-Bourboulis, A. C. Aschenbrenner, M. Bauer *et al.*, "The pathophysiology of sepsis and precision-medicine-based immunotherapy," *Nature immunology*, vol. 25, no. 1, pp. 19–28, 2024.

[4] C. W. Seymour, J. N. Kennedy, S. Wang *et al.*, "Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis," *Jama*, vol. 321, no. 20, pp. 2003–2017, 2019.

[5] Z. Xu, C. Mao, C. Su *et al.*, "Sepsis subphenotyping based on organ dysfunction trajectory," *Critical Care*, vol. 26, no. 1, p. 197, 2022.

[6] C. Yin, R. Liu, D. Zhang, and P. Zhang, "Identifying sepsis subphenotypes via time-aware multi-modal auto-encoder," in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020, pp. 862–872.

[7] Y. Ren, J. Pu, Z. Yang *et al.*, "Deep clustering: A comprehensive survey," *IEEE transactions on neural networks and learning systems*, 2024.

[8] B. Yang, X. Fu, N. D. Sidiropoulos, and M. Hong, "Towards k-means-friendly spaces: Simultaneous deep learning and clustering," in *international conference on machine learning*. PMLR, 2017, pp. 3861–3870.

[9] S. Van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *Journal of statistical software*, vol. 45, pp. 1–67, 2011.

[10] K.-H. Liu, C.-Y. Chiang, H.-Y. Wang, and Y.-J. Tseng, "Temporal phenotype matrix engineering for electronic health records–enhancing coronary artery disease prediction," in *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2023, pp. 1–4.

[11] A. E. Johnson, L. Bulgarelli, L. Shen *et al.*, "Mimic-iv, a freely accessible electronic health record dataset," *Scientific data*, vol. 10, no. 1, p. 1, 2023.

[12] S. Lambden, P. F. Laterre, M. M. Levy, and B. Francois, "The sofa score—development, utility and challenges of accurate assessment in clinical trials," *Critical Care*, vol. 23, pp. 1–9, 2019.

[13] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.