Patient Clustering for Vital Organ Failure Using ICD Code With Graph Attention

Zhangdaihong Liu, Ying Hu, Xuan Wu, Gert Mertes, Yang Yang, and David A. Clifton

Abstract*—*Objective: **Heart failure, respiratory failure and kidney failure are three severe organ failures (OF) that have high mortalities and are most prevalent in intensive care units. The objective of this work is to offer insights into OF clustering from the aspects of graph neural networks and diagnosis history.** ***Methods:*** **This paper proposes a neural network-based pipeline to cluster three types of organ failure patients by incorporating embedding pre-train using an ontology graph of the International Classification of Diseases (ICD) codes. We employ an autoencoder-based deep clustering architecture jointly trained with a K-means loss, and a non-linear dimension reduction is performed to obtain patient clusters on the MIMIC-III dataset.** ***Results:*** **The clustering pipeline shows superior performance on a public-domain image dataset. On the MIMIC-III dataset, it discovers two distinct clusters that exhibit different comorbidity spectra which can be related to the severity of diseases. The proposed pipeline is compared with several other clustering models and shows superiority.** ***Conclusion:*** **Our proposed pipeline gives stable clusters, however, they do not correspond to the type of OF which indicates these OF share significant hidden characteristics in diagnosis. These clusters can be used to signal possible complications and severity of illness and aid personalised treatment.** ***Significance:*** **We are the first to apply an unsupervised approach to offer insights from a biomedical engineering perspective on these three types of organ failure, and publish the pre-trained embeddings for future transfer learning.**

***Index Terms*—Artificial neural networks, clustering methods, graph attention, ICD ontology, organ failure.**

I. Introduction

ORGAN failure (OF) is the main reason for admitting patients to Intensive Care Units (ICU) and the main cause of death in ICU [1], [31]. The mortality rate remains high for OF patients and is significantly higher for patients with multiple OFs [3]. [2] showed that the most common hospital admission diagnosis of patients that had unplanned transfer to ICU was heart failure (HF) (12%). Moreover, HF is reported to affect over 26 million people globally and has a growing prevalence, especially with an ageing population [29]. The most common diagnosis of unplanned ICU transfers was respiratory failure (RF) (27%). It also has the highest incidence rate in ICU and is associated with high short-term mortality and long ICU stays [27]. A population-based cohort study showed that kidney failure (KF) had the highest one-year mortality rate (18.2%) among all OFs that were investigated [28]. The overall mortality of acute KF is around 20%, rising to over 50% for patients who require dialysis [23].

Patients with these OFs have very poor quality of life, and the cost burden of these OFs results in huge health expenditures for countries [32]. It is crucial for a nation's health system to better understand the underlying relationships between these OFs so that precautionary measures can be taken and early intervention can be achieved more effectively to improve the mortality and treatment. However, identifying patients with vital organ failure timely and correctly can be challenging due to the broad diagnosis associated with presenting symptoms and variations in patient presentations. Furthermore, the pathological and clinical complexities of those OFs are high.

Electronic health records (EHR) store rich information of patients' hospital admissions including medical histories, demographics, and symptoms, etc. The International Classification of Diseases (ICD) code is a globally used clinical tool for recording patients' diagnosis and procedures undergone in hospitals and is widely stored (with little missingness) in most EHR systems. The ninth revision of ICD contains over ten thousand different codes [7]. These diagnostic codes can be collapsed into a smaller number of clinically meaningful concepts to form an ontology.

Manuscript received 20 January 2022; revised 18 June 2022, 1 October 2022, 14 October 2022, and 21 January 2023; accepted 30 January 2023. Date of publication 8 February 2023; date of current version 19 July 2023. The work of Zhangdaihong Liu was supported by the Jiangsu Provincial Double Innovation Talent Programme. The work of Ying Hu was supported by the Shanghai Municipal Health Commission General Project under Grant 202040083. Yang Yang was supported by the Startup Fund for Young Faculty at SJTU under Grant BJ1-3000-22-0066. David A. Clifton was supported by the NIHR Oxford Biomedical Research Centre, an NIHR Research Professorship, an RAEng Research Chair, the InnoHK Centre for Cerebro-cardiovascular Engineering, and the Pandemic Sciences Institute at the University of Oxford. *(Corresponding author: Yang Yang.)*

Zhangdaihong Liu is with the Oxford-Suzhou Centre for Advanced Research, China, and also with the Department of Engineering Science, University of Oxford, U.K.

Ying Hu is with the Shanghai Xuhui Central Hospital, Xuhui Hospital, Fudan University, China.

Xuan Wu is with the National Key Laboratory for Novel Software Technology, Nanjing University, China.

Gert Mertes is with the Nuffield Department of Population Health and the Department of Engineering Science, University of Oxford, U.K.

Yang Yang is with the School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China (e-mail: emma002@sjtu.edu.cn).

David A. Clifton is with the Department of Engineering Science, University of Oxford, U.K., and also with the Oxford-Suzhou Centre for Advanced Research (OSCAR), University of Oxford, China.

This article has supplementary downloadable material available at https://doi.org/10.1109/TBME.2023.3243311, provided by the authors.

Digital Object Identifier 10.1109/TBME.2023.3243311

Such structured ontology trees can help us to present more descriptive statistics for easier analysis and interpretation [6]. One popular ontology is created by the Clinical Classifications Software (CCS) [36]. The ICD codes with closer relationships are likely to fall under the same lower-level parent medical concept. Such ontology structure may help machine learning models to better learn the representation of rare conditions and thus, boost model performance.

The diagnosis ICD code is one of the most accessible and standardised modalities in EHR systems and the diagnosis history records rich and vital information for patients' health status. Moreover, it is the reference for or is related to many other medical modalities such as medication and procedure. A large amount of literature has shown the power of diagnosis information in various clinical tasks [5], [17], [48]. Thus, in this paper, we focus on using the diagnostic information to offer insights into OF patient clustering, an unsupervised task that no research has attempted. We hypothesise that the disease complexities are embedded in the ICD codes assigned to patients during their hospital visits.

In this work, we adopted the attention mechanism and ICD initialisation approach proposed in [6], in which they showed superiority of applying such mechanism to the ICD ontology in prediction tasks. Additionally, we turned the supervised prediction task into an unsupervised setting for OF patient clustering by employing an auto-encoder (AE) based model architecture. We applied this pipeline to the MIMIC-III dataset on patients with the aforementioned OFs, in order to learn patient groups from the diagnosis histories and having more insights on these OFs from an unsupervised point of view. The experiment pipeline can be divided into three stages: (1) pre-train ICD embeddings from the ontology tree; (2) pre-train an AE embedded with attentions and with layer-wise construction loss only; (3) joint-train the AE with a clustering loss added; (4) apply UMAP to further improve the clustering performance.

The main contributions of this work include: (1) introducing an OF patient clustering pipeline, where the inputs are ICD embeddings pre-trained with ontology; (2) we discovered two distinct clusters that can be used for complication signalling and potentially are related to disease severity and aid personalised treatment; (3) we publish the pre-trained ICD embeddings[1] which have strong power in identifying OF types with supervised learning. To our knowledge, we are the first to apply pre-trained ICD embeddings to cluster OF patients. The clustering pipeline composition is novel for this biomedical engineering task. We are also the first to publish the pre-trained ICD embeddings for the convenience of future transfer learning.

II. Related Work

ICD embedding learning: Learning embeddings for ICD codes, which are typically represented by dense vectors, using machine learning methods has been a popular research topic [14], [22]. The learnt embeddings are often used as features for supervised tasks such as predictions and classifications since they contain rich information for patients' medical histories. Natural language processing (NLP) techniques are suitable tools to aid the learning because ICD codes are often contained in free text parts of the EHR (e.g. charted clinician notes). [34] used long short-term memory to automatically perform ICD coding given the diagnosis descriptions; [16] combined convolutional neural network and 'Document to Vector' to achieve text multi-label classification and automated ICD coding; in recent work [18] used state-of-the-art NLP model BERT to joint learn embeddings for ICD and age to predict diseases.

ICD ontology and graph neural network: The ICD ontology graph is beneficial for ICD embedding learning. There are over 10,000 codes in the ICD Ninth Revision and significantly more in the later revisions. These codes have clinical hierarchies which can be for better understanding of the relationships between diseases and easier analysis. There are a few widely-accepted ontology schemes such as the one mentioned above (CCS) and SNOMED-CT [24]. Incorporating the ontology tree into ICD embedding learning could enhance the relationships between ICD codes and help to learn embeddings for rare codes. [6] were the first to embed ICD ontology graph into deep neural networks to predict diseases. Later, [33] also employed this ICD ontology graph and updated the attention training to improve the ICD embedding learning. Finally, they used BERT for medication recommendation.

Deep Clustering: Clustering is an unsupervised method in machine learning, and has been a fundamental tool to learn data structures in an exploratory fashion when no label is given. Since the development of DEC (Deep Embedded Clustering [37]) which combines deep neural network with clustering, deep clustering algorithms have drawn much attention from machine learning researchers. Many deep clustering methods have emerged and they can be categorised into different types based on the loss function. The loss function generally consists of a network loss to learn the latent representations and a clustering loss applied to these representations to achieve the clustering goal [21]. The network loss determines the architecture of the neural network. AE (with reconstruction loss) is the most common architecture for deep clustering models. DEC, DCN (Deep Clustering Network) [38], SR-K-means (Soft Regularized K-means) [9] and K-Autoencoders [44] all adopted this architecture. There are also generative models such as Generative Adversarial Network and Variational AE that are used as the network architecture [10], [45], [46]. There are also methods such as JULE (Joint Unsupervised Learning) [39] and DAC (Deep Adaptive Image Clustering) [4] which are CNN-based and involve only a well-designed clustering loss to extract discriminative features to achieve clustering purpose specifically for images. The clustering can be achieved by using an AE with reconstruction loss only [30]. Those with an additional clustering loss, which is normally added after a pre-train of the network, can simultaneously preserve the structure of the data and achieve clustering. Variations of K-means and KL divergence are most widely applied clustering losses in the discriminative and generative network setting respectively.

[1] https://github.com/lzdh/MIMIC-III-ICD9-Pretrained-Embeddings

TABLE I
DATA SUMMARY TABLE. 2216 IS THE NUMBER OF PATIENTS WITH ONLY ONE ORGAN FAILURE AND HAS AT LEAST TWO HOSPITAL VISITS

	HF	RF	KF	Total
Number of patients	889	523	804	2216
Number of ICD codes/medical concepts	3266/729			

Many of the aforementioned methods showed competitive performance compared with supervised tasks on public-domain datasets such as MNIST [15]. However, we found little work that applies the above techniques to medical applications, and much less to cluster OF patients.

III. DATA

MIMIC-III [11] is a public dataset which contains around 60,000 ICU admissions and over 650,000 diagnoses, recorded using International Classification of Diseases, Ninth Revision (ICD-9). Organ failures are often the main reasons to admit a patient to ICU, which makes MIMIC-III an ideal dataset for our analysis.

To identify the patients with HF, RF and KF, we carefully selected the following ICD codes: all end-level ICD codes under 428 (HF), 518.81 (acute RF), 518.83 (chronic RF), 518.84 (acute and chronic RF), 518.51 (acute RF following trauma and surgery), 518.53 (acute and chronic RF following trauma and surgery), 770.84 (RF of newborn), 584 (acute KF), 669.3 (acute KF following labour and delivery) and 586 (renal failure). In total there are 24 ICD codes. In MIMIC-III, a single patient may have multiple visits (admissions). We included all visits of a patient. Notably, the median and 95th percentile for the number of visits are 2 and 4 respectively.

The data summary is shown in Table I. We selected all patients with one of the three OFs. It is clinically interesting to investigate patients with distinct OFs, and as a consequence of this, the task is simplified. Furthermore, to better learn from the diagnostic history and reduce the randomness/noise in the data, we removed patients with fewer than two hospital visits. This yielded 2216 unique patients in this study, and we named this cohort the *target cohort*. The data pre-processing pipeline is shown in Supplementary Appendix Fig. 11.

For the ICD ontology, we adopted the scheme created by the Clinical Classifications Software (CCS) [36] which is widely recognised and applied in the literature. There are single-level (ICD codes have only one overall category) and multi-level (ICD codes have hierarchical ancestral categories) CCS categories. We used multi-level CCS to construct the ontology tree. In the ontology tree, the leaf nodes are the billable ICD codes that are actually stored in the EHR system and the upper level ancestors are more general medical concepts. This is illustrated in Fig. 1. For example, 'Heart Failure' has three ancestors (apart from the root node which is shared by all ICD codes). Both 'Heart Failure' and 'Atrial Fibrillation' are end-level ICD codes. They are under a different 'parent' node, but belonging to the same 'grandparent' node. In this study, there are 3266 unique end-level ICD codes appeared in the diagnoses of the OF patients, and they have 729 ancestor nodes (medical concepts) in the CCS ontology tree.

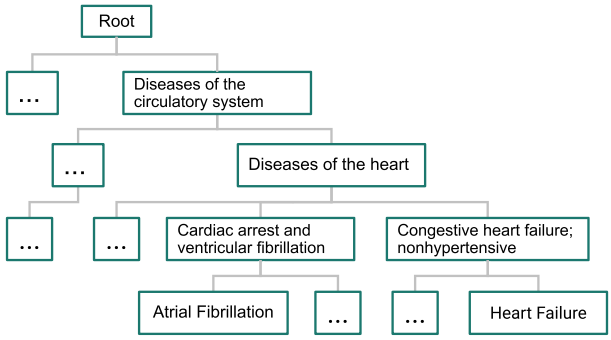

Fig. 1. A snapshot of the CCS ontology tree.

IV. METHODS

A. Analysis Pipeline

Fig. 2 A lists the key components from in the clustering pipeline and Fig. 2 B shows the details of the end-to-end deep clustering model. The architecture is based around an AE that learns the latent representations of the model input. Clustering is achieved by adding a clustering loss to the bottleneck latents of the AE.

To incorporate ICD ontology into the pipeline, we adopted the approach proposed in GRAM [6]. In brief, [6] imposed an attention mechanism to the ontology tree to establish connections between the leaf ICD nodes and their ancestor medical concepts. The goal is to construct an embedding matrix E for the leaf ICD nodes where each embedding in E is a weighted linear combination of the ancestor embeddings and the original ICD embedding itself. More specifically, let us assume a randomly initialised embedding matrix G containing the initial embedding for all nodes in the ontology tree. The 'attended' embedding matrix E contains embeddings for ICD codes only and is constructed by 1.

$$e_i = \sum_{j \in \mathcal{A}(i)} \alpha_{ij} g_j, \quad \text{where} \quad \alpha_{ij} = \frac{\exp(f(g_i, g_j))}{\sum_{k \in \mathcal{A}(i)} \exp(f(g_i, g_k))}, \tag{1}$$

where e_i is the ith column in E, corresponding to the ith ICD code, g_i the ith column in G, representing the ancestors of targeted ICD code, α_{ij} the attention weights, $\mathcal{A}(i)$ the set for ICD code i and all of its ancestors, and $f(\cdot)$ in this case represents a two-layer MLP. Taking Fig. 1 as an example again, the 'attended' embedding of 'Heart Failure' would be a linear combination of the initial embeddings of itself, 'Congestive hear failure; non-hypertensive,' 'Diseases of the heart' and 'Diseases of the circulatory system'.

The 'attended' ICD embedding matrix E (concatenation of all MLP outputs in Fig. 2 B) was then further mapped (taking inner product) with the patient-diagnosis encoding matrix M to serve as the input of a stacked AE.

As shown in Fig. 2 B, the encoding matrix M contains the patients' diagnosis information. M is a $N \times C$ matrix where N is the number of patients and C is the total number of ICD codes in the dataset. m_{ij} represents the counts of ICD code j from all visits of patient i. Fig. 3 illustrates how this map is generated.

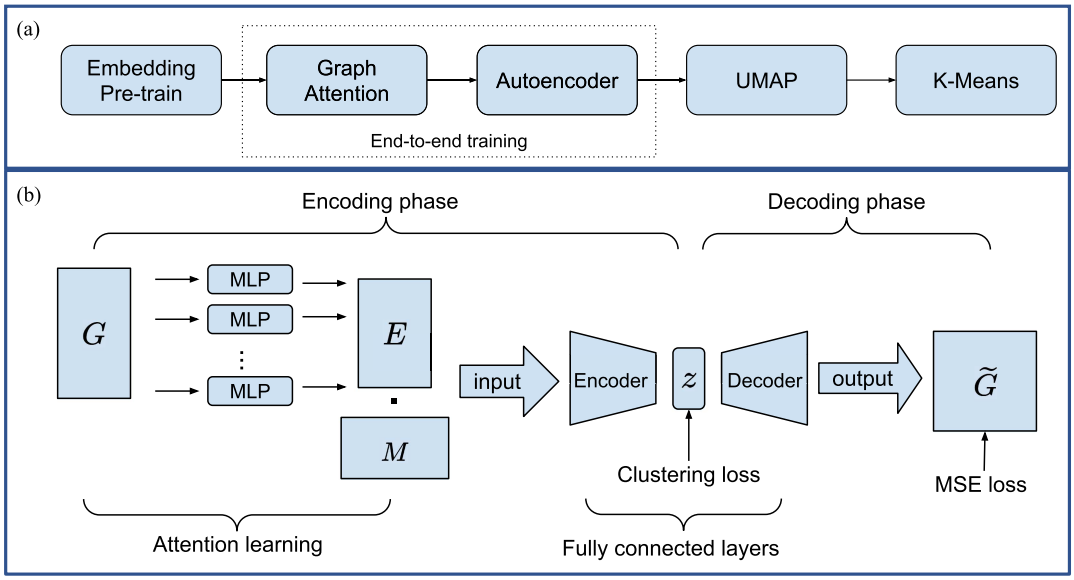

Fig. 2. Box A presents the overall clustering pipeline with the key components listed. The dotted box shows the end-to-end training components. Box B shows the details of the end-to-end training. The input of the model is the product of the patients multi-hot ICD encoding matrix M and the pre-trained ICD embeddings G (trained by GloVe). Each ICD GloVe embedding goes through a multi-layer perceptron (MLP) to learn the attention weights. The concatenation of all MLP outputs is the patient-wise updated ICD embedding matrix E. Taking the inner product between E and a patient-ICD encoding matrix M to serve as the input of a stacked AE. The reconstruction loss is applied between the model input and final reconstruction $\widetilde{MG}$; the clustering loss is imposed on the bottleneck latents z.

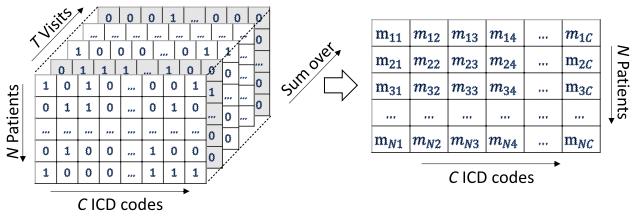

Fig. 3. The 3D matrix is a N (number of patients) by C (number of ICD codes) by T (maximum number of visits among all patients) binary matrix indicating whether a patient acquires an icd code in a visit. The visit does not have temporal order, i.e. the tth visit for different patients may be at different real-world time. The matrix is padded up to the maximum number of visits using 0 to fill the visits that do not exist for a patient. The multi-hot encoding matrix M is the 2D matrix on the right which is summed over the 'visit' axis in the 3D matrix.

The initialisation of G can be random, however, [6] showed that initialising G with GloVe can boost the model performance (details of GloVe initialisation can be found in [28] and [6]). Therefore, we initialised G with the same GloVe training, and kept the GloVe embedding dimension as 128. In brief, GloVe is trained on sequences containing all leaf ICD nodes obtained within each visit of a patient and all their corresponding ancestor nodes in the ontology tree. Therefore, GloVe is able to incorporate the hierarchical relationship embedded in the ontology as well as the co-occurrence information between the ICD leaf nodes.

We used layer-wise mean-squared error (MSE) as the reconstruction loss, summing the reconstruction loss between each encoder and decoder layer; K-means loss as the clustering loss and it was applied to the bottleneck latents z. The joint loss function is expressed in (2).

$$\begin{aligned}\min_{\mu} \quad & ||G - \widetilde{G}||_F^2 + \lambda_1 \sum_{l=1}^{k-1} ||H^l - \tilde{H}^l||_F^2 \\ & + \lambda_2 \sum_{i=1}^{k} \sum_{j=1}^{n} w_{ij} ||z_j - \mu_i||^2, \qquad (2) \\ \text{s.t.} & \sum_{i=1}^{k} w_{ij} = 1; \quad w_{ij} \in \{0,1\} \forall i, j\end{aligned}$$

where $||\cdot||_F$ indicates the Frobenius norm, k the number of clusters, H^l and $\tilde{H}^l$ the output of the lth encoder and decoder layer respectively, w_{ij} the binary cluster assignment parameter assigning point j to cluster i, z_j the jth row of the bottleneck latents, and μ_i is the centroid for the ith cluster. The end-to-end training algorithm is described in Supplementary Appendix Fig. 1.

B. Model Training and Assessment

The training procedure consisted of two pre-trains, a separate ICD embedding pre-train using GloVe (which we name *ICD pre-train*) and an end-to-end *AE pre-train*, and a *joint-train*. Notably, the *AE pre-train* and *joint-trainjoint-train* shared the same end-to-end training pipeline (illustrated in Fig. 2 B) with only loss function different. The *AE pre-train* was trained with MSE as loss only (first two terms in (2)); the *joint-trainjoint-train* has the loss function shown in (2), sum of reconstruction loss and clustering loss. We used Pytorch [26] to train both of the pre-trains and joint-train. The training details can be found in Supplementary Appendix B.

Regarding the model assessment, we applied the converged model to the dataset, and extracted the bottleneck latents for clustering. Before applying clustering algorithms, we further applied UMAP [20] on the bottleneck latents to reduce the dimensionality to 2. We will show in Section V-A that adding UMAP significantly improves the clustering performance. Moreover, [49] applied several manifold learning methods to embeddings extracted from AE and showed that UMAP is able to discover the most clusterable latent representations. This application of UMAP does not complicate the original model and facilitates visualisation. For inference, we investigated the ICD codes as well as their single-level CCS categories belonging to different clusters.

Although we have no ground truth of cluster assignment for patients, it is reasonable to assume the number of clusters is 3 since we are considering three types of organ failure. In this study, we start by setting $k = 3$ for K-means loss and also experimented for k equal to 2 and 4. To assess the clustering results, we calculated the Silhouette score [43].

To further assess our pipeline, we compared it with DCN [39], an extensively applied benchmark clustering model, and converted an off-the-shelf model Med-BERT [49] for clustering. We used the same hyper-parameters as in the original works where we can. Since DCN does not incorporate any attention mechanism, we used the average of the code embeddings as input. For Med-BERT, we pre-trained BERT on the whole MIMIC cohort using the masked language model (MLM) and fine-tuned the model on the OF target cohort with a classification task using the OF type as labels. We extracted the encoder embeddings and averaged them for each patient to serve as input to K-means to cluster the OF cohort. The rest of the setting stays the same with Med-BERT [49].

C. Validation of the Deep Clustering Pipeline on MNIST

To validate the above loss function and training scheme, we tested the deep clustering model (without attention learning) on MNIST [16], an image dataset with 10 classes. The model was trained with the same neural network architecture (AE based), training scheme (a pre-train plus joint-trainjoint-train) and loss function (expressed in (2)). The final clusters are obtained by applying K-means to the 2-dimensional UMAP latents reduced from the bottleneck latents extracted from the converged model. The joint-train algorithm starts from step 11 in Supplementary Appendix Fig. 1 in this case.

To assess the clusters, we calculated the standard normalised mutual information (NMI [47]) between cluster outcomes and the true labels and plotted a confusion matrix to visualise the results. Higher NMI indicates better alignment between the clustering outcomes and the true labels.

V. Results

In this section, we present first the results of testing the clustering model on MNIST, then we move to the main results obtained from MIMIC-III. This part of the results is presented according to the pipeline training order: GloVe ICD embedding training results are shown first; AE pre-train results come second; finally, there are the clustering results and inference obtained from the joint-training.

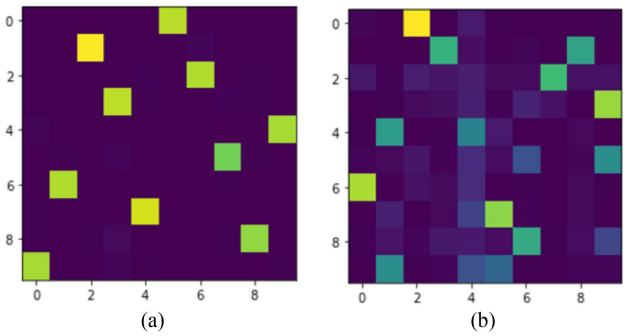

Fig. 4. Confusion matrices for clustering results on MNIST with (a) and without (b) UMAP applied. More yellow blocks (relative to green) indicates more instances that fall under the corresponding cluster and class. Note that the numbers on the axis do not indicate the 10-digit classes since this is an unsupervised setting. (a) With UMAP, NMI = 0.929. (b) Without UMAP, NMI = 0.512.

A. Testing Deep Clustering Model on MNIST

Fig. 4(a) shows the confusion matrix obtained by applying the pipeline introduced in Section IV-C to MNIST. This result, NMI= 0.929, is higher than all related models considered in Section II (the highest NMI is 0.917 and given by DEPICT [8]) and is at a competitive level with the state-of-the-art clustering models ([8], [10]). Notably, confusion matrices in Fig. 4 show how unsupervised cluster assignments align with the true labels. Since the clusters do not match any specific classes, the blocks are randomly distributed in the matrix. However, if the clusters align well with the classes, one should expect k (k equals to the number of true classes) distinguished blocks, and distributed in k different rows and columns (like shown in Fig. 4(a)). We also present the result without applying UMAP before cluster acquirement, and found much worse performance Fig. 4(b). This may be explained by the suffering of 'curse of dimensionality' of K-means. This is, therefore, the reason why we added UMAP to the analysis of MIMIC-III data.

B. OF Patients Deep Clustering on MIMIC-III

1) Interpretation of ICD Pre-Train: As introduced in Section III, the OF patient cohort comprises over 3266 end-level ICD codes and 729 ancestral medical concepts. GloVe initialisation training gives a dataset-specific embedding matrix for all of the 3995 (3266 + 729) nodes in the ontology tree. It supposes to reveal the co-occurrence information in the OF dataset between the different ICD codes as well as between ICD codes and their ancestors. To be able to interpret the embeddings, we visualised them by applying UMAP to the 128-dimensional GloVe embeddings and reduced the dimensionality to 2. In Fig. 5, we observed clusters where all similar medical concepts gather together such as gynaecological diagnoses (Box B), and diabetes-related diagnoses (Box A). There are also clusters on related diseases such as lung cancer, respiratory diseases and pleurisy in Box D of Fig. 5. Meanwhile, we observed some

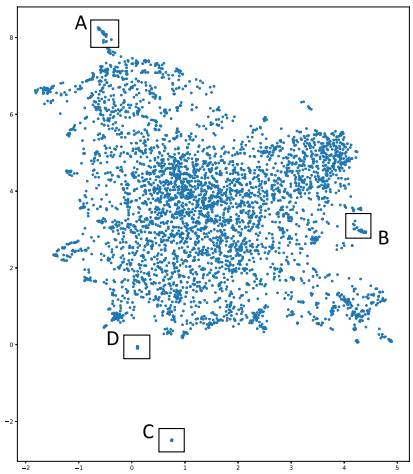

Fig. 5. 2-dimensional UMAP visualisation of 128-dimensional ICD and medical ancestor embeddings trained by GloVe. Box A contains codes related to medical concepts of diabetes (e.g. Diabetes mellitus with complications) and some other diseases such as 'Pulmonary heart disease' and 'Genitourinary symptoms and ill-defined conditions'. Box B contains all maternal related medical concepts. Box C is a mixture of diseases including 'Secondary malignancies,' 'Cancer of bronchus; lung,' 'Osteoarthritis,' 'Other hereditary and degenerative nervous system conditions,' 'Other nutritional; endocrine; and metabolic disorders'. Codes in Box D are 'Hypertension with complications and secondary hypertension,' 'Cancer of bronchus; lung,' 'Skin and subcutaneous tissue infections,' 'Pleurisy; pneumothorax; pulmonary collapse,' 'Other lower respiratory disease' and 'Inflammatory diseases of female pelvic organs'.

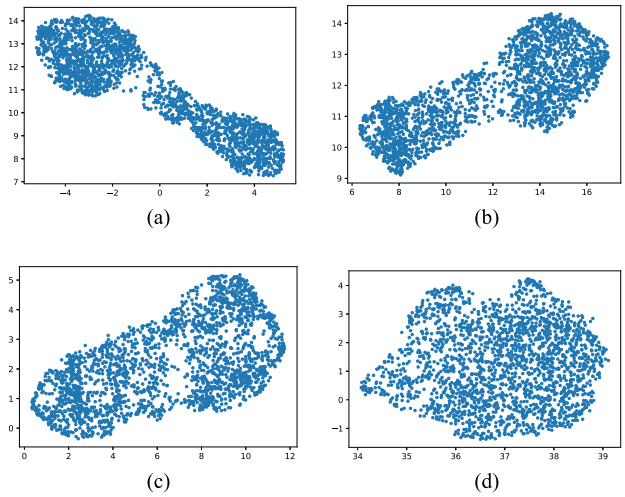

Fig. 6. UMAP visualisations for the bottleneck latents after pre-train. The latents are extract from applying the converged model to all of the OF patients. (a), (b) and (c) are trained by the model with 3, 4 and 2 hidden layers receptively. (a) and (b) show a rough representation of two possibly three clusters whereas (c) is hardly showing any structure. (a) 3 hidden layers with widths 128, 64, 32. (b) 3 hidden layers with widths 256, 128, 64. (c) 4 hidden layers with widths 256, 128, 64, 32. (d) 2 hidden layers widths 64, 32.

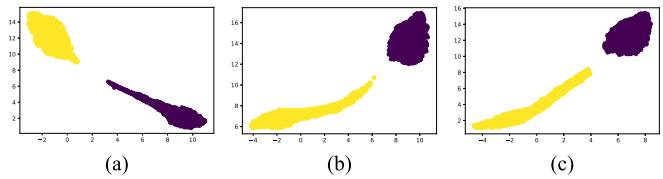

Fig. 7. UMAP visualisations for the bottleneck latents after joint-train. The latents are extract from applying the converged model to all of the OF patients. (a), (b) and (c) are trained with 2, 3 and 4 clusters ($k = 2, 3$ and 4) respectively. The two colours are clustering labels assigned by HDBSCAN. (a) k = 2. (b) k = 3. (c) k = 4.

seemingly unrelated diseases clustered together: skin infection appeared with pulmonary diseases (Box D). The points in Box C are far away from the rest of the points, but they represent a combination of a variety of diseases, e.g., nerve system diseases, nutritional disorders, osteoarthritis and tumour related diseases. The emergence of these clusters might be caused by the co-occurrence of the diseases in this specific dataset or outliers.

2) AE Pre-Train: The 128-dimensional GloVe embeddings of all ICD codes and their ancestral concepts were then fed into an AE with reconstruction loss only for pre-train (no clustering loss imposed in Fig. 2 B). Based on the experiments we carried out on MNIST, the visualisation of bottleneck latents after pre-train should be close to the one after joint-train. This statement is also supported by clustering literature [30] where no clustering loss was applied; only AE with reconstruction loss was used for clustering and visualisation. Therefore, visualising the bottleneck latents after pre-train would help us having some ideas on the number of clusters. Since this is a pure unsupervised setting, the selection of the number of clusters should be aligned with the data structure.

As with visualising the GloVe embeddings, we applied UMAP to the bottleneck latents extracted from applying the converged pre-train model to the whole OF cohort. We show this visualisation in Fig. 6(a). It looks like the data are moving towards two clusters, and it is possible that the larger cluster may further divide.

To see the effects of model architecture, we tested with different numbers of hidden layers and layer widths. We observed similar pattern in UMAP visualisation with the same number of hidden layers (with different widths, Figs. 6(b)) or more layers Fig. 6(c). However, this pattern cannot be learnt with fewer layers Fig. 6(d). This experiment gave us an estimate of the number of clusters to explore during the joint-train. Since Fig. 6(a) gives the clearest structure and with the simplest model architecture, we used this architecture to carry out the analysis.

3) Clustering Results From the Join-Train Stage: We ran the joint-train stage for several repetitions of each $k \in \{2, 3, 4\}$. The clustering visualisations for all ks look very stable Fig. 7. All cases in Fig. 7 display two distinguished clusters. We also observed that as the training proceeds, the longer-shaped cluster tend to be stretched out even further, with a long tail (an example is shown in Supplementary Appendix C) and no further distinguished division appeared. This kind of visualisation might be a sign of over-training. Therefore, we stopped the training before the cluster became too stretched out. We then moved to interpreting the two clusters presented in Fig. 7.

The Silhouette scores for the three sets of clustering results shown in Fig. 7 are 0.800, 0.748 and 0.740 for $k = 2$, $k = 3$

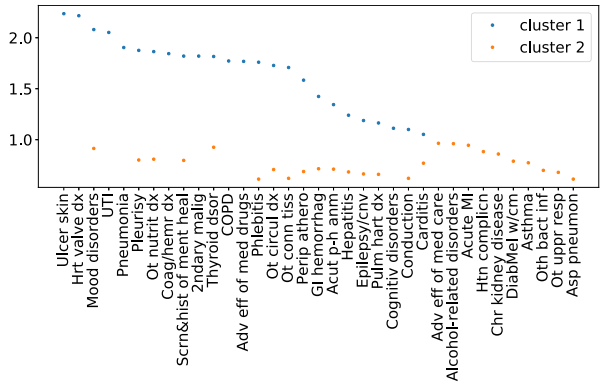

Fig. 8. CCS single-level categories with the top 10 to 20 percent occurrence in the two clusters. The occurrence is normalised by cluster size and ordered in decreasing orders. The label ticks are abbreviated category names. The full category names can be found in Supplementary Appendix F.

and $k = 4$ respectively. We focused on the $k = 2$ case since it has the highest Silhouette score. Moreover, all ks gave very similar results (results for $k = 3$ and 4 can be found in Supplementary Appendix D). We first investigated the ICD codes in the two clusters. Due to the large number of ICD codes involved, we applied CCS single-level category to represent the ICD codes to aid interpretation and visualisation. We visualised the CCS single-level category by a range of percentiles based on occurrence frequencies, and focused on interpreting the CCS categories with the top 10-20% occurrence (Fig. 8). This is due to the fact that the most occurring ICD codes/CCS categories are generally diseases with high prevalence in the population which is not helpful in distinguishing the clusters. We present the results for the top 10% CCS categories and top 20% most-occurring ICD codes (split into 4 ranges) in Supplementary Appendix D.

From Fig. 8, we can see that the two clusters exhibit different comorbidity spectra, and the spectrum of Cluster 1 has higher disease frequencies that the spectrum of Cluster 2. For cluster 1, the most commonly occurring CCS category is 'Chronic ulcer of skin' which can be a complication for all OFs. We also observed other OF related CCS categories belonging only to cluster 1 in this range such as 'Heart valve disorders,' 'Urinary tract infections,' 'Pneumonia' and 'COPD' which are related to HF, KF and RF respectively. The unique and OF-related categories belonging to cluster 2 in this range include 'Acute myocardial infarction,' 'Chronic kidney disease,' 'Other upper respiratory disease' which can also be related to HF, KF and RF, respectively. Therefore, the clusters are not grouped by failing organs, but by severity of the diseases in some way: the occurrence frequency for cluster 1 is larger than cluster 2 (normalised by cluster sizes), i.e. the patients in cluster 1 have more diagnoses than those in cluster 2. Moreover, some diagnoses that are unique to cluster 1, such as 'Chronic ulcer of skin,' 'Coagulation and hemorrhagic disorders' and 'Secondary malignancies,' are signs of more severe deterioration in patients with organ failure [13], [25]. The same figures for the cases of $k = 3$ and $= 4$ are shown in Supplementary Appendix D.

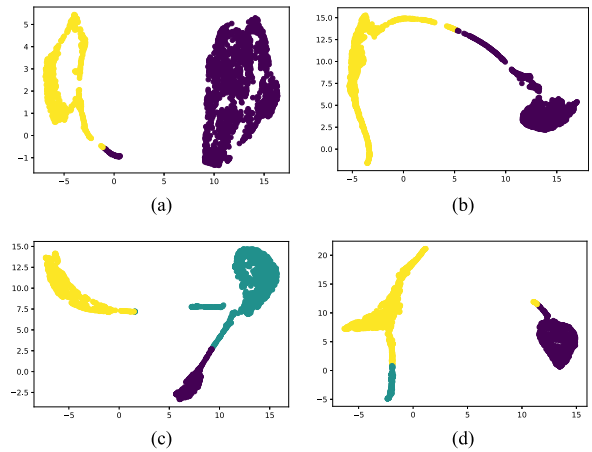

Fig. 9. UMAP visualisation for DCN setting $k = 2$ (subplots (a) and (b)) and $k = 3$ (subplots (c) and (d)). We ran each setting two times. The colors are predicted labels given by DCN. (a) k = 2, run 1. (b) k = 2, run 2. (c) k = 3, run 1. (d) k = 3, run 2.

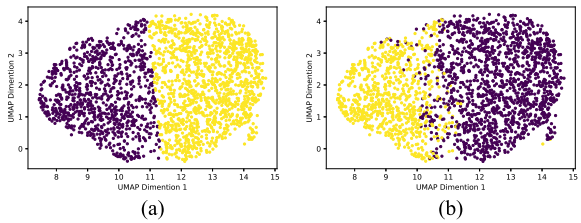

Fig. 10. UMAP visualisation for the Med-BERT embeddings. The subjects in (a) are coloured by K-means (K = 2) labels where K-means was applied to the 2 d UMAP representations; the subjects in (b) are coloured by K-means (K = 2) labels where K-means was applied to the extracted Med-BERT embeddings directly.

C. Comparison with DCN and Med-BERT

We ran DCN for $k = 2$ and $k = 3$ (in K-means), and found that although it also displayed a pattern of two clusters as shown in Fig. 9, the stability between different ks and different runs is much worse compared with our method.

DCN did the same MNIST task as we presented in Section V-A and the NMI reported in [39] (0.81) is significantly lower than using our proposed method (0.93).

For Med-BERT, we pre-trained the model for 200 epochs and fine-tuned for another 100 epochs. The classification accuracy reached 98.11% during fine-tune. We furthered applied UMAP to the extracted BERT embeddings for visualisation and improving clustering performance. K-means was applied to the UMAP reduced embeddings as well as directly to the BERT embeddings with K set to 2. The visualisation is shown in Fig. 10. The Silhouette scores for Fig. 10(a) and (b) are 0.468 and 0.167, respectively.

VI. Discussions

The investigation of multiple types of vital organ failures especially under an unsupervised setting is very challenging and understudied both clinically and in the field of machine learning.

To achieve the clustering purpose, we combined the best of several published works such as using layer-wise reconstruction

loss, two-stage training and adding graph attention. At the same time, we also tried to make the architecture stay low complexity such as using the discriminative AE plus UMAP architecture. Therefore, the model training does not require extensive computation power. Moreover, we explored different clustering loss functions such as adding regularisation terms and replacing K-means loss with fuzzy c-means loss, however, these modifications did not bring extra gain in NMI or Silhouette scores. Therefore, we carried out this clustering pipeline and added the attention mechanism to cluster the OF patients in an end-to-end training fashion.

We further investigated the impact of input features on the clusters with the following experiments: 1) feeding the product of E and M in Fig. 2 B directly to K-means; 2) applying UMAP to the product of E and M, and then feeding the reduced features to K-means; 3) applying K-means directly to the bottleneck latents without UMAP. We set $k = 2$ for all the above scenarios. The Silhouette scores are 0.5267, 0.5739 and 0.5685, respectively, for the three cases, which are significantly lower than our proposed pipeline (Silhouette score $= 0.8005$). Furthermore, we carried out a side classification task to prove that our 'attended' ICD embeddings have superior performance in classifying the OF labels (Supplementary Appendix Section E).

Compared with the previous deep clustering works which were mostly proposed in other areas such as computer vision, our proposed pipeline is designated for application in clinical settings - in particular, we extended the encoder in the AE to integrate the diagnosis codes and their ontology using graph attention to learn more stable representations. As a potential consequence, we observed better stability of our pipeline compared with a stereotypical benchmark model, DCN in the clinical task. For the common MNIST task which has true labels and was implemented in almost all deep clustering works. Our pipeline gives higher NMI than other deep clustering methods considered in this work ([10], [38], [45], [46]). We attribute this partially to our application of the non-linear dimension reduction method, UMAP.

From the adaptation of an off-the-shelf BERT model proposed in [49], we did not discover clusterable embeddings. We conjecture that this poor performance may be attributed to the small data size, and therefore the model is over-parameterized. BERT may show strengths better in supervised tasks. A better BERT model tailored specifically for clinical tasks with relatively small sample sizes can be investigated in the future.

Overall, this paper presents a pure exploratory analysis, and the unsupervised setting poses several challenges to the analysis which are common issues of unsupervised learning. First of all, it made the interpretation of the results challenging. Since we had no ground truth, it is difficult to choose the assessment measure and inference approach. Apart from the Silhouette score, we investigated the disease frequencies in different clusters. However, we are aware that there are other ways to interpret the clusters and we may get different information by investigating different measures. Secondly, to demonstrate the efficacy of the clustering method, we tested the clustering part of the pipeline (without attention) on MNIST. We acknowledge that this is not a perfect validation, but this is a representative task given the popularity of MINIST in computer vision and signal processing. We further studied the stability of our current results including exploring different AE architectures (number of layers and layer widths) and assigning different number of clusters. Our results showed fair robustness. Nonetheless, other factors such as the attention architecture including the number and type of hidden layers can affect the model performance. These aspects are valuable future directions to explore, ideally under supervised settings since it can provide more quantifiable assessment measures. Moreover, this work only considered patients recorded with only one organ failure. It can be served as a starting point for studies that include patients with multiple OFs in real-world scenarios. One other limitation of this work is that it only uses ICD information; integrating other data modalities such as demographics, procedures and vital signs will be valuable future work.

VII. Conclusion

This paper proposes an unsupervised learning pipeline for clustering OF patients in MIMIC-III using ICD diagnosis code which has been rarely studied so far, via graph ontology learning and deep neural networks.

We tested this deep clustering model on the public-domain dataset, MNIST, and found that if we add UMAP, a non-linear dimension reduction method, to the bottleneck latents before clustering, the model performance can be improved significantly. It achieved an NMI of 0.929 – a competitive level of performance with the state-of-the-art clustering algorithms even without the use of convolutional layers.

We discovered two clusters for the OF patients from the model. This discovery is stable to the AE architecture when there are enough hidden layers (3 in this case) to learn such structure, and is robust to the number of clusters to which we assigned K-means during the joint-train learning. The clusters produced by the model did not correspond to the three OFs well. Instead, the two clusters rather related to the severity of the patients, one group having considerably more diagnoses than the other and focusing on different sets of diseases. This outcome may suggest that these three OFs are not separable only based on the ICD information; these groups of patients share similar underlying characteristics or the complexity of the underlying structure is too high to be learnt in this way. However, this model can potentially be used in clinics as a severity identification tool for patients with these OFs and to flag the possible complications that may arise from organ failure.

To our knowledge, we are the first to use GloVe embeddings to perform clustering on heart failure, respiratory failure and kidney failure patients.

Acknowledgment

DAC is an Investigator in the Pandemic Sciences Institute, University of Oxford, Oxford, U.K. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, the Department of Health, InnoHK – ITC, or the University of Oxford.

References

[1] J. Orban et al., "Causes and characteristics of death in intensive care units: A prospective multicenter study," *Anesthesiology*, vol. 126, no. 5, pp. 882–889, 2017.

[2] S. R. Bapoje et al., "Unplanned transfers to a medical intensive care unit: Causes and relationship to preventable errors in care," *J. Hosp. Med.*, vol. 6, no. 2, pp. 68–72, 2011.

[3] V. Bul et al., "Multiorgan failure predicts mortality in emphysematous pancreatitis: A case report and systematic analysis of the literature," *Pancreas*, vol. 46, no. 6, pp. 825–830, 2017.

[4] J. Chang et al., "Deep adaptive image clustering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5879–5887.

[5] E. Choi et al., "Gram: Graph-based attention model for healthcare representation learning," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2017, pp. 787–795.

[6] T. H. Cost and U. Project, "Clinical classifications software (CCS) for ICD-9-cm," 2017. [Online]. Available: https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

[7] N. C. for Health Statistics, "International classification of diseases,ninth revision, clinical modification (ICD-9-cm)," 2021. [Online]. Available: https://www.cdc.gov/nchs/icd/icd9cm.htm

[8] K. G. Dizaji et al., "Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5736–5745.

[9] M. Jabi et al., "Deep clustering: On the link between discriminative models and K-means," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 1887–1896, Jun. 2021.

[10] Z. Jiang et al., "Variational deep embedding: An unsupervised and generative approach to clustering," 2016, in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, doi: 10.24963/ijcai.2017/273.

[11] A. E. Johnson et al., "MIMIC-III, A freely accessible critical care database," *Sci. Data*, vol. 3, 1, pp. 1–9, 2016.

[12] D. P. Kingma and J. Ba, "Adam: A method for Stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.

[13] D. R. Kuypers, "Skin problems in chronic kidney disease," *Nature Rev. Nephrol.*, vol. 5, no. 3, pp. 157–170, 2009.

[14] L. S. Larkey and W. B. Croft, "Combining classifiers in text categorization," in *Proc. 19th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1996, pp. 289–297.

[15] Y. LeCun, "The MNIST database of handwritten digits," 1998. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[16] M. Li et al., "Automated ICD-9 coding via a deep learning approach," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 4, pp. 1193–1202, Jul./Aug. 2019.

[17] Y. Li et al., "Behrt: Transformer for electronic health records," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, 2020.

[18] A. L. et al., "Rectifier nonlinearities improve neural network acoustic models," in *Proc. Int. Conf. Mach. Learn.*, 2013, vol. 30, pp. 1–6.

[19] L. McInnes, J. Healy, and S. Astels, "HDBSCAN: Hierarchical density based clustering," *J. Open Source Softw.*, vol. 2, no. 11, pp. 205–206, 2017.

[20] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection," *J. Open Source Softw.*, vol. 3, 29, 2018, doi: 10.21105/joss.00861.

[21] E. Min et al., "A survey of clustering with deep learning: From the perspective of network architecture," *IEEE Access*, vol. 6, pp. 39501–39514, 2018.

[22] E. Moons et al., "A comparison of deep learning methods for ICD coding of clinical records," *Appl. Sci.*, vol. 10, no. 15, 2020, Art. no. 5262.

[23] K. Nash, A. Hafeez, and S. Hou, "Indication for dialysis initiation and mortality in patients with chronic kidney failure: A retrospective cohort study," *Amer. J. Kidney Dis.*, vol. 69, no. 1, pp. 41–50, 2017.

[24] NIH, "Snomed ct," 2019. [Online]. Available: https://www.nlm.nih.gov/healthit/snomedct/index.html

[25] M. Nimah and R. J. Brilli, "Coagulation dysfunction in sepsis and multiple organ system failure," *Crit. Care Clin.*, vol. 19, no. 3, pp. 441–458, 2003.

[26] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Adv. Neural Inf. Process. Syst.*, pp. 8024–8035, 2019.

[27] P. B. Pedersen et al., "Prevalence of organ failure and mortality among patients in the emergency department: A population-based cohort study," *BMJ Open*, vol. 9, no. 10, 2019, Art. no. e032692.

[28] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[29] P. Ponikowski et al., "Heart failure: Preventing disease and death worldwide," *ESC Heart Failure*, vol. 1, no. 1, pp. 4–25, 2014.

[30] S. Saito and R. T. Tan, "Neural clustering: Concatenating layers for better projections," in *proc. 4th Int. Conf. Learn. Representations*, 2017. [Online]. Available: https://openreview.net/pdf?id=r1PyAP4Yl

[31] Y. Sakr et al., "Patterns and early evolution of organ failure in the intensive care unit and their relation to outcome," *Crit. Care*, vol. 16, no. 6, 1–9, 2012.

[32] G. Savarese and L. H. Lund, "Global public health burden of heart failure," *Cardiac Failure Review*, vol. 3, no. 1, pp. 7–11, 2017.

[33] J. Shang et al., "Pre-training of graph augmented transformers for medication recommendation," in *Proc. of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019, pp. 5953–5959.

[34] H. Shi et al., "Towards automated ICD coding using deep learning," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, pp. 1066–1076, 2018, doi: 10.18653/v1/P18-1098.

[35] J. T. Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," in *Proc. 4th Int. Conf. Learn. Representations*, 2016.

[36] M. Q. Stearns et al., "Snomed clinical terms: Overview of the development process and project status," in *Proc. AMIA Symp.*, 2001, pp. 662–666.

[37] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 478–487.

[38] B. Yang et al., "Towards K-means-friendly spaces: Simultaneous deep learning and clustering," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3861–3870.

[39] J. Yang, D. Parikh, and D. Batra, "Joint unsupervised learning of deep representations and image clusters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5147–5156.

[40] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995.

[41] T. K. Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, 1995, pp. 278–282.

[42] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 785–794.

[43] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[44] Y. Opochinsky et al., "K-autoencoders deep clustering," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 4037–4041.

[45] A. Lin et al., "Mixture model auto-encoders: Deep clustering through dictionary learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 3368–3372.

[46] S. Chazan, S. Gannot, and J. Goldberger, "Deep clustering based on a mixture of autoencoders," in *Proc. IEEE 29th Int. Workshop Mach. Learn. Signal Process.*, 2019, pp. 1–6.

[47] D. Cai, X. He, and J. Han, "Locally consistent concept factorization for document clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 6, pp. 902–913, Jun. 2011.

[48] L. Rasmy et al., "Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *NPJ Digital Med.*, vol. 4, no. 1, pp. 1–13, 2021.

[49] R. McConville et al., "N2D:(not too) deep clustering via clustering the local manifold of an autoencoded embedding," in *Proc. IEEE 25th Int. Conf. Pattern Recognit.*, 2021, pp. 5145–5152.