MD-Manifold: A Medical Distance Based Manifold Learning Approach for Heart Failure Readmission Prediction

Shaodong Wang
Iowa State University
shaodong@iastate.edu

Qing Li
Iowa State University
qlijane@iastate.edu

Wenli Zhang
Iowa State University
wlzhang@iastate.edu

Abstract

Dimension reduction is considered as a necessary technique in Electronic Healthcare Records (EHR) data processing. However, no existing work addresses both of the two points: 1) generating low-dimensional representations for each patient visit; and 2) taking advantage of the well-organized medical concept structure as the domain knowledge. Hence, we propose a new framework to generate low-dimensional representations for medical data records by combining the concept-structure based distance with manifold learning. To demonstrate the efficacy, we generated low-dimensional representations for hospital visits of heart failure patients, which was further used for a 30-day readmission prediction. The experiments showed a great potential of the proposed representations (AUC = 60.7%) that has comparative predictive power of the state-of-the-art methods, including one hot encoding representations (AUC = 60.1%) and PCA representations (AUC = 58.3%), with much less training time (improved by 99%). The proposed framework can also be generalized to various healthcare-related prediction tasks, such as mortality prediction.

1. Introduction

Electronic Healthcare Records (EHR) data, an electronic version of patients' medical history, has been widely used to improve healthcare quality in a variety of ways. There are a large number of unique medical concepts in EHR systems, such as 17,000 International Classification of Diseases (ICD) -9 codes [1] and 360,000 National Drug Codes (NDC). These unique medical concepts are one of the fundamental causes of high dimensionality in EHR data. In each visit of a patient in the EHR data, there could be one or more ICD codes that represent the health condition of the patients. For visit-wise machine learning tasks, such as the prediction of mortality and readmission for each patient visit, processing these ICD-9 codes in each visit as categorical data with One Hot encoding leads to the dimensionality of 17,000. The high dimensionality could bring the problem of overfitting, and higher cost of training time and storage. Therefore, it is necessary to generate low-dimensional representations for patient visits that contain medical concepts, which is the first goal of this study.

In addition, the well-organized hierarchical structure is the nonnegligible characteristic of the medical concepts in the EHR data. Many medical concepts like ICD codes were arranged in a hierarchical structure based on their relationship with each other, which was determined by the experts of healthcare. For example, heart disease is one of the circulatory system diseases, and thus the ICD-9 code of heart disease ('420-429') belongs to the circulatory system disease ('390-459'). The patient visits that contain ICD codes with close relationships in the concept hierarchy reflect similar health conditions of the patients, the low-dimensional representations of which should also be close. Taking the hierarchy as domain knowledge into consideration, the generated low-dimensional representations align well with the medical knowledge and have a great potential to help machine learning models achieve better performance. Therefore, the second goal of the study is to incorporate the established domain knowledge into the low-dimensional representations of patient visits.

Although the representation of a single medical concept is widely studied [2], a informative representation for each set of medical concepts remains unknown. A straightforward solution is to implement dimension reduction techniques, such as Singular Value Decomposition (SVD), on the One Hot encoded sets of concepts. However, it does not take advantage of the well-defined concept hierarchy as mentioned above. In light of these limitations, we propose a new framework, Medical-Distance-Manifold (MD-Manifold), to utilize the domain knowledge in the hierarchical structure of medical concepts and generate low dimensional representations for the sets of concepts in a patient visit. We first calculate the distance between medical concepts based on their hierarchical structure, with which we generate the distance between sets of concepts (visits). With the obtained set-level distance as the distance between visits, we implement

URI: https://hdl.handle.net/10125/71209
978-0-9981331-4-0

manifold learning models to produce low-dimensional representations for patient visits.

To evaluate the proposed framework, MD-Manifold, we use heart failure patients' readmission prediction as a research case. Readmission is defined as an event when a patient is admitted again within a specific time interval after the last hospitalization. The readmission prediction for heart failure patients has a significant meaning in practice. In the US, heart failure is one of the main causes of medical institution admissions [3]. Within 30 days after the hospital discharge, approximately 24% heart failure patients would experience all-cause readmission, which costs around $17 billion every year [3]. The readmission is an indicator of disease progression and a source of the economic burden to the medical system [3]. Therefore, the early identification of patients at risk of readmission is a crucial step for enhancing disease management and patient control.

The contributions of this study are significant. Theoretically, the proposed framework takes advantage of the domain knowledge in the concept hierarchy for the low-dimensional representations. We examine two concept-level distance metrics, four set-level distance metrics, and two manifold learning models, including Laplacian Eigenmap (LE) [6], and Uniform Manifold Approximation and Projection (UMAP) [7]. One of the two concept-level metrics developed by us outperforms the state-of-the-art distance metric of [4] in predicting readmission of heart failure patients. Our experiments show the great potential of the proposed low-dimensional representations in the medical machine learning field. From the perspective of readmission prediction, the proposed framework can improve the patient control and decrease the healthcare cost by identifying heart failure patients with high risk of readmission. Other visit-wise machine learning studies, such as mortality prediction, can also benefit from our work by embedding the low-dimension representations into their models.

2. Related work

In this section, we present the existing related studies of dimension reduction, manifold learning, distance metrics, and readmission prediction and introduce the idea-forming process.

Dimension reduction in EHR: By regarding medical concepts in each EHR record as words in a sentence, many researchers learned low dimensional representations (embeddings) for each medical concept [2] with techniques in natural language processing. Furthermore, [5] considered the hierarchical structure of ICD codes as the domain knowledge when generating the low dimensional embeddings. **However**, the representation for each individual medical concept could be inappropriate for visit-wise classical machine learning models, when each patient visit contains multiple concepts in the EHR data. Classical machine learning models require input samples of the same dimensionality, while various numbers of medical concepts in each visit lead to the unfixed dimensions for visits. Therefore, it is the representation of each visit (set of medical concepts), instead of each individual concept, that is in need for visit-wise machine learning tasks. Nevertheless, the representations for visits are still not sufficiently understood.

Manifold learning: We find manifold learning, which is an approach of non-linear dimensionality reduction, a great tool to fill the abovementioned gap. With the distances between data points as the inputs, the manifold learning generates low-dimensional representations that keep the geometry of the original data points. If we set up a distance metric between sets of concepts based on the hierarchical structure of the medical concepts, then the generated representations from manifold learning can incorporate the domain knowledge naturally. Therefore, the manifold learning can tackle these types of problems as long as we set up a meaningful distance between visits. There are various manifold learning algorithms, including Isomap, Locally Linear Embedding, tSNE, LE [6], and UMAP [7]. We adapt LE (classical method) and the UMAP (state-of-the-art method) in this study. Notice tSNE has been widely used in the dimension reduction before the invention of UMAP. We do not adopt it because tSNE takes much more time to generate the representations compared with UMAP [7].

Distance metrics: To construct the distance between visits that include multiple medical concepts, there are two steps, concept-level distance and set-level distance [8]. The concept-level distance measures the distance between medical concepts, based on which the set-level distance measures the distance between visits. As summarized by [8], the most appropriate concept-level distance was proposed by [4]. On the other hand, there are four set-level distance metrics that are equally good at separating visits [8]. We introduce them in detail in Section 3.

Readmission prediction: Readmission prediction is a critical research area in improving patient care. LACE index was first developed to evaluate the likelihood of patient readmission [9]. Then, machine learning models were widely implemented for higher accuracy [10]. With the recent boost of deep learning algorithms, historical visits of patients were used in readmission prediction with sequential models [3]. However, most of the experiments showed that

sequential deep learning models barely outperformed classical machine learning models, which also indicated the necessity of the abovementioned representations of patient visits for classical machine learning models.

3. Research design and the proposed framework: MD-Manifold

As shown in Figure 1, the proposed framework, Medical-Distance-Manifold (MD-Manifold), including three steps: concept-level distance calculation, set-level distance calculation, and manifold learning. The fundamental idea is to melt the medical-concept-hierarchy as domain knowledge into the distance between patient visits, and extract the representation of each visit from the defined distance with manifold learning. In the first step, we measure the distance between medical concepts based on the concepts' relationships in the hierarchy, which is the key step to take advantage of the medical knowledge outside of the dataset. In the second step, based on concept-level distances, we develop the set-level distances to measure the distances between patient visits. In the third step, we generate the low-dimensional representations for patient visits by extracting information from the measured distances between patient visits with manifold learning.

The patient visits in the dataset are represented by $V = \{V_i\}_{i=1,2,\dots,r}$, where r is the number of visits in the dataset. Each visit contains a set of concepts as the indicator of the patient's health condition, which are denoted as a, b, and c, etc., for example, $V_1 = \{a, b\}, V_2 = \{a, c\}$. Other features of the patients, such as age and gender, are not considered in this study. We assume the concepts in the data have a hierarchical structure in the form of a parent-child relationship as shown in Figure 2, for example, the ICD codes in EHR data.

3.1 Step 1: concept-level distance (CD) calculation

Concept-level distance is the crucial step where we take advantages of the well-organized medical concept hierarchy. The concept-level distance of two medical concepts, a and b, are measured by their positions in the concept hierarchy. We introduce a widely used distance metric, CD_{WP}, and our new distance metric, CD_{new}.

Given the concept structure as shown in Figure 2, if two concepts are connected, then the concept in the upper level is called a parent, and the one in the lower level is called a child. For example, in Figure 2, c is the parent of d and d is the child of c. Intuitively, CD_{WP} considers a and b as distant if their least common ancestor (LCA) is much closer to the root of the concept tree compared with a and b. Specifically, $CD_{WP}(a, b) = 1 - \frac{2IC(c)}{IC(a)+IC(b)}$, where c is the LCA, and Information Content (IC) is defined as the concept level in the concept tree. A concept is considered to have more IC if it is farther from the root because it is more specific. Particularly, the IC of the root (level 1) is defined as 1, the IC of the concept that is connected with the root (level 2) is

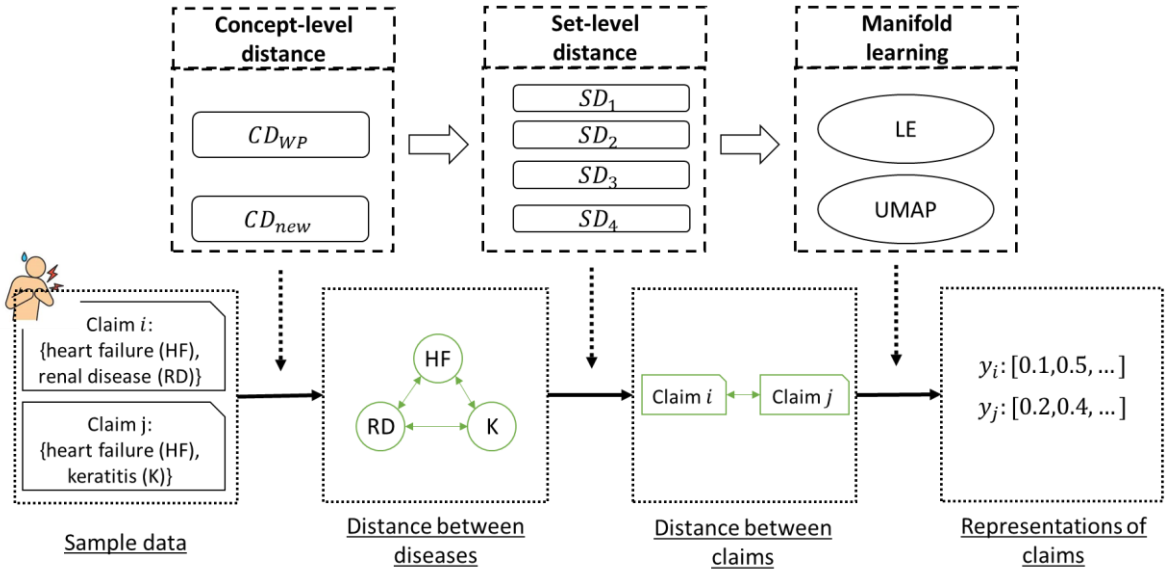


***Figure 1**: the MD-Manifold framework.*

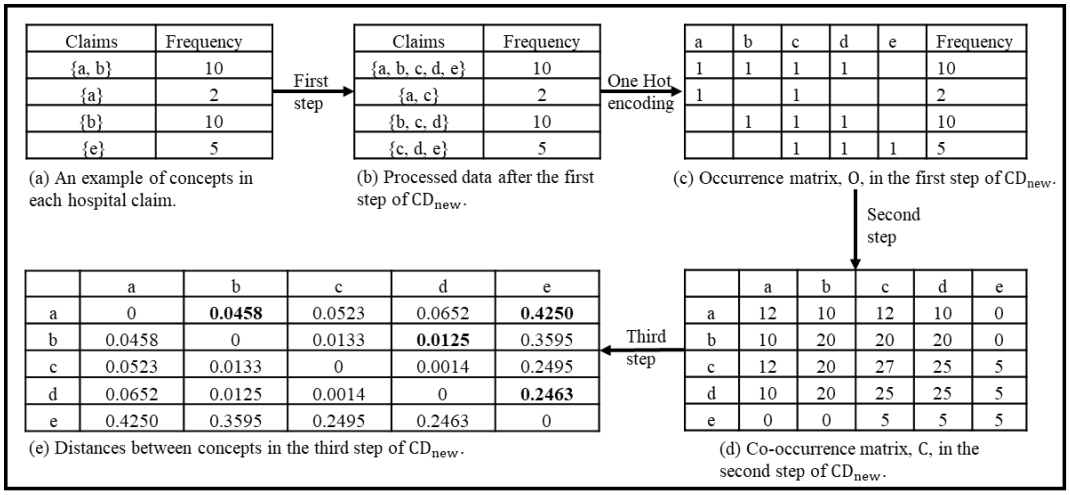


***Figure 3**: An example of CD_{new}.*

defined as 2, and so on. If IC(c) is much smaller than IC(a) and IC(b), this indicates that c is far from a and b; consequently, a and b are also distant with a large $CD_{WP}(a, b)$, and vice versa. For example, as shown in Figure 2, suppose a is a level-4 concept, b is a level-5 concept, and their LCA, c, is a level-3 concept, then the distance between a and b is $1 - \frac{2\times3}{4+5} = \frac{1}{3}$.

However, the method, CD_{WP} has its limitations that the distance is fully determined by the concept structure regardless of the concept co-occur frequency in practice. For example, two distant concepts in the structure co-occurring frequently tend to relate closely with each other, which is not reflected in the concept structure. Moreover, it is also likely that a concept occurs more frequently than its siblings. Thus, it is possible that the concept might have a closer relationship with its parent than its

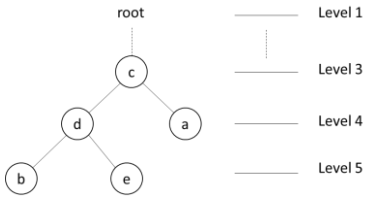


***Figure 2**: An example of concept hierarchy.*

siblings. Nevertheless, the distance between a parent and each child is equal in CD_{WP}. For example, in Figure 2, $CD_{WP}(b, d) = CD_{WP}(e, d)$, regardless of the frequency of b and e in practice.

To address the abovementioned limitation, we propose a new concept level distance metric, CD_{new}, that considers both the structure of the concept hierarchy and the frequency of concepts. The calculation of the proposed concept level distance, CD_{new}, consists of three steps. The first step considers the hierarchical structure of concepts by inserting concept ancestors. In the second and third steps, we set up the concept level distance based on the co-occurrence of concepts in the dataset. (1) For each concept in V_i, we add all the ancestors that the concept belongs to into the dataset. (2) We construct a co-occurrence matrix, C, with the number of co-occurrences of two concepts as its element. Specifically, $C = O^T O$, where O is the occurrence matrix in the first step. (3) We consider each row of the co-occurrence matrix as a feature of the corresponding concepts and generate a cosine distance for each pair of rows as a concept level distance. Explicitly, $CD_{new}(a, b) = 1 - \frac{C_a \cdot C_b}{\sqrt{C_a \cdot C_a}\sqrt{C_b \cdot C_b}}$, where C_a and C_b correspond to rows of a and b on C, respectively. Suppose we have a dataset, as shown in Figure 3 (a), with the concept structure as in Figure 2. The left column in Figure 3 (a) is the concepts that belong to each patient visit, and the right column is the corresponding frequency. For example, there are 10 patient visits in the dataset that contain both a and b. Through the first step, we insert the ancestors as shown in Figure 3 (b), whose occurrence matrix, O, is shown in Figure 3 (c). Afterwards, we can generate the co-occurrence matrix in the second step as shown in Figure 3 (d). In the end, CD_{new} of all pairs of concepts are measured through cosine distance, as shown in Figure 3 (e). Notice that the concept b occurs more than e in Figure 3 (a). After the three proposed steps, as we expected, (b, d) has a smaller distance than (d, e) with $CD_{new}(b, d) = 0.0125$ and $CD_{new}(b, e) = 0.2463$. Moreover, due to the higher co-occurrence frequency of (a, b) than

(a, e), $CD_{new}(a, b) = 0.0458$ is smaller than $CD_{new}(a, e) = 0.425$, in spite of the equal distant relationship in the concept hierarchy.

3.2 Step 2: set-level distance (SD) calculation

Based on concept-level distances, we are able to develop four set-level distance metrics [8] to measure the distances between visits, as shown below. Note the cardinality of the two sets of concepts, V_i and V_j, was denoted as $|V_i|$ and $|V_j|$, respectively.

(1) The first metric uses the average distance of the most similar concept pairs. $SD_1(V_i, V_j) = \frac{1}{|V_i|+|V_j|}(\sum_{a\in V_i} \min_{b\in V_j} CD(a, b) + \sum_{b\in V_j} \min_{a\in V_i} CD(b, a))$.

(2) The second metric considers the average distance of all concept pairs that are not in the union of two

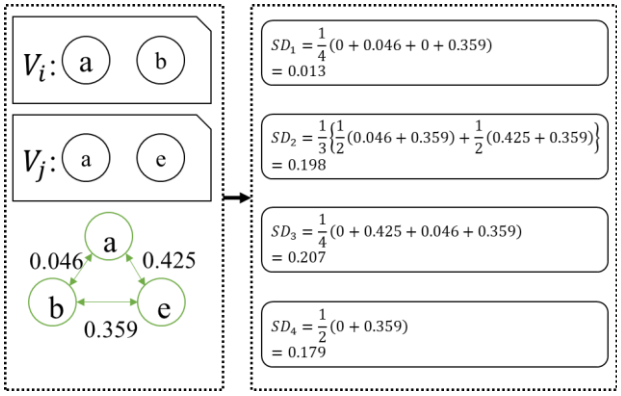


Figure 4: *An example of four set-level distances.*

sets. Specifically,

$SD_2(V_i, V_j) = \frac{1}{V_i \cup V_j}(\sum_{a\in V_i \setminus V_j} \frac{1}{|V_j|}\sum_{b\in V_j} CD(a, b) + \sum_{b\in V_j \setminus V_i} \frac{1}{|V_i|}\sum_{a\in V_i} CD(b, a))$.

(3) The third metric takes the average of the distances of all concept pairs.

$SD_3(V_i, V_j) = \frac{1}{|V_i|\cdot|V_j|}\sum_{a\in V_i, b\in V_j} CD(a, b)$.

(4) The fourth metric regards the two sets of concepts V_i, V_j as a bipartite undirected graph $G = (V_i, V_j)$ with the concept-level distance CD as a weighting function, where all pairs of concepts of V_i are connected to all concepts of V_j [8]. However, no concepts within a set are connected. The Kuhn-Munkres algorithm [11] finds the minimum weighted bipartite matching (MWBM), which is a subset of edges with a minimum sum of weights and at most one edge is incident to each node in G. For hospital records, MWBM is the most similar ICD pairs from patient visit V_i and V_j. Lastly, the set-level distance can be measured by averaging all weights in MWBM.

$SD_4(V_i, V_j) = \frac{1}{|MWBM|}\sum_{(a,b)\in MWBM} CD(a, b)$.

Figure 4 shows the example of the four set-level distances when $V_i = \{a, b\}$ and $V_j = \{a, e\}$ with Figure 3 (e) as their concept-level distance. Four set-level distance metrics lead to different distances between visits, where SD_1 gives a relatively smaller distance and SD_3 generates a larger distance.

Notice that the concept-level distance measures are the groundwork of the set-level distance. Their combination will result in various distance measures for sets. In total, there are $2 \times 4 = 8$ combinations. We evaluated the efficiency of all combinations in the dimension reduction algorithms.

3.3 Step 3: manifold learning

As the last step of our proposed framework, we extract the information in the defined distance between patient visits with manifold learning and produce a low-dimensional representation for each visit. Considering the computational speed, we adopt LE and UMAP in this study.

Before applying LE and UMAP, we construct a graph for the dataset. Regarding each data point (i.e., each set of concepts), V_i, as a vertex in the graph, $G(V, E)$, we connect two vertices as an edge depending upon their k-nearest neighbors. Note that If vertex V_i is a k-nearest neighbor to V_j, but V_j is not a k-nearest neighbor toV_i, the vertices V_i and V_j still forms an edge. The LE and UMAP would generate a d-dimensional representation, y_i, for each data point $V_i \in V$. d is a small number relative to the original data dimensionality.

Laplacian Eigenmap (LE): Laplacian Eigenmap is a classical manifold learning technique that preserves local geometrical information in datasets. Simply, the generated low-dimensional representations will be similar if data points are close in the original dataset. We incorporate the defined distance as the domain knowledge into LE's weighting function. Given a connected graph, G(V, E), LE assigns a weight, W_{ij}, to the edge using the distance between two connected vertices, V_i and V_j. Specifically, $W_{ij} = e^{-\frac{distance^2(V_i,V_j)}{2\sigma^2}}$ if V_i, V_j are connected, otherwise $W_{ij} = 0$, where $V_i, V_j \in V$ and σ is a heat kernel parameter. Usually, the distance metric can be

Euclidean distance or Mahalanobis distance, etc. in many applications [12]. Here, we induce the above-defined set-level distance as the domain knowledge, which results in $W_{ij} = e^{-\frac{SD^2(V_i,V_j)}{2\sigma^2}}$.

The LE generates low dimensional representation by minimizing the loss function, $loss_{LE} = \sum_{ij} W_{ij} \cdot ||y_i - y_j||^2$, where y_i, y_j are d-dimensional representations of vertices V_i, V_j.

Uniform Manifold Approximation and Projection (UMAP): Similar to LE, the UMAP optimizes the low dimensional graph to be as geometrically similar as possible to the high dimensional graph, G, which was constructed from the original dataset. If vertex V_i, V_j are connected, the weight of their edge will be $W_{ij} = W_{j|i} + W_{i|j} - W_{j|i}W_{i|j}$, where $W_{j|i} = e^{(-distance(V_i,V_j)-\rho_i)/\sigma_i}$ and ρ_i is the distance to the nearest neighbor of V_i. σ_i is the normalizing factor, which is chosen by $\sum_{j=1}^{k} \exp\left(\frac{-\max(0, distance(V_i,V_j)-\rho_i)}{\sigma_i}\right) = \log_2 k$. Similar to LE, Euclidean distance and Mahalanobis distance can also be used in the weighting function of UMAP [7]. To take advantage of the domain knowledge, we apply the distance between sets of concepts. As a result, a new weighting function for each edge would be built from our set-level distance. Using stochastic gradient descent as the optimization process, the UMAP minimizes its loss function: $loss_{UMAP} = \sum_{i \neq j} W_{ij} \log \frac{W_{ij}}{\left(1+a\|y_i-y_j\|_2^{2b}\right)^{-1}} + \left(1 - W_{ij}\right) \log \frac{1-W_{ij}}{1-\left(1+a\|y_i-y_j\|_2^{2b}\right)^{-1}}$, where a and b are positive values, and y_i and y_j are the d-dimensional representations for V_i and V_j, respectively.

To summarize, the proposed framework, MD-Manifold, takes advantage of the well-organized medical concept hierarchy so that the generated low-dimensional representations align well with the medical knowledge outside of the patients' hospital-visits dataset. The representations can be further implemented in the visit-wise machine learning tasks, including readmission prediction, as shown in the experiments.

4. Experiments

To show the supremacy of the proposed framework, we took the readmission prediction for heart failure patients as a research case. Due to the huge amount of readmission cases of heart failure patients and their significant amount of cost, developing a predictive model for heart failure readmission is of increasing interest [13]. We generated the low-dimensional representations for each visit under the proposed framework, MD-Manifold. Then the generated low-dimensional representations will be used to predict readmission for heart failure patients.

4.1 Data description

We extracted the dataset of patients with heart failure in 2014 from the Healthcare Cost and Utilization Project (HCUP), Nationwide Readmission Database (NRD), issued by the Agency for Healthcare Research and Quality (AHRQ) [14]. Each patient may have multiple visits in the record. To maintain the consistency and the quality of the dataset, we extracted the records from the large, private, non-profit, and teaching hospitals in a single large metropolitan area, stratified by the NRD (NRD_STRATUM = 109). We labeled the visit as a readmission visit if the patient was readmitted within 30 days of the discharge from the last hospitalization. The visits in December were removed due to the lack of data in the next year. Finally, the dataset of the experiments consisted of 26,358 visits from adult patients (age >= 18) whose primary disease were the heart failure, among which there were 6,553 (25%) readmission cases.

The experiments were conducted on the patient diagnosis in each visit, which is a set of International Classification of Disease, Version 9, Clinical Modification (ICD-9-CM) codes [1], including a primary code. There are 17,000 ICD codes in total, which leads to a high dimension of 17,000 for each visit with one-hot encoding. As shown in Figure 5, the ICD codes have a tree structure with specific diseases in the low level and ambiguous concept in the upper level. For example, the ICD '428' (Heart failure) is the child of '420-429' (Other Forms of Heart Disease), which belongs to '390-459' (Diseases of The Circulatory System). On the other hand, the ICD '428' further has some more specific diseases in

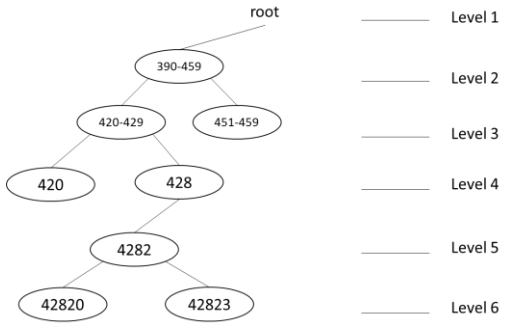


Figure 5: *Part of the structure of the ICD-9-CM.*

the lower level as its descendants, such as '4282' (Systolic heart failure) and '42823' (Acute on chronic systolic heart failure).

4.2 Results of the concept-level distance and set-level distance calculation

Concept-level distance. As displayed in Figure 6, the distributions of all generated distances of CD_{WP} and CD_{new} in the first step of MD-Manifold, show different shapes. Most of the CD_{WP} distances are greater than 0.8, while CD_{new} gather around $0.2 - 0.6$. Also, as mentioned in the methodology section, the CD_{WP} distance was fixed regardless of the data we were using, and CD_{new} incorporated information from both data and the domain knowledge (i.e., ICD hierarchy). For example, the ICD codes '5856' and '40391' existed in 1,834 and 1,600 records, respectively, among which 1,469 records included both '5856' and '40391'. Due to their high co-occurrence frequency, it was reasonable to believe they had a close relationship. In CD_{new}, '5856' and '40391' were the nearest neighbors with each other with a distance of 0.0026. However, in CD_{WP}, their distance is $1 - \frac{2}{5+6} = 0.8182$, which was almost the longest distance among all ICD pairs. Besides, under CD_{new} the parents and children concepts were still close. For example, $CD_{new}('5856', '585') = 0.0676$ and $CD_{new}('40391', '4039') = 0.0699$, where '4039' and '585' were the concept parents of '40391' and '5856', respectively.

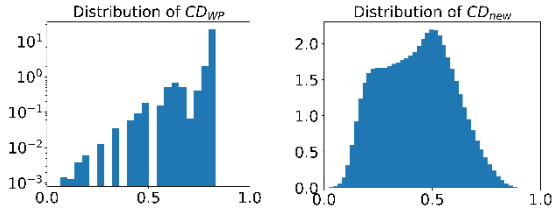


***Figure 6**: Distribution of CD_{WP} (left) and CD_{new} (right).*

Set-level distance. We calculated the Pearson Correlation Coefficients (PCC) for each pair of the 8 distances between patient visits, as shown in Figure 7. Large PCC values indicate the high correlation or similarity between the two distance metrics. The PCC varies from 0.43 to 0.99, as shown in Figure 7. The larger circle with a darker color indicates a higher correlation. The set-level distances, SD_{1-4}, with our new concept-level distance, CD_{new}, were highly correlated with each other. Their PCCs were all above 0.69, half of which were greater than 0.9. On the other hand, the set-level distances with CD_{WP} displayed more discrepancy with PCCs, the highest one being 0.80. Also, the PCCs of the combinations across CD_{WP} and CD_{new} were all below 0.77, which reflects the large difference between two concept-level distances.

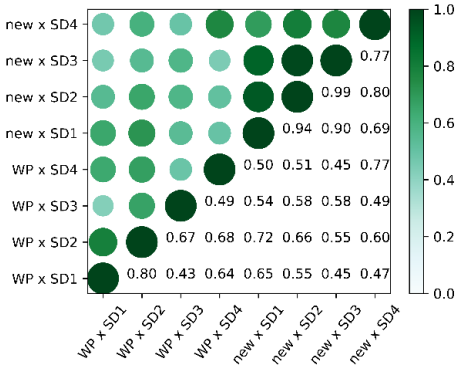


***Figure 7**: PCC of the distances between patient visits.*

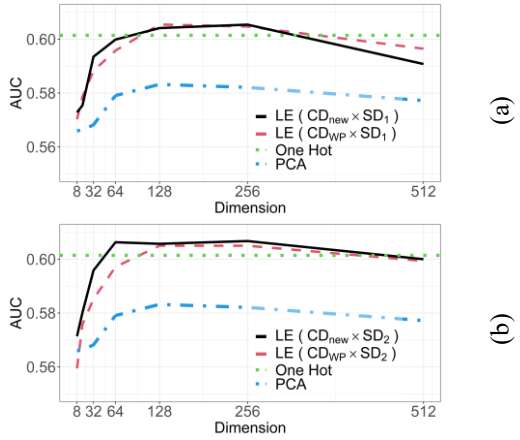
4.3 Dimension reduction and readmission prediction

With the obtained distances between sets of ICD codes in the patient visits, we generated low-dimensional representations for the visits with LE and UMAP. With One Hot encoding, the ICD codes in the diagnosis of each visit would need a vector of 17,000 dimensions to represent. In our experiments, we reduced the dimension to 8, 16, 32, 64, 128, 256, and 512, consecutively. In the end, we evaluated the low dimensional representations with a readmission prediction task, which is a critical problem in practice. We examined the information being preserved by the area under the receiver operating characteristics curve (AUC) scores in the five-fold cross-validation [15]. The more information preserved in the representations, the higher the AUC score is in the readmission prediction task. Also, the training time of the classifier was recorded to show the computational cost saved during the training process.

In the readmission prediction task, we selected a linear classifier, Logistic Regression with $l1$ penalty (LR) as the discriminative model. The LR had been proven to have an equivalent performance with many advanced Recurrent Neural Network models in the readmission prediction of the 2013 HCUP dataset [3]. We set the $l1$ penalty to 0.1 based on the cross-validation, which was consistent with [3].

Besides the low dimensional representations from the proposed method, we generated representations using One Hot encoding and Principal Component Analysis (PCA) on the One Hot encoding as two baselines. We implemented PCA on the One Hot encoded representations and decreased the dimension to 8, 16, 32, 64, 128, 256, and 512.

The AUC scores of the prediction task of all representations through LE are shown in Figure 8. The results are separated according to the four set-level distance metrics, as shown in Figure 8 (a)-(d). The x-axis represents the dimension of the representations, the y-axis represents the AUC scores, and the colors indicate different dimension reduction methods. The green dotted horizontal line indicates One Hot encoding, the blue dotted line indicates PCA, and the black solid and red dash line indicate LE with CD_{new} and CD_{WP}, respectively. As dimension increases, the AUC scores also increase, which reflects the higher information content in the representations. The higher AUC scores indicate that the representations through LE are more informative than PCA representations in this readmission prediction. Notice the AUC score of One Hot encoded representations, whose dimensionality is 17,000, is 0.601. Most importantly, the representations from the combinations of CD_{new} and SD_2 exceed One Hot encoded representations in terms of AUC when their dimensionality increases to 64, which means our representations can be more informative in the machine learning tasks than the original data. Also, the highest AUC, 0.607, of the proposed representations were reached by LE with CD_{new} and SD_2 at dimension 256, which exceeded the maximal AUC of PCA, 0.583. The outperformance comes from the domain knowledge in the hierarchy of the ICD codes when we construct the distance metrics for the manifold learning. Besides, the new concept level distance CD_{new} usually achieves the higher AUC scores than CD_{WP}, which means CD_{new} defined more proper distance between ICD codes for the prediction task. Among the four set-level distance metrics, SD_2 performed best because the AUC scores of LE (black and red lines) in Figure 8 (b) are higher than that in Figure 8 (a), (c), (d).



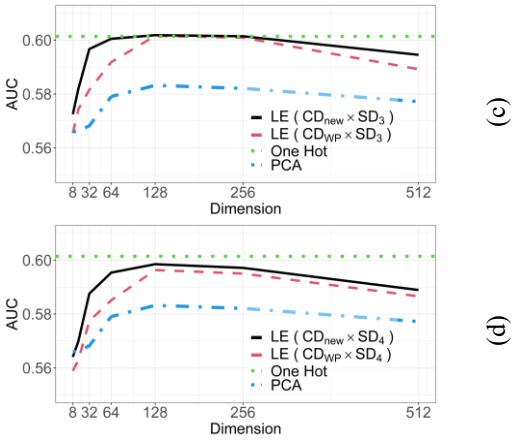
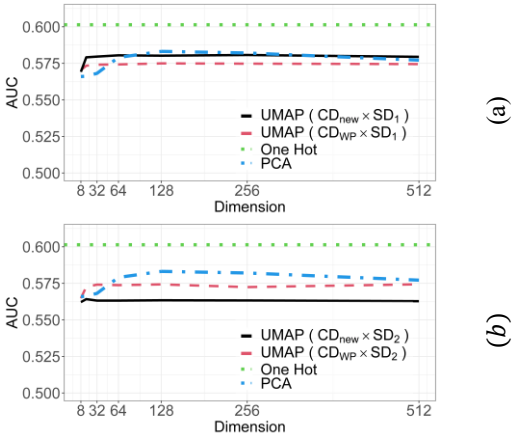


***Figure 8**: The AUC scores of the prediction task of representations through LE. The results of representations using SD_1, SD_2, SD_3, and SD_4 are separated to (a), (b), (c), and (d).*

Surprisingly, the representations through UMAP did not perform as well as LE in the experiments, as shown in Figure 9. First, none of the representations by UMAP outperform the two baselines. The AUC score of One Hot encoding (0.601) is above all UMAP representations. PCA behave similarly to the UMAP with the first and second set-level distance, SD_1 and SD_2, while PCA outperform the UMAP with SD_3 and SD_4. Second, unlike LE, the AUC scores of UMAP are stable across the dimensions. When the dimension of representations increases from 8 to 512, the AUC scores of UMAP vary within 0.01. Third, in the UMAP, the proposed concept-level distance, CD_{new}, outperforms the CD_{WP} when combined with the set-level distance, SD_3, as shown in Figure 9 (c). The possible reasons for the unsatisfying performance of UMAP could be that the UMAP did not capture enough global structure between visits in our experiments, as mentioned in [7]. Also, most applications of the UMAP are supervised tasks, such as visualization [7]. The UMAP may not be the best choice for supervised tasks like our experiments.



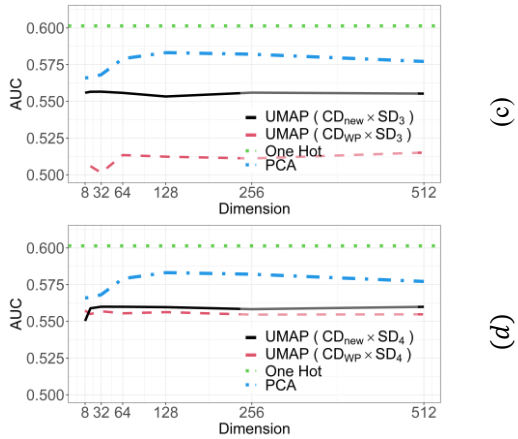
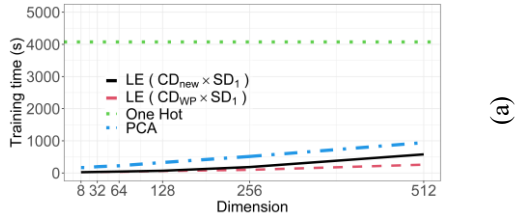


***Figure 9**: The AUC scores of the prediction task of representations through UMAP. The results of representations using SD_1, SD_2, SD_3, and SD_4 are separated to (a), (b), (c), and (d).*

The low dimensional representations save much time on model training. Figure 10 shows the total training time used in the five-fold cross-validation of LR on the representations from baselines and LE, where the green dotted horizontal line indicates One Hot encoding, the blue dotted line indicates PCA, and the black solid and red dash line indicate LE with CD_{new} and CD_{WP}, respectively. Intuitively, as the dimensionality of the manifold representations increases, the training time goes up. More importantly, most of the representations end training in 500 seconds, while it takes 4,075 seconds to train the 17,000 dimensional One Hot encoded representations. At dimension 64, where our representations achieve higher AUC than One Hot encoding, our representations ($30\ s$) spend 99% less time on training than One Hot encoding. Notice that the blue lines (PCA) are always above the solid black lines (LE with CD_{new}) and the red dash lines (LE with CD_{WP}) in Figure 10 (a)-(d), which means the LR is relatively easier to converge on our representations compared with PCA. This reflects the better quality of our representations than the PCA from another point of view. Due to the unsatisfying AUC of UMAP, we do not present the training time of UMAP.



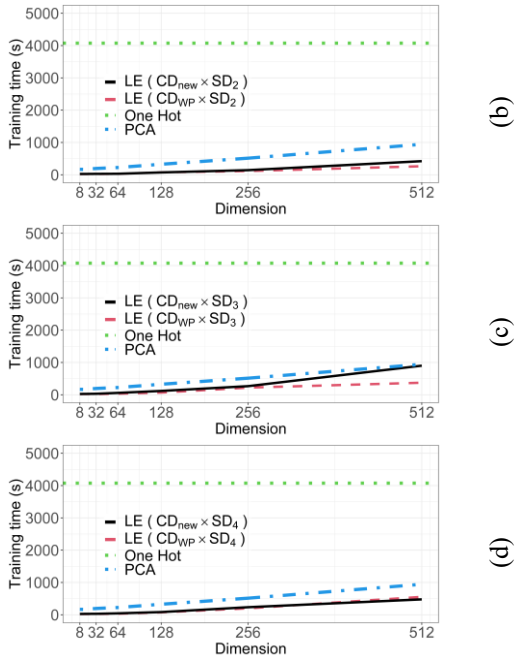


***Figure 10**: The training time of the LR on representations through LE. The results of representations using SD_1, SD_2, SD_3, and SD_4 are separated to (a), (b), (c), and (d).*

To conclude, by incorporating the domain knowledge, the proposed low dimensional representations through LE preserved more information than PCA, which even exceeded the high dimensional One Hot encoding. Through LE, the new concept-level distance, CD_{new}, outperforms the previous metric, CD_{WP}. Combined with either of the two concept-level distances, SD_2 produced the most informative representations among the four set-level distance metrics. The representations through UMAP did not perform well in the task of readmission prediction in terms of AUC. Furthermore, the generated low-dimensional representations saved much time for model training. Due to the promising performance of our representations in low dimensions, our framework showed its great potential in medical and machine learning fields.

5. Discussion and Conclusion

One advantage of the proposed method is that the generated low-dimension representations are robust to low-quality data where similar but inaccurate concepts are documented (e.g., health providers may record the parent or siblings of the accurate disease code). Since the proposed method is based on the concept hierarchical structure, substituting a concept with a similar concept in the data records does not affect the measured distance between concepts significantly. Thus the generated representations still

retain accurate information. On the other hand, One Hot encoding and PCA do not take the medical concept hierarchy into consideration, and thus two similar concepts are regarded as completely different. Therefore, One Hot encoding and PCA are sensitive to the quality of data.

In the experiments section, we only include features of patient diagnosis in the LR to ensure the fair evaluation of the produced representations. There might be multicollinearity between the representations of diagnosis and other features, such as demography information. In that case, AUC scores of LR that is trained on the mixed features do not reflect the true information content in the representations. On the other hand, [3] conducted a readmission prediction study for heart failure patients on the NRD 2013 dataset (ours is NRD 2014). Although used almost all features in the database and many complex deep learning models, the best AUC in the study of [3] is 0.643, only 0.035 higher than ours, which reflects the effectiveness of the low-dimension representation generated by our framework in the readmission prediction.

Considering the computational efficiency, we selected LE and UMAP in the third step of the proposed framework, which is not thorough. We plan to explore more manifold learning algorithms in the future, such as Isomap and Locally Linear Embedding. Moreover, different machine learning tasks, such as mortality prediction, are in need in the coming work for the comprehensive evaluation of the representations. Surprisingly, we also found that as a state-of-art manifold learning algorithm, Uniform Manifold Approximation and Projection (UMAP) did not perform as well as Laplacian Eigenmap (LE) in our experiments. Therefore, we also plan to conduct more experiments to investigate the insights of the unexpected and unsatisfying performance of UMAP.

To sum up, in this study, we proposed a new framework to generate low-dimensional representations for patient hospital visits by combining the medical concept-structure based distance and manifold learning. By considering the well-organized hierarchy of the medical concepts when constructing the distance metrics between visits, we incorporated medical domain knowledge into the representations. In the experiments of readmission prediction for heart failure patients, we showed the great potential of the proposed framework-the generated representations can be more informative than the original data. Not only exceed PCA, our representations also reached higher AUC in the low dimensionality than the high-dimensional One Hot encoding. Moreover, our proposed concept-level distance metric, which is the first step in our framework, outperforms the existing metric in the experiments. From the perspective of applications, our framework could boost the readmission study, as shown in the experiments, and improve other machine learning studies in the research area of healthcare, such as mortality prediction.

6. Reference

[1] D. J. Cartwright, “ICD-9-CM to ICD-10-CM Codes: What? Why? How?,” *Adv. Wound Care*, 2013.

[2] Y. Choi, C. Y.-I. Chiu, and D. Sontag, “Learning Low-Dimensional Representations of Medical Concepts.,” *AMIA Jt. Summits Transl. Sci. proceedings. AMIA Jt. Summits Transl. Sci.*, 2016.

[3] A. Allam, M. Nagy, G. Thoma, and M. Krauthammer, “Neural networks versus Logistic regression for 30 days all-cause readmission prediction,” *Sci. Rep.*, vol. 9, no. 1, pp. 1–11, 2019.

[4] Z. Wu and M. Palmer, “Verbs semantics and lexical selection,” 1994.

[5] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, “GRAM: Graph-based attention model for healthcare representation learning,” 2017.

[6] M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” *Neural Comput.*, 2003.

[7] L. McInnes, J. Healy, N. Saul, and L. Großberger, “UMAP: Uniform Manifold Approximation and Projection,” *J. Open Source Softw.*, 2018.

[8] Z. Jia, X. Lu, H. Duan, and H. Li, “Using the distance between sets of hierarchical taxonomic clinical concepts to measure patient similarity,” *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, pp. 1–11, 2019.

[9] C. Van Walraven *et al.*, “Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community,” *CMAJ*, 2010.

[10] S. Yu, F. Farooq, A. van Esbroeck, G. Fung, V. Anand, and B. Krishnapuram, “Predicting readmission risk with institution-specific prediction models,” *Artif. Intell. Med.*, 2015.

[11] H. W. Kuhn, “The Hungarian method for the assignment problem,” *Nav. Res. Logist. Q.*, 1955.

[12] D. Aouada, Y. Baryshnikov, and H. Krim, “Mahalanobis-based adaptive non-linear dimension reduction,” 2010.

[13] J. D. Frizzell *et al.*, “Prediction of 30-Day All-Cause Readmissions in Patients Hospitalized for Heart Failure,” *JAMA Cardiol.*, 2017.

[14] Agency for Healthcare Research and Quality (AHRQ), “Overview of the Nationwide Readmissions Database (NRD),” *Healthcare Cost and Utilization Project (HCUP)*, 2016. .

[15] J. Huang and C. X. Ling, “Using AUC and accuracy in evaluating learning algorithms,” *IEEE Trans. Knowl. Data Eng.*, 2005.