WellFactor: Patient Profiling using Integrative Embedding of Healthcare Data

Dongjin Choi*, Andy Xiang†, Ozgur Ozturk†, Deep Shrestha†,
Barry Drake‡, Hamid Haidarian†, Faizan Javed†, and Haesun Park*§

*School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA
†Kaiser Permanente, USA
‡Georgia Tech Research Institute, Atlanta, GA, USA
*jin.choi@gatech.edu, hpark@cc.gatech.edu
†{Andy.X.Xiang, Ozgur.X.Ozturk, deep.x.shrestha, Hamid.Haidarian, Faizan.X.Javed}@kp.org
‡drakeleeb@gmail.com

Abstract—In the rapidly evolving healthcare industry, platforms now have access to not only traditional medical records, but also diverse data sets encompassing various patient interactions, such as those from healthcare web portals. To address this rich diversity of data, we introduce WellFactor: a method that derives patient profiles by integrating information from these sources. Central to our approach is the utilization of constrained low-rank approximation. WellFactor is optimized to handle the sparsity that is often inherent in healthcare data. Moreover, by incorporating task-specific label information, our method refines the embedding results, offering a more informed perspective on patients. One important feature of WellFactor is its ability to compute embeddings for new, previously unobserved patient data instantaneously, eliminating the need to revisit the entire data set or recomputing the embedding. Comprehensive evaluations on real-world healthcare data demonstrate WellFactor's effectiveness. It produces better results compared to other existing methods in classification performance, yields meaningful clustering of patients, and delivers consistent results in patient similarity searches and predictions.

Index Terms—Patient profiling, Healthcare, Nonnegative matrix factorization, Recommendation systems

I. INTRODUCTION

The digital revolution and the rise of healthcare web portals have significantly changed patient interactions within the healthcare domain. While these portals primarily serve purposes such as accessing personal clinical records or scheduling appointments, they also store extensive data on patient activities, including searches for clinical information and browsing patterns [1]–[4]. This surge in healthcare big data comes with challenges. One such challenge is the infrequent nature of user interactions on these portals. While these portals experience infrequent user interactions, unlike generic web domains, they maintain a more accurate record of user demographics and offer detailed clinical diagnosis information.

Profile-based models, leveraging user interaction histories, have been effectively used for recommendation across various domains [5]–[8]. Studies on embedding often focus on using the learned vector in downstream machine learning models
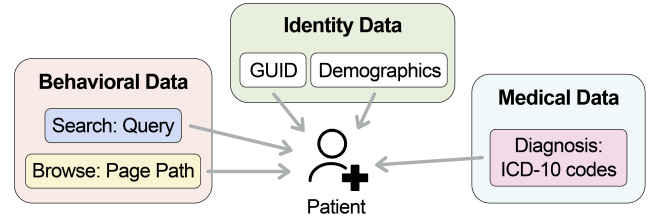
Fig. 1. Illustration of the diverse patient data sources collected from interactions on the web portal and with medical professionals.

regardless of their numerical value [9]. In contrast, our approach, akin to profile-based models, recognizes the latent significance of each dimension in the learned profiles for further recommendations. This is demonstrated in our *Cluster Analysis* (subsection V-D), where embedding dimensions directly correspond to key patient cluster characteristics.

This study aims to create comprehensive patient profiles on healthcare web portals and utilizes demographic data for refined classification and recommendation tasks. These profiles integrate interaction data with diagnostic information, offering a robust foundation for various applications, including personalized healthcare recommendations. While many digital platforms aim solely to enhance user engagement, healthcare portals have additional considerations. Consequently, every patient embedding or subsequent recommendation must reflect both patient preferences and prospective health advantages. For instance, demographic nuances, such as age, can significantly influence user interaction patterns. Younger users, notably those in their 20s and 30s, may display a pronounced affinity for mobile applications, making them more receptive to app-based health recommendations [10]–[13]. Conversely, older users, particularly those in their 50s and above, may also require more mental wellness support due to chronic illnesses and related pains [14]–[16].

In this study, we introduce WellFactor, an algorithm based on constrained low rank approximation (CLRA) that particularly employs nonnegativity constraints. This approach facilitates the integration of diverse user feature vectors directly within the objective function. Several innovative aspects define WellFactor:

- **Integration of data**: The incorporation of integrative objective function and nonnegativity constraints, WellFactor seamlessly blends various data sources, creating a comprehensive representation of patient profiles. Moreover, our method handles missing or unobserved data domains, addressing few-shot scenarios.
- **Efficient embedding computation**: WellFactor utilizes an alternating block coordinate descent algorithm optimized for unique characteristics of the objective function. It avoids materializing extensive data matrices. Moreover, WellFactor predicts embeddings for previously unseen patients without the need to re-examine the entire data.
- **Task-specific embeddings**: With the incorporation of semi-supervision, WellFactor is tailored to produce high-quality embeddings, particularly optimized for certain tasks. The enhanced embeddings offer a more refined view of patients.

We evaluated WellFactor using real-world datasets obtained from Kaiser Permanente's web portal. Our evaluation process has shown the effectiveness of WellFactor compared to other methods. The results consistently showed that WellFactor outperformed other existing methods, particularly in terms of classification accuracy. The method demonstrated its capacity to generate meaningful patient clusters with the potential to tailor personalized healthcare recommendations. Moreover, its efficacy in patient similarity searches and predictive tasks showcases its comprehensive capability to utilize and represent patient data. Our results indicate the effectiveness of WellFactor in handling healthcare data and its potential use for healthcare professionals and researchers.

II. RELATED WORK

As described in Section I, the recommendation of wellness apps in healthcare web portals is related to several areas such as content-based recommendation, prediction in the healthcare domain, clustering, and nonnegative matrix factorization. We offer a brief literature review on each of these topics related to our proposed approach.

A. Content-based Recommendation

Our goal is to produce personalized recommendations tailored to individual user preferences and needs. Such methodologies have been explored under content-based recommendation. Techniques that rely on the representation of item contents that align with user interests are the basis of this approach. Profile-based models emphasize the formation of user preferences and interests as vectors or lists. Some models construct user profiles based on users' search or browsing history within a specific time window [7], [8]. A clustering-based approach groups user trajectories into distinct clusters, suggesting different recommendations for each group [17]. Another approach treats recommendations as a classification problem [18].

Several popular techniques are based on Artificial Neural Networks (ANNs), with some incorporating aspects of the Markov Decision Process [19]. For instance, some methods employ reinforcement learning techniques to track the evolving interests of users [20], [21]. Others, such as the self-attention-based method [22], utilize a time embedding model.

B. Contextual Embedding Methods and Clinical Domain-Specific Embeddings

Contextual embedding methods generate vector representations of text that capture semantic nuances. GPT-2 (Generative Pre-trained Transformer 2) [23] is a method used for capturing long-term textual dependencies. SentenceBERT (SBERT) [24] produces a fixed-size embedding vector for an entire text input. Beyond generic embeddings, the healthcare domain demands specialized embedding techniques to understand the semantics of medical text. BioSentVec is one such approach which focuses on matrix factorization-based embeddings for biomedical sentences [25].

C. Prediction in the Healthcare Domain

The method we introduce is related to recommendations within healthcare web portals. Previous research, such as KETCH [26], has explored recommending relevant threads on healthcare forums based on user symptoms or conditions. The active incorporation of user diagnosis data in our study is inspired by such insights. Our research direction is motivated by representation techniques, such as Metacare++ [27], that leverage the hierarchical structure of ICD-9 diagnosis codes.

D. Clustering and Constrained Low Rank Approximation

Our profiling technique is inspired by clustering methods, especially those focused on constrained low rank approximation [28], [29]. The proposed method yields low-dimensional representations, interpretable as probability distributions. Clustering methods have been applied in recommendation systems, including session clustering for web page recommendations [30] and user clustering based on time-framed data [31]. Clustering based on evolving browsing data has also been used for recommendations [32]. Recommendations have been improved using multi-view clustering techniques [33]. Matrix factorization-based techniques have been employed to produce transparent recommendations [34]. The MEGA model [35] uses nonnegative matrix factorization with a wide range of hypergraphs.

III. HEALTHCARE WEB PORTAL DATA

In this study, our primary focus is on the integration of clinical data with internet activity data within the healthcare domain. Given the proprietary nature of our primary dataset, we detail its characteristics to enhance reproducibility and discuss its potential applicability to analogous public datasets. The data we used in our research, sourced from Kaiser Permanente, has been anonymized in accordance with the Health Insurance Portability and Accountability Act (HIPAA)[1], ensuring the privacy and security of patients' information.

---

[1] Health Insurance Portability and Accountability Act (https://www.hhs.gov/hipaa/)
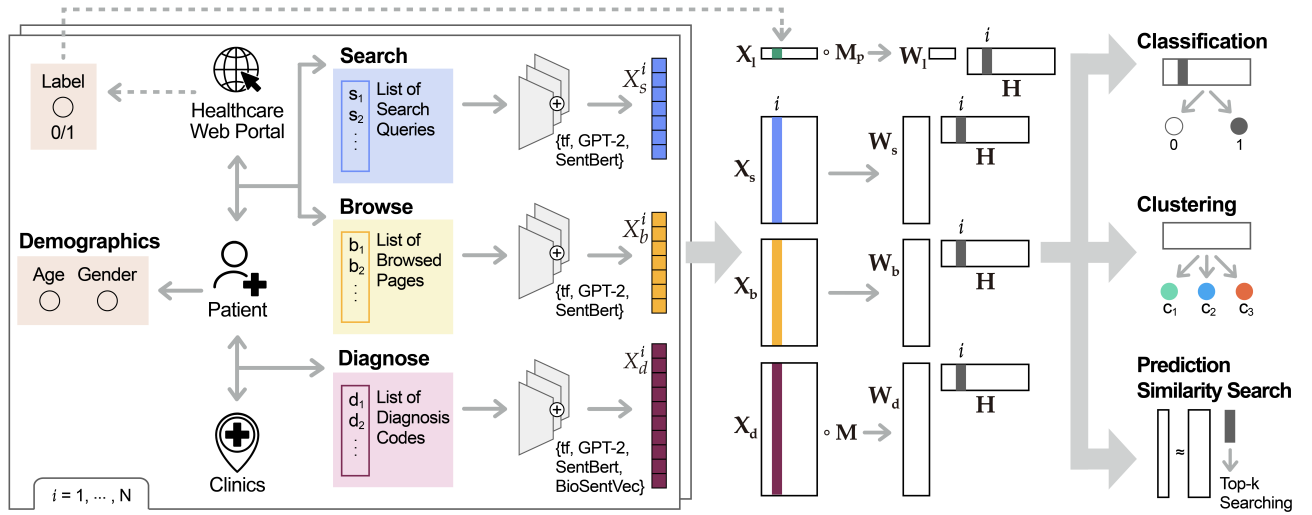
Fig. 2. Graphical overview of the proposed WellFactor patient profiling framework.

TABLE I
STATISTICS ON THE DATA SET UTILIZED IN OUR STUDY

	Diagnosis	Search Histories	Browsing Activities	Demographics
# Patients	599,499	298,574	636,150	1,177,031
# Instances	6,268,788	1,220,210	299,099,939	1,177,031
# Fields	3	1	1	2

TABLE II
OVERVIEW OF DATA FIELDS AND EXAMPLES

Data Used	Field Name	Examples
Diagnosis	ICD-10	R80.9
	official text	*PROTEINURIA*
	friendly text	*PROTEINURIA (PROTEIN IN URINE)*
Search Histories	query	*How to eat low carb*
Browsing Activities	site-path	*'kporg:health-wellness:healtharticle.40-positive-affirmations'*
Demographics	age	48
	gender	male

The data capture patient interactions on the Kaiser Permanente Digital (KPD) website[2]. As depicted in Figure 1, this dataset offers a detailed view of patient information. It encompasses medical diagnoses based on ICD-10 codes [36] gathered during interactions with medical professionals, and a broad spectrum of digital user interactions: search histories, browsing activities, and demographic particulars such as age and gender. All information, especially medical diagnoses, is anonymized to ensure confidentiality. For a more granular overview of the dataset, refer to Table I.

Table II provides an overview of the data sets, their fields,

and examples for each data set. Within the diagnosis data, the primary field is the ICD-10 codes. These codes are derived from treatments patients receive at Kaiser Permanente (KP) and its affiliated clinics. Confirmed diagnoses from these interactions are archived in KP's electronic health record systems. Each diagnosis in KP's electronic health record systems includes an "official text" (medical terminology) and a "friendly text" (more colloquial name) annotated by KP specialists for internal reference.

The data set also contains patients' search histories on the KP web portal. This platform allows users to input queries to locate appointments, articles, and medical providers. These search results are presented similarly to a conventional search engine results page (SERP), and we specifically collected the query expressions patients entered. Moreover, the browsing activity data set contains the patients' web page interaction information on both desktop and mobile platforms. This extends beyond search results to include pages displaying personal data, clinician specifics, and appointment-booking locations within the KP framework. We gathered information on web pages accessed by users, which are marked with the path signifying its position in KP's internal web hierarchy. Furthermore, essential demographics, such as age and gender, were also recorded in the system. For all data, we limited the data period of 2022 (from 2022-01-01 to 2022-12-31)

A. Label Collection for Semi-supervision and Evaluation

For semi-supervision and evaluation, we sourced labels based on users' interactions with a mental health app banner displayed on KP's home page during 2022. These apps include Calm, Ginger, and myStrength. A cohort of randomly selected users during this timeframe was exposed to this banner, which directed them to download the self-care apps. Our labels specifically identify whether a user clicked on this banner. After label collection, we found that a mere 15,382 users interacted with the banner, a small portion when compared

with the total number of patients (1.1 M). Given this significant imbalance, undersampling will be adopted in subsequent experiments, such as classification.

B. Relation to Public Datasets

Our research dataset integrates clinical and digital domain records for over a million individuals, differing from datasets like MIMIC-IV [37], which primarily offers patient anonymized clinical notes and demographics. The integrative nature of our dataset underscores its value and the opportunities it presents, even when public data is favored for reproducibility. As public datasets evolve, aligning them with diverse data sources will make them more valuable. A future direction involves integrating textual data with public clinical datasets like MIMIC-IV.

IV. METHODOLOGY

Our proposed patient profiling method integrates multiple sources of data, distinguishing it from existing methods. By integrating information from patients' digital interactions and medical records using our algorithms, our approach aims to provide a broader understanding of patients. This is further enhanced by the employment of semi-supervision techniques. The architecture of our method is presented in Figure 2.

Leveraging the comprehensive dataset detailed in section III, our methodology makes full use of all user data present in the digital healthcare platform. Digital healthcare platforms typically provide user interaction data with digital interfaces, akin to the e-commerce sector, and diagnostic data from user interactions with medical experts. We propose algorithms that can effectively integrate multiple sources of heterogeneous data domains, producing more accurate patient profiles. While the details of our algorithms and implementations are demonstrated using the search, browsing, and diagnosis data views from section III, our method can be adapted to settings with any number of heterogeneous data domains.

In this section, we introduce a new approach for learning user profiles that utilizes multiple data sources simultaneously. The features obtained from the proposed method provide a richer representation of user behavior and preferences, as they integrate information from various sources, such as content, semantic relationships, and domain-specific information. Some of the commonly utilized methods for information fusion include early fusion which merges raw data at the data representation level [38] and late fusion which solves a given problem applying solution methods separately to each data set and then merges the results [39]. Our method integrates the objective functions from all data sets into a single objective function.

The objective function level information fusion method we introduce here does not require an input matrix that contains all merged raw data as in early fusion. Instead, each part of the merged objective function takes each representation of a data set as its input. Then the goal of the overall merged objective function is to find one common lower-dimensional embedding that captures the essential information from all *views* of the data by computing a common low rank factor.

A. Feature Processing

In order to provide effective recommendations, we consider various user features derived from collected data. We utilize the content information representing the content in the standard Term Frequency (TF) encoding. In addition, we incorporate features that capture semantic relationships between words and phrases within the user's textual data. We use GPT-2 [23], a generative language model, and sentenceBERT [24], a variation of the BERT model, optimized for sentence-level representations, to process all data types, including search, browsing, and diagnosis records. For the diagnosis data, we additionally utilize BioSentVec [25], a sentence embedding model specifically trained on biomedical texts, to capture the domain-specific semantic information more effectively. Since the range of elements in these three matrices from different data sets may vary significantly, we use Min-Max scaling [40] for each matrix.

B. Learning User Profiles

To illustrate the details, we assume that we have the three different views for a set of n users, i.e., the users' web search data, browsing data, and diagnosis data. Then we generate three feature-by-user matrices, which are search-by-user matrix $\mathbf{Y}_s \in \mathbf{R}^{m_s \times n}$, browsing-by-user matrix $\mathbf{Y}_b \in \mathbf{R}^{m_b \times n}$, and diagnosis-by-user matrix $\mathbf{Y}_d \in \mathbf{R}^{m_d \times n}$. Given a matrix $\mathbf{Y}_i$ where $i \in \{s, b, d\}$, the scaled matrix $\mathbf{X}_i$ is obtained as

$$\mathbf{X}_i = (\mathbf{Y}_i - \min(\mathbf{Y}_i))/(\max(\mathbf{Y}_i) - \min(\mathbf{Y}_i)),$$

where $\min(\mathbf{Y}_i)$ denotes the matrix of the same size as $\mathbf{Y}_i$ where all elements are identically set to the minimum of $(\mathbf{Y}_i)$, and $\max(\mathbf{Y}_i)$ is defined analogously using the maximum value of the elements in $\mathbf{Y}_i$.

If we were to find a low rank approximation for just one of the data sets $\mathbf{X}_i$ via nonnegative matrix factorization (NMF) [41], then the objective function would be

$$\min_{\{\mathbf{W_i}, \mathbf{H}\} \geq 0} \|\mathbf{Y}_i - \mathbf{W}_i \mathbf{H}\|_F.$$

Now we merge three objective functions and produce a common embedding. Then the objective function for this method is as follows:

$$\min_{(\mathbf{W}_s, \mathbf{W}_b, \mathbf{W}_d, \mathbf{H}) \geq 0} \alpha_s \|\mathbf{X}_s - \mathbf{W}_s \mathbf{H}\|_F^2 + \alpha_b \|\mathbf{X}_b - \mathbf{W}_b \mathbf{H}\|_F^2$$
$$+ \alpha_d \|\mathbf{X}_d - \mathbf{W}_d \mathbf{H}\|_F^2, \quad (1)$$

where α_s, α_b, α_d denote balancing factors for each low-rank approximation term, and $\mathbf{X}_s$, $\mathbf{X}_b$, $\mathbf{X}_d$ represent the feature by data matrices from each domain: search, browsing, and diagnosis, respectively. Note that the factor H is common across all domains, which provides an embedding in k dimensional space for the data items that reflects their relationships with search, browsing, and diagnosis, simultaneously. The factors $\mathbf{W}_s$, $\mathbf{W}_b$, and $\mathbf{W}_d$ represent the basis matrices in the reduced k dimensional space for each domain.

The matrix $\mathbf{H}$, resulting from solving our merged objective function, serves a dual purpose: soft clustering and embedding. Each column within $\mathbf{H}$, specifically the ith column H^i, encapsulates an integrated embedding of the ith data item. Furthermore, it can also be understood as a probability distribution that illustrates how the ith data item spans across clusters. This approach facilitates a clearer interpretation of results, as the learned embedding can be understood in terms of the contributions of different cluster representatives present in the columns of the basis matrices $\mathbf{W}_s$, $\mathbf{W}_b$, and $\mathbf{W}_d$. This methodology not only uses user profiles for downstream tasks but also provides a latent factor representation, unifying characteristics across all domains. For a more in-depth exploration of the interpretation of the factor matrix $\mathbf{H}$ within the context of the soft clustering, we refer to [41].

To optimize the objective function in Eqn. 1, we adopt a Block Coordinate Descent (BCD) approach. In each iteration of our proposed BCD method, we alternate updating one of the matrices $\mathbf{W}_s$, $\mathbf{W}_b$, $\mathbf{W}_d$, and $\mathbf{H}$ while fixing the other three matrices by solving the following subproblems until a stopping criteria is satisfied:

$$\min_{\mathbf{W}_i \geq 0} \|\mathbf{X}_i - \mathbf{W}_i \mathbf{H}\|_F, \text{ for } i = s, b, d$$

$$\min_{\mathbf{H} \geq 0} \left\| \begin{bmatrix} \sqrt{\alpha_s} \mathbf{X}_s \\ \sqrt{\alpha_b} \mathbf{X}_b \\ \sqrt{\alpha_d} \mathbf{X}_d \end{bmatrix} - \begin{bmatrix} \sqrt{\alpha_s} \mathbf{W}_s \\ \sqrt{\alpha_b} \mathbf{W}_b \\ \sqrt{\alpha_d} \mathbf{W}_d \end{bmatrix} \mathbf{H} \right\|_F. \quad (2)$$

Each of the four subproblems is a nonnegativity-constrained least squares (NLS) problem and there are several methods that can effectively solve these NLS problems. We utilize the BPP (Block Principal Pivoting) method as it has been shown to produce the best performance in previous extensive studies and for various possible stopping criteria, see [41]. Assuming that each subproblem has a unique solution, the limit point of the iteration will be a stationary point [41], [42].

In fact, the same objective function can be expressed as

$$\min_{(\tilde{\mathbf{W}}, \mathbf{H}) \geq 0} \left\| \begin{bmatrix} \sqrt{\alpha_s} \mathbf{X}_s \\ \sqrt{\alpha_b} \mathbf{X}_b \\ \sqrt{\alpha_d} \mathbf{X}_d \end{bmatrix} - \tilde{\mathbf{W}} \mathbf{H} \right\|_F, \quad (3)$$

which is the same as applying an NMF to the $\mathbf{X}_i$'s scaled by $\sqrt{\alpha_i}$ and stacking them up. Then by scaling and partitioning the computed factor $\tilde{\mathbf{W}}$, we obtain the $\mathbf{W}_i$'s. However, there are added advantages of the proposed objective function level fusion, in terms of its generalizability. The first is when some view of the data is in the form of data-data relationships. For example, suppose we have an additional view of the data representing the relationships or interactions among the users, which is represented in a similarity matrix $\mathbf{A}$. Then we can add an additional term of Symmetric NMF [43], [44], $\alpha_a \left\| \mathbf{A} - \mathbf{H}^T \mathbf{H} \right\|_F$ to Eqn. 1 and compute the common $\mathbf{H}$ factor that represents all four views of the data. This data-data relationship information cannot be represented in the form in Eqn. 3. In addition, when there are some missing elements in a data matrix, we can compute the common factor $\mathbf{H}$ bypassing the missing elements, i.e., not letting the missing elements

influence the result. We discuss this in detail in the next section for Eqn. 1.

C. Unobserved Data: Assumptions and Matrix Masking

Data and features in each domain are not always fully observed. It is essential to develop a method to handle unobserved or missing data and features. This approach aids in producing accurate recommendations based on the available information. For search and browsing data, we assume a closed-world assumption [45], meaning that unobserved matrix entries indicate no existing relationship. This is due to users having the freedom to search and browse, and they can also choose not to engage in such activities based on their intentions. On the other hand, for diagnosis data, we utilize an open-world assumption [46], i.e., unobserved matrix entries are considered to represent an *unknown* relationship. This is because the absence of a diagnosis does not always reflect a user's intention. Instead, it may indicate that the user has not received a diagnosis from a medical expert.

In order to handle unobserved entries in the diagnosis data, we introduce a masking matrix $\mathbf{M} \in \{0,1\}^{m \times n}$, where its entry is 1 when the corresponding entry in the diagnosis matrix $\mathbf{X}_d$ is observed and 0 when it is not observed. We modify the objective function in Eqn. 1 to incorporate the masking matrix as follows

$$\min_{(\mathbf{W}_s, \mathbf{W}_b, \mathbf{W}_d, \mathbf{H}) \geq 0} \alpha_s \|\mathbf{X}_s - \mathbf{W}_s \mathbf{H}\|_F^2 + \alpha_b \|\mathbf{X}_b - \mathbf{W}_b \mathbf{H}\|_F^2$$
$$+ \alpha_d \|\mathbf{M} \circ (\mathbf{X}_d - \mathbf{W}_d \mathbf{H})\|_F^2. \quad (4)$$

As in the previous section, we use the BCD framework to solve Eqn. 4, updating the four factor matrices in each iteration. Updating of $\mathbf{W}_s$ and $\mathbf{W}_b$ can be done in the same way as in Eqn. 2, respectively. However, the updating of $\mathbf{H}$ and $\mathbf{W}_d$ will be different due to the masking matrix. Considering the effects of the masking matrix, the subproblems in Eqn. 2 change as follows:

$$\min_{\mathbf{W}_d \geq 0} \|\mathbf{M} \circ (\mathbf{X}_d - \mathbf{W}_d \mathbf{H})\|_F,$$

$$\min_{\mathbf{H} \geq 0} \left\| \begin{bmatrix} \sqrt{\alpha_s} \mathbf{X}_s \\ \sqrt{\alpha_b} \mathbf{X}_b \\ \mathbf{M} \circ \left(\sqrt{\alpha_d} \mathbf{X}_d\right) \end{bmatrix} - \begin{bmatrix} \sqrt{\alpha_s} \mathbf{W}_s \mathbf{H} \\ \sqrt{\alpha_b} \mathbf{W}_b \mathbf{H} \\ \mathbf{M} \circ \left(\sqrt{\alpha_d} \mathbf{W}_d \mathbf{H}\right) \end{bmatrix} \right\|_F.$$

Accordingly, $\mathbf{W}_d$ and $\mathbf{H}$ can be updated row by row and column by column, respectively, using the following rules:

$$\min_{\mathbf{W}_d(j,:) \geq 0} \|\mathbf{X}_d(j,:) D(\mathbf{M}(j,:)) - \mathbf{W}_d(j,:) \mathbf{H} D(\mathbf{M}(j,:))\|_F,$$
$$(5)$$

$$\min_{\mathbf{H}(:,j) \geq 0} \left\| \begin{bmatrix} \sqrt{\alpha_s} \mathbf{X}_s(:,j) \\ \sqrt{\alpha_b} \mathbf{X}_b(:,j) \\ \sqrt{\alpha_d} D(\mathbf{M}(:,j)) \mathbf{X}_d(:,j) \end{bmatrix} - \begin{bmatrix} \sqrt{\alpha_s} \mathbf{W}_s \\ \sqrt{\alpha_b} \mathbf{W}_b \\ \sqrt{\alpha_d} D(\mathbf{M}(:,j)) \mathbf{W}_d \end{bmatrix} \mathbf{H}(:,j) \right\|_F,$$
$$(6)$$

where $D(\mathbf{z})$ denotes the diagonal matrix whose diagonal entries are defined by the given row or column vector $\mathbf{z}$, i.e., $D(\mathbf{z}) = [d_{ij}]_{n \times n}$ and $d_{ii} = z_i$ where $\mathbf{z} \in \mathbb{R}^n$ and z_i is the i-th component of z.

D. Semi-Supervised Embedding

Incorporating partially known prior information can enhance the quality of patient profiles. Our semi-supervised learning approach, grounded in data-level supervision, exploits the partially observed labels of the data items to refine the embedding quality. This semi-supervised approach optimizes the patient profiling framework's efficacy, potentially leading to improved outcomes in applications.

Given the availability of partial label information for data items, this prior knowledge can be seamlessly integrated into the embedding algorithm using the label matrix $\mathbf{X}_l$. The matrix $\mathbf{X}_l \in \{0,1\}^{p \times n}$ denotes the partially observed labels where p represents the number of distinct label classes for users. An entry $x_l^{ij} = 1$ indicates that the jth object is part of the ith class.

In our application context, a data item represents a user. As highlighted in Section III, we have label information indicating whether a user has accessed any mental wellness support apps, particularly those focused on self-care such as Calm, Ginger, and myStrength. It is worth noting that while we exemplify the semi-supervision process using labels of mental wellness support app download attempts in this context, the label could be representative of various patient statuses. This might include particular diagnoses or interactions with specific medical resources, emphasizing the flexibility of our approach.

By integrating this partially observed label information into our semi-supervised learning framework, the algorithm makes use of the available information to capture the underlying structure of the user. Consequently, this contributes to more informed patient profiles. To implement this semi-supervision, we extend the objective function by including a new term that minimizes the Frobenius norm of the difference between the observed labels $\mathbf{X}_l$ and their approximated values $\mathbf{W}_l\mathbf{H}$, where $\mathbf{W}_l$ is an additional basis matrix for labels:

$$\min_{(\mathbf{W}_s, \mathbf{W}_b, \mathbf{W}_d, \mathbf{W}_l, \mathbf{H}) \geq 0} \alpha_s \|\mathbf{X}_s - \mathbf{W}_s\mathbf{H}\|_F^2 + \alpha_b \|\mathbf{X}_b - \mathbf{W}_b\mathbf{H}\|_F^2$$
$$+ \alpha_d \|\mathbf{M} \circ (\mathbf{X}_d - \mathbf{W}_d\mathbf{H})\|_F^2 + \alpha_l \|\mathbf{M}_l \circ (\mathbf{X}_l - \mathbf{W}_l\mathbf{H})\|_F^2 ,$$

where α_l is a regularization parameter that controls the trade-off between fitting the observed labels and the other components of the objective function, and $\mathbf{M}_l$ is an entry-wise masking matrix that indicates whether an entry in $\mathbf{X}_l$ is observed or not.

The update process for $\mathbf{W}_l$ is analogous to the columnwise update of $\mathbf{W}_d$ as shown in Eqn. 5. Therefore, the detailed procedure will be skipped for brevity. The columnwise update of $\mathbf{H}$ can be extended from Eqn. 6 and is expressed as follows:

$$\min_{\mathbf{H}(:,j) \geq 0} \left\| \begin{bmatrix} \sqrt{\alpha_s}\mathbf{X}_s(:,j) \\ \sqrt{\alpha_b}\mathbf{X}_b(:,j) \\ \sqrt{\alpha_d}D(\mathbf{M}(:,j))\mathbf{X}_d(:,j) \\ \sqrt{\alpha_p}D(\mathbf{M}_p(:,j))\mathbf{P}(:,j) \end{bmatrix} - \begin{bmatrix} \sqrt{\alpha_s}\mathbf{W}_s \\ \sqrt{\alpha_b}\mathbf{W}_b \\ \sqrt{\alpha_d}D(\mathbf{M}(:,j))\mathbf{W}_d \\ \sqrt{\alpha_P}D(\mathbf{M}_p(:,j))\mathbf{W}_p \end{bmatrix} \mathbf{H}(:,j) \right\|_F$$

E. Embedding of Previously Unseen Data Items

Given the multi-domain information about users, we can predict the embedding for a previously unseen patient based on the already computed bases vectors for search, browsing,

and diagnosis, i.e., $\mathbf{W}_s$, $\mathbf{W}_b$, and $\mathbf{W}_d$. In the following, we show how to achieve this using the integrated nonnegative least squares (NNLS) method. Suppose a new patient signs up and we have information on the person such as search histories, browsing activities, and diagnosis. Using the results we have already computed from the WellFactor method from known patients, we can determine a profile for this new patient. Let q be the patient's ID to be entered as a query. The patient's observed search, browse, and diagnosis records, once processed analogously as described in subsection IV-A, are represented as column vectors X_s^q, X_b^q, and X_d^q. Consequently, computing for $\mathbf{H}^q$ in the equation below offers the patient's profile representation across three subspaces represented in $\mathbf{W}_s$, $\mathbf{W}_b$, and $\mathbf{W}_d$:

$$\min_{H^q \geq 0} \alpha_s \|X_s^q - \mathbf{W}_s H^q\|_F^2 + \alpha_b \|X_b^q - \mathbf{W}_b H^q\|_F^2 \tag{7}$$
$$+ \alpha_d \|X_d^q - \mathbf{W}_d H^q\|_F^2 .$$

The objective function in Eqn. 7 computes the patient embedding, $\mathbf{H}^q$. This embedding succinctly represents the patient based on their individual records, placing them within the established embedding subspace.

To foster reproducibility and encourage further developments in patient profiling, the code for our WellFactor framework has been made publicly available. Interested researchers and developers can access and utilize the codebase via our GitHub repository[3].

V. EVALUATION AND RESULTS

In this section, we evaluate our method across various tasks related to patient profiling and recommendation. Specifically, our evaluation concentrates on three primary tasks: classification (identifying which group a patient belongs to), clustering (grouping similar patients), predicting patient embeddings, and similarity search (finding patients analogous to a given example).

A. Evaluation Setting

1) Competing Methods: We selected several baseline algorithms for user embedding as comparisons for our proposed method:

- **HashGNN [47]:** Given that our problem setting can be conceptualized as graphs, incorporating HashGNN is a logical step. It involves constructing a graph where nodes signify patients, diagnoses, queries, and pages. This method was run with standard parameters.
- **Text Embeddings:** *GPT-2* and *SentenceBERT* are prominent text embeddings. To create user embeddings with them, we computed embeddings for each domain-specific text associated with users and then derived their average.
- **Biomedical-Domain Embedding:** *BioSentVec*, specifically designed for the biomedical domain, was applied only to diagnosis domain texts to compute average embeddings.

---

[3]https://github.com/skywalker5/wellfactor

TABLE III
COMPARISON OF VARIOUS EMBEDDING METHODS EVALUATED USING XGBOOST: METRICS INCLUDE ROC-AUC, ACCURACY, RECALL, PRECISION, AND F1-SCORE. RESULTS ARE PRESENTED IN TERMS OF MEAN PERCENTAGES WITH STANDARD DEVIATIONS. THE **BOLD** VALUES REPRESENT THE BEST SCORES AND <u>UNDERLINED</u> VALUES SIGNIFY THE SECOND-BEST SCORES FOR EACH METRIC.

Method	ROC-AUC	Accuracy	Recall	Precision	F1-score
HashGNN	65.95 ± 0.63	62.09 ± 0.44	65.91 ± 3.84	61.60 ± 1.20	63.61 ± 1.56
GPT-2 (PCA)	76.73 ± 0.52	69.74 ± 0.58	69.92 ± 3.53	70.11 ± 1.79	69.94 ± 1.26
SentBERT (PCA)	75.46 ± 0.58	68.86 ± 0.56	71.96 ± 4.29	67.88 ± 1.39	69.77 ± 1.56
BioSentVec (PCA)	65.37 ± 1.04	61.41 ± 0.82	68.45 ± 6.19	60.44 ± 1.82	64.01 ± 2.15
WellFactor$_{NS}$	<u>81.47</u> ± 0.56	<u>73.68</u> ± 0.70	<u>74.70</u> ± 4.27	**73.64** ± 2.98	<u>74.01</u> ± 0.81
WellFactor$_{S}$	**81.65** ± 0.55	**73.98** ± 0.55	**76.13** ± 1.91	<u>73.11</u> ± 0.97	**74.57** ± 0.79

These competitors were selected based on their common use and relevance to our application. Since GPT-2, SentenceBert, and BioSentVec were utilized to constitute our initial data features, our intent is to assess if an integrative embedding approach with the same embedding size can improve upon performance of these individual methods.

2) Experimental Details: For a fair comparison, we standardized the embedding length at 128 dimensions. If the inherent dimensionality of any method exceeded this, we employed Principal Component Analysis (PCA) to reduce the dimension to 128. In our XGBoost model, we appended patients' age and gender information to the embedding for training and testing, acknowledging age and gender's potential impact on health-related outcomes. We treated gender as a categorical feature, exercising the flexibility XGBoost offers in handling diverse data types. To ensure the reliability and stability of our results, we averaged metrics over 10 experiments and also computed the standard deviations. This iterative approach provides a broad overview of the method's robustness.

B. Classification for App Recommendations

In applying our method, our goal is to predict if a user will engage with a mental health support app among the available options. This task bears resemblance to link prediction or ranking challenges commonly encountered in the realm of information retrieval. To assess the quality of our app recommendations, we collected visitation logs for the relevant download pages of the apps: Calm, Ginger, and myStrength as described in subsection III-A. A binary representation was employed to capture user engagement. Users visiting the download page at least once were assigned a value of 1, and all others were assigned a value of 0. We employed XGBoost for our evaluations, due to its reputation for efficiency and accuracy. XGBoost, an implementation of gradient boosted decision trees, is known for its flexibility in handling various data types, including categorical variables like gender. Consequently, we incorporated patients' age and gender information in concatenation with the computed embeddings, widening data's breadth and depth.

1) Performance Metrics: To evaluate the classification performance, we compare the output predictions with actual labels. We compute the count of categories: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Then, fundamental metrics such as recall (or true positive

rate, TPR), precision, and the F1 score are computed. The classifier's outcome is a list of probabilities of patients being positive. Setting the optimal threshold of the probability will modify the number of predicted positives and metrics like precision and recall. The ROC curve visualizes this by plotting the TPR against the false positive rate (FPR) across thresholds between 0 and 1. The area under this curve, known as the ROC-AUC, gives us a singular metric indicative of our model's performance across all thresholds.

To address the problem that metrics such as recall, precision or F1 score are sensitive to threshold selections, the Youden index offers a way to determine an optimal threshold. This index, mathematically represented as $J = TPR - FPR$, computes the threshold where the balance between sensitivity (or recall) and specificity is maximized. Thus, by optimizing the Youden index, we identify a threshold that best translates our continuous prediction scores into binary outcomes, enabling meaningful comparisons using metrics such as Precision, Recall, and F1 score.

C. Classification results

Our proposed methods, namely WellFactor$_{NS}$ (without semi-supervision) and WellFactor$_{S}$ (with semi-supervision), have showcased remarkable efficacy when contrasted with other baseline techniques. Specifically, the WellFactor$_{S}$ method achieved an impressive ROC-AUC of 81.65%, accuracy of 73.98%, recall of 76.13%, and an F1-score of 74.57%. Its counterpart, WellFactor$_{NS}$, although slightly trailing in some metrics, presented a praiseworthy ROC-AUC of 81.47% and an F1-score of 74.01%, reflecting its competitive performance.

D. Cluster Analysis of Patients

Our approach inherently provides information on both embedding and soft clustering results. One of the unique attributes of WellFactor is the imposition of nonnegative constraints [48]. The dimension with the highest value in a patient's embedding can be interpreted as the predominant cluster that best describes that patient.

The basis matrices, represented by $\mathbf{W}_i (i = s, b, d)$, serve as the outputs of our model, characterizing representative patients for each cluster within specific domains (search, browse, and diagnose). It is important to note that while part of each $\mathbf{W}_i$ matrix is derived from term-frequency matrix decomposition, other rows correspond to other text embeddings. Every row

Data Domain	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Search 1	cpap	autopay	general	autopay	pharmacy
2	apria	auto	optometrist	auto	refill
3	supplies	taxes	mandarin	payments	refills
4	department	1095a	indian	payonline	receipts
5	wang	payonline	appoitment	send	phone
Browse 1	sleep	premium	task	html5	technical
2	problems	estimates	whoops	reprom	dispose
3	durable	billing	ns	pb	unwanted
4	equipment	getting	videovist	vb	termsconditions
5	apnea	manager	footer	autopay	drugs
Diagnose 1	sleep	bloating	nose	resuscitate	meningioma
2	apnea	salivary	dysphagia	myasthenia	myopathy
3	snoring	leiden	idiopathic	gravis	manic
4	cpap	splints	prolactinoma	macrocytosis	fall
5	rhythm	shin	hands	dependent	ll

in $\mathbf{W}_i$ represents either a specific keyword from the term-frequency decomposition or an embedding from the other mentioned models.

Using WellFactor, we derived a 128-dimensional embedding. Out of these 128 clusters, we selectively emphasized the five largest clusters. Then, the top five keywords are selected within each cluster that exhibited the highest values in the corresponding columns of $\mathbf{W}_i$ matrices for each domain. The results are summarized in Table IV. For example, in Cluster 1 in particular, we observe recurring keywords such as 'sleep' and 'apnea' across various domains, demonstrating the robustness of our method. Such findings underline the efficacy of our approach and also suggest its potential application in facilitating group-based recommendations.

E. Prediction and Similarity Search

The subsequent analyses focus on the evaluation of our embedding method for predictability and similarity search. Given the multifaceted nature of healthcare data, it is critical to assess whether our embedding technique can capture the intrinsic relationships between patients and reliably predict their future interactions or diagnoses. Our primary objective is to evaluate the predictability of our embedding technique. Here is how we approached this:

a) Selection of Diseases for Analysis: As presented in Table VI, the diseases chosen for this study were selected based on worldwide prevalence, profound impact on mortality, and significant associations with modifiable risk factors. Hypertension and Diabetes, for instance, are directly influenced by dietary habits, sedentary lifestyles, and obesity [49]. COPD (Chronic Obstructive Lung Disease) emerges primarily due to smoking and environmental pollutants [50], while Chronic Kidney Disease is frequently a complication of other conditions, illustrating the intricate interrelations of these diseases [51]. Depression, exacerbated by modern-day stressors and societal pressures, is becoming increasingly prevalent [52]. Cancers, on the other hand, span a vast range of conditions, with a mix of genetic, environmental, and lifestyle origins [53].

Additionally, the study addresses obesity, with recent findings indicating alarming health effects of being overweight and obesity in a majority of countries over the past 25 years [54]. Osteoarthritis, a degenerative joint disease particularly of the hip and knee, represents another significant global health burden, and its occurrence is expected to increase with an aging population [55].

b) Data Preparation: We initially established cohorts of individuals diagnosed with major diseases in the 4th quarter (Q4) of 2022. The underlying assumption is that people with similar embedding vectors are more likely to share certain health outcomes, including the emergence of specific diseases.

c) Embedding Calculation: Using data from the first three quarters (Q1, Q2, and Q3), we computed the embedding for each patient. This ensures that our embedding results are derived solely from historical data, allowing us to make predictions about events in the upcoming quarter (Q4).

d) Similarity Search: For each patient diagnosed with a major disease in Q4, we calculated the vector distance between their embedding (from Q1-Q3) and the embedding results of all other patients. This allowed us to rank other patients based on their similarity.

e) Performance Assessment: To evaluate the efficacy of our embedding method in identifying similar patients, we measured its accuracy using the "precision@k" metric. This metric indicates the proportion of retrieved relevant patients among the top k predictions. Given a patient diagnosed with a disease in Q4, another patient is considered "relevant" if they were also diagnosed with the same disease in Q4. Formally:

$$\text{Precision@k} = \frac{\text{\# of relevant items in the top k predictions}}{k}.$$

As a baseline, we also compute the theoretical precision when matching patients randomly by ranking patient similarities without any consideration for their embedding vectors. Thus, we aim to demonstrate the absolute baseline performance one might expect without any predictive modeling. Comparing the precision values obtained from our embedding approach with

TABLE V
Performance of the proposed embedding method for predicting major diseases in Q4 based on data from Q1, Q2, and Q3. The evaluation is done by finding the top similar patients using the vector distance in the embedding space and checking how many of them are in the same cohort. The precision@k values indicate the proportion of true positive predictions among the top k predictions.

Disease	Number of Patients in Cohort	precision@10	precision@20	precision@50	Random Precision
Hypertension	3,244	1.25%	1.23%	1.18%	0.51%
Diabetes	5,313	2.78%	2.65%	2.49%	0.83%
COPD	363	0.34%	0.25%	0.20%	0.06%
Chronic Kidney Disease	2,949	2.67%	2.64%	2.41%	0.46%
Depression	2,945	1.28%	1.29%	1.27%	0.46%
Cancers	2,136	0.95%	0.91%	0.91%	0.34%
Obesity	6,159	2.35%	2.32%	2.31%	0.97%
Osteoarthritis	2,269	0.78%	0.80%	0.83%	0.36%

TABLE VI
Selected Diseases and their ICD-10 Codes

Disease	ICD-10 Code
Hypertension	I10
Diabetes	E11, E12, E13, E14
Chronic Obstructive Pulmonary Disease (COPD)	J44
Chronic Kidney Disease	N18
Depression	F32, F33
Cancers	C00-C97
Obesity	E66
Osteoarthritis	M15-M19

the theoretical random baseline offers a clear picture of our method's predictive capability.

f) Discussion of Results: The results, as summarized in Table V, clearly demonstrate the potential of our embedding approach for predicting major diseases for patients based on historical data. For every disease under consideration, our method consistently outperforms the theoretical random baseline, highlighting its efficacy in capturing meaningful patient similarities. Furthermore, even in instances where the precision values might seem modest, such as in COPD, the difference between our method's precision and the random baseline is still marked, reinforcing the utility of our approach for identifying patients with similar health outcomes.

VI. Conclusion

We propose WellFactor, a method for patient profiling using integrative embedding of healthcare data. Our method seamlessly integrates information from multiple sources to create comprehensive patient profiles and has been shown to outperform existing methods in classification, clustering, and similarity searches. It efficiently handles missing or unobserved data and can compute embeddings for previously unseen patients. WellFactor is a versatile tool for deriving meaningful insights from diverse healthcare data sources, enabling personalized healthcare recommendations and improved patient care.

Acknowledgment

References

[1] S. K. Tanbeer and E. R. Sykes, "Myhealthportal–a web-based e-healthcare web portal for out-of-hospital patient care," *Digital Health*, vol. 7, p. 2055207621989194, 2021.

[2] S. S. Coughlin, J. J. Prochaska, L. B. Williams, G. M. Besenyi, V. Heboyan, D. S. Goggans, W. Yoo, and G. De Leo, "Patient web portals, disease management, and primary prevention," *Risk management and healthcare policy*, pp. 33–40, 2017.

[3] D. C. Chou and A. Y. Chou, "Healthcare information portal: a web technology for the healthcare community," *Technology in Society*, vol. 24, no. 3, pp. 317–330, 2002.

[4] L. E. Moody, "E-health web portals: delivering holistic healthcare and making home the point of care," *Holistic nursing practice*, vol. 19, no. 4, pp. 156–160, 2005.

[5] A. Pretschner and S. Gauch, "Ontology based personalized search," in *Proceedings 11th International Conference on Tools with Artificial Intelligence*. IEEE, 1999, pp. 391–398.

[6] X. Shen, B. Tan, and C. Zhai, "Implicit user modeling for personalized search," in *Proceedings of the 14th ACM international conference on Information and knowledge management*, 2005, pp. 824–831.

[7] K. Sugiyama, K. Hatano, and M. Yoshikawa, "Adaptive web search based on user profile constructed without any effort from users," in *Proceedings of the 13th international conference on World Wide Web*, 2004, pp. 675–684.

[8] B. Tan, X. Shen, and C. Zhai, "Mining long-term search history to improve search accuracy," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 718–723.

[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, Eds., 2013, pp. 3111–3119. [Online]. Available: https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

[10] K.-H. Hwang, S. M. Chan-Olmsted, S.-H. Nam, and B.-H. Chang, "Factors affecting mobile application usage: exploring the roles of gender, age, and application types from behaviour log data," *International Journal of Mobile Communications*, vol. 14, no. 3, pp. 256–272, 2016.

[11] S. Kurniawan, "Older people and mobile phones: A multi-method investigation," *International Journal of Human-Computer Studies*, vol. 66, no. 12, pp. 889–901, 2008.

[12] T. L. Mitzner, J. B. Boron, C. B. Fausset, A. E. Adams, N. Charness, S. J. Czaja, K. Dijkstra, A. D. Fisk, W. A. Rogers, and J. Sharit, "Older adults talk technology: Technology usage and attitudes," *Computers in human behavior*, vol. 26, no. 6, pp. 1710–1721, 2010.

[13] A. K. Hall, J. M. Bernhardt, V. Dodd, and M. W. Vollrath, "The digital health divide: evaluating online health information access and use among older adults," *Health Education & Behavior*, 2015.

[14] A. Fiske, J. L. Wetherell, and M. Gatz, "Depression in older adults," *Annual review of clinical psychology*, vol. 5, pp. 363–389, 2009.

[15] H. Chhabra, S. Sharma, and S. Verma, "Smartphone app in self-management of chronic low back pain: a randomized controlled trial," *European Spine Journal*, vol. 27, pp. 2862–2874, 2018.

[16] A. B. Irvine, H. Russell, M. Manocchia, D. E. Mino, T. C. Glassen, R. Morgan, J. M. Gau, A. J. Birney, and D. V. Ary, "Mobile-web app to self-manage low back pain: randomized controlled trial," *Journal of medical Internet research*, vol. 17, no. 1, p. e3130, 2015.

[17] G. Linden, B. Smith, and J. York, "Amazon. com recommendations: Item-to-item collaborative filtering," *IEEE Internet computing*, vol. 7, no. 1, pp. 76–80, 2003.

[18] C. Basu, H. Hirsh, and W. W. Cohen, "Recommendation as classification: Using social and content-based information in recommendation," in *Proceedings of the Fifteenth National Conference on Artificial Intelligence and Tenth Innovative Applications of Artificial Intelligence Conference, AAAI 98, IAAI 98, July 26-30, 1998, Madison, Wisconsin, USA*, J. Mostow and C. Rich, Eds. AAAI Press / The MIT Press, 1998, pp. 714–720. [Online]. Available: http://www.aaai.org/Library/AAAI/1998/aaai98-101.php

[19] R. S. Sutton, A. G. Barto *et al.*, *Introduction to reinforcement learning*. MIT press Cambridge, 1998, vol. 135.

[20] G. Liao, X. Shi, Z. Wang, X. Wu, C. Zhang, Y. Wang, X. Wang, and D. Wang, "Deep page-level interest network in reinforcement learning for ads allocation," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 2292–2296.

[21] X. Zhao, C. Gu, H. Zhang, X. Yang, X. Liu, J. Tang, and H. Liu, "Dear: Deep reinforcement learning for online advertising impression in recommender systems," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 1, 2021, pp. 750–758.

[22] L. Chen, N. Yang, and P. S. Yu, "Time lag aware sequential recommendation," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 212–221.

[23] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[24] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," 2019.

[25] Q. Chen, Y. Peng, and Z. Lu, "Biosentvec: creating sentence embeddings for biomedical texts," in *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 2019, pp. 1–5.

[26] L. Cui and D. Lee, "Ketch: Knowledge graph enhanced thread recommendation in healthcare forums," in *Proceedings of the 45th international acm sigir conference on research and development in information retrieval*, 2022, pp. 492–501.

[27] Y. Tan, C. Yang, X. Wei, C. Chen, W. Liu, L. Li, J. Zhou, and X. Zheng, "Metacare++: Meta-learning with hierarchical subtyping for cold-start diagnosis prediction in healthcare data," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 449–459.

[28] R. Du, B. Drake, and H. Park, "Hybrid clustering based on content and connection structure using joint nonnegative matrix factorization," *Journal of Global Optimization*, vol. 74, pp. 861–877, 2019.

[29] H. Kim, D. Choi, B. Drake, A. Endert, and H. Park, "Topicsifter: Interactive search space reduction through targeted topic modeling," in *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 2019, pp. 35–45.

[30] G. N. Demir, A. Ş. Uyar, and Ş. Gündüz-Öğüdücü, "Multiobjective evolutionary clustering of web user sessions: a case study in web page recommendation," *Soft Computing*, vol. 14, pp. 579–597, 2010.

[31] F.-H. Wang and H.-M. Shao, "Effective personalized recommendation based on time-framed navigation clustering and association mining," *Expert systems with applications*, vol. 27, no. 3, pp. 365–377, 2004.

[32] M. Li, L. Wen, and F. Chen, "A novel collaborative filtering recommendation approach based on soft co-clustering," *Physica A: Statistical Mechanics and its Applications*, vol. 561, p. 125140, 2021.

[33] G. Guo, J. Zhang, and N. Yorke-Smith, "Leveraging multiviews of trust and similarity to enhance clustering-based recommender systems," *Knowledge-Based Systems*, vol. 74, pp. 14–27, 2015.

[34] A. B. Melchiorre, N. Rekabsaz, C. Ganhör, and M. Schedl, "Protomf: Prototype-based matrix factorization for effective and explainable recommendations," in *Proceedings of the 16th ACM Conference on Recommender Systems*, 2022, pp. 246–256.

[35] J. J. Whang, R. Du, S. Jung, G. Lee, B. Drake, Q. Liu, S. Kang, and H. Park, "Mega: Multi-view semi-supervised clustering of hypergraphs," *Proceedings of the VLDB Endowment*, vol. 13, no. 5, pp. 698–711, 2020.

[36] W. H. Organization *et al.*, *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*. World Health Organization, 1992.

[37] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark, "Mimic-iv," *PhysioNet. Available online at: https://physionet.org/content/mimiciv/1.0/(accessed August 23, 2021)*, 2020.

[38] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[39] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, pp. 345–379, 2010.

[40] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern recognition*, vol. 38, no. 12, pp. 2270–2285, 2005.

[41] J. Kim, Y. He, and H. Park, "Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework," *J. Glob. Optim.*, vol. 58, no. 2, pp. 285–319, 2014. [Online]. Available: https://doi.org/10.1007/s10898-013-0035-4

[42] D. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999.

[43] D. Kuang, H. Park, and C. H. Q. Ding, "Symmetric nonnegative matrix factorization for graph clustering," in *Proceedings of the Twelfth SIAM International Conference on Data Mining, Anaheim, California, USA, April 26-28, 2012*. SIAM / Omnipress, 2012, pp. 106–117. [Online]. Available: https://doi.org/10.1137/1.9781611972825.10

[44] D. Kuang, S. Yun, and H. Park, "Symnmf: nonnegative low-rank approximation of a similarity matrix for graph clustering," *J. Glob. Optim.*, vol. 62, no. 3, pp. 545–574, 2015. [Online]. Available: https://doi.org/10.1007/s10898-014-0247-2

[45] R. Reiter, "On closed world data bases," in *Logic and Data Bases, Symposium on Logic and Data Bases, Centre d'études et de recherches de Toulouse, France, 1977*, ser. Advances in Data Base Theory, H. Gallaire and J. Minker, Eds. New York: Plemum Press, 1977, pp. 55–76. [Online]. Available: https://doi.org/10.1007/978-1-4684-3384-5_3

[46] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich, "A review of relational machine learning for knowledge graphs," *Proc. IEEE*, vol. 104, no. 1, pp. 11–33, 2016.

[47] Q. Tan, N. Liu, X. Zhao, H. Yang, J. Zhou, and X. Hu, "Learning to hash with graph neural networks for recommender systems," in *Proceedings of The Web Conference 2020*, 2020, pp. 1988–1998.

[48] B. Drake, J. Kim, M. Mallick, and H. Park, "Supervised raman spectra estimation based on nonnegative rank deficient least squares," in *Proceedings of the 13th International Conference on Information Fusion*, Edinburgh, Scotland, UK, July 6-9, 2010, pp. 698–711.

[49] W. H. Organization *et al.*, *Global status report on noncommunicable diseases 2014*. World Health Organization, 2014, no. WHO/NMH/NVI/15.1.

[50] D. Singh, A. Agusti, A. Anzueto, P. J. Barnes, J. Bourbeau, B. R. Celli, G. J. Criner, P. Frith, D. M. Halpin, M. Han *et al.*, "Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease: the gold science committee report 2019," *European Respiratory Journal*, vol. 53, no. 5, 2019.

[51] A. Levin, M. Tonelli, J. Bonventre, J. Coresh, J.-A. Donner, A. B. Fogo, C. S. Fox, R. T. Gansevoort, H. J. Heerspink, M. Jardine *et al.*, "Global kidney health 2017 and beyond: a roadmap for closing gaps in care, research, and policy," *The Lancet*, 2017.

[52] W. H. Organization *et al.*, "Depression and other common mental disorders: global health estimates," World Health Organization, Tech. Rep., 2017.

[53] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[54] G. . O. Collaborators, "Health effects of overweight and obesity in 195 countries over 25 years," *New England journal of medicine*, vol. 377, no. 1, pp. 13–27, 2017.

[55] M. Cross, E. Smith, D. Hoy, S. Nolte, I. Ackerman, M. Fransen, L. Bridgett, S. Williams, F. Guillemin, C. L. Hill *et al.*, "The global burden of hip and knee osteoarthritis: estimates from the global burden of disease 2010 study," *Annals of the rheumatic diseases*, vol. 73, no. 7, pp. 1323–1330, 2014.