Disease Inference from Health-Related Questions via Sparse Deep Learning

Liqiang Nie, Meng Wang, Luming Zhang, Shuicheng Yan, *Member, IEEE*,
Bo Zhang, *Senior Member, IEEE*, Tat-Seng Chua, *Senior Member, IEEE*

Abstract—Automatic disease inference is of importance to bridge the gap between what online health seekers with unusual symptoms need and what busy human doctors with biased expertise can offer. However, accurately and efficiently inferring diseases is non-trivial, especially for community-based health services due to the vocabulary gap, incomplete information, correlated medical concepts, and limited high quality training samples. In this paper, we first report a user study on the information needs of health seekers in terms of questions and then select those that ask for possible diseases of their manifested symptoms for further analytic. We next propose a novel deep learning scheme to infer the possible diseases given the questions of health seekers. The proposed scheme comprises of two key components. The first globally mines the discriminant medical signatures from raw features. The second deems the raw features and their signatures as input nodes in one layer and hidden nodes in the subsequent layer, respectively. Meanwhile, it learns the inter-relations between these two layers via pre-training with pseudo-labeled data. Following that, the hidden nodes serve as raw features for the more abstract signature mining. With incremental and alternative repeating of these two components, our scheme builds a sparsely connected deep architecture with three hidden layers. Overall, it well fits specific tasks with fine-tuning. Extensive experiments on a real-world dataset labeled by online doctors show the significant performance gains of our scheme.

Index Terms—Community-based Health Services, Question Answering, Disease Inference, Deep Learning

1 Introduction

The greying of society, escalating costs of healthcare and burgeoning computer technologies are together driving more consumers to spend longer time online to explore health information. One survey in [1] shows that 59% of U.S. adults have explored the internet as a diagnostic tool in *2012*. Another survey in [2] reports that the average U.S. consumer spends close to 52 hours annually online to find wellness knowledge, while only visits the doctors three times per year in *2013*. These findings have heightened the importance of online health resources as springboards to facilitate patient-doctor communication.

The current prevailing online health resources can be roughly categorized into two categories. One is the reputable portals run by official sectors, renowned organizations, or other professional health providers. They are disseminating up-to-date health information by releasing the most accurate, well-structured, and formally presented health knowledge on various topics. WebMD[1] and MedlinePlus[2] are the typical examples. The other category is the community-based health services, such as HealthTap[3] and HaoDF[4]. They offer interactive platforms, where health seekers can anonymously ask health-oriented questions while doctors provide the knowledgeable and trustworthy answers. Figure 1 illustrates one question answer (QA) example. However, the community-based health services have several intrinsic limitations. First of all, it is very time consuming for health seekers to get their posted questions resolved. The time could vary from hours to days [3]. Second, doctors are having to cope with an ever-expanding workload, which leads to decreased enthusiasm and efficiency. Taking HealthTap as an example, as of January *2014*, it had gathered 50 thousand doctors and accumulated more than 1.1 billion answers, i.e., on average each doctor has online replied approximately 23 thousand times since its foundation in *2010*. Third, qualitative replies are conditioned on doctors' expertise, experiences and time, which may result in diagnosis conflicts among multiple doctors and low disease coverage of individual doctor [4]. It is thus highly desirable to develop automatic and comprehensive wellness systems that can instantly answer all-round questions of health seekers and alleviate the doctors' workload.

The biggest stumbling block of automatic health

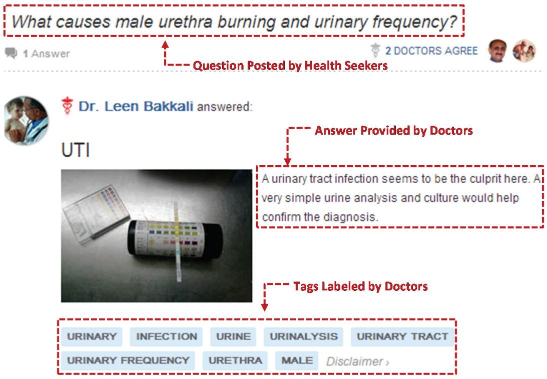
Liqiang Nie, Luming Zhang and Tat-Seng Chua are with the School of Computing, National University of Singapore, Singapore. (email: nieliqiang@gmail.com; zglumg@gmail.com; chuats@comp.nus.edu.sg).

Shuicheng Yan is now with the Department of Electrical and computer Engineering at National University of Singapore, Singapore. (email: eleyans@nus.edu.sg).

Meng Wang is with the Hefei University of Technology, China.(email:eric.mengwang@gmail.com).

Bo ZHANG is with Department of Computer Science and Technology, Tsinghua University, Beijing, China. (email: dcszb@tsinghua.edu.cn).

[1] http://www.webmd.com/
[2] http://www.nlm.nih.gov/medlineplus/
[3] https://www.healthtap.com
[4] http://www.haodf.com



What causes male urethra burning and urinary frequency?

Dr. Leen Bakkali answered:

UTI

A urinary tract infection seems to be the culprit here. A very simple urine analysis and culture would help confirm the diagnosis.

Fig. 1. An illustrative example of a QA pair from community-based health services (HealthTap). One question may receive multiple answers. Here we only display one due to the space reason.

system is disease inference. According to our user study on $5,000$ questions that will be detailed in Section 4, health seekers frequently ask for: (1) supplemental cues of their diagnosed diseases; (2) preventive information of their concerned diseases; and (3) possible diseases of their manifested signals. Table 1 displays three corresponding question examples. The former two genres usually involve the exact disease names and expected sub-topics or sub-problems of the given diseases, such as the side effects of specific medications, and treatments. They can be automatically and precisely answered by either directly matching the questions in the archived repositories or syntactic information extraction from the structured health portals. The existing automatic question answering techniques are applicable here [5], [6]. The third genre conveys parts of the health seekers' demographic information, physical and mental symptoms, as well as medical histories, in which they do not know what conditions they might have and expect the doctors to offer them some forts of online diagnosis. If the diseases are correctly inferred, these questions are naturally transferred to the first genre. Hence a robust disease inference approach is the key to break the barrier of automatic wellness systems.

However, little research has been dedicated to disease inference in the community-based health services. Disease inference is different from topics or tags assignment to short questions [7], where topics or tags are direct summarizations of given data instances and they may explicitly appear in the questions. While disease inference is a reasoning consequence based on the given question, this task is non-trivial due to following reasons. First, vocabulary gap between diverse health seekers makes the data more inconsistent, as compared to other formats of health data. For example, "shortness of breath" and "breathless" were used by different health seekers to refer to the same semantic "dyspnea". Second, health seekers describe their problems in short questions, containing 14.5 terms per question on average. The incompleteness hinders the effective similarity estimation based on shared contexts. Third, medical attributes such as age, gender and symptoms, are highly correlated and do not unusually appear as compact patterns to signal the health problems. For example, "tight chest", "wheezing", and "dyspnea" frequently co-occur to hint of "asthma". In addition, it is expensive to construct the ground truth for various diseases. These factors limit the disease inference performance that can be obtained by general shallow learning methods. Shallow learning methods refer to the most modern learning algorithms such as decision trees, and support vector machine (SVM), where the output of a learning scheme is directly followed by a classifier as if the system has only one layer.

This paper aims to build a disease inference scheme that is able to automatically infer the possible diseases of the given questions in community-based health services. We first analyze and categorize the information needs of health seekers. As a byproduct, we differentiate questions of this kind that require disease inference from other kinds. It is worth emphasizing that large-scale data often leads to explosion of feature space in the lights of n-gram representations [8], [9], especially for the community generated inconsistent data. To avoid this problem, we utilize the medical terminologies to represent our data. Our scheme builds a novel deep learning model, comprising two components, as demonstrated in Figure 2. The first globally mines the latent medical signatures. They are compact patterns of inter-dependent medical terminologies or raw features, which can infer the incomplete information. The raw features and signatures respectively serve as input nodes in one layer and hidden nodes in the subsequent layer. The second learns the inter-relations between these two layers via pre-training. Following that, the hidden nodes are viewed as raw features for more abstract signature mining. With incremental and alternative repeating of these two components, our scheme builds a sparsely connected deep learning architecture with three hidden layers. This model is generalizable and scalable. Fine-tuning with a small set of labeled disease samples fits our model to specific disease inference. Different from conventional deep learning algorithms, the number of hidden nodes in each layer of our model is automatically determined and the connections between two adjacent layers are sparse, which make it faster. Extensive experiments on real-world dataset labeled by online doctors were conducted to validate our scheme.

The main contributions of this work are threefold:

1) To the best of our knowledge, this is the first work on automatic disease inference in the community-based health services. Distinguished from the conventional sporadic efforts that

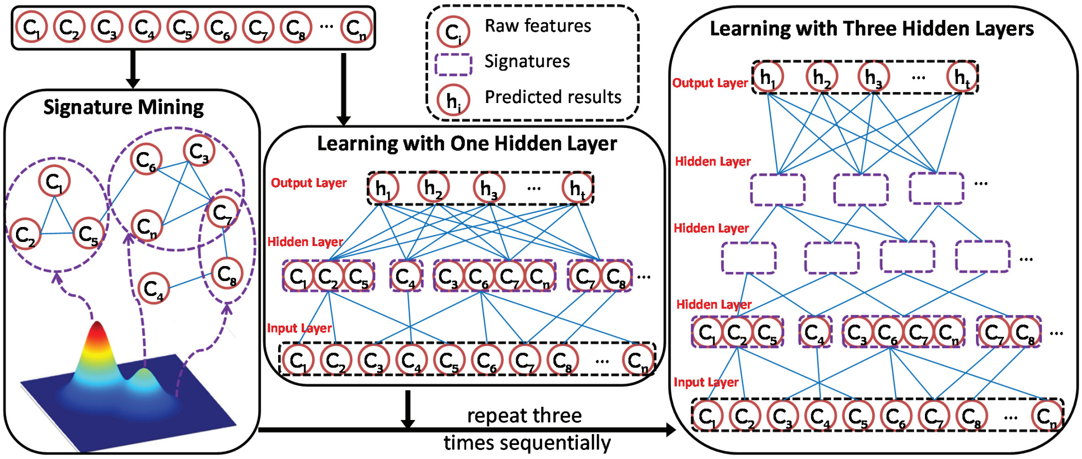

C_1 C_2 C_3 C_4 C_5 C_6 C_7 C_8 ··· C_n
C_i Raw features
Signatures
h_i Predicted results
Signature Mining
Learning with One Hidden Layer
Output Layer
h_1 h_2 h_3 ··· h_t
Hidden Layer
Input Layer
Learning with Three Hidden Layers
repeat three times sequentially

Fig. 2. The illustrative process of our sparsely connected deep learning construction. We incrementally added hidden layers until it satisfies the predefined convergence criterion. The convergent criterion is defined as the accuracy of deep learning model with $(n+1)$ hidden layers will not significantly outperform that with n hidden layers based on 5 fold based t-test.

generally focus on only a single or a few diseases based on the hospital generated records with structured fields, our scheme benefits from the volume of unstructured community generated data and it is capable of handling various kinds of diseases effectively.

2) It investigates and categorizes the information needs of health seekers in the community-based health services and mines the signatures of their generated data.
3) It proposes a sparsely connected deep learning scheme to infer various kinds of diseases. This scheme is pre-trained with pseudo-labeled data and further strengthened by fine-tuning with online doctor labeled data.

The remainders are structured as follows. Section 2 briefly reviews the related work. The information needs of health seekers are then analyzed in Section 3. The two components of our scheme are respectively introduced in Section 4 and 5. Section 6 details the experimental results and analysis, followed by our concluding remarks in Section 7.

2 Related Work

Research on healthcare is literally the most vital part of science for humans, as none of us are immune to physical ailments. The existing literatures are diverse and roughly follow four lines of research: information extraction [10], [11], [12], [13], [14], disease inference [15], [16], [17], [18], preventive medicine [4], [19], [20], [21], as well as medical search [22], [23], [24], [25], [26], [27], [28], [29].

Information extraction from medical text is the basis for other higher-order analytics, such as representation, classification, and clustering. The work in [12] utilized SVM to recognize the medication related entities in hospital discharge summaries, and classified these atomic elements into pre-defined categories, such as treatments and conditions. Beyond extraction, Sondhi et al. [30] constructed entity graphs by exploring their co-occurrence relations and studied how to leverage such graphs to convert raw entity mentions into more useful knowledge, which is helpful for feature expansion. These efforts only consider the explicitly present medical entities, while they overlook the temporal aspect of data as well as the latent discriminative patterns across patient records [14]. To deal with these two problems, Wang et al. [10], [11] proposed a nonnegative matrix factorization based framework to mine common and individual shift-invariant temporal patterns from heterogeneous events over different patient groups, which is able to handle sparseness and scalability problems. As a complementary work, a simple yet effective tool for visualizing the temporal associations among multiple records was designed in [13].

Researchers have been increasingly attracted to use machine learning techniques to assist health professionals in the diagnosis of diseases. Shouman et al. [17] and Ghumbre et al. [31] have respectively explored decision tree and SVM in the inference of heart disease, which is the leading cause of death in the world over the past 10 years according to the report from world health organization. A learning framework was presented in [16] that focused on Alzheimer disease inference from magnetic resonance images by integrating visual similarities and user feedback. Instead of building single disease related model, Zhang et al. [18] trained an infectious disease model with the sentence-level semantic features, and obtained promising performance. Fakoor et al. [15] in *2013* figured out the scalability and generality problems of these inference models, and utilized unsupervised feature learning method to enhance cancer types classification.

TABLE 1
Categorization of health seeker needs. Answer parts are not listed due to the limited space.

Categories	Specific Motivations	Question Examples
Disease diagnosed, ask for supplemental information	Alternative medical treatment, Side effect of specific medications, Resolution of conflict diagnosis, Second opinion, ...	I have been diagnosed with polyarteritis nodosa and am currently taking 32mg of steroids and 125mg of cyclophosphamide. Is this the correct medicine ? ...
Currently healthy, ask for prevention	Consequences of current habit, Lifestyle, Diet, Nutrition, Precaution, Exercises, ...	I'm a healthy and active 38 year-old Asian male whose non-smoking mother died from lung cancer 4 years ago. Am I at greater risk ? How can I pre-empt it ? ...
Disease undiagnosed, ask for possible diseases of their manifested signals	Feeling ill with symptoms, Other potential complication of diagnosed diseases, ...	What disease or illness should I look into if I feel tired, sleepy all the times, muscular and joint sored, decrease in memory and concentration ? ...

Most of the current healthcare is reactive, triggered by the emerging symptoms of diseases. The irreversible consequences of reaction such as death have motivated the drive towards preventive medicine, where the primary concern is recognizing disease risk and taking action at the earliest signs. Khosla et al. [20] presented an integrated machine learning scheme to predict stroke for early intervention and treatment, which achieved improved performance on the cardiovascular health study dataset. More practically, a novel system named CARE was designed in [4], which combined collaborative filtering methods with clustering to predict each patient's greatest disease risks based on their own medical history and those of similar patients. Understanding how the disease progresses is of essential importance for proactive healthcare. Zhou et al. recently modeled the progression of Alzheimer disease with multi-task learning [19] and fused sparse group lasso [21], respectively.

Medical retrieval is the dominant way for knowledge exchanging and sharing. Huang et al. [27] proposed a re-ranking model for promoting diversity in medical search. Query-adaptive weighting methods that can dynamically aggregate and score various search results have been presented in [22], [23]. These three approaches were all evaluated on the TREC dataset [32]. A more systematic evaluating framework for medical record search was developed in [26] using BLULab de-identified medical corpus. In addition, Yang et al. [33] studied the consumer healthcare retrieval problem from novel points. They examined how the same topic is expressed differently in lay and expert languages, and how topics are shifted according to domain technicality.

The aforementioned approaches were designed for and validated on hospital or lab generated patient records, which are well-organized with structured fields. These approaches are not applicable to online health data due to two main reasons. From the perspective of data property, they have different data structure, quality and number of training samples. From the point of techniques, most of the previous efforts are unable to take advantages of other data types beyond the targeted ones, and hence are not scalable or generalizable. However, the research on online health data is relatively rare. Luo et al. [28] in *2008* built a medical Web search engine called iMed, which employed medical knowledge and an interactive questionnaire to help searchers form queries. After that, health seeker needs were explored from multiple aspects, including intentions and attention [25], onset and persistence of medical concerns [24], as well as the comprehensive wellness search [29]. The existing work was monotonously conducted on retrieval.

3 Data Collection

We collected more than 900 popular disease concepts from EveryoneHealthy[5], WebMD and MedlinePlus. They cover a wide range of diseases, including endocrine, urinary, neurological and other aspects. With these disease concepts as queries, we crawled more than 220 thousand community generated QA pairs from HealthTap. Each QA pair consists of three parts as shown in Figure 1: question from health seekers, answers and multiple tags associated with the answers provided by the doctors. According to our statistics, each question in our dataset has 3.16 answers and each answer is tagged with 7.12 tags, on average. By the way, the disease concepts are naturally viewed as pseudo-labeled categories for their searched QA pairs. For the QA pairs which were retrieved by multiple queries, they were assigned to the most relevant one.

To comprehensively validate our scheme, two groups of diseases and their corresponding QA pairs were selected from our whole dataset. As listed in Table 2, the diseases in the first group were randomly sampled, while those in the second group were manually and deliberately selected to ensure they are semantically and syntactically similar, e.g., they may share partial overlapping symptoms. The purpose of the second disease group is to make the disease inference harder and hence to validate the robustness of our scheme. For each disease in these two groups, we only keep the QA pairs satisfying the following two conditions via semi-auto rule-based approach:

[5] http://www.everyonehealthy.com/

TABLE 2
Lists of two selected disease groups for fine-tuning and evaluation.

ID	First Disease Group Dataset II	Second Disease Group Dataset III
I	Parkinson	Asthma
II	Polycystic Kidney	Lung Cancer
III	Lyme Disease	Breast Cancer
IV	Liver Cancer	Fibrocystic Breast
V	Dental Cavities	Diabetes Insipidus
VI	Bladder Cancer	Diabetic Nephropathy
VII	Hearing Loss	Diabetes Type 1
VIII	Crohn's Disease	Diabetes Type 2
IX	Raynaud's Syndrome	Congestive Heart Failure
X	Influenza	Heart Attack

TABLE 3
Statistical information of our datasets.

Dataset ID	Utility	Disease #	QA Pairs #
Dataset I	Pre-training	945	225,245
Dataset II	Fine-tuning & Evaluation	10	1,674
Dataset III	Fine-tuning & Evaluation	10	1,587

1) The question part does not explicitly mention the disease name, or mentions the disease name but contextualized in interrogative scenarios such as this question "*I get breast pain the 2 weeks before my period, worse pain in the right one, plus it itches sometimes. Does that mean breast cancer?*".
2) The answer or tag part must explicitly mention the disease name. Figure 1 illustrates a good example.

This kind of QA pairs is used as ground truth to perform the tasks of fine-tuning and evaluation. Table 3 shows the detailed statistics of our datasets. Notably, the ground truth construction procedure ensures that the fine-tuning data fall into the third category in Table 1, i.e., disease undiagnosed, asking for possible diseases of their manifested signals. On the other hand, the pre-training dataset is a mixture of all the three categories illustrated in Table 1. This is due to many reasons. First, it is difficult to collect large-scale training samples belonging to the third category. According to our statistics, the third category is not the majority, and hence it is not easy, if it is not impossible to sufficiently select such training samples. Second, one objective of pre-training is to learn the general model and comprehensive patterns for wellness domain, while fine-tuning with a small set of labeled disease samples fits this model to specific disease inference. If we only utilize the third category QA pairs, we may somehow miss some key information in terms of patterns.

4 Health Seeker Needs Analytic

To make more informed decisions towards better health, health seekers are getting increasingly savvy with their information needs [34]. Specifically, each health seeker has very specific needs and knows what they expect when they look into the Internet. This leads to diverse, sophisticated and complex motivations and needs of online health seeking.

To gain insights into health seeker needs, we randomly collected $5,000$ QA pairs from HealthTap, which cover a wide range of topics, including cancer, endocrine and pregnancy. We carefully went over all these QA pairs and observed that the health seeker needs can be abstracted into three main categories, as shown in Table 1. Specific motivations and question examples are also provided to enhance the understanding of this categorization. From this table, it can be seen that the three categories do not mutually overlap and cover all the possible cases. This is because the health seeker with respect to a concerned health problem can only be in one state out of the three at one time: healthy status, suffering from diagnosed disease or undiagnosed disease.

It is worth mentioning that there exist some questions where the health seekers ask about one undiagnosed disease but who already have been diagnosed with another disease. Take this question as an example[6], "*I had breast cancer a year ago and had a right mastectomy. Now they have found pleural nodules in right and left lower lobes 5mm, 4, 8mm, 2, 4mm, 4.5mm should I be worried about lung cancer?*", which is a real question selected from HealthTap. On the face of it, such questions do not belong to any one of the three categories. However, in our work, our categorization targets at the health seeker needs rather than health seekers themselves, and other information conveyed in the questions is regarded as contexts. For instance, in the above question, the health seeker is asking for possible "lung cancer". Even though "breast cancer" is known and diagnosed, it is viewed as the medical history. Hence this question falls into the third category without ambiguity.

We also conducted a user study to investigate the health seeker needs. Three volunteers were invited to manually classify each of the $5,000$ QA pairs into one of the three pre-defined categories. It is worth noting that each volunteer was pre-trained with the definitions of category types as well as corresponding examples. We performed a voting method to establish the final classification of each QA pair. For cases where each class equally receiving one vote, a discussion was carried out among the volunteers to obtain the final decision. According to our statistics, the distributions of QA pairs over the three categories are 79%, 6% and 15%, respectively. Even though the third category is not the majority, it greatly increases the bottlenecks of the automatic health system as we have analyzed before.

However, automatically categorizing this community generated health data is somewhat difficult because of the negated language and

[6] https://www.healthtap.com/user_questions/280838

vocabulary gap.

Regarding the negated language, negated identifiers are frequently used by medical practitioners to indicate that patients do not have given conditions. Some traditional approaches do not distinguish between the positive and negative contexts of medical concepts in medical records, which may prevent the learning/retrieval performance from being effective. Take the following two short medical records as an example. Intuitively, their contexts are totally different, while a learning or search system may inaccurately consider such medical records to be equivalent.

1) *"A heart disease patient with no medical history of lung cancer"*;
2) *"A heart disease patient with evidence of lung cancer previously"*.

Inspired by the work in [35], we filtered the medical terms/concepts having a negative context during the feature extraction process in sentence level. The following example explicitly explains our approach:

1) *"A heart disease patient with no medical history of lung cancer"* ↦ *"a heart disease patient"*.

Such negation filtering approach is able to improve the recall of finding medical records containing non-filtered medical concepts with the correct contexts. However, it suffers from two limitations: (1) it alters the original contexts, and hence it may leads to serious information loss; and (2) it does not necessarily prevent matching with those containing filtered medical concepts with opposite contexts.

In the health communities, users with diverse backgrounds do not necessarily share the same vocabulary [36]. Sometimes, the same medical subjects may be colloquially expressed with various medical concepts. For example, "birth control" and "family planning" are commonly used by individuals to refer to the same medical terminology "contraception". The traditional context representations such as n-gram are unable to capture the variants of medical concepts and may lead to an explosion of feature dimension. To tackle these problems, we employed the MetaMap tool [37] to detect medical attributes that are noun phrases in health domain, and then normalize them to standardized terminologies in the SNOMED CT Metathesaurus[7]. In our previous work, we have detailed this procedure [38]. The semantic types of these terminologies span from symptom, treatment, medication, body parts, to demographics. In this paper, we utilize these normalized medical attributes to represent the community generated health data. We represent the QA pairs with these terminologies and study the health seeker needs via QA pair classification. We conduct experiments to validate this study in Section 7.2.

5 Signature Mining

The main challenging problem in health domain is the inter-dependent medical attributes, which is named as signature in this paper. As compared to individual raw feature, signatures are essential cues for diseases. For example, "urinary frequency", "excessive thirst" and "blurry vision" together likely signal hints to diabetes. While lonely "blurry vision" may be the result of abnormalities present at birth such as near or far sightedness. It can also be a symptom of numerous conditions that do not directly involve the eyes, such as migraine, stroke and side effects of medications. Therefore, the medical signatures are more descriptive than raw features and will significantly reduce the dimension of feature space. However, it is difficult to extract such signatures from individual data instances, as their structures are usually implicitly distributed over a large-scale dataset.

In our work, the latent signatures are viewed as overlapping dense subgraphs $\mathcal{G}_{\mathcal{S}} = \{G_1, G_2, ..., G_u\}$ embedded in a global graph G. The graph construction is straightforward, where vertices and edges are the normalized medical attributes and their co-occurrence relations, respectively. Let $\mathcal{Q} = \{q_1, q_2, ..., q_m\}$ denotes our QA pairs associated with k disease types $\mathbf{y} = \{y_1, y_2, ..., y_k\}$. And let $\mathcal{C} = \{c_1, c_2, ..., c_n\}$ stands for the ordered set of normalized medical attributes extracted from $\mathcal{Q}$. Thus each QA pair q_i can be represented as a tuple of n-dimension feature and label $(\mathbf{x}^i, \mathbf{y}^i)$, where only one entry in $\mathbf{y}^i$ equals to one, and the remainders are zeros. The medical attribute to QA pair matrix, $\mathbf{H} \in \mathbf{R}^{n \times m}$, can be defined as,

$$H_{ij} = \begin{cases} 1 & \text{if attribute } i \text{ occur in QA pair } j; \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Then the adjacency matrix $\mathbf{A} \in \mathbf{R}^{n \times n}$ of normalized medical attributes over the whole dataset can be computed,

$$\mathbf{A} = \mathbf{H}\mathbf{H}^T. \quad (2)$$

Here self-loops are meaningless, so all the diagonal elements of $\mathbf{A}$ are set to 0. To avoid the common medical attributes overwhelming the non-common ones, matrix normalization is necessary. Some efforts utilized the asymmetric normalization approach, as they consider that the relations among medical symptoms do not always obey symmetric property [30], [39]. In our work, to facilitate the dense subgraph mining, we pick the widely used symmetric normalizing approach,

$$\widetilde{\mathbf{A}} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}, \quad (3)$$

where $\mathbf{D}$ is a diagonal matrix with its (i,i)-element equal to the sum of the i-th row of $\mathbf{A}$.

Let $\triangle = \{\mathbf{s} \in \mathbf{R}^n : \forall i \quad s_i > 0 \; and \; ||\mathbf{s}||_1 = 1\}$ be a simplex, where $||\mathbf{s}||_1 = \sum_{i=1}^{n} |s_i|$ is l_1 norm of

[7] http://www.ihtsdo.org/snomed-ct/

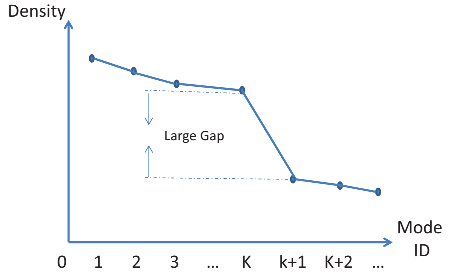


Fig. 3. Automatic strategy for the K selection.

$\mathbf{s} = (s_1, s_2, ..., s_n)^T$. The indices of non-zero entries of $\mathbf{s}^k \in \triangle$ constitute its support, denoted as $\delta_k = \{j | s_j^k \neq 0\}$, where s_j^k describes the probability that medical attribute a_j falls into its corresponding subgraph G_k. We define the density of G_k as,

$$g(\mathbf{s}^k) = \sum_{i=1}^{n} \sum_{j=1}^{n} s_i^k \widetilde{a}_{ij} s_j^k = (\mathbf{s}^k)^T \widetilde{\mathbf{A}} (\mathbf{s}^k), \tag{4}$$

where $g(\mathbf{s}^k)$ reflects the strength of overall internal pairwise relations in the subgraph G_k. Then the dense subgraph mining problem is converted to solving the following standard quadratic optimization problem,

$$\arg\max_{\mathbf{s}} g(\mathbf{s}) = \mathbf{s}^T \widetilde{\mathbf{A}} \mathbf{s}, \tag{5}$$

where the constraint condition is $\mathbf{s} \in \triangle$. Any local maximizer $\mathbf{s}^*$ of g($\mathbf{s}$) indicates a potential dense subgraph. We follow the solution in [40], [41] to optimize this problem. They first define the concept "mode" as the local maximizer of Eqn (5). Each mode has a density value to indicate how coherent it is. They then align each mode to a subgraph. The vertices involved in a mode with large density form a dense subgraph. We proposed a method to determine the number of modes as illustrated in Figure 3. In particular, we do a tri-stage procedure,

1) Sort all the detected modes in a decreasing order based on their densities.
2) Calculate the difference between two adjacent values in the sorted order.
3) Find the largest drop which is a boundary of the leading modes and other noise modes.

The results have some key properties that well meet the requirements of our task. The first is overlapping that fit the scenarios in healthcare domain, where some medical attributes may belong to multiple signatures. For example, "female" is an important part of the signatures related to pregnancy and breast cancer. The second is that some nodes may not be involved in dense subgraphs. This keeps the possible noise and outliers out of the signatures. The third one is that the each local maxima of the function corresponds to one dense subgraph signature. The number of local maximizers indicates the number of nodes in the hidden layers in the deep learning architecture.

Another benefit of dense subgraph mining is that it naturally infers the incomplete information. With very few exceptions, research on community generated data that concerns the organization, retrieval and recommendation will inevitably contain an element of incomplete information [42]. This seems more prominent for analytic of healthcare domain, because health seekers tend to describe their health problems in short questions. For example, the non-obvious symptoms or sense of privacy may cause health seekers to pose incomplete questions. Also, some key medical attributes such as gender and age, which probably matter in doctors' decisions, may be arbitrarily judged as unimportant cues and hence are not included in the questions. A range of methods have been proposed to minimize the damaging effects of incomplete information in various domains, including leveraging internal context [7], [43] and external knowledge [44]. For the community-based question answering services, researchers typically utilize the corresponding answers to compensate the information limitations [5], [7]. However, it is not reasonable in our work to incorporate the answer part for disease inference. This is because the answer parts are not available for new incoming questions. On the other hand, dense subgraph is a natural extension for many given questions with incomplete attributes. Take this question "*what causes urinary frequency in young woman?*" as an example. A small set of normalized medical attributes "urinary frequency" and "young woman" can be extracted from the question itself independently. This set can be expanded with "urinary urgency", "pelvic pain" and "blood in urine". This is because they are globally detected as a compact pattern with "urinary frequency".

6 Disease Inference

As aforementioned, vocabulary gap, incomplete information, inter-dependent medical attributes and limited ground truth have greatly hindered the performance of classic shallow machine learning approaches. To tackle these problems, we propose a novel deep learning scheme to infer the possible diseases given the questions of health seekers. Compared to shallow learning, deep learning has several advantages. First, it is able to learn representative and scalable features from other disease types [45]. Take the lung cancer inference learning as an example. When building its classifier, the training data can be liver cancer or other disease samples rather than strictly constrained to lung cancer. This addresses the limited ground truth and necessity of disease-aware feature extraction [15]. Second, inherited from its deep architectures, it repeatedly learns the more abstract compact patterns layer by layer. This enables the system to mine the underlying connections among medical attributes. Third, deep learning can seamlessly integrate signatures as hidden nodes. As analyzed previously, signatures infer the incomplete information. Most importantly, with deep learning, each data instance will be ultimately represented by a mixture of very high-level abstract patterns, which are semantic descriptors and thus are more robust of data inconsistency caused by vocabulary gap.

6.1 Sparsely Connected Deep Learning

Our sparsely connected deep learning model has L layers with $d_l (1 \leq l \leq L)$ nodes in each layer. To be more specific, the first layer contains the input n-dimension raw features and the L-th layers denotes the output disease types, while the intermediate layers are hidden layers, which are unseen from the data. Unlike general deep learning architectures, in our work, nodes in the higher layer are the signatures of and connect to the nodes in its adjacent lower layer, rather than fully connected. These relations are explicitly indicated by the affinity matrix W_{ij}^l, which stands for the relation strength between node i in layer l and node j in its subsequent higher layer. If node i belongs to the signature node j, W_{ij}^l is initialized to s_i^j that is the probability of node i belonging to signature j; otherwise W_{ij}^l is permanently set to 0. This is the so-called sparsely connected deep learning, where only the last hidden layer and the output layer are fully connected. Meanwhile, b_i^l represents the bias term with node i in layer $l+1$. z_i^l and o_i^l denote the sum of weighted inputs and activated output value of node i in layer l, respectively. A notable case is the first layer with $\mathbf{o}^1 = \mathbf{x}$. We can thus formulate the matrix-vectorial notation of input and output in each layer as,

$$\begin{cases} \mathbf{z}^{l+1} = \mathbf{W}^l \mathbf{o}^l + \mathbf{b}^l, \\ \mathbf{o}^{l+1} = \mathbf{f}(\mathbf{z}^{l+1}), \\ \mathbf{h}_{\mathbf{W},\mathbf{b}}(\mathbf{x}) = \mathbf{o}^L = \mathbf{f}(\mathbf{z}^L), \end{cases} \tag{6}$$

where $\mathbf{f}(\cdot)$ and $\mathbf{h}_{\mathbf{W},\mathbf{b}}(\mathbf{x})$ are respectively the activation function and objective function, both in an element-wise fashion. The sigmoid function with output ranging between 0 and 1 is chosen as the activation function. It is formulated as,

$$f(z) = \frac{1}{1 + exp(-z)}. \tag{7}$$

Inspired by [46], the overall cost function of our proposed sparsely connected deep learning, denoted as $J(\mathbf{W}, \mathbf{b})$, can be defined as,

$$\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\left|\left|\mathbf{h}_{\mathbf{W},\mathbf{b}}(\mathbf{x}^i) - \mathbf{y}^i\right|\right|^2\right) + \frac{\mu}{2}\sum_{l=1}^{L-1}\sum_{i=1}^{d_l}\sum_{j=1}^{d_{l+1}}(W_{ij}^l)^2. \tag{8}$$

The first term estimates the average of empirical loss; The second is the weight decay term, which keeps the value of the weights small and can improve generalization in a feed-forward neural network [47]; The regularization parameter μ balances the effects of these two terms. The optimal model can be trained via minimizing the cost function $J(\mathbf{W}, \mathbf{b})$.

6.2 Optimization

To reduce our cost function $J(\mathbf{W}, \mathbf{b})$, we adopt the gradient descent to update the parameters,

$$\begin{cases} W_{ij}^l(t+1) = W_{ij}^l(t) - \lambda\frac{\partial}{\partial W_{ij}^l}J(\mathbf{W}, \mathbf{b}), \\ b_i^l(t+1) = b_i^l(t) - \lambda\frac{\partial}{\partial b_i^l}J(\mathbf{W}, \mathbf{b}), \end{cases} \tag{9}$$

where λ is the learning rate. This parameter determines how fast or slow we will move towards the optimal values. If the λ is very large we may skip the optimal solution. If it is too small we will need too many iterations to converge to the optimal values. So using a good λ adjuster is crucial during training. According to the formulation of $J(\mathbf{W}, \mathbf{b})$, its partial derivatives of W_{ij}^l and b_i^l can be derived as,

$$\begin{cases} \frac{1}{m}\sum_{i=1}^{m}\frac{\partial}{\partial W_{ij}^l}\frac{1}{2}\left|\left|\mathbf{h}_{\mathbf{W},\mathbf{b}}(\mathbf{x}^i) - \mathbf{y}^i\right|\right|^2 + \mu W_{ij}^l, \\ \frac{1}{m}\sum_{i=1}^{m}\frac{\partial}{\partial b_i^l}\frac{1}{2}\left|\left|\mathbf{h}_{\mathbf{W},\mathbf{b}}(\mathbf{x}^i) - \mathbf{y}^i\right|\right|^2. \end{cases} \tag{10}$$

To compute the partial derivatives, we define an "error term", denoted as φ_i^l, which measures the accountability the node i in layer l should hold for any errors in the final output. For the nodes in the output layer, φ_i^L is defined as,

$$\begin{aligned} \varphi_i^L =& \frac{\partial}{\partial z_i^L}\frac{1}{2}\left|\left|\mathbf{h}_{\mathbf{W},\mathbf{b}}(\mathbf{x}) - \mathbf{y}\right|\right|^2 \\ =& -(y_i - o_i^L)f^{'}(z_i^L) \\ =& -(y_i - o_i^L)o_i^L(1 - o_i^L). \end{aligned} \tag{11}$$

This definition takes the difference between the activation values of the final layer and the target values into consideration. While for the hidden nodes, they are unable to directly compare with the target values. To deal with this problem, the φ_i^l is empirically defined as the average of the weighted error terms of the nodes that uses o_i^l as an input. It is mathematically stated as,

$$\varphi_i^l = f^{'}(z_i^L)\sum_{j=1}^{d_{l+1}}W_{ij}^l\varphi_j^{l+1} = o_i^l(1 - o_i^l)\sum_{j=1}^{d_{l+1}}W_{ij}^l\varphi_j^{l+1}. \tag{12}$$

The desired partial derivatives are given as,

$$\begin{cases} \frac{\partial}{\partial W_{ij}^l}\frac{1}{2}\left|\left|\mathbf{h}_{\mathbf{W},\mathbf{b}}(\mathbf{x}) - \mathbf{y}\right|\right|^2 = o_j^l\varphi_i^{l+1}, \\ \frac{\partial}{\partial b_i^l}\frac{1}{2}\left|\left|\mathbf{h}_{\mathbf{W},\mathbf{b}}(\mathbf{x}) - \mathbf{y}\right|\right|^2 = \varphi_i^{l+1}. \end{cases} \tag{13}$$

Based on the gradient descent, we iteratively update the parameters with initializations until the pre-defined termination criterions are met.

TABLE 4
The inter-volunteer agreement evaluation in terms of Kappa metric.

Case #	Category #	Volunteer #	Overall Agreement	Fixed-marginal Kappa	Free-marginal Kappa
$5,000$	3	3	84.0%	0.759	0.760

6.3 Pre-training and Fine-tuning

The strength of unsupervised pre-training to deep learning has been proven [48]. Our model holds three sparsely hooked hidden layers and they are pre-trained sequentially. The construction process is illustrated in Figure 2. To be specific, we first construct a graph over our whole dataset based on the underlying connections among the raw features. We next mine the signatures from this graph via dense subgraph detection. These signatures are viewed as hidden nodes and placed in the first hidden layer. Before stacking up the next hidden layer, we treat the current architecture as a learning model with only one hidden layer, and learn the initialized $\mathbf{W}^1$ and $\mathbf{b}^1$ with pseudo-labeled samples. Following that, we view the hidden nodes in the first hidden layer as the raw features, and define the co-occurrence relations between two hidden nodes as the sum of $o_i^1 o_j^1$ for all training data points $\mathbf{x}^i (1 \leq i \leq m)$. Naturally, a graph can be constructed with hidden nodes as vertices and their relations as edges. We again perform dense subgraph detection over this graph and the mined signatures as the second hidden layer nodes. We still deem the current model as a learning model with only one hidden layer. This is because the output of first hidden layer can be derived with fixed $\mathbf{W}^1$ and $\mathbf{b}^1$. Based on this setting, we can estimate the parameters $\mathbf{W}^2$ and $\mathbf{b}^2$. Analogously, we have the third hidden layer with first two hidden layers fixed.

Equally important to pre-training, the supervised fine-tuning is the key step to fit our model to specific tasks. When performing fine-tuning, our model is an indivisible unit. Compared to pre-training stage which learns the underlying signatures and their connections from various diseases, fine-tuning is strictly conducted on the specific diseases that we expect to model. However, when it comes to the parameter updates, either in pre-training or fine-tuning, W_{ij}^l is always ignored and permanently set as 0, if nodes i and j are disconnected. This boosts the learning efficiency.

7 Experiments

7.1 Feature Extraction

With the aid of MetaMap tool, we identified $18,334$ unique medical attributes. After removal of low frequency (lower than 5) and normalization, we obtained $6,003$ terminologies (sometimes, multiple various medical concepts may be mapped to the same terminology.). Each data instance can thus be represented by $6,003$-dimensional bag of medical terminology histograms. On the other hand, if the traditional term-based approach was employed to represent Dataset I, the feature dimension is up to $259,003$ without low frequency removal and $63,590$ even with low frequency removal (lower than 5), respectively.

Notably, as aforementioned, negative qualifiers are commonly used by health seekers and doctors to indicate the absence of medical signals in community-based health services[35]. Give this real-world question *"what could cause breast pain in a lady of* 37 *years of age occasionally without a lump and without breastfeeding?"* as an example. So during feature extraction, we look at the surrounding text of each medical attribute, and heuristically filter out those co-occurring with negation related words such as "not", "deny", "without", at sentence level.

The feature extraction process including the negated language handling is the same for fine-tuning and pre-training dataset.

7.2 On Health Seeker Needs Analytic

We first randomly selected $5,000$ QA pairs from dataset I. After carefully reading and thoroughly understanding, we summarized the health information needs with three high-level categories: (1) disease diagnosed but asking for supplement, (2) healthy status but asking for preventive knowledge, and (3) disease undiagnosed and asking for possible diseases of their manifested signals. As aforementioned, we then invited three volunteers to perform a user study. Each volunteer independently assigned each QA pair to any one of these categories. The study results capture the underlying distribution of health seeker needs: 79%, 6% and 15% for each of the three categories, respectively.

Here we evaluated the inter-volunteer agreement with the Kappa method introduced in [49]. The Kappa metric is a chance corrected statistic to quantitatively measure the degree of inter-volunteer agreement. It can be interpreted as expressing the extent to which the observed amount of agreement among volunteers exceeds what would be expected if all inter-volunteers made their labeling results randomly. Kappa result ranges from 0 to 1. The higher the value of Kappa, the stronger the agreement is. Kappa value of more than 0.7 typically indicates that the agreement is strong. Table 4 displays the analytical result for our work, where the Kappa value is much greater than 0.7. The result demonstrates that there are sufficient inter-volunteer agreements.

We regarded the categorization of health seeker needs as QA pairs classification. The manually labeled

TABLE 5
Categorization performance of health seeker needs.

Classifiers	Questions	Questions+ Answers	Questions+ Answers+ Tags
KNN	87.4%	82.6%	83.4%
SVM	**90.4**%	83.6%	84.8%
Naive Bayes	90.2%	85.4%	84.4%
Decision Tree	89.4%	83.8%	85.2%

$5,000$ QA pairs were split into training subset (70%) and testing subset (30%). Four well-known supervised classification approaches were respectively trained on the training subset and evaluated on the testing subset. A free software packaging popular machine learning algorithms, named Weka[8], was employed here to accomplish the classification task. The comparative evaluation results are displayed in Table 5. From this table, it is noticed that all classifiers achieve the best performance when trained and tested on questions only. This is distinctly different from other QA pair classification tasks [5], where incorporating more information cues usually results in improved performance. In our work, these results make sense. Because only the questions are posted by health seekers and they convey their needs. While answers and tags are provided by doctors to reply the corresponding questions, and if we merge them together with questions, the first and third categories are difficult to distinguish. This can be interpreted as follows: the answers and tags might be viewed as diagnosis reports for the undiagnosed diseases in the third category, and hence the questions with answers and tags are naturally shifted to the first category where diseases are diagnosed. From the angle of performance comparison among various approaches, SVM slightly outperforms the others.

We intentionally did not apply deep learning architectures to categorize the health seeker needs. This is because the number of categories is fixed and there is no necessity to generalize. Besides, when the training samples are sufficient, learning features from other data types is to paint the lily.

7.3 On Model Architecture Construction

Our sparsely connected deep learning model contains five layers including the input and output layers. The nodes in the input layer represent our $6,003$ raw features, and nodes in the output layer denote the inference results that are utilized to approximate the real disease types. In the pre-training stage using Dataset I, the number of output nodes is 945. While during fine-tuning stage with separate Dataset II and III, the number of output nodes is the same, i.e., 10.

The three hidden layers were constructed incrementally, alternating between subgraph mining and pre-training. Initially, we regarded our learning model with only one hidden layer. Each node in the hidden layer is corresponding to a signature obtained via dense subgraph mining from a large graph, where the $6,003$ raw features and their co-occurrences over Dataset I were assumed as nodes and edges, respectively. The graph altogether holds $1,446,253$ edges with normalized weights. We identified $1,965$ unique dense subgraphs from this graph. The number of raw features in each dense subgraph ranges from 2 to 10. And on average, each dense subgraph involves 2.42 raw features. That is to say, each node in the first hidden layer, on average, links to 2.42 input nodes. These dense subgraphs cover $4,376$ raw features, i.e., $1,627$ raw features are independent and don't belong to any dense subgraphs. To ensure information completeness, the independent raw features were also considered as signatures containing only one raw feature. Therefore, the first hidden layer has $3,592$ hidden nodes, which are linked to nodes in the input layer by $6,381$ edges. The current model with only one hidden layer was pre-trained on Dataset I to learn $\mathbf{W}^1$ and $\mathbf{b}^1$. To facilitate the understanding of signatures, we list some representative examples in Table 6.

We then fixed the first input layer and first hidden layer, as well as the connections between them. We viewed the $3,592$ hidden nodes in the first hidden layer as raw features to again construct a graph over Dataset I. This graph has $1,126,572$ edges with weight normalization, where $1,271$ dense subgraphs were extracted. Each dense subgraph contains 3.14 vertices on average. On the other hand, $1,318$ vertices in the graph were not assigned into any dense subgraphs, while they were also viewed as signatures as the first hidden layer construction. Hence the second hidden layer has $2,589$ hidden nodes and $5,309$ links to the first hidden layer. $\mathbf{W}^2$ was pre-trained over Dataset I with $\mathbf{W}^1$ fixed.

Similarly, with previous layers fixed, we constructed the third hidden layer. The number of hidden nodes in this layer is $1,671$ including 976 dense subgraphs as signatures and 695 independent individuals. However, it is worth emphasizing that the third hidden layer and the output layer are fully connected. The overall information of our deep architecture is displayed in Table 7.

7.4 On Inference Performance Comparison

To demonstrate the effectiveness of our proposed disease inference scheme, we compare it against three state-of-the-art techniques. All of them can benefit from unlabeled data, which ensures fair comparison.

- **MTSVM** Multi-switch Transductive SVM using L2-SVM-MFN. Transductive SVM (TSVM) is well-known for linear semi-supervised classification on large and sparse datasets. Sindhwani et al. [50] improved TSVM training

[8] http://www.cs.waikato.ac.nz/ml/weka/

TABLE 7
Overview information of our proposed sparsely connected deep learning architecture.

Layer ID	Layer I	Layer II	Layer III	Layer IV	Layer V
Layer Types	Input Layer	First Hidden Layer	Second Hidden Layer	Third Hidden Layer	Output Layer
Number of Nodes in This Layer	$6,003$	$3,592$	$2,589$	$1,671$	945 for Pre-training; 10 for Fine-tuning & Evaluation
Number of Links to Its Lower Layer	Nil	$6,381$	$5,309$	$4,599$	$1,579,095$ for Pre-training; $16,710$ for Fine-tuning & Evaluation

TABLE 6
Illustration of two types of signature examples.

Dense Subgraphs as Signatures	Independent Raw Features as Signatures
Tight Chest, Wheezing, Dyspnea	Abdominal Bloating
Diarrhea, Night Sweats, Fever Decreased in Appetite	Dicloxacillin
Pain of Skin, Itching of Skin Eruption of Skin	Diaphragmatic Hernia
Muscle Spasms of Head, Poor Balance, Joint Pain Headache, Muscle Pain	Cholestasis of Pregnancy
Chest Pain, Cough, Fatigue, Dyspnea, Weight Loss finding	Discolored Teeth

by an order of magnitude via incorporating the proposed multiple switching heuristic. Based on finite Newton technique, this implementation significantly enhances the training speed of TSVM over existing methods such as SVM^{light}.

- **DAS^{3}VM** Deterministic Annealing Semi-supervised SVM using L2-SVM-MFN. Inspired by deterministic annealing approach, this algorithm was firstly implemented in [50]. It is able to alleviate the problem of local minimum in the TSVM optimization procedure while also being computationally attractive.
- **SASR** Stacked Auto-encoder + Softmax Regression [51]. Auto-encoder is an unsupervised learning algorithm that applies back-propagation and constrains the target values to approximate the inputs. We learned three hidden layers with random initializations incrementally. Softmax regression classifier was employed as the output layer. This deep architecture is fully connected.
- **SCDL** our Sparsely Connected Deep Learning.

The implementations of **MTSVM** and **DAS^{3}VM** currently only support binary classification, and the codes are available here[9]. We reduced the single multi-class problem into multiple binary classification problems and refined the codes to fit our task. We empirically set the number of hidden nodes in **SASR** to $4,000$, $3,000$, and $2,000$ for the first to third hidden layer, respectively. Both **SASR** and **SCDL** were fine-tuned before evaluation to better match the requirements of specific inference task. In addition, these two approaches were implemented with the help of Pylearn2[10]. To train the first three models, Dataset I served as unlabeled data and other two served as labeled data. Dataset I together with 80% of labeled Dataset II and III respectively were used to train the semi-supervised SVM. Each sample in Dataset I includes question, answers and tags. While samples in Dataset II and III only contains question, where answers and tags were utilized to extract the disease names as ground truth.

Early stopping was adopted to prevent overfitting during fine-tuning. Dataset II and III were respectively divided into three parts: (a) 60% that we actually trained on; (b) 20% that we used to see how well the model is generalizing to new data; and (c) 20% for testing. The idea behind early stop is to return the model that does the best at classifying the validation set, rather than the model that assigns the highest probability to the training set. A monitor-based termination criterion implemented in Pylearn2 was adopted to determine stop point when less than 1% is being made at reducing the mislabeling examples on the validation set. Also, early stop was utilized to the pre-training.

Besides, to provide a better picture about the tradeoff between modeling complexity and performance, we compared our proposed model with dumber machine learning models, such as SVM and KNN. The QA pairs in Dataset II and III were split into 80% for training and the rest for testing. Similar to **MTSVM** and **DAS^{3}VM**, the input of dumber models was the normalized medical attributes based features, rather than medical signatures.

The disease inference performance is comparatively displayed in Table 8. From this table, we can see that **SASR** and **SCDL** significantly outperform the other two semi-supervised SVM approaches on the two datasets. This may be due to the fact the two deep learning approaches can capture the main compact and semantic factors of variation in the datasets. On the other hand, our proposed **SCDL** shows consistently superior performance over **SASR**. This may be caused by the tunable node number in each hidden layer of **SASR**, for which it is hard to obtain an optimal value. Also, the fully connected relations may link irrelevant medical attributes together.

In addition, we performed five-fold t-test between our proposed model and each of the baselines on the Dataset II and Dataset III, respectively. Table 9 illustrates the significance test results. From this table, it can be seen that all the p-values are much smaller than 0.05, which indicates that the improvements of

[9] http://vikas.sindhwani.org/svmlin.html
[10] http://deeplearning.net/software/pylearn2/

TABLE 8
Display of performance comparison among six approaches on two datasets. (Question Only)

Approaches	Performance on Dataset II	Performance on Dataset III
SVM	85.07%	77.01%
KNN	83.58%	76.65%
MTSVM	93.73%	84.54%
DAS^{3}VM	94.32%	86.44%
SASR	96.71%	89.27%
SCDL	**98.21**%	**91.48**%

TABLE 9
The paired significance test based on five fold.

Pairwise Significance Test	p-value on Dataset II	p-value on Dataset III
SCDL vs SVM	$1.47E-5$	$1.22E-5$
SCDL vs KNN	$9.72E-6$	$4.39E-6$
SCDL vs MTSVM	$5.35E-4$	$1.70E-4$
SCDL vs DAS^{3}VM	$1.34E-3$	$5.21E-4$
SCDL vs SASR	$4.41E-2$	$5.37E-3$

our proposed model are statistically significant.

7.5 On the Number of Hidden Layers

The most simple deep learning structure just has the input and output layers, which is equivalent to shallow learning framework. In our work, we incrementally added the hidden layers between the input and output layer until it satisfies the predefined convergence criterion. The convergent criterion is defined as the accuracy of deep learning model with $(n+1)$ hidden layers will not significantly outperform that with n hidden layers based on five-fold t-test. We utilized such strategy to balance the accuracy and interpretability. Table 10 displays the performance comparison with various hidden layers on Dataset II. From this table, it can be seen that the deep learning network structure with three hidden layer achieves the best performance. This convergent scenario also occurs on the Dataset III.

7.6 On the Sensitivity of Parameters

TABLE 10
Disease inference performance on Dataset II with various numbers of hidden layers.

Number of Layers	Performance on Dataset II
Structure with One hidden layer	89.00%
Structure with Two hidden layers	93.13%
Structure with Three hidden layers	**98.21**%
Structure with Four hidden layers	98.21%

The intuitive solution for the choice of learning rate is to have a constant rate. Another simple rule of thumb is to decrease the learning rate over time: $\frac{\lambda_0}{1+\tau}$, where λ_0 and τ are respectively the initial learning rate and number of epoches. However, they all suffer from the sensitivity of initializations. In our work, we employed the adaptive learning rate adjuster provided by Pylearn2 to monitor and adjust the the learning rate λ. This adjuster is triggered on each epoch. It will shrink the learning rate if the objective goes up. The idea is that in this case the learning algorithm is overshooting the bottom of the objective function. On the other hand, the adjuster will increase the learning rate if the objective decreases too slowly. This makes our learning rate parameter less important to the initialized value. The initial value is set as 0.1. Though it is not a very mathematically principled approach, it works well in practice.

The other parameter is the decay constant. It balances the importance between weight decay and cost function. A fundamental problem with weight decay is that different types of weights in the deep architecture will require different decay constants for good generalization. In our work, we have three constant to times different weight matrixes: input-hidden, hidden-hidden and hidden-output. With the support of Pylearn2, these three parameters were empirically set as 0.00005, 0.0001 and 0.00005.

7.7 On Complexity Analysis

From the efficiency perspective, **MTSVM** is the faster one among the two semi-supervised SVM approaches, even though **DAS^{3}VM** achieved better performance slightly. Among the two deep architectures, **SCDL** is many times faster than **SASR** during pre-training and fine-tuning. The computational complexity of deep learning mainly comes from weight updates. In fully connected architectures with L layers and d_l nodes in each layer, the update number of each epoch should be $\sum_{l=1}^{L-1} d_l d_{l+1}$. While in our approach, the total update number per epoch is $(\eta \sum_{l=1}^{L-2} d_l) + d_L d_{L-1}$, where η is the average number of links each node in l layer connected to the $l+1$ layer. In practice, η is usually smaller than 2. This intuitively explains why our approach is much efficient than the other fully connected deep learning architectures.

8 Conclusions and Future Work

This paper first performed user study to analyze the health seeker needs. This provides the insights of community-based health services. It then presented a sparsely connected deep learning scheme that is able to infer the possible diseases given the questions of health seekers. This scheme is constructed via alternative signature mining and pre-training in an incremental way. It permits unsupervised feature learning from other wide range of disease types. Therefore, it is generalizable and scalable as compared to previous disease inference using shallow learning approaches, which are usually trained on hospital generated patient records with structured fields. Classical deep learning architectures are densely connected and the node number in each hidden layers are tediously adjusted. In contract, our model is sparsely connected with improved learning efficiency, and the number of hidden nodes is automatically determined.

Our current model are unable to identify discriminant features for each specific disease. In the future, we will pay more attention on that.

REFERENCES

[1] S. Fox and M. Duggan, "Health online 2013," Pew Research Center, Survey, 2013.

[2] "Online health research eclipsing patient-doctor conversations," Makovsky Health and Kelton, Survey, 2013.

[3] T. C. Zhou, M. R. Lyu, and I. King, "A classification-based approach to question routing in community question answering," in *The International World Wide Web Conference*, 2012.

[4] D. A. Davis, N. V. Chawla, N. Blumm, N. Christakis, and A.-L. Barabasi, "Predicting individual disease risk based on medical history," in *The International Conference on Information and Knowledge Management*, 2008.

[5] L. Nie, M. Wang, Z. Zha, G. Li, and T.-S. Chua, "Multimedia answering: Enriching text qa with media information," in *Proceedings of the International ACM SIGIR Conference*, 2011.

[6] L. Nie, M. Wang, Y. Gao, Z.-J. Zha, and T.-S. Chua, "Beyond text qa: Multimedia answer generation by harvesting web information," *Multimedia, IEEE Transactions on*, 2013.

[7] L. Nie, Y.-L. Zhao, X. Wang, J. Shen, and T.-S. Chua, "Learning to recommend descriptive tags for questions in social forums," *Acm Transactions on Information System*, 2014.

[8] D. Zhang and W. S. Lee, "Extracting key-substring-group features for text classification," in *The ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2006.

[9] M. Gallé, "The bag-of-repeats representation of documents," in *Proceedings of the International ACM SIGIR Conference*, 2013.

[10] F. Wang, N. Lee, J. Hu, J. Sun, S. Ebadollahi, and A. Laine, "A framework for mining signatures from event sequences and its applications in healthcare data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[11] F. Wang, N. Lee, J. Hu, J. Sun, and S. Ebadollahi, "Towards heterogeneous temporal clinical event pattern discovery: A convolutional approach," in *The ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2012.

[12] S. Doan and H. Xu, "Recognizing medication related entities in hospital discharge summaries using support vector machine," in *Proceedings of the International Conference on Computational Linguistics*, 2010.

[13] T. D. Wang, C. Plaisant, A. J. Quinn, R. Stanchak, S. Murphy, and B. Shneiderman, "Aligning temporal data by sentinel events: Discovering patterns in electronic health records," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2008.

[14] I. Batal, L. Sacchi, R. Bellazzi, and M. Hauskrecht, "A temporal abstraction framework for classifying clinical temporal data," in *Proceedings of the American Medical Informatics Association*, 2008.

[15] R. Fakoor, F. Ladhak, A. Nazi, and M. Huber, "Using deep learning to enhance cancer diagnosis and classification," in *Proceedings of the International Conference on Machine Learning*, 2013.

[16] C. B. Akgül, D. Ünay, and A. Ekin, "Automated diagnosis of alzheimer's disease using image similarity and user feedback," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2009.

[17] M. Shouman, T. Turner, and R. Stocker, "Using decision tree for diagnosing heart disease patients," in *Proceedings of the Australasian Data Mining Conference*, 2011.

[18] Y. Zhang and B. Liu, "Semantic text classification of disease reporting," in *Proceedings of the International ACM SIGIR Conference*, 2007.

[19] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in *The ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2011.

[20] A. Khosla, Y. Cao, C. C.-Y. Lin, H.-K. Chiu, J. Hu, and H. Lee, "An integrated machine learning approach to stroke prediction," in *The ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2010.

[21] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via fused sparse group lasso," in *The ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2012.

[22] D. Zhu and B. Carterette, "An adaptive evidence weighting method for medical record search," in *Proceedings of the International ACM SIGIR Conference*, 2013.

[23] N. Limsopatham, C. Macdonald, and I. Ounis, "Learning to combine representations for medical records search," in *Proceedings of the International ACM SIGIR Conference*, 2013.

[24] R. W. White and E. Horvitz, "Studies of the onset and persistence of medical concerns in search logs," in *Proceedings of the International ACM SIGIR Conference*, 2012.

[25] M.-A. Cartright, R. W. White, and E. Horvitz, "Intentions and attention in exploratory health search," in *Proceedings of the International ACM SIGIR Conference*, 2011.

[26] B. Koopman, P. Bruza, L. Sitbon, and M. Lawley, "Evaluating medical information retrieval," in *Proceedings of the International ACM SIGIR Conference*, 2011.

[27] X. Huang and Q. Hu, "A bayesian learning approach to promoting diversity in ranking for biomedical information retrieval," in *Proceedings of the International ACM SIGIR Conference*, 2009.

[28] G. Luo and C. Tang, "On iterative intelligent medical search," in *Proceedings of the International ACM SIGIR Conference*, 2008.

[29] L. Nie, T. Li, M. Akbari, J. Shen, and T.-S. Chua, "Wenzher: Comprehensive vertical search for healthcare domain," in *Proceedings of the International ACM SIGIR Conference*, 2014.

[30] P. Sondhi, J. Sun, H. Tong, and C. Zhai, "Sympgraph: A framework for mining clinical notes through symptom relation graphs," in *The ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2012.

[31] S. Ghumbre, C. Patil, and A. Ghatol, "Heart disease diagnosis using support vector machine," in *Proceedings of the International Conference on Computer Science and Information Technology*, 2011.

[32] E. M. Voorhees and W. Hersh, "Overview of the the text retrieval conference 2012 medical records track," in *The Text REtrieval Conference*, 2012.

[33] S.-h. Yang, S. P. Crain, and H. Zha, "Bridging the language gap: Topic adaptation for documents with different technicality," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011.

[34] R. Cline and K. Haynes, "Consumer health information seeking on the internet: the state of the art," *Health Education Research*, 2001.

[35] B. King, L. Wang, I. Provalov, and J. Zhou, "Cengage learning at trec 2011 medical track," in *Proceedings of TREC*, 2011.

[36] L. Nie, M. Akbari, T. Li, and T.-S. Chua, "A joint local-global approach for medical terminology assignment," in *Proceedings of the International ACM SIGIR Workshop*, 2014.

[37] A. R. Aronson and F.-M. Lang, "An overview of MetaMap: historical perspective and recent advances," *Journal of the American Medical Informatics Association*, 2010.

[38] L. Nie, Y.-L. Zhao, M. Akbari, J. Shen, and T.-S. Chua, "Bridging the vocabulary gap between health seekers and healthcare knowledge," *IEEE Transactions on Knowledge and Data Engineering*, 2014.

[39] L. Nie, S. Yan, M. Wang, R. Hong, and T.-S. Chua, "Harvesting visual concepts for image search with complex queries," in *Proceedings of the International ACM Multimedia Conference*, 2012.

[40] H. Liu, L. Latecki, and S. Yan, "Fast detection of dense subgraphs with iterative shrinking and expansion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[41] H. Liu and S. Yan, "Robust graph mode seeking by graph shift," in *Proceedings of the International Conference on Machine Learning*, 2010.

[42] K. I. Penny and I. Atkinson, "Approaches for dealing with missing data in health care studies," *Journal of Clinical Nursing*, 2012.

[43] Y.-L. Zhao, L. Nie, X. Wang, and T.-S. Chua, "Personalized recommendations of locally interesting venues to tourists via cross region community matching," *ACM Transactions on Intelligent Systems and Technology*, 2013.

[44] Y. Chen, Z. Li, L. Nie, X. Wang, T.-S. Chua, and X. Zhang, "A semi-supervised bayesian network model for microblog topic classification," in *Proceedings of the International Conference on Computational Linguistics*, 2012.

[45] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[46] A. Ng, "Unsupervised feature learning and deep learning," Stanford, Tutorial, 2013.

[47] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," in *Advances in Neural Information Processing Systems*, 1992.

[48] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *Journal of Machine Learning Research*, 2010.

[49] M. Warrens, "Inequalities between multi-rater kappas," *Advances in Data Analysis and Classification*, 2010.

[50] V. Sindhwani and S. S. Keerthi, "Large scale semi-supervised linear svms," in *Proceedings of the International ACM SIGIR Conference*, 2006.

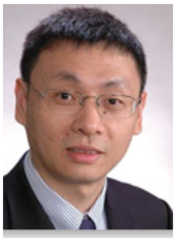
[51] Y. Bengio, "Learning deep architectures for ai," *Foundations and Trends in Machine Learning*, 2009.

Dr. Liqiang Nie is currently a research fellow in the School of Computing, National University of Singapore. He respectively received the B.E. degree from Xi'an Jiaotong University of China, Xi'an, in 2009, and the Ph.D. degree from National University of Singapore, in 2013. His research interests include information retrieval and healthcare analytics. Various parts of his work have been published in top forums including ACM SIGIR, ACM SIGMM, TOIS, TIST and TMM. Dr. Nie has been served as reviewers for various journals and conferences.

Dr. Meng Wang is a professor in the Hefei University of Technology, China. He received the B.E. degree and Ph.D. degree in the Special Class for the Gifted Young and the Department of Electronic Engineering and Information Science from the University of Science and Technology of China (USTC), Hefei, China, respectively.

His current research interests include multimedia content analysis, search, mining, recommendation, and large-scale computing. He received the best paper awards successively from the 17th and 18th ACM MM.

Dr. Luming Zhang Luming Zhang received his Ph.D. degree in computer science from Zhejiang University, China. Currently he is a Postdoctoral Research Fellow at the School of Computing, National University of Singapore. His research interests include multimedia analysis, image enhancement, and pattern recognition. His research interests mainly include Weakly supervised learning, image enhancement, and multimedia applications. He has authored/co-authored more than 40 scientific articles at various top venues, including IEEE T-IP, T-MM, T-CYB, CVPR, and ACM MM. He served as a Guest editor of Neurocomputing, Signal Processing, Multimedia tools and applications, Multimedia systems, and JVCI.



Dr. Shuicheng Yan is currently an Associate Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, and the Founding Lead of the Learning and Vision Research Group. Dr. Yan's research areas include computer vision, multimedia and machine learning, and he has authored or coauthored over 300 technical papers over a wide range of research topics, with Google Scholar citation 12,396 times and H-index-49. He is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, and ACM TRANSACTIONS ON INTELLIGENT SYSTEMS AND TECHNOLOGY, and has been serving as the Guest Editor of the special issues for TMM and CVIU. He received the Best Paper Awards from ACM MM in 2012 (demo), PCM in 2011, ACM MM in 2010, ICME in 2010, and ICIMCS in 2009, the winner prizes of the classification task in PASCAL VOC from 2010 to 2012, the winner prize of the segmentation task in PASCAL VOC in 2012, the Honourable Mention Prize of the detection task in PASCAL VOC in 2010, the 2010 TCSVT Best Associate Editor Award, the 2010 Young Faculty Research Award, the 2011 Singapore Young Scientist Award, and the 2012 NUS Young Researcher Award.

Dr. Bo Zhang received the B.E. degree from the Department of Automatic Control, Tsinghua University, China, in 1958. From 1980 to 1982, he visited the University of Illinois at Urbana-Champaign, USA, as a Scholar. He is currently a Professor with the Department of Computer Science and Technology, Tsinghua University. His research interests include artificial intelligence, machine learning, pattern recognition, knowledge engineering, intelligent robotics, and intelligent control. He was a recipient of the Prize of European Artificial Intelligence in 1984, the 1st-Class of Science and Technology Progress Prize three times in 1987, 1993, and 1998, and the 3rd-Class Prize two times in 1995 and 1999. He is a member of the Chinese Academy of Sciences.

Dr. Tat-Seng Chua is the KITHCT Chair Professor at the School of Computing, National University of Singapore (NUS). He was the Acting and Founding Dean of the School of Computing during 1998-2000. He joined NUS in 1983, and spent three years as a research staff member at the Institute of Systems Science (now I2R) in the late 1980s. Dr Chua's main research interests are in multimedia information retrieval, multimedia question-answering, and the analysis and structuring of user-generated contents. He works on several multi-million-dollar projects: interactive media search, local contextual search, and real-time live media search.

Dr. Chua has organized and served as program committee member of numerous international conferences in the areas of computer graphics, multimedia and text processing. He is the conference co-chair of ACM Multimedia 2005, CIVR (Conference on Image and Video Retrieval) 2005, and ACM SIGIR 2008, and the Technical PC Co-Chair of SIGIR 2010. He serves in the editorial boards of: ACM Transactions of Information Systems (ACM), Foundation and Trends in Information Retrieval (NOW), The Visual Computer (Springer Verlag), and Multimedia Tools and Applications (Kluwer). He sits in the steering committee of ICMR (International Conference on Multimedia Retrieval), Computer Graphics International, and Multimedia Modeling conference series; and serves as member of International Review Panels of two large-scale research projects in Europe. He is the independent director of 2 listed companies in Singapore.