Information Sciences 499 (2019) 12–24

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Diabetic complication prediction using a similarity-enhanced latent Dirichlet allocation model

Shuai Ding[a,b], Zhenmin Li[a,b], Xiao Liu[c,*], Hui Huang[a,b], Shanlin Yang[a,b]

[a] School of Management, Hefei University of Technology, Anhui, Hefei, China
[b] Key Laboratory of Process Optimization and Intelligent Decision-Making (Ministry of Education), Hefei University of Technology, Anhui, Hefei, China
[c] School of Information Technology, Deakin University, Melbourne, Australia

ARTICLE INFO

Article history:
Received 27 December 2018
Revised 30 April 2019
Accepted 12 May 2019
Available online 20 May 2019

Keywords:
Diabetic complication prediction
Similarity enhancement
Latent Dirichlet allocation
Topic mining
Multilabel classification

ABSTRACT

Diabetes and its complications have been recognized worldwide as a major public health threat. Predicting diabetic complications is regarded as a highly effective technique for increasing the survival rate of diabetic patients. While many studies currently use medical images and structured medical records, very limited efforts have been dedicated to applying data mining techniques for unstructured textual medical records, such as admission and discharge records. Moreover, the similarities among medical records that are overlooked by existing approaches could potentially improve the accuracy of prediction models. In this paper, we propose an approach for diabetic complication prediction based on a similarity-enhanced latent Dirichlet allocation (seLDA) model. Specifically, we first estimate the similarity between textual medical records after data preprocessing, and then we perform seLDA-based diabetic complication topic mining based on similarity constraints. Finally, we construct a prediction model by solving a multilabel classification problem with support vector machines (SVMs). The experimental results show that our approach outperforms the conventional LDA-based approach in similarity indices by 22.49%. Additionally, our approach shows significant improvements in prediction accuracy over four other representative seLDA-based approaches, including random forests (RF), k-nearest neighbors (KNN), logistic regression (LR) and deep neural networks (DNNs).

1. Introduction

Diabetes and its complications have been identified as one of the most serious public health diseases worldwide. According to statistics published by the International Diabetes Federation (IDF), approximately 451 million adults (aged 18 to 99 years old) worldwide were suffering from diabetes in 2017, while approximately 5 million people (aged 20 to 99 years old) have died of diabetes, constituting 9.9% of all global mortalities [6]. Diabetes is a chronic metabolic disorder, and long-term uncontrolled diabetes can lead to multiple complications, including diabetic cerebrovascular disease, coronary heart disease and endocrine function autonomic neuropathy, which are the main causes of death in patients with diabetes. Therefore, effective prevention, early prediction and long-term control of diabetic complications are the keys to saving the lives of patients and improving their quality of life.

[*] Corresponding author.
E-mail address: xiao.liu@deakin.edu.au (X. Liu).

https://doi.org/10.1016/j.ins.2019.05.037
0020-0255/© 2019 Published by Elsevier Inc.

In recent years, the development of artificial intelligence (AI) techniques, including artificial neural networks, data mining, machine learning and natural language processing, has significantly changed the research of diabetes detection and prediction. These techniques, often developed as parts of clinical decision supporting systems, have effectively improved the quality of clinical diagnosis and treatment and hence have increased the survival rate of patients with diabetes. For instance, IBM Watson Health has collaborated with the American Diabetes Association in conducting an in-depth study of diabetes utilizing big data technology [5]. The IDx company has successfully developed IDx-DR, an AI diagnostic system for screening diabetic retinopathy, which has been successfully commercialized [31]. The latest Chinese national standards of hospital informatization emphasize the importance of applying AI technology in diabetes [39].

In academia, analysis of the risk factors, predictions, and diagnosis of diabetes and diabetic complications have become popular research topics. Lee et al. [16] investigated the correlation between hemoglobin levels and risk of diabetic retinopathy (DR) based on a cross-sectional study of a population. Hunt et al. [12] used the connective tissue growth factor of blood plasma to predict the risk of myocardial infarction for patients with diabetes. Piri et al. [20] constructed a clinical decision support system for predicting diabetic retinopathy using ensemble learning to solve the problem of low compliance with diabetic retinopathy screening for patients. However, there are very few studies on diabetic complication prediction using unstructured textual medical records.

Textual medical records, such as the admission diagnoses, discharge diagnoses and progress notes, contain abundant medical information which can potentially help us to achieve better precision for predicting diabetic complications. So far, however, chronicled unstructured medical records have often been overlooked by most diabetic complication prediction approaches. Clearly, efficient data mining techniques are crucial for capturing the features of patients and diseases from massive unstructured textual medical records. For example, topic models are typically employed to study the relationship between disease-medications and disease-symptoms hidden in textual medical records [36]. The latent Dirichlet allocation (LDA) model, one of the most successful topic models, has been extensively used for information retrieval in various data mining applications such as document summarization [1]. It has been observed that the occurrence and development of diabetic complications in patients with similar characteristics, such as body diathesis, gender and age, are also similar. There are obvious similarities between the diagnostic measures and clinical presentations of similar patients. Therefore, it is undoubtedly imperative to take into account the similarities between the medical records of different patients when performing LDA-based topic mining for diabetic complication prediction.

In this paper, we propose an innovative approach for diabetic complication prediction based on a similarity enhanced latent Dirichlet allocation model. Our approach is built upon an improved data mining technique considering similarity constraints for unstructured textual medical records, as well as a multilabel classification technique for constructing the final prediction model. Our major contributions can be summarized as follows:

- We propose a similarity enhanced latent Dirichlet allocation (seLDA) model for latent topic mining to improve the accuracy of our diabetic complication prediction model. The similarity estimations between medical records are first calculated before performing diabetic complication topic mining.
- We developed a complete seLDA-based approach for predicting diabetic complications which includes a novel algorithm employing support vector machines (SVMs) as the underlying technique for solving a multilabel classification problem. We use a singular value decomposition technique to extract topic characteristics from multidimensional time-series data.
- We employ practical medical records of diabetic inpatients to evaluate our proposed approach. During the preprocessing step, a customized word library is used to perform text preprocessing steps, such as text segmentation and synonym conversion, followed by diabetic complication topic mining using seLDA. The experimental results show that our approach, named SVM-seLDA, is significantly better than other representative seLDA-based approaches in prediction quality.

The remainder of this paper is organized as follows. Section 2 reviews the research work of studies of diabetes and its complications. Section 3 introduces the overall flow of our proposed approach. Section 4 presents the details of diabetic complication topic mining with similarity constraints. Section 5 describes our approach for diabetic complication prediction and describes the procedure to construct the prediction model. Section 6 demonstrates the experimental results in detail. Finally, Section 7 concludes this paper and points out our future work.

2. Related work

The rapid development in artificial intelligence and machine learning techniques has promoted research on clinical decision support systems (CDSS) that target diabetes and diabetic complications. A number of approaches that effectively apply machine learning techniques to analyze diabetic complications have been proposed [15,19,29]. Yu et al. [38] proposed a nondestructive approach based on feature extraction and machine learning algorithms to predict proliferative diabetic retinopathy (PDR). In 2018, Dashtbozorg et al. [10] described an approach for detecting retinal microaneurysms using a local convergence filter and a random undersampling boosting classifier named RUSBoost which can provide decision support for the early diagnosis of diabetic retinopathy. However, these approaches do not focus on massive unstructured medical data with hidden similarities among different patients.

The diversity of diabetic complications, together with the complexity of their contributing factors, makes predicting them highly difficult. Therefore, it is essential to clarify the risk factors as well as the data source that contains adequate information for a prediction. To identify the clinical presentations associated with diabetic complication prediction, as shown in

Table 1
Summary of Diabetic Complications and Cause Analysis.

Ref.	Diabetic Complications	Risk Factors	Data Source
[17,24,33]	Foot disease	Address, socioeconomic status, gender, age, degree of atherosclerosis, duration of diabetes, smoking history, foot pressure, skin condition	Infrared thermal images of the foot
[30,32]	Muscle infarction	Gender, changes in coagulation fibrinolysis system	MRI images, clinical text records
[8,21,22]	Retinopathy, neuropathy, cardiovascular and kidney complications	Gender, age, race, family medical history, level of hypertension, dietary habits, obesity, cholesterol, low density lipoprotein (LDL), high density lipoprotein (HDL), glycosylated hemoglobin, creatine and albumin	Retinal and iris images
[13,23,34]	Retinopathy, cardiovascular disease, pregnancy complications, nephropathy, nerve injury, vascular injuries and foot syndrome	Gender, living environment, blood sugar level during pregnancy, level of cholesterol, LDL, HDL, glycosylated hemoglobin, creatine and albumin	clinical text records, electrocardiograms, heart rate variability records
[11,28]	Neuropathy	Gender, age, duration of diabetes, body mass index, level of blood sugar, blood pressure, blood lipid, genetic susceptibility, smoking history and dietary protein content	Clinical text records, electrocardiograms
[7,27]	Oral complications	Metabolic level, oral health behavior, oral care habits	Clinical text records

Table 1, we summarized the categories, risk factors and data sources from a series of works studying diabetic complications. It is evident from the table that textual medical records can contain abundant information which is highly conducive to constructing a more accurate model for predicting diabetic complications. As a major source of patient information, electronic medical records such as progress notes can contain abundant information of great significance. For instance, the chief complaint, admission and discharge diagnoses, the treatment process and clinical presentations. They can be utilized to effectively facilitate the prediction, analysis, diagnosis and risk prediction of various diseases. Dai et al. [9] extracted supervision information from patient progress notes and used it to identify potential microaneurysm areas through image-text mapping in the feature space to further enable the detection of diabetic retinopathy. It can also be employed to solve the problem of unbalanced microaneurysm predictions. Buchan et al. [3] proposed an ontology-guided feature extraction approach to improve the performance of predicting coronary artery disease. Miotto et al. [18] proposed an unsupervised deep learning approach that can deduce a general representation of a disease from electronic health data to improve clinical prediction modeling. However, these techniques which extract information from electronic medical records ignored the noticeable similarities in the diagnostic measures, clinical manifestations, and the development of complications among diabetic patients with similar characteristics.

Text mining techniques are widely used for the analysis of textual data. The topic model is a type of statistical model for text mining, which is used to comprehend massive textual information according to the topic distribution of documents [2]. The LDA topic model has been widely used in text analysis and data mining for medical applications. Rekatsinas et al. [25] presented SourceSeer, a novel algorithmic framework integrating spatiotemporal topic models and source-based anomaly detection techniques to effectively predict the occurrence and development of rare infectious diseases. Rumshisky et al. [26] proposed a readmission prediction approach by using LDA to perform topic mining of inpatient psychiatric discharge narrative notes. Chang et al. [4] tracked the symptoms of diabetic complications in the context of EMR. However, the above approaches mainly focused on single diabetic complication and overlooked the similarities among clinical characterizations in patients' medical records. In addition, they are not suitable for the prediction of multiple diabetic complications.

3. Overview: diabetic complication prediction based on seLDA

To improve the quality of diabetic complication prediction, it is pivotal to capture the similarities among medical records of different patients. Moreover, generating the prediction results is essentially a multilabel classification problem which requires a classifier with high performance.

In this paper, we propose an approach for diabetic complication prediction, named seLDA-based diabetic complication prediction (seDCP), which consists of two major steps as shown in Fig. 1. The similarity-enhanced LDA approach is capable of performing diabetic complication topic mining with similar constraints, which effectively improves the similarity estimation of the resulting topic models. Furthermore, the combination of singular value decomposition (SVD) and support vector machine (SVM) techniques efficiently solves the multilabel classification for multidimensional time-series data to obtain the final prediction model. The overall framework of the seDCP approach is depicted in Fig. 1.

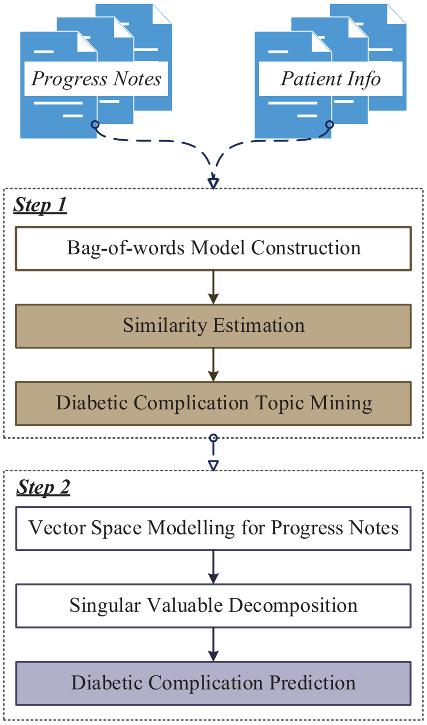

Progress Notes
Patient Info
Step 1
Bag-of-words Model Construction
Similarity Estimation
Diabetic Complication Topic Mining
Step 2
Vector Space Modelling for Progress Notes
Singular Valuable Decomposition
Diabetic Complication Prediction

Fig. 1. The process of diabetic complication prediction based on seLDA.

Step 1: The proposed seDCP approach preprocesses a medical records dataset, which includes extraction of the progress notes and segmentation of words using a bag-of-words model. The similarity estimation for each medical record pair is then obtained. After that, seDCP performs seLDA-based diabetic complication topic mining with the similarity constraints. The resulting potential topics of each medical record are acquired.

Step 2: The proposed seDCP approach constructs the vector space for the progress notes using their potential topics. The SVD technique is then employed to gather the characteristics of multidimensional time-series data within the vector space. Afterwards, seDCP generates the prediction model by solving a multilabel classification problem using the SVM technique according to the labeling sequence of diabetic complications.

Since our proposed approach is capable of handling both continuous and discrete input data, it is general enough to be applied to a wide range of textual medical data sets.

4. Diabetic complication topic mining with similarity constraints

This section describes the process of topic mining using seLDA for diabetic complication prediction is described. First, we introduce the approach of similarity constraint estimation based on a distance calculation between any two medical records. Second, we elaborate on diabetic complication topic mining with similarity constraints which relies on the seLDA model.

4.1. Similarity constraint estimation

The similarity estimation of patients' medical records is used for selecting analogous medical records, so that the distance between any two medical records is less than a threshold. Various medical records can be generated during hospitalization, such as admission diagnoses, discharge diagnoses and progress notes. The complexity of similarity estimation increases exponentially with the expanding scope of the considered medical records. Therefore, it is crucial to confine the scope of the similarity analysis. Otherwise, the process will not be scalable. In this work, we only consider admission diagnoses to estimate the similarity because diabetes can lead to a variety of complications, and thus a series of diabetic complications often appear in the admission diagnoses of patients with diabetic history [32].

Patients with different personal attributes, such as occupation, age and gender, often exhibit diverse diabetic symptoms and complications. Moreover, differences in body diathesis also result in dissimilarities in drug tolerance, affecting medications and other aspects of the process of clinical diagnosis and treatment. Therefore, it is necessary to consider the basic information (personal attributes) of patients when calculating the similarity of medical records. These attributes can be classified into two types: continuous and discrete. We discretize continuous attributes according to their piecewise relationship. Furthermore, the degree of similarity decreases with an increasing difference level. We define a distance between two attributes so that there is a negative correlation between the distance and the level of difference. The distance between continuous attributes can be calculated by the following formula:

$$d(cattr_i, cattr_j) = 1 - \frac{|seg_i - seg_j|}{n-1}, \tag{1}$$

where $cattr_i$ and $cattr_j$ represent two distinct attributes, seg_i and seg_j, respectively, to denote their corresponding segment indices. The number of segments is n. For instance, considering that there is a piecewise relationship between different age groups, there is no significant difference in symptoms and medications among the same age group. Thus, we classify the age groups into four levels according to the international population age structure: level 1 for children aged 0–17, level 2 for youths aged 18–45, level 3 for middle-aged individuals of 46–59 years old, and level 4 for the elderly aged above 59.

In the distance calculation of discrete attributes, since there is no obvious piecewise relationship among the same type of discrete attributes, we consider the distance of the same attribute to be 1 and the distance of distinct attributes to be 0.

$$d(dattr_i, dattr_j) = \begin{cases} 1 & dattr_i = dattr_j \\ 0 & dattr_i \neq dattr_j \end{cases} \tag{2}$$

where $dattr_i$ and $dattr_j$ represent the attribute data of the i-th and j-th entities, respectively. For two-valued data, considering that there is no evident level relationship between each pair of data, we assign the distance to 1 if both values are identical and 0 otherwise.

Since the attributes of admission diagnosis fundamentally differ from other discrete attributes, we employ the Jaccard distance to calculate the distance between diagnostic results, formulated as follows:

$$d(dia_i, dia_j) = \frac{dia_i \cap dia_j}{dia_i \cup dia_j}, \tag{3}$$

where dia_i and dia_j represent the boolean vectors of the i-th and j-th patients, respectively. For instance, when $dia_i = 01001$ and $dia_2 = 11101$, the Jaccard distance is $d(dia_i, dia_j) = 2/4 = 0.5$.

To properly balance the influences of age, sex, occupation and admission diagnosis, weighting parameters μ_1, μ_2, μ_3 and μ_4 are employed to calculate the similarity between medical records as follows:

$$\begin{aligned} sim(T_i, T_j) = {} & \mu_1 \cdot d(sex_i, sex_j) + \mu_2 \cdot d(age_i, age_j) + \mu_3 \cdot d(job_i, job_j) + \mu_4 \cdot d(dia_i, dia_j) \\ & \mu_1 + \mu_2 + \mu_3 + \mu_4 = 1 \\ & 0 \le \mu_1, \mu_2, \mu_3, \mu_4 \le 1 \end{aligned} \tag{4}$$

After calculating the similarity for each pair of medical records, we select those with similarity values larger than a threshold τ into a target set D, given by:

$$D = \left\{ (T_i, T_j) | i, j \in [1, M] \right\} \tag{5}$$

where M is the number of patients.

4.2. seLDA-based diabetic complication topic mining

The LDA model is a highly modular generative statistical model capable of searching for potential topics in a document through large amounts of data. The LDA model contains three layers: documents, topics and words. LDA defines a document as a series of combinations of topics, and words with a corresponding topic. The topics and words of a document are generated by the probability distributions of document-topic and topic-word. These two distributions can be obtained using the Dirichlet distribution. Each document has its own topic-related probability distribution, and the words in the document are sampled from different topic distributions.

$$\sum p(word|doc) = \sum (word|topic) * p(topic|doc) \tag{6}$$

In the LDA model, the goal of iteration is to maximize the occurrence probability $p(Z, W|\alpha, \beta)$ to find the topic distribution of the text. However, conventional LDA models do not take into account the similarity of a medical record, which leads to a sharp difference in the topic distribution for similar medical records. This, in turn, affects the construction of diabetic complication models. Our goal is to establish a topic model that satisfies the similarity constraint of medical records. Therefore, we use the Gibbs-EM algorithm to solve seLDA and achieve our goal by altering the objective function in the EM step.

Since there can be multiple chronological progress notes in a single medical record, the similarity estimation should consider the similarity between any two different sets of progress notes; that is, the document-topic distribution of different sets of progress notes in the target set D should be as similar as possible. Let p be the index of medical records for a patient, of which there are L_m progress notes. Hence, the vector space of progress note topics is given by $\theta r^p = \{\theta_{m,1}, \theta_{m,2}, \ldots, \theta_{m,L_m}\}$.

Let θr^p and θr^q denote two sets of progress note topics; we use the average distance of topic distributions to calculate the similarity constraints, formulated by:

$$dis(\theta r^p, \theta r^q) = \frac{d(\sum_{lm=1}^{L_m} \sum_{ln=1}^{L_n} d(\theta_{m,L_m}, \theta_{m,L_n}))}{L_m * L_n}, \tag{7}$$

where $d(\theta_{m,L_m}, \theta_{m,L_n})$ represents the Euclidean distance between these two vectors. The larger $dis(\theta r^p, \theta r^q)$ is, the lower the similarity it represents. We apply the similarity constraint into the target function $L(\alpha, \beta) = log(p(Z, W|\alpha, \beta))$, resulting in the following equation:

$$L(\alpha, \beta) = \log(p(Z, W|\alpha, \beta)) - \gamma \sum_{(\theta r^p, \theta r^q) \in D} dis(\theta r^p, \theta r^q) \tag{8}$$

where γ is a hyperparameter used to adjust the weight of the similarity constraint. The Gibbs-EM algorithm differs from the Gibbs sampling approach because it modifies the document-topic distribution of document m, denoted by $\theta_m = \{\theta_{m,1}, \theta_{m,2}, \ldots, \theta_{m,k}\}$, as follows:

$$\theta_{m,k} = \frac{e^{\mu_{m,k}}}{\sum_{i=1}^{K} e^{\mu_{m,i}}} \tag{9}$$

where $\theta_{m,k}$ represents the probability of k being the topic of document m. Since $\mu_m = \mu_{m,1}, \mu_{m,2}, \ldots, \mu_{m,k}$ follows a normal distribution, the maximum objective function can be further rewritten as:

$$L(\mu) = \log(p(Z, W|\mu\theta)) + \log p\big(\mu|(0, N(0, 1))\big) - \gamma \sum_{(\theta r^p) \in D} dis(\theta r^p, \theta r^q) \tag{10}$$

The procedure of Gibbs-EM model training is described using the pseudocode in Algorithm 1.

Algorithm 1 Gibbs-EM Model Training for seLDA.

Input: A tuple $\langle \mathcal{M}, \mathcal{K}, \mathcal{V} \rangle$, where $\mathcal{M}$ is the number of documents, $\mathcal{K}$ is the number of topics, $\mathcal{V}$ is the number of words in the document, constant $\beta = 0$, and the maximum number of iterations for Gibbs-EM, E-step and M-step is tEM, tE and tM, respectively.

Output: A tuple $\langle \theta, \varphi \rangle$, where θ represents the document-topic distribution, and φ represents the topic-word distribution.

```
procedure TRAIN($\langle \mathcal{M}, \mathcal{K}, \mathcal{V} \rangle$)
    Randomly initialize the topic index number $z$ for each word of every document, as well as the topic parameter $\mu$
    while $t < tEM$ do
        E-step:
          Perform Gibbs sampling on topic words, iterate $tE$ times for the document
        M-step:
          Calculate objective function $L(\mu)$
          Iteratively calculate $\mu_{m,k}$ using stochastic gradient descent approach
        while $n < tM$ do
            for $m = 1 : M$ do
                for $k = 1 : K$ do
                    $\mu^t_{m,k(n+1)} = \mu^t_{m,k(n)} - \sigma \frac{\partial(L(\mu))}{\mu^t_{m,k(n)}}$
                end for
            end for
        end while
    end while
    return $\langle \theta, \varphi \rangle$
end procedure
```

5. seLDA-based diabetic complication prediction

In this section, we describe the second step of the proposed approach, i.e., the seLDA-based diabetic complication prediction approach. A series of progress notes is accumulated during the hospitalization of a patient. A timed topic series in the form of a multidimensional vector space of the progress note is produced after performing the diabetic complication topic mining using the seLDA model. Traditional classification techniques require that each data entry should be unidimensional and of equal length, whereas each timed topic series consists of a series of unequal topic sequences, which leads to the inapplicability of traditional machine learning classification techniques such as k-nearest neighbors, random forest, logistic regression, and support vector machines. Due to the characteristics of multidimensional time-series data, it is essential to perform feature extraction and selection to gather the characteristics of multidimensional time-series data before the classification is carried out. We exploit the SVD technique to collect the characteristics, as it has been widely used in solving clustering and prediction problems with multidimensional time-series data in various scenarios [35,37]. The basic SVD can be defined as:

$$A = U\sum V^T, \tag{11}$$

where A denotes the matrix to be decomposed, corresponding to the timed topic series. U and V are termed unitary matrices. Σ represents a rectangular diagonal matrix. The diagonal entries σ_i are known as the singular values of A. The singular values on the diagonal σ_i form eigenvectors. Since the length of the timing dimension of each patient's time-series is not necessarily identical, we perform zero-padding for all the eigenvectors, leading to the same length in each dimension. Thereby, the final eigenvector x can be obtained.

Since there are usually multiple diseases in diabetic complication predictions, the final step is a multilabel classification problem. Label-based transformations [40] are a common technique to transform the original multilabel classification problem into a set of single-label problems. Due to the efficient employment of support vector machines in disease predictions [14], we select SVM as the underlying classification technique. Specifically, the number of SVMs is determined by the number of diabetic complications to solve the multilabel classification problem.

The proposed approach for diabetic complication prediction is described by the pseudocode in Algorithm 2, which is termed the seLDA-based Diabetic Complication Prediction Algorithm (seDCPA). In Algorithm 2, P represents the serial number sequence of all the patients and X is the matrix formed by the eigenvectors λ_p obtained from the SVD of the multidimensional time-series data of all patients. Y_i is the complication label matrix obtained from the discharge diagnosis for disease i, which is used as the gold standard for classification. s represents the index sequence of diabetic complication labels in Y_i. t represents the classification result of i, and MOD signifies the final diabetic complication prediction model.

Algorithm 2 seLDA-Based Diabetic Complication Prediction Algorithm.

```
Input: A tuple $\langle\theta, \varphi\rangle$, where $\theta$ represents the topic distribution, and $\varphi$ represents the topic-word distribution.
Output: Diabetic complication prediction model $MOD$
 1: procedure PREDICTION($\langle\theta, \varphi\rangle$)
 2:     Perform diabetic complication topic mining using Algorithm 1
 3:     for each $p$ in $P$ do
 4:         Construct vector space for progress notes
 5:         $\lambda_p \leftarrow SVD(\theta r^p)$
 6:     end for
 7:     Generate matrix $X$ and corresponding $Y_i$ from all $\lambda_p$
 8:     for each $i$ in $s$ do
 9:         $mod_i, t \leftarrow SVM(X, Y_i)$
10:         $MOD \leftarrow mod_i$
11:     end for
12:     return $MOD$
13: end procedure
```

A description of the pseudo code is given as follows. First, we extract each progress note's topic from the data set (line 1). Second, we construct the topic vector space of each progress note of each patient, followed by feature extraction (lines 2–5). Third, we construct matrix X from eigenvectors λ_p and then generate the complication label matrix Y_i according to the sequence of patients appearing in X (line 6). Fourth, we perform a multilabel classification using the SVM according to the labeling sequence in Y_i and store the result in MOD (lines 7–11). Note that the number of classifiers equals the number of topics. Finally, the diabetic complication prediction model is obtained (line 12).

6. Evaluation

In this section, we present an experimental evaluation of the proposed approach on a set of real medical records. This section is organized as follows. The description of the medical data used for the evaluation is presented first, followed by an introduction of the evaluation metrics. Next we compare the similarity estimations obtained from our seLDA and conventional LDA approaches. To evaluate the performance of diabetic complication predictions, we then compare our SVM-seLDA approach with other seLDA-based approaches that use representative classification algorithms, including random forests (RF), k-nearest neighbors (KNN), logistic regression (LR) and deep neural networks (DNNs).

6.1. Data description

We utilize authentic medical records of diabetic patients from a local general hospital. All records are anonymized and retrieved with the patient's consent and hospital's authorization for release of information. We randomly select the progress notes of 500 patients, among which there are 265 males and 235 females. Their ages range from 9 to 87. We select 70% of the total number of progress notes for model training and use the remaining 30% for testing. Note that the number of progress notes is equal to the number of days of hospitalization for a patient. There are a total of 2855 progress notes for 500 patients who are used for experiments, or an average of 5.71 progress notes per patient. Moreover, the diabetic complication prediction is fundamentally a multilabel classification problem, and thus, it is necessary to preprocess the dataset so that it can be applied to conventional classification algorithms.

In the evaluation, we perform the following steps for data preprocessing: (1) extract the effective textual data, such as progress notes, ages and jobs of patients, from the original medical records using regularization method, (2) perform text segmentation of the progress notes using a dictionary and stop words customized for analyzing diabetic complications, and (3) represent the segmented progress notes using a bag-of-words model to obtain the input data for applying the seLDA model. We also quantize patient information, such as ages and jobs, to facilitate the similarity estimation.

6.2. Evaluation metrics

To evaluate the performance of the seLDA model, we introduce the similarity estimation index (SIMD) as a metric for similarity estimations. We calculate the SIMD value for a set of progress note pairs obtained by LDA and seLDA. The definition of SIMD is given below.

$$\begin{aligned} SIMD &= \sum_{(\theta r^m, \theta r^n)\in D} dis(\theta r^m, \theta r^n) \\ &= \sum_{(\theta r^m, \theta r^n)\in D} \frac{d(\sum_{lm=1}^{L_m}\sum_{ln=1}^{L_n} d(\theta_{m,L_m}, \theta_{m,L_n}))}{L_m * L_n} \end{aligned} \tag{12}$$

Here, *SIMD* represents the similarity estimation of the entire vector set of similar medical record pairs, denoted by *D*. Note that $d(\theta_{m,L_m}, \theta_{m,L_n})$ represents the Euclidean distance between these two medical records. We use accuracy and Hamming loss as evaluation metrics to gauge the performance of the approach for diabetic complication prediction. Several terms must be introduced before defining these metrics.

Definition 6.1. Let *PR*(*m*) and *GS*(*m*) represent the set of diabetic complications of a prediction and discharge diagnosis for patient *m*, respectively. Furthermore, we define four sets representing four cases of prediction correctness.

$$\begin{aligned} TP(m) &= \{\, d_{m,i} \mid d_{m,i} \in PR(m) \wedge d_{m,i} \in GS(m)\} \\ TN(m) &= \{\, d_{m,i} \mid d_{m,i} \notin PR(m) \wedge d_{m,i} \notin GS(m)\ \} \\ FP(m) &= \{\, d_{m,i} \mid d_{m,i} \in PR(m) \wedge d_{m,i} \notin GS(m)\} \\ FN(m) &= \{\, d_{m,i} \mid d_{m,i} \notin PR(m) \wedge d_{m,i} \in GS(m)\} \end{aligned} \tag{13}$$

The prediction of diabetic complications is in the form of a set of diseases for each patient. The outcome can be categorized into four cases of prediction correctness by comparing patients' prediction results with their discharge diagnoses. The accuracy is fundamentally a description of prediction errors and a measure of statistical bias. We choose accuracy over precision because the numbers of true positive and false positive predictions are comparable. The accuracy is determined by the ratio of the sum of true positive and false positive predictions to the total number of predictions, as defined below.

$$ACC = \frac{1}{N}\sum_{m=1}^{M} \frac{|TP(m)| + |TN(m)|}{|PR(m)|} \tag{14}$$

The Hamming loss is a loss function that denotes the fraction of wrong predictions to the total number of predictions. HL is widely used as an evaluation metric for multilabel classification problems, as defined below:

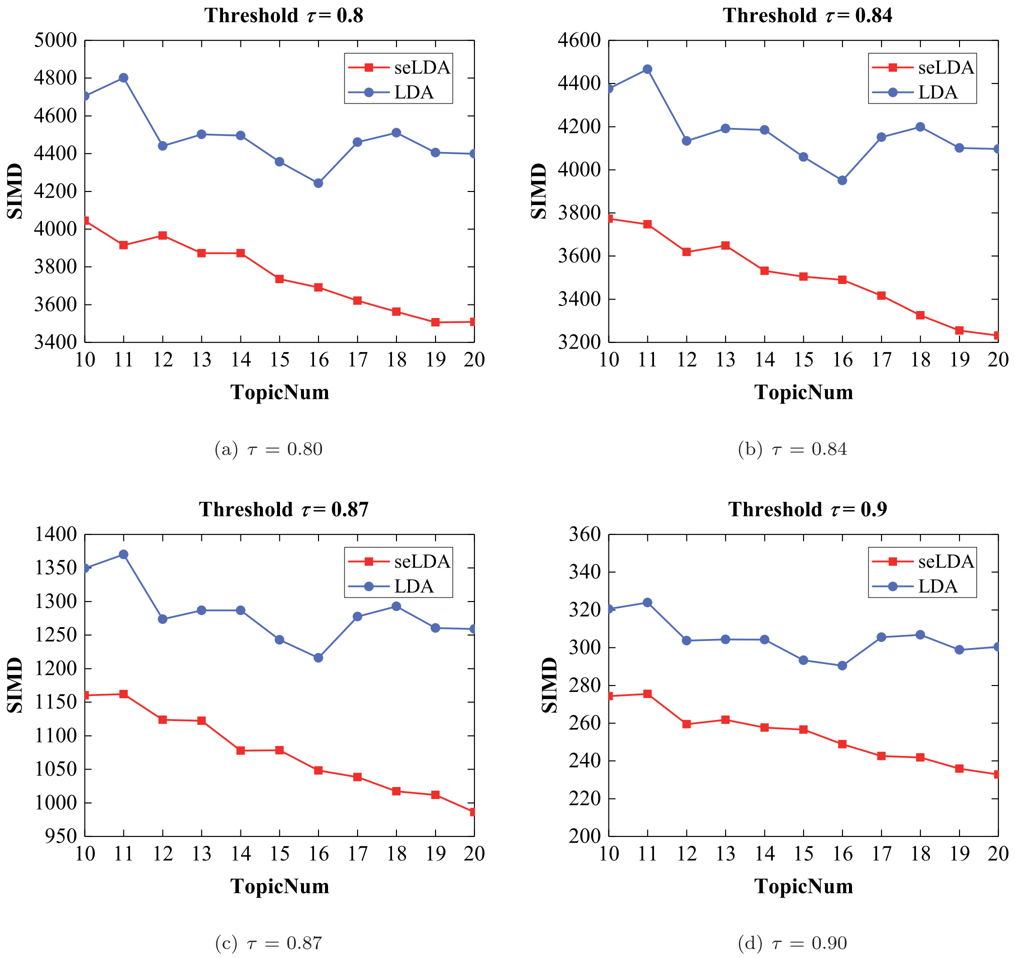
$$HL = \frac{1}{M \cdot L}\sum_{l=1}^{L}\sum_{m=1}^{M} y_{m,l} \oplus z_{m,l}, \tag{15}$$

where $y_{m,j}$ and $z_{m,j}$ are diabetic complications in *TP*(*m*) and *TP*(*m*), respectively.

6.3. Comparison of similarity estimations

In this subsection, we evaluate the similarity estimations resulting from the LDA and seLDA approaches. The SIMD is employed as the evaluation metric, and a lower SIMD value indicates a higher similarity. Considering the influence of the number of topics on topic modeling and the number of similar progress notes on the similarity estimation, we take the threshold of the similarity constraints (defined in subsection 4.1) and the number of topics as varying parameters. In the experiments we set the similarity thresholds (τ) to 0.80, 0.84, 0.87 and 0.90 respectively. The number of topics (K) varies from 10 to 20, with a step length of 1.

Evaluation results of the similarity estimations in terms of SIMD for the seLDA and the conventional LDA approaches are summarized in Fig. 2. As shown in the figure, the SIMD results decrease with increasing number of topics since more



(a) $\tau = 0.80$

(b) $\tau = 0.84$

(c) $\tau = 0.87$

(d) $\tau = 0.90$

Fig. 2. Comparison of the similarity estimation indices using seLDA- and LDA-based topic mining approaches with varying τ.

similarities can be extracted when there are more topics in the document. The seLDA approach constantly outperforms the conventional LDA approach for all similarity threshold settings, with an average improvement of 16.35%. The maximum improvement reaches 22.49%, which occurs when τ and the number of topics are 0.90 and 20, respectively. Moreover, compared to LDA, the seLDA approach achieved better linearity for the improvement with a rising number of topics, which indicates that the robustness and stability of our approach are ensured. In general, the improvement in similarity estimations rises as the similarity constraint threshold τ increases, because a higher threshold leads to fewer eligible progress notes with more similarities. The gap between similarity estimations obtained by the two approaches increases with the rising number of topics for each case because more similarity can be extracted using seLDA with more topics.

6.4. Similarity constraint threshold determination

As indicated by the previous experiment, the similarity constraint threshold τ has a significant impact on the similarity estimation index, and in turn affects the quality of diabetic complication prediction. The correlation between the number of medical record pairs and τ is summarized in Fig. 3. As seen from the figure, when the threshold τ is below 0.8, the number of medical record pairs decreases exponentially from 58,131 to 56,240 with increasing τ. A large number of medical record pairs leads to high complexity for acquiring an accurate diabetic complication prediction model. In contrast, the number of medical record pairs rapidly declines to zero with τ above 0.8. The shortage of medical record pairs with an adequate similarity is also detrimental to predicting diabetic complications. Based on the previous analysis, it is reasonable to set τ to 0.8 in the following experiments.

6.5. Comparison with seLDA-based approaches

In this subsection, we evaluate the quality of diabetic complication predictions based on seLDA using accuracy and Hamming loss as evaluation metrics. To better evaluate our SVM-seLDA approach, we implement four other seLDA-based ap-

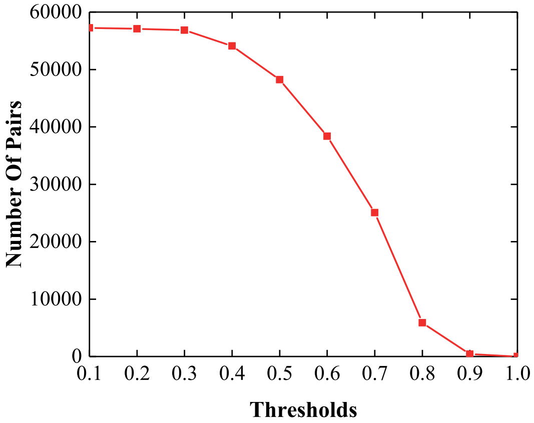
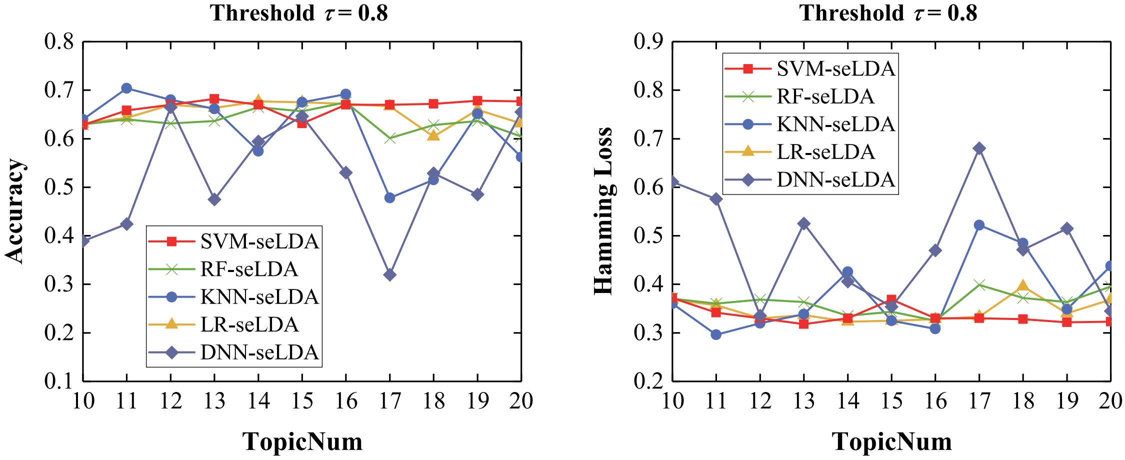


Fig. 3. Correlation between the number of medical record pairs and τ.



(a) Comparison of accuracy

(b) Comparison of Hamming loss

Fig. 4. Comparison of seLDA-based approaches with varying numbers of topics when $\tau = 0.8$.

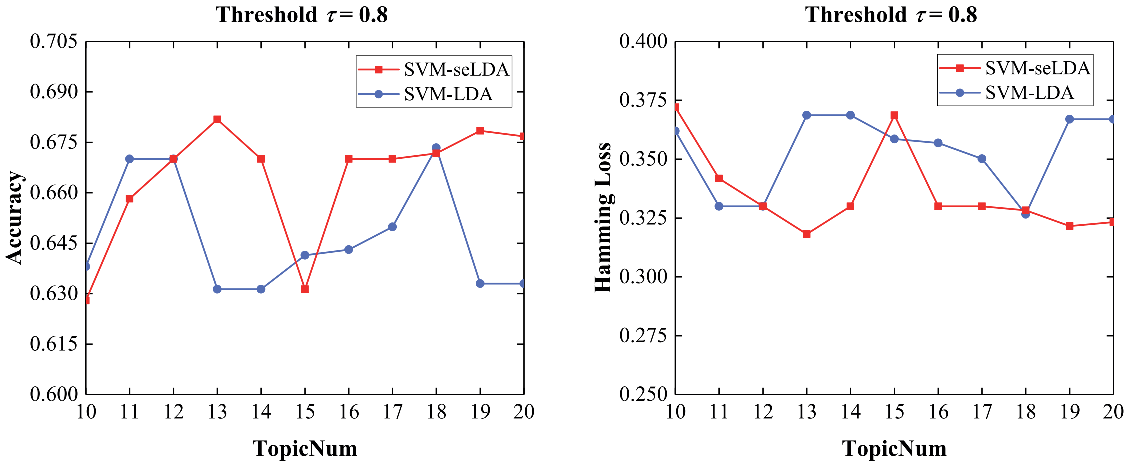
proaches to compare, including RF, KNN, LR and DNN as multilabel classifiers. For each approach, a series of combinations of parameter settings are evaluated manually, from which we select the one with the best performance in terms of accuracy and Hamming loss.

For KNN, the number of the nearest neighbors is set to 4, and the Euclidean distance is employed as the distance metric. For LR, we use the *liblinear* library to optimize the loss function, and the number of iterations is set to 100. The L2 Regularization method is also used with the penalty factor set to 1. For RF, GINI importance is employed as the impurity metric. The number of estimators is set to 50. The minimum number of samples required to split an internal node is set to 2, while the minimum number of samples required to be at a leaf node is set to 1. The minimum weighted fraction of the sum total of weights required to be at a leaf node is set to 0.15. The DNN employed in the evaluation is a multilayer perceptron consisting of an input layer, a hidden layer and an output layer. The input layer contains the vector set of similar medical record pairs. There are 19 subnet layers in the hidden layer, and the rectified linear unit activation function is used by each layer. The output layer uses a softmax function as the activation function, and the number of epochs is set to 220. For SVM, the linear kernel function was used with the parameter C set to be 10^{-3} and γ to be the inverse of the number of features.

Comprehensive experimental results are shown in Fig. 4. From the figure, the SVM-seLDA approach obtains the best average performance in terms of both accuracy and Hamming loss. The results obtained by both the KN-seLDA and DNN-seLDA approaches exhibit drastic fluctuations in both accuracy and Hamming loss. The SVM-seLDA approach outperforms the RF-seLDA and LR-seLDA approaches in both accuracy and Hamming loss except when the number of topics is 14 or 15. The KN-seLDA approach achieves the best accuracy and Hamming loss when the number of topics is 11, 12, 15 or 16. However, SVM-seLDA still outperforms KN-seLDA by 9.14% on average. Note that the SVM-seLDA approach is less sensitive to the number of topics, which means it has better stability. The corresponding accuracy and Hamming loss results stabilize at 0.68 and 0.33, respectively, despite the increasing numbers of topics. The SVM-seLDA approach achieved worse results

Table 2
The average performance of each approach and the improvement of SVM over other seLDA-based approaches.

	SVM	RF	KN	LR	DNN	Average Improvement
Average accuracy Value	**0.6642**	0.6367	0.6212	0.6540	0.5193	–
Improvement	–	4.33%	6.92%	1.57%	27.91%	10.18%
Average Hamming Loss Value	**0.3358**	0.3633	0.3788	0.3460	0.4807	–
Improvement	–	7.58%	11.35%	2.96%	30.15%	13.01%



(a) Comparison on accuracy

(b) Comparison on Hamming loss

Fig. 5. Comparison with the LDA-based approach using SVM with varying numbers of topics when $\tau = 0.8$.

in both accuracy and Hamming loss when the number of topics reached 15 due to the unpredicted effects of specific topic numbers on the classification results. Note that the DNN-seLDA achieves the worst performance in general, which implies that the DNN technique is not suitable in this scenario.

The average accuracy and Hamming loss values, as well as the improvements of SVM-seLDA compared with other approaches, are shown in Table 2. From the table, we see that the SVM-seLDA approach is superior to all other approaches in both metrics on average with the improvement ranging from 4.33% to 30.15%. The performance of RF-seLDA is the closest to SVM-seLDA, but its results deteriorate when the number of topics is greater than 16.

6.6. Comparison with the LDA-based approach using SVM

To validate the effectiveness of our seLDA-based approach, we further compare the performance of diabetic complication predictions between seLDA- and LDA-based approaches. The SVM algorithm is employed as the classification algorithm for this experiment because it has been proven to be one of the best methods for multilabel classification. The results comparing the accuracy and Hamming loss are shown in Fig. 5. For both accuracy and Hamming loss, the results obtained by SVM-seLDA improve at the beginning, and stabilize at 0.675 and 0.325, respectively.

Referring to Fig. 5, there is no evident correlation between the performance of diabetic complication predictions and the number of topics. There is a sudden performance drop for SVM-seLDA when the number of topics reaches 15. There is a performance improvement when the number of topics reaches 18 for SVM-LDA. When the number of topics increases from 10 to 11, there is a notable performance increase for both the SVM-seLDA and SVM-LDA methods. These performance variations are caused by both characteristics of the medical records employed in these experiments, and our methods of topic mining and classification. On average, SVM-seLDA outperforms SVM-LDA by 2.7% and 5.2% for accuracy and Hamming loss, respectively. This improvement is mainly achieved though the advantage of seLDA over LDA in topic mining.

7. Conclusions

Due to the severe harm to human health caused by diabetes and diabetic complications, high-quality predictions of diabetic complications are of great significance for the prevention and treatment of diabetic complications, which will effectively improve the survival rate of diabetic patients. In this paper a novel approach for diabetic complication prediction based on a similarity-enhanced latent Dirichlet allocation (seLDA) model is presented. We first calculate the similarity of each medical record pair after data preprocessing, and then seLDA-based diabetic complication topic mining is performed using the obtained similarity estimations as constraints. Afterwards, we construct the vector space for the progress notes using their latent topics. After that, the prediction model is acquired by solving the multilabel classification problem using the support vector machines technique. The experimental results show that our proposed SVM-seLDA approach consistently outperforms the conventional LDA approach and other seLDA-based approaches in both similarity estimations and diabetic complication predictions. In the future we will incorporate the time-series information contained within medical records when applying the seLDA model based approach for diabetic complication prediction. Specifically, time stamps of medical records will be used during topic mining with the aim of further improving the quality of diabetic complication predictions.

Declaration of Competing Interest

I would like to declare on behalf of my co-authors that: a. This manuscript is the authors' original work and has not been published nor has it been submitted simultaneously elsewhere. b. All authors have checked the manuscript and have agreed to the submission.

Acknowledgment

This work is fully supported by the National Natural Science Foundation of China [Nos. 91846107, 71571058, and 71690235], Anhui Provincial Science and Technology Major Project [Nos. 17030801001, and 18030801137], and the Fundamental Research Funds for the Central Universities [No. PA2019GDQT0021].

References

[1] D.M. Blei, Probabilistic topic models, Commun. ACM 55 (4) (2012) 77–84.
[2] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (Jan) (2003) 993–1022.
[3] K. Buchan, M. Filannino, Ö. Uzuner, Automatic prediction of coronary artery disease from clinical narratives, J. Biomed. Inform. 72 (2017) 23–32.
[4] N.-W. Chang, H.-J. Dai, J. Jonnagaddala, C.-W. Chen, R.T.-H. Tsai, W.-L. Hsu, A context-aware approach for progression tracking of medical concepts in electronic medical records, J. Biomed. Inform. 58 (2015) S150–S157.
[5] Y. Chen, J.E. Argentinis, G. Weber, Ibm watson: how cognitive computing can be applied to big data challenges in life sciences research, Clin. Therap. 38 (4) (2016) 688–701.
[6] N. Cho, J. Shaw, S. Karuranga, Y. Huang, J. da Rocha Fernandes, A. Ohlrogge, B. Malanda, Idf diabetes atlas: global estimates of diabetes prevalence for 2017 and projections for 2045, Diab. Res. Clin. Pract. 138 (2018) 271–281.
[7] H.-L. Collin, L. Niskanen, M. Uusitupa, J. Töyry, P. Collin, A.-M. Koivisto, H. Viinamäki, J.H. Meurman, Oral symptoms and signs in elderly patients with type 2 diabetes mellitus: a focus on diabetic neuropathy, Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endodontol. 90 (3) (2000) 299–305.
[8] L. Crossland, D. Askew, R. Ware, P. Cranstoun, P. Mitchell, A. Bryett, C. Jackson, Diabetic retinopathy screening and monitoring of early stage disease in australian general practice: tackling preventable blindness within a chronic care model, J. Diab. Res. 2016 (2016).
[9] L. Dai, R. Fang, H. Li, X. Hou, B. Sheng, Q. Wu, W. Jia, Clinical report guided retinal microaneurysm detection with multi-sieving deep learning, IEEE Trans. Med. Imag. 37 (5) (2018) 1149–1161.
[10] B. Dashtbozorg, J. Zhang, F. Huang, B.M. ter Haar Romeny, Retinal microaneurysms detection using local convergence index features, IEEE Trans. Image Process. 27 (7) (2018) 3300–3315.
[11] J.L. Gross, M.J. De Azevedo, S.P. Silveiro, L.H. Canani, M.L. Caramori, T. Zelmanovitz, Diabetic nephropathy: diagnosis, prevention, and treatment, Diab. Care 28 (1) (2005) 164–176.
[12] K.J. Hunt, M.A. Jaffa, S.M. Garrett, D.K. Luttrell, K.E. Lipson, M.F. Lopes-Virella, L.M. Luttrell, A.A. Jaffa, V. Investigators, et al., Plasma connective tissue growth factor (ctgf/ccn2) levels predict myocardial infarction in the veterans affairs diabetes trial (vadt) cohort, Diab. Care (2018) dc172083.
[13] R. Kaaja, Vascular complications in diabetic pregnancy, Thromb. Res. 123 (2009) S1–S3.
[14] F. Lauer, Y. Guermeur, Msvmpack: a multi-class support vector machine package, J. Mach. Learn. Res. 12 (Jul) (2011) 2293–2296.
[15] I. Lazar, A. Hajdu, Retinal microaneurysm detection through local rotating cross-section profile analysis, IEEE Trans. Med. Imag. 32 (2) (2013) 400–407.
[16] M.-K. Lee, K.-D. Han, J.-H. Lee, S.-Y. Sohn, J.-S. Jeong, M.-K. Kim, K.-H. Baek, K.-H. Song, H.-S. Kwon, High hemoglobin levels are associated with decreased risk of diabetic retinopathy in korean type 2 diabetes, Sci. Rep. 8 (1) (2018) 5538.
[17] Y. Liu, A. Polo, M. Zequera, R. Harba, R. Canals, L. Vilcahuaman, Y. Bello, Detection of diabetic foot hyperthermia by using a regionalization method, based on the plantar angiosomes, on infrared images, in: Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the, IEEE, 2016, pp. 1389–1392.
[18] R. Miotto, L. Li, B.A. Kidd, J.T. Dudley, Deep patient: an unsupervised representation to predict the future of patients from the electronic health records, Sci. Rep. 6 (2016) 26094.
[19] C. Pereira, D. Veiga, J. Mahdjoub, Z. Guessoum, L. Gonçalves, M. Ferreira, J. Monteiro, Using a multi-agent system approach for microaneurysm detection in fundus images, Artifi. Intellig. Med. 60 (3) (2014) 179–188.
[20] S. Piri, D. Delen, T. Liu, H.M. Zolbanin, A data analytics approach to building a clinical decision support system for diabetic retinopathy: developing and deploying a model ensemble, Decis. Support Syst. 101 (2017) 12–27.
[21] A. Prayitno, A.D. Wibawa, M.H. Purnomo, Early detection study of kidney organ complication caused by diabetes mellitus using iris image color constancy, in: Information & Communication Technology and Systems (ICTS), 2016 International Conference on, IEEE, 2016, pp. 146–149.
[22] P. Prentasic, S. Loncaric, Z. Vatavuk, G. Bencic, M. Subasic, T. Petkovic, L. Dujmovic, M. Malenica-Ravlic, N. Budimlija, R. Tadic, Diabetic retinopathy image database (dridb): a new database for diabetic retinopathy screening programs research, in: Image and Signal Processing and Analysis (ISPA), 2013 8th International Symposium on, IEEE, 2013, pp. 711–716.
[23] L. Quintero, S. Wong, R. Parra, J. Cruz, N. Antepara, D. Almeida, F. Ng, G. Passariello, Stress ecg and laboratory database for the assessment of diabetic cardiovascular autonomic neuropathy, in: Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE, IEEE, 2007, pp. 4339–4342.
[24] P. Rani, B. Aliahmad, D.K. Kumar, A novel approach for quantification of contour irregularities of diabetic foot ulcers and its association with ischemic heart disease, in: Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE, IEEE, 2017, pp. 1437–1440.
[25] T. Rekatsinas, S. Ghosh, S.R. Mekaru, E.O. Nsoesie, J.S. Brownstein, L. Getoor, N. Ramakrishnan, Sourceseer: Forecasting rare disease outbreaks using multiple data sources, in: Proceedings of the 2015 SIAM International Conference on Data Mining, SIAM, 2015, pp. 379–387.
[26] A. Rumshisky, M. Ghassemi, T. Naumann, P. Szolovits, V. Castro, T. McCoy, R. Perlis, Predicting early psychiatric readmission with natural language processing of narrative discharge summaries, Translat. Psychiatry 6 (10) (2016) e921.
[27] R. Saini, S.A. Al-Maweri, D. Saini, N.M. Ismail, A.R. Ismail, Oral mucosal lesions in non oral habit diabetic patients and association of diabetes mellitus with oral precancerous lesions, Diab. Res. Clin. Pract. 89 (3) (2010) 320–326.

[28] K.J. Schjoedt, K. Rossing, T.R. Juhl, F. Boomsma, P. Rossing, L. Tarnow, H.-H. Parving, Beneficial impact of spironolactone in diabetic nephropathy, Kidney Int. 68 (6) (2005) 2829–2836.

[29] S.A.A. Shah, A. Laude, I. Faye, T.B. Tang, Automated microaneurysm detection in diabetic retinopathy using curvelet transform, J. Biomed. Opti. 21 (10) (2016) 101404.

[30] F. Shaik, A.K. Sharma, S.M. Ahmed, A novel approach for detection and analysis of abnormalities in mri images related to diabetic myonecrosis, Int. J. Mod. Electron. Commun. Eng. (IJMECE) 4 (1) (2016).

[31] D.A. Sim, P.A. Keane, A. Tufail, C.A. Egan, L.P. Aiello, P.S. Silva, Automated retinal image analysis for diabetic retinopathy in telemedicine, Curr. Diab. Rep. 15 (3) (2015) 14.

[32] A. Trujillo-Santos, Diabetic muscle infarction: an underdiagnosed complication of long-standing diabetes, Diab. Care 26 (1) (2003) 211–215.

[33] S.B. Vali, A.K. Sharma, S.M. Ahmed, Implementation of modified chan vase algorithm to detect and analyze diabetic foot ulcers, in: Recent Trends in Electrical, Electronics and Computing Technologies (ICRTEECT), 2017 International Conference on, IEEE, 2017, pp. 36–40.

[34] A.I. Vinik, D. Ziegler, Diabetic cardiovascular autonomic neuropathy, Circulation 115 (3) (2007) 387–397.

[35] H.-W. Wang, H. Gu, Z.-L. Wang, Fuzzy prediction of chaotic time series based on svd matrix decomposition, in: Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on, 4, IEEE, 2005, pp. 2493–2498.

[36] T. Wang, Z. Huang, C. Gan, On mining latent topics from healthcare chat logs, J. Biomed. Inform. 61 (2016) 247–259.

[37] Y. Xie, A. Wulamu, Y. Wang, Z. Liu, Implementation of time series data clustering based on svd for stock data analysis on hadoop platform, in: Industrial Electronics and Applications (ICIEA), 2014 IEEE 9th Conference on, IEEE, 2014, pp. 2007–2010.

[38] S. Yu, D. Xiao, Y. Kanagasingam, Machine learning based automatic neovascularization detection on optic disc region, IEEE J. Biomed. Health Inform. 22 (3) (2018) 886–894.

[39] H. Zhang, B.T. Han, Z. Tang, Constructing a nationwide interoperable health information system in china: the case study of sichuan province, Health Policy Technol. 6 (2) (2017) 142–151.

[40] M.-L. Zhang, Z.-H. Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (8) (2014) 1819–1837.