ARTICLE IN PRESS
DECSUP-12847; No of Pages 16
Decision Support Systems xxx (2017) xxx–xxx
Contents lists available at ScienceDirect
Decision Support Systems
journal homepage: www.elsevier.com/locate/dss

A data analytics approach to building a clinical decision support system for diabetic retinopathy: Developing and deploying a model ensemble

Saeed Piri [a], Dursun Delen [b,*], Tieming Liu [a], Hamed M. Zolbanin [c]

[a] Department of Industrial Engineering and Management, College of Engineering, Architecture and Technology, Oklahoma State University, Stillwater, OK 74078, United States
[b] Department of Management Science and Information Systems, Spears School of Business, Oklahoma State University, Tulsa, OK 74106, United States
[c] Department of Information Systems and Operations Management, Miller School of Business, Ball State University, Muncie, IN 47306, United States

ARTICLE INFO

Article history:
Received 8 September 2016
Received in revised form 22 April 2017
Accepted 7 May 2017
Available online xxxx

Keywords:
Diabetic retinopathy
Data analytics
Predictive modeling
Clinical decision support systems
Model ensembles
Variable importance

ABSTRACT

Diabetes is a common chronic disease that may lead to several complications. Diabetic retinopathy (DR), one of the most serious of these complications, is the most common cause of vision loss among diabetic patients. In this paper, we analyzed data from more than 1.4 million diabetics and developed a clinical decision support system (CDSS) for predicting DR. While the existing diagnostic approach requires access to ophthalmologists and expensive equipment, our CDSS only uses demographic and lab data to detect patients' susceptibility to retinopathy with a high accuracy. We illustrate how a combination of multiple data preparation and modeling steps helped us improve the performance of our CDSS. From the data preprocessing aspect, we aggregated the data at the patient level and incorporated comorbidity information into our models. From the modeling perspective, we built several predictive models and developed a novel "confidence margin" ensemble technique that outperformed the existing ensemble models. Our results suggest that diabetic neuropathy, creatinine serum, blood urea nitrogen, glucose serum plasma, and hematocrit are the most important variables in detecting DR. Our CDSS provides several important practical implications, including identifying the DR risk factors, facilitating the early diagnosis of DR, and solving the problem of low compliance with annual retinopathy screenings.

1. Introduction

An estimated 29 million Americans, aged 20 years or older, either have been diagnosed or remain undiagnosed with diabetes. In 2012, the United States incurred about $245 billion in direct and indirect medical costs related to disability, work loss, and premature mortality associated with diabetes [1]. Over the course of this disease, the blood glucose level imbalance may lead to several complications, including neurological disorders, heart problems, kidney disease, and hyperosmolarity. A major complication of diabetes that has not received enough attention is diabetic retinopathy. This complication is caused by damage to the blood vessels of retina, the light-sensitive tissue at the back of the eye. At its early stages, diabetic retinopathy may be asymptomatic or only show mild vision problems, but if it is not diagnosed and treated in time, it can eventually cause blindness. This unnoticed progress of diabetic retinopathy has turned it into the most common cause of vision loss among diabetics and a leading cause of blindness for American adults. According to the *2014 National Diabetes Statistics Report*, between the years 2005 and 2008, 4.2 million of American diabetics aged 40 or older suffered from diabetic retinopathy and this number is expected to grow to about 16 million by 2050. Among patients who have had diabetes for up to 20 years, almost all type I and >60% of type II diabetics develop retinopathy [2].

Although retinopathy is preventable and existing treatments can slow down the disease progress, vision loss that happens in the late stages of retinopathy cannot be restored. Thus, it is critical to diagnose this complication as early as possible. The current method for diagnosing diabetic retinopathy is a comprehensive eye examination in which after a patient's eye is dilated, an ophthalmologist examines the retina with an indirect ophthalmoscope and a special lens. Unfortunately, and despite the intimidating statistics about the high prevalence of retinopathy, the annual diabetic retinopathy evaluation has one of the lowest rates of patient compliance for several reasons. First, many patients do not seek proper medical attention because this disease is asymptomatic at the early stages; second, availability of ophthalmologists is low or even nonexistent in many areas, especially in rural communities; and third, many patients find the necessary eye dilation unpleasant. Because of this low compliance rate, about 50% of patients with diabetic retinopathy are undiagnosed (National Eye Institute report, 2015). Therefore,

* Corresponding author at: Department of Management Science and Information Systems, Spears School of Business, Oklahoma State University, 700 N. Greenwood Ave., Tulsa, OK 74106, United States.
E-mail addresses: piri@ostatemail.okstate.edu (S. Piri), dursun.delen@okstate.edu (D. Delen), tieming.liu@okstate.edu (T. Liu), hmzolbanin@bsu.edu (H.M. Zolbanin).
URL: (S. Piri), http://spears.okstate.edu/delen (D. Delen), (T. Liu), (H.M. Zolbanin).

http://dx.doi.org/10.1016/j.dss.2017.05.012
0167-9236/

the rising prevalence of diabetes, coupled with barriers to ophthalmological screenings that lead to a high rate of undiagnosed diabetic retinopathy patients, create an urgent need for a tool to detect this complication. To be useful, this tool should be non-invasive, readily available to diabetic patients, validated on a large number of cases, and eliminate the need for specialized equipment that is not universally available. This study sets out to employ an analytics approach on data collected during routine primary care visits to fills this gap. Specifically, we build a clinical decision support system (CDSS) for prediction of diabetic retinopathy that satisfies the aforementioned requirements for diagnostic tools.

Our CDSS has several advantages over the existing diagnostic systems. First and foremost, it only uses the results of a simple blood test and demographic data to predict the risk of diabetic retinopathy. Therefore, unlike the dominant approach in the extant literature that uses image processing on images of retina, it does not require eye exams, thereby addressing the low rates of compliance with annual ophthalmologic tests for diabetic patients. Equally important, our CDSS eliminates the need to have access to specialists, which is particularly critical for patients living in remote areas. Second, our decision support system is based on a large database of clinical encounters that span over several years and across several states of America. The decisions of this system are far more generalizable and valid compared to those of other systems that employ a similar approach but only use data from a few hundred patients. Finally, our CDSS uses a greater number of risk factors to predict the outcome. This not only improves the prediction results, but also sheds more light on contribution and importance of different risk factors on diabetic patients' susceptibility to retinopathy.

We expect that the CDSS we develop in this effort will be able to detect diabetic retinopathy at its early stages with a high degree of accuracy. This CDSS, which relies exclusively on lab data, not only helps overcome one of the major barriers to the early diagnosis of diabetic retinopathy, but also provides a new standard of care that will improve quality and increase compliance in healthcare without raising costs. Therefore, this work contributes to the decision support and medical informatics literatures from three perspectives: methodology, data management, and application. In the methodological aspect, we develop and evaluate a novel approach in building ensemble models. This approach aggregates the predictions of individual models by calculating a weighted confidence margin across all models. However, in contrast to the existing weighted averaging ensembles that assign weights to individual models based on their overall prediction performance, our confidence margin ensemble assigns varying weights to the constituting models. These weights are calculated for each observation in the data and are based on the distance between the estimated probabilities of records and the decision cut off point. We show how this approach improves the accuracy of decisions made by our CDSS. From the data management perspective, we processed a very large transactional database of clinical encounters and aggregated the observations at the patient level. This enabled us to create a single data set containing comorbid conditions of patients to develop an accurate picture of their health status. Consequently, our CDSS is able to consider a larger number of risk factors and provides a more realistic depiction of the coexistence of chronic diseases. Finally, in the application aspect, we develop an accessible, easy-to-implement, and inexpensive solution to the currently high proportion of undiagnosed retinopahty among diabetics. This CDSS reduces direct and indirect medical costs of the healthcare system in the United States and more importantly, saves eyesight for a large number of citizens.

The rest of this paper is organized as follows. In Section 2, we provide a brief review of the related literature in the field of healthcare analytics. In Section 3, we explain the methodology we employed in the study. The results of the predictive models are provided in Section 4, and Section 5 includes discussion and conclusion of the research.

2. Literature review

The extensive availability of healthcare data in the past decade, as well as advances in the area of data mining and machine learning, has generated the interesting field of healthcare analytics. The development of CDSSs by data analysts with the aid of clinical experts' knowledge has eased the burden on physicians and clinicians and smoothed clinical procedures. Analyzing healthcare data and applying machine learning techniques in this area have several benefits: patients can be stratified based on the severity of a particular disease or condition and, consequently, suitable treatments can be provided for each group; risk factors of different diseases can be identified, leading potentially to better health management; and diseases can be detected at early stages, allowing for appropriate interventions and treatments. For a comprehensive discussion about healthcare analytics, its promises and its potentials we refer the readers to [3].

Even though CDSSs based on EMR data have been broadly used by practitioners in recent years, the field of ophthalmology has received limited attention [4]. This dearth exists while several researchers have studied the relationship between diabetic retinopathy and different potential risk factors. For instance, Karma, et al. [5] studied the existence of diabetic retinopathy in 328 diabetic patients using ophthalmoscopy and wide field fundus photography and tried to identify the association between diabetes duration and other risk factors, such as nephropathy and coronary disease. In another study, Klein, et al. [6] measured the relationship between retinopathy and hyperglycemia by studying 1878 diabetics.

Most of the existing CDSSs for diabetic retinopathy use image processing algorithms. While these algorithms facilitate early detection of diabetic retinopathy, they require an image of the retina. Therefore, although they ease the burden of assessing the images of retina, they fail to address the evident barrier of patients' access to specialists. Examples of studies that belong to this category are (Kahai, et al. [7], Paunksnis, et al. [8], Marsolo, et al. [9], Tsai, et al. [10], Noronha, et al. [11] Bursell, et al. [12], Kumar and Madheswaran [13], and Xiao, et al. [14]). We refer the readers to Mookiah, et al. [15] for a comprehensive review of research in this category.

The other category of CDSSs for diabetic retinopathy includes those matched with lenses or an ophthalmoscope that can be used on a smartphone. Prasanna, et al. [16] proposed a portable smartphone-based CDSS that requires attaching an ophthalmoscope to a smartphone to capture fundus images, and captured images will be processed by the algorithm installed on the smartphone. Bourouis, et al. [17] also proposed a smartphone-based algorithm integrated with microscopic lenses used to capture retinal images. Their CDSS uses a neural network model to analyze such images and provide the results. Despite all the benefits of these algorithms, additional equipment is still required for retinal imaging, which, for many diabetics and primary care providers, may be cost-prohibitive or unavailable.

Many research projects have studied the association of retinopathy and different lab tests. For instance, the association of retinopathy and hemoglobin A1c has been shown in several studies ([18,19,20]). Researchers have also studied the relationship between cholesterol and retinopathy and have found the two to be related ([21,22]). The Diabetes Control and Complications Trial (DCCT) and the U.K. Prospective Diabetes Study (UKPDS) have shown that controlling the glucose level could reduce the risk of retinopathy [23]. Other studies have shown that retinopathy and hypertension are associated [24]. Besides blood tests, some urine tests such as proteinuria are shown to be associated with retinopathy [25].

While these studies show the potential for developing tools that can detect or predict retinopathy using lab results, only a few studies have used lab and demographic data to detect diabetic retinopathy without requiring retinal imaging. Skevofilakas, et al. [26] developed a CDSS using data from 55 type I diabetic patients to predict the risk of diabetic retinopathy. They applied classification-based Rule Induction with C5.0,

Please cite this article as: S. Piri, et al., A data analytics approach to building a clinical decision support system for diabetic retinopathy: Developing and deploying a mode..., Decision Support Systems (2017), http://dx.doi.org/10.1016/j.dss.2017.05.012