Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

doi:10.3233/FAIA230683

Challenges in the Exploitation of Historical Clinical Data for the Classification of Diabetic Retinopathy Patients

Jordi PASCUAL-FONTANILLES [a,1] Aida VALLS [a,c] Antonio MORENO [a,c]
Pedro ROMERO-AROCA [b,c]

[a] *ITAKA, Dept. Enginyeria Informàtica i Matemàtiques*
Universitat Rovira i Virgili, Tarragona, Catalonia, Spain
[b] *Servei d'Oftalmologia, Hospital Universitari Sant Joan de Reus, Catalonia, Spain*
[c] *Institut d'Investigació Sanitària Pere Virgili, Tarragona, Catalonia, Spain*
ORCiD ID: Jordi Pascual-Fontanilles https://orcid.org/0000-0002-7528-5819, Aida Valls https://orcid.org/0000-0003-3616-7809, Antonio Moreno https://orcid.org/0000-0003-3945-2314, Pedro Romero-Aroca https://orcid.org/0000-0002-7061-8987

Abstract.
Analysing medical data stored in Electronic Health Records is of great interest to build clinical decision support systems. There is a lot of hidden knowledge in these databases, obtained from the continuous work of medical practitioners when attending and diagnosing patients. However, it is not easy for Machine Learning methods to exploit these data. Their use always requires a careful and complex pre-processing stage. In this paper, we study the case of diagnosis of diabetic retinopathy, which is made by ophthalmologists. The characteristics of a 12-year dataset about Type-2 diabetic people are analysed. Several numerical and categorical variables were selected by experts as relevant risk factors for this disease. We explain the challenges that are being faced in order to generate a dataset composed by time series with the same length and intervals. The final aim of the research is to build a clinical decision support system that can make a personalised prediction of the evolution of the disease.

Keywords. Time series classification, Clinical decision support systems, Classification, Diabetic retinopathy

1. Introduction

This short paper analyses the characteristics of a real dataset about Type-2 diabetic people. The dataset contains different clinical and analytical variables of different types (numerical and categorical), which have been extracted from the Electronic Health Records (EHR) of diabetic patients. It contains data collected between 2010 and 2021. The final goal of our research project is to construct a classifier that distinguishes between several

[1] Corresponding Author. Email: jordi.pascual@urv.cat.

levels of the diabetic given the historical data collected in his/her successive visits. The purpose of this short paper is to present the complexities of the available temporal data, which will guide the pre-processing stage of data preparation, which is usually a crucial step to build proper classifiers.

2. Diabetic retinopathy classification

Diabetic retinopathy (DR) is an ocular complication produced by the increase of blood sugar levels due to diabetes, which damage the back of the eye (retina). The lesions it can produce cause vision loss and even blindness if not detected and treated at an early stage [1]. According to the medical ETDRS standard classification [2], the following categories are considered: healthy ($DR = 0$), mild ($DR = 1$), moderate ($DR = 2$) and severe ($DR = 3$). In 2021, the number of diabetic patients worldwide was estimated at 537 million people. In Spain it is expected to affect about 11.1% people by 2030, reaching 3.8 million inhabitants [3]. The prevalence of DR among type-2 diabetic people in Spain in 2022 was 15.28%, including 1.92% for the severe level ($DR = 3$) [1].

Computer methods to assist in DR diagnosis are being developed. Many works detect DR with the analysis of eye fundus images. Because they are costly to obtain, other clinical decision support systems (CDSS) based solely on EHR data are nowadays on the focus of research. Machine learning methods are trained with EHR to build DR classifiers. Several studies on the literature analyse how different kinds of classifiers perform on EHR-based DR datasets [4,5,6,7,8,9]. They use techniques such as random forests, XGBoost, logistic regression, support vector machines or k-nearest neighbours. There is not a consensus in the selection of the best classifier, as there are varied results on each study. In [10][11], Retiprogram was presented as a CDSS to assess Type-2 Diabetes Mellitus patients' risk of developing DR. It takes into account the current patient's conditions, including the analytical data from the last blood analysis. It is an ensemble method consisting on a Fuzzy Random Forest classifier, with an accuracy of 81%, sensitivity of 80%, and specificity over 84%. First, it was developed as a binary CDSS [11]. Later, it was extended to deal with the ordinal multi-class case, being able to detect the levels of DR severity [12].

3. Challenges of the DR Data Available in Electronic Health Records

The goal of our research is to improve DR severity detection by means of a time series classifier. Therefore, we need to work with multi-variate series data representing the changes of several clinical and analytical risk factors, stored in the EHR.

For this research, we have obtained a dataset with 231,064 diabetic patients from Catalonia (Spain), with their medical data from 2010 to 2021. Figure 1 depicts the frequency of patient's visits by ophthalmologists. Only patients with at least 2 visits are considered. It can be observed that patients are usually visited with a frequency over 18 months; therefore, the diagnosis must be made with relatively short sequences of data.

Usual machine learning techniques for time series classification assumes a dataset with a homogeneous time series (i.e. with equal length and intervals). However, the data extracted from EHR is not homogeneous. Considering the 12-year dataset on diabetic patients, we present here a list of challenges that need to be tackled:

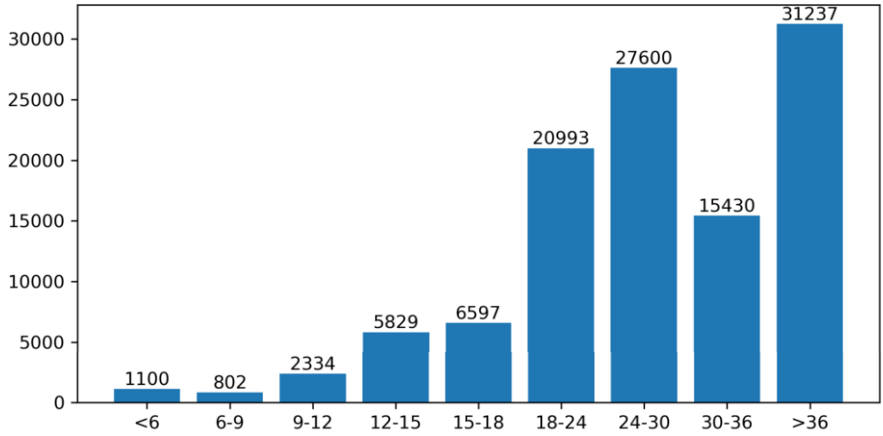

Figure 1. Diabetic patients visits mean frequency in months

- **Short sequences**: around half of the patients have a single visit on their EHR, and must be discarded. Moreover, just 1% of them have at least 6 visits in 12 years, which we consider the minimum amount of information needed to construct a reliable temporal sequence for a patient. Thus, from the whole dataset, only 2108 patients could be considered for training and testing. Moreover, the length of the sequences is still unusually short for time series classifiers, which usually manage long samples.
- **Irregular visits frequency**: most time series algorithms require data spaced at regular time intervals. This is not the case here, where we can find two situations when establishing a period of 1 year. First, we can find a few cases of patients having multiple records for the same period. Second, in some sequences there is a lack of a visit in a certain year, whose data should be interpolated to fill in the gap.
- **Different data alignment**: all the time sequences should be aligned on certain time points with the same number of occurrences. However, the time span for different patients varies, as it depends on the date of diagnosis of diabetes. For the trends analysis, we need a sequence of the same number of consecutive years, even if they do not start and end at the same date.
- **Missing data**: the values of the risk factors must be extracted from the EHR for each time point. These values may be stored in different sources (blood test reports, doctor's notes, etc.) and may have been recorded at different dates. Thus, each patients' visit cannot be found in the EHR as a single element, and has to be reconstructed from the multiple sources that are considered on a visit. If any of those values is missing, the visit data cannot be included in the time series, unless we apply some imputation method.
- **Labelling mistakes**: some mistakes may be found on the diagnosis data at EHR, due to human errors. For instance, we found that for some patients the level of DR decreases over time, which is not a common situation. Similarly, in the case of treatment type, we found some cases in which no medication was reported for a patient treated with insulin before, which is very unusual. Therefore, data values must be carefully curated taking into account medical expertise before being used.

- **Data imbalance**: the DR dataset is highly imbalanced towards the negative class. This is expected due to the low incidence of this disease. The patients distribution is: 1730 patients in $DR = 0$ (82.1%), 209 in $DR = 1$ (9.9%), 133 in $DR = 2$ (6.3%) and 36 in $DR = 3$ (1.7%). Consequently, the availability of examples with a high degree of DR is scarce. Because of this over-representation, the classification models have more difficulty to correctly identify and distinguish the positive classes. Some oversampling technique could be used to minimise this problem.

The presented challenges require tailored solutions that take into account the medical knowledge related to the DR disease. Time series classification methods could only be used if satisfactory pre-processing techniques are applied on the raw ERH data. These preparation methods are our next line of work. Moreover, these challenges should also be analysed for other diseases diagnoses that have a temporal evolution.

Acknowledgements

This study is funded by Instituto de Salud Carlos III and the European Union with project PI21/00064, and by URV projects 2022PFR-URV-41 and 2021PFR-B2-103. The first author has a pre-doctoral grant (2022 FI_B1 00036) from Generalitat de Catalunya and Fons Social Europeu.

References

[1] Romero-Aroca P, López-Galvez M, Martinez-Brocca MA, Pareja-Ríos A, Artola S, Franch-Nadal J, et al. Changes in the Epidemiology of Diabetic Retinopathy in Spain: A Systematic Review and Meta-Analysis. Healthcare (Switzerland). 2022 7;10:1318.

[2] Wilkinson CP, Ferris FL, Klein RE, Lee PP, Agardh CD, Davis M, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology. 2003;110:1677-82.

[3] Shaw JE, Sicree RA, Zimmet PZ. Global estimates of the prevalence of diabetes for 2010 and 2030. Diabetes research and clinical practice. 2010;87(1):4-14.

[4] Zhao Y, Li X, Li S, Dong M, Yu H, Zhang M, et al. Using Machine Learning Techniques to Develop Risk Prediction Models for the Risk of Incident Diabetic Retinopathy Among Patients With Type 2 Diabetes Mellitus: A Cohort Study. Frontiers in Endocrinology. 2022 5;13:885.

[5] Tsao HY, Chan PY, Su ECY. Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms. BMC Bioinformatics. 2018 8;19:111-21.

[6] Ogunyemi OI, Gandhi M, Tayek C. Predictive Models for Diabetic Retinopathy from Non-Image Teleretinal Screening Data. AMIA Summits on Translational Science Proceedings. 2019;2019:472.

[7] Sun Y, Zhang D. Diagnosis and Analysis of Diabetic Retinopathy Based on Electronic Health Records. IEEE Access. 2019;7:86115-20.

[8] Ogunyemi O, Kermah D. Machine Learning Approaches for Detecting Diabetic Retinopathy from Clinical and Public Health Records. AMIA Annual Symposium Proceedings. 2015;2015:983.

[9] Reddy GT, Bhattacharya S, Ramakrishnan SS, Chowdhary CL, Hakak S, Kaluri R, et al. An Ensemble based Machine Learning model for Diabetic Retinopathy Classification. International Conference on Emerging Trends in Information Technology and Engineering, ic-ETITE 2020. 2020 2.

[10] Romero-Aroca P, Valls A, Moreno A, Sagarra-Alamo R, Basora-Gallisa J, Saleh E, et al. A Clinical Decision Support System for Diabetic Retinopathy Screening: Creating a Clinical Support Application. Telemedicine and e-Health. 2019;25:31-40.

[11] Saleh E, Błaszczyński J, Moreno A, Valls A, Romero-Aroca P, de la Riva-Fernández S, et al. Learning ensemble classifiers for diabetic retinopathy assessment. AI in Medicine. 2018;85:50-63.

[12] Pascual-Fontanilles J, Lhotska L, Moreno A, Valls A. Adapting a Fuzzy Random Forest for Ordinal Multi-Class Classification. Frontiers in Artificial Intelligence and Applications. 2022 10;356:181-90.