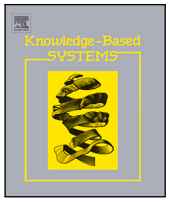
Knowledge-Based Systems 300 (2024) 112154

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



Multivariate data binning and examples generation to build a Diabetic Retinopathy classifier based on temporal clinical and analytical risk factors

Jordi Pascual-Fontanilles [a,d,*], Aida Valls [a,c,d], Pedro Romero-Aroca [b,c]

[a] ITAKA, Dept. Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Av.Paisos Catalans 26, 43007 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
[d] Center of Environmental, Food and Toxicological Technology: TecnATox, Reus, Spain

ARTICLE INFO

Keywords:
Time series classification
Clinical decision support systems
Classification
Class imbalance
Fuzzy logic
Diabetic Retinopathy

ABSTRACT

In this paper, we explore the possibility of exploiting retrospective clinical data from Electronic Health Records (EHR) for classification tasks in chronic patients. The different intervals, short length and high class imbalance make it unfeasible to use traditional time series techniques. The first contribution of the paper is a preprocessing method to construct a multivariate time series dataset using EHR data, which infers missing data and regularizes the data frequency. The second contribution addresses class imbalance by using domain knowledge and existing short EHR series. We synthetically extrapolate patients' data by using similar long time series and a fuzzy-based approach. The paper addresses the problem of detection of Diabetic Retinopathy (DR). Expert domain knowledge from ophthalmologists has been used in the proposed techniques to guide the processing of time series. The novelty in that case study consists in not using eye-fundus image analysis. Instead, the proposed methods are based solely on EHR data. Several multivariate multiclass time series classifiers are used to detect the four levels of DR severity from the pre-processed data sequences. Experiments prove the quality of the sequence preprocessing techniques proposed for EHR data. Results indicate that the TapNet classifier is the best one for DR grading. Despite being tested for DR detection, the proposed data preparation methods are applicable to other diseases with similar characteristics.

1. Introduction

Chronic conditions are the ones that last for a long period (at least 12 months) and can be controlled but not cured, such as arthritis, asthma, cancer, diabetes, or osteoporosis, among others [1]. Such patients undergo continuous monitoring of their health conditions, mainly by the family physician, because these conditions may be degenerative and cause the appearance of other secondary diseases. In this work, we propose a novel method to perform the risk assessment of patients to suffer a certain disease derived from a chronic condition using the historical data stored in the Electronic Health Records (EHR). In that way, we exploit the knowledge about the medical state of the patient in the history of previous visits to the clinician to make the risk assessment, instead of simply relying on the data of the current state. The assessment is done by using a time series classifier with ordinal multiple classes, which indicate different degrees of severity of the disease. The characteristics of the analytical and clinical information available in the EHR for chronic patients require designing appropriate preprocessing methods for the transformation of the patient's data into homogeneous time series. This is a crucial step in order to build good-performing classifiers to help in the diagnosis of these secondary diseases.

There are several challenges that need to be considered when applying machine learning techniques to the prediction of secondary diseases due to chronic conditions [2]. First, regardless of continuous monitoring, short sequences of data are found for a considerable number of patients, as the periodicity of the monitoring is highly variable from one patient to another. Without a sufficient number of visits, most data sequences are not usable for a retrospective study. Additionally, irregular visit frequency poses another challenge, as most time series classifiers require data spaced at regular intervals, which is not the case here, where we can find patients having multiple records for the same years or gaps in certain years. Moreover, different data alignment and missing values in some variables are also difficulties to be addressed. Finally, labelling mistakes may occur due to incorrect diagnosis data or human error when introducing the data, which could lead to errors

* Corresponding author at: ITAKA, Dept. Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Av.Paisos Catalans 26, 43007 Tarragona, Catalonia, Spain.
E-mail address: jordi.pascual@urv.cat (J. Pascual-Fontanilles).

https://doi.org/10.1016/j.knosys.2024.112154
Received 11 April 2024; Received in revised form 31 May 2024; Accepted 16 June 2024
Available online 26 June 2024
0950-7051/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

when building the models and making predictions. Furthermore, the management of multivariate time series is required since each variable will be transformed into a different time series.

Due to the requirement of most state-of-the-art time series classifiers of having sequences with the same length and same time period intervals, this paper first presents a new method to adjust the time period as well as to interpolate some few missing data points. Patients with too short history are discarded in this step to avoid inferring too much data to avoid generating incorrect series.

As a second contribution, we define a novel fuzzy-based method to compensate for class imbalance, which is common because of the low prevalence of most chronic diseases. The proposed data generation method is applied to very short time series of real patients, boosting them with fictive data to obtain additional minority class examples to be used during the training stage. A fuzzy approach has been used during the generation of new data values, because doctors reason qualitatively on the attribute values when assessing the patients' conditions (e.g., age: child/young/old; body mass: underweight/normal/overweight; hypertension: good control/bad control, etc.). For health treatment, a difference of one year in age, or of one kilogram makes no difference in the diagnosis, as it is done at a more general level (with labels representing more general states). Fuzzy logic is a well-known paradigm to reason qualitatively [3]. In the literature, several fuzzy-based clinical decision support systems can be found. Ahmadi et al. examined the use of fuzzy logic methods in disease diagnosis [4]. They found that fuzzy logic approaches were used to diagnose 38 different diseases such as heart, kidney, thyroid or pulmonary diseases. They found that fuzzy logic had a positive effect in 91% of the analysed cases. Therefore, in this work, we take advantage of the fuzzy linguistic model that represents the domain knowledge of the medical specialists to generate different numerical values for fictive patients, which correspond to the same labels of real patients. We generate synthetic values making use of fuzzy linguistic variables in order to introduce some degree of variability to the new examples, without assigning unrealistic values.

As a third contribution, we address the problem of the classification of the different degrees of severity of Diabetic Retinopathy. Once the multivariate time series dataset is created, we train several standard time series classifiers to study their performance in solving this ordinal classification problem. A comparative study is performed among some different types of time series classifiers to determine how they perform when using the preprocessed series generated from EHR real data. With the presented methods, we achieve high-quality performance, with an accuracy of 93.7% and a quadratic weighted kappa value of 0.868.

Although the task we aim to solve could also be foreseen as forecasting, we did not take that approach for the following reasons. Forecasting techniques would allow predicting the status of the patient at several points in the future, but our main objective is to estimate the current risk level of a patient, given his/her retrospective data, being a classification task rather than forecasting. Moreover, because of the short historic periods available, forecasting techniques may not be appropriate.

The methods defined in this paper have been tested on a real case study, which is the one that motivated this research. The disease studied is named Diabetic Retinopathy, and it is introduced in the next subsection.

1.1. DR risk assessment

This work has been focused on improving Diabetic Retinopathy (DR) risk assessment. DR is an ocular complication produced by the increase in blood sugar levels due to suffering from the diabetes condition. Its progression leads to the eye's blood vessels breaking and may also generate small blood spots, haemorrhages and exudates. These lesions produce vision loss and may even cause blindness if they are not detected and treated at an early stage. In fact, DR is the main cause of low vision and blindness in adults throughout the world. Although the prevalence of blindness worldwide has consistently shown a decrease between 1990 and 2020, it is not the case of diabetic retinopathy, which has increased in the world [5]. The prevalence of DR among type-2 diabetic people in Spain in 2022 was 15.28% [6].

Screening programs exist in many countries to detect DR with the use of non-mydriatic cameras that obtain images of the eye's fundus. Various medical societies recommend an annual screening to appropriately detect DR in its early stages. Despite the involvement of general practitioners, endocrinologists and ophthalmologists, annual screening has proven to be very difficult to carry out. Young adults (18–34) also express many inconveniences for following this yearly screening [7]. Due to costs and personnel availability, patients are usually screened on average every 2–3 years, not annually. Therefore, it is important to develop diagnostic systems that can support diagnosis without the need of eye fundus images, relying instead on the individual's clinical risk factors.

The most common computer-based approach to assess DR is based on the analysis of eye fundus images with computer vision techniques [8–10]. However, these images can only be taken at some of the patient's visits because there are not enough resources to perform this test as frequently as it should be. As deterioration of the eye can be rapid, this image-based screening procedure is not sufficient to prevent vision loss in some cases.

In [11,12], a decision support system, named Retiprogram, was developed to assess Type-2 diabetes mellitus patients' risk of developing DR using only clinical data from the EHR. Retiprogram is used when a patient is visited by medical doctors, and it only takes into account the current patient's conditions (including some analytical data given in the last blood analysis). The goal of Retiprogram is to help doctors determine the DR risk, and so, to determine who needs an eye fundus test in order to focus their use only on critical patients. The Retiprogram system is based on a fuzzy random forest classifier, and it has an accuracy of 80% (with a sensitivity of 80% and a specificity of 84%) [13]. It was first developed as a binary classification system, able to distinguish the negative ($DR = 0$) and positive ($DR = 1$) classes [12]. In a second version, it was extended to deal with ordinal multiclass classification [14]. In that case, the positive class was subdivided into more accurate categories, according to the ETDRS standard classification [15]: mild ($DR = 1$), moderate ($DR = 2$) and severe ($DR = 3$). They are ordered from the best to the worst medical conditions.

The main causes of miss-classifications are the general ambiguity of the problem (very similar patients can belong to different classes), and the high imbalance between classes. More than 80% of diabetic patients do not develop DR, and consequently, the availability of diabetic patient's data with DR is scarce. Because of this over-representation, the classification models have more difficulty to correctly identify and distinguish the positive classes. Furthermore, when a patient is diagnosed of DR, he/she usually starts some treatments in order to improve some clinical factors, therefore, for the patients under treatment it is more challenging to distinguish their DR grade only observing the values of a unique visit. Our hypothesis is that a retrospective analysis could be more adequate to have an overall view of the patient's conditions evolution and could improve the grading of DR, especially for those long-term patients where classical machine learning models have more difficulties.

For this research, we have worked using a dataset with a total of 231,064 Type-2 diabetic patients from Catalonia (Spain), with medical records from 2010 to 2021. We observed that patients are usually visited by doctors every 6 to 24 months, therefore, diagnosis must be made with rather short sequences of data. This will also bring the focus on detecting DR on long-term diabetic patients. This approach could seamlessly integrate into the existing Diabetic Retinopathy detection process at primary health centers, enhancing the DR detection for long-term diabetic patients and resulting in cost savings for the health center. Furthermore, it would improve patients' quality of life by minimizing unnecessary eye screening tests.

1.2. Contributions

This paper defines appropriate methods for solving these particular problems of EHR medical series data obtained in the regular visits of chronic patients. Contributions are the following

- A binning technique to merge data in the same interval in order to adjust the time periods.
- A new method to interpolate missing data points by double iteration for series extension.
- A novel fuzzy-based method to compensate for class imbalance based on the linguistic vocabulary of the domain of medical professionals.
- Validation of the pre-processing techniques with the use of several standard time series classifiers to distinguish the different degrees of severity of Diabetic Retinopathy.

While this paper specifically focuses on the case Diabetic Retinopathy detection, the methods proposed extend beyond this domain. Any clinical scenario that involves the use of EHR data for assessing the risk of disease development may potentially benefit from the proposed algorithms. The methods have been defined in general terms so that they can be easily used in any other ordinal multi-classification problem with short and irregular sequences of data.

The rest of the paper is organized as follows. Section 2 presents the time series classification problem, and briefly reviews how DR classification and the imbalance problem have been studied in the literature. In Section 3, we introduce the proposed method for pre-processing the EHR data into time series data. The short time series problem is also presented and undertaken in this section. Next, Section 4 presents the proposed approach for boosting short time series to compensate for class imbalance. Section 5 presents some state-of-the-art multivariate time series classifiers. In Section 6, the time series classifiers are compared and its obtained results are discussed. Finally, Section 7 presents the conclusions and the future work.

2. Related work

The Time Series Classification (TSC) problem is attracting a lot of interest since every day more time series data is being produced. Technology advances facilitate the collection and storing of data values along the time. TSC can be used in many different areas and has a broad range of possible applications. One important field are biomedical and health-care applications, such as the case of diabetic retinopathy classification.

Wang et al. [16] reviewed TSC in the field of biomedical applications from 2016 to 2021. According to their study, the two main types of temporal data being used are electroencephalogram (EEG) and electrocardiogram (ECG), which are both signal series. The most common pre-processing method for EEG and ECG is filtering, which is used to remove artefacts and noise. Once the signal has been processed, some features are extracted, and the best ones are fed into a machine learning classifier. In this survey study, temporal data from EHR is in the fifth position, with half the articles than the former types. EHR might contain any kind of medical information, sometimes also including signals such as EEG or ECG. Authors also identified that the scarcity of data (small dataset) is a common problem in biomedical applications. In this domain, a small dataset is considered when the amount of patients is low (e.g. less than 20). However, in many of those cases, the low number of patients is compensated with the availability of long data sequences, so the classifiers performance is good because of the use of signal processing and feature engineering techniques on top of a classifier.

Pasos et al. [17] reviewed how several algorithms performed on multivariate time series classification (MTSC) problems. Dynamic Time Warping (DTW) was used as a baseline classifier, as it is still competitive in comparison to more recent proposed alternatives. According to their experiments, the ROCKET classifier [18] was the best performing overall in many applications. However, there is no classifier that outperforms the rest in all domains. Their suggestion is that ROCKET and DTW are good enough classifiers to use as an initial technique.

Focusing now on DR diagnosis systems, this problem is commonly approached as an image analysis problem, consisting on classifying eye fundus images. Dubey and Dixit [9] and Atwany et al. [8] reviewed the recent developments on these kinds of systems. Even DR risk classification task is the most frequent, some studies also perform segmentation tasks, such as the identification of some eye structures relevant for DR (f.i. optical disc, optical nerve or blood vessels), or the location of DR lesions (f.i. microaneurisms or exhudates). In the majority of cases, machine learning or deep learning methods are used for the image analysis. For instance, in [10] they applied transfer learning to a Convolutional Neural Network to improve the grading of DR using fundus images.

Instead of using eye fundus images, in this paper we want to use data collected in the Electronic Health Record (EHR) by doctors during the visits and tests to diabetic patients. Because obtaining eye fundus images with non-mydriatic cameras is costly and time-consuming, EHR based clinical decision support systems are nowadays starting to be considered. In the literature, some approaches using EHR information can be found. Sun and Zhang used the first hospitalization EHR data of diabetic patients to create a DR dataset. They filtered the available variables to preserve the relevant medical ones for the DR. Some of the most sensitive variables they found were unsaturated iron binding capacity, bilirubin, and glycosilated serum protein. They compared how several machine learning binary classifiers performed whether feature engineering is applied or not to the dataset [19]. They obtained the best results when applying feature engineering using a random forest classifier. Some other studies in the literature analyse how different kinds of classifiers perform on EHR-based DR datasets [20–23]. They use techniques such as random forests, XGBoost, logistic regression, support vector machines or k-nearest neighbours. There is not a consensus in which is the best classifier for the diabetic retinopathy disease, as there are different results on each study. Moreover, up to our best knowledge, no one employs temporal datasets.

Temporal data has been used in other health-care problems. However, the particular characteristics of the available EHR data lead to time series that do not have the common structure found in other time series data. Temporal data usually consists of a long sequence of equally-spaced signals, whereas we have series of EHR data coming from visits made at different time intervals (for each patient and for different patients). Some works in the literature already approached the problems of making predictions using the EHR data of patients as temporal data. Itzhak et al. predict acute hypertensive episodes by measuring four vital signs on patients in intensive care units (ICU) [24]. They use temporal abstraction and mine time-intervals-related patterns to extract features for a classifier. Sheikhalishahi et al. propose a novel ante-hoc interpretable neural network to provide a prediction of the onset of delirium to prioritize critically ill-patients, which is common in the ICU [25]. Regarding DR, Rabhi et al. modelled the evolution of HbA1c as an irregular variable-length sequence, and tested several deep learning methods to predict whether Type-1 diabetic patients have DR using this single sequential variable [26].

In contrast to patients in constant monitoring at ICUs, a Type-2 diabetic patient is typically visited with an average frequency of once a year. Collecting a sufficient series of data needs at least 5–6 years, after the diagnosis of diabetes. Therefore, the classifier system we want to build must work with short sequences. In time series related works, small time series datasets have usually a low number of patients with long signals series [16]. In this work, the situation is the opposite, with a considerable number of patients, but the length of their time series is short. Moreover, we need to include several variables in the classifier, which according to specialists are required to properly diagnose this illness, and to differentiate among the several risk levels

of DR (i.e. multi-class). In Section 3, a data pre-processing procedure for this kind of sort EHR time series is proposed.

Another problem we need to face is the inherent class imbalance on DR. Most of the diabetic people will not suffer from DR, as the world prevalence was estimated to be about 22.27% in 2020 [27]. Patients with progression to the worst classes are also a minority in comparison with the ones that have a mild degree. Most state-of-the-art time series classifiers are not suited to solve problems with imbalanced class distributions, thus, methods to balance the class distribution at the data level are commonly used. Several approaches can be used to compensate time series for the imbalance on the minority classes:

1. **Sampling methods:** some techniques are applied to the data on the original dataset to oversample and/or undersample it. For instance, in random oversampling, examples of the positive (minority) classes are replicated to balance the class distribution. On the contrary, undersampling consists on randomly removing examples from the majority class.
2. **Synthetic data generation:** the examples introduced to compensate the class imbalance are artificially generated from the existing data. The most common method is SMOTE (Synthetic Minority Over-Sampling Technique) [28]. Synthetic data points are generated by taking one of the k-nearest neighbours of a sample, and randomly choosing one point of the vector that unites the sample and the selected nearest neighbour. On the literature, several variations or methods based on the methodology of SMOTE can be found. For instance, T-SMOTE [29] is a variation for time series which takes into account the temporal characteristics of the data to select the nearest neighbours. T-SMOTE can be used on both univariate and multivariate time series.
3. **Data augmentation:** slightly modified copies of the data or synthetic examples created from the existing data are introduced to compensate for the class imbalance. Methods are highly dependent on the data types that have to be augmented. In the time series case, Iwana and Uchida [30] analysed over 50 data augmentation methods for time series, and they proposed a taxonomy with 4 families of methods: Random transformation methods apply a transformation function with some randomness to the time series; Pattern mixing combines patterns to generate new ones, which overcomes the assumption that all random transformations are possible on the data; Generative models use either statistical or neural network models to sample time series from feature distributions; Time series decomposition uses feature extraction techniques to extract features or underlying patterns, which are then used to generate new examples.

In medical diagnosis, the patients' values of the different risk factors are not totally independent. Although doctors know that there are some underlying relations, they are not usually completely defined. For instance, doctors may know that some combinations of values are not possible. Consequently, it is important that the balancing method used does not generate examples that may not be not real, as this may hamper the quality of the classifier built. This paper proposes a new method that combines both synthetic data generation and random transformation data augmentation. On the one hand, short series are extended by synthetically generating the missing data. On the other hand, we are also conditioning the generated data to be similar to other existing examples (i.e. to a real patient), which is something that cannot be assured when using interpolation without introducing further pre-processing. In Section 4, the proposed method is explained.

3. Time series pre-processing for EHR data

Most state-of-the-art classification techniques require time series of equal length and with regular time intervals. That is, the difference between each two consecutive time points is always the same. When data comes from sensors, this requirement can be easily satisfied. On the contrary, when collecting data from patient's visits from the EHR, we face the problem of different time spans between consecutive visits, as it may depend on the patients' health state or the availability of doctors in certain periods. This is the case of diabetic retinopathy monitoring.

The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Ethics Committee CEIM (Comite de etica en investigaciones medicas del Insitut de Investigacions Sanitaries Pere Virgili (IISPV)) of Tarrragona, Spain, approval number Ref. CEIM: 028/2018.

3.1. Diabetic retinopathy series data

At each visit, ophthalmologists collect some clinical and analytical data about the diabetic patient, which are stored in his/her EHR. For the construction of Retiprogram, six relevant risk factors for DR diagnosis were selected by experts, they consist of six numerical and three categorical variables [11]. Numerical variables include the current age, body mass index (BMI), duration of Type-2 diabetes (EVOL), HbA1c, CKDEPI and microalbuminuria (MA). Categorical variables include gender, treatment of Type-2 diabetes (TTM) and control of arterial hypertension (HTAR). With this information and with the analysis of the eye fundus image, the ophthalmologist determines the degree of DR.

After some years, each patient has a sequence of values for that variables, which are stored in his/her EHR, together with the DR diagnosis value assigned by the ophthalmologist in each visit ($DR = \{0, 1, 2, 3\}$). The DR diagnosis is included as a categorical variable in the training dataset, except for the last entry of the sequence, since this is the value the system must predict. This way, we can use the previous DR evolution to train the time series classifier. Moreover, we can use the last DR value as ground truth to validate the output value.

From the dataset of series collected, we discard the ones with a single visit. The frequency of visits in months is deployed in Fig. 1. Its can be seen that the frequency of consecutive visits is not homogeneous. We observe also that most patients have a visit frequency of 18 months or higher, needing many years to collect a sufficiently long series of data.

After confirming the complexity of the Catalan diabetic population dataset, in terms of both short length and irregular frequency, an appropriate pre-processing procedure is proposed in this paper. It is explained in the following subsections.

3.2. Data binning

In order to determine the length of the patients' time series in the DR dataset, a study was conducted on the number of visits per patient. The counts are shown in Table 1. As it can be observed, the majority of patients have a short number of visits. High-risk patients should be visited at most annually (some of them even in 6 months or less, Fig. 1), and low risk patients can be visited with a lower frequency (18 to 30 months).

According to the knowledge of medical experts in DR, series must have intervals of 1 year. Thus, for patients with at least two visits, a binning has been applied with 1-year bins. In the cases where multiple visits have occurred in the same year, they are aggregated in a single visit that represents the status of the patient that year.

Let $P = (p_1, p_2, \dots, p_n)$ be the set of patients with at least two visits in their EHR, and let $V = (v_1, v_2, \dots, v_m)$ be the set of relevant variables for DR diagnosis. Each patient $p \in P$ has a vector of data for each variable $v \in V$, denoted p_v. The length of this vector ranges from $[1, t_p]$, where t_p is the number of visits of patient p.

We define a set of bins B representing natural years, for example $B = \{2000, 2001, \dots\}$. Let us assume a function $year(p_{v,i})$ that returns the year in which the $p_{v,i}$ value was collected.

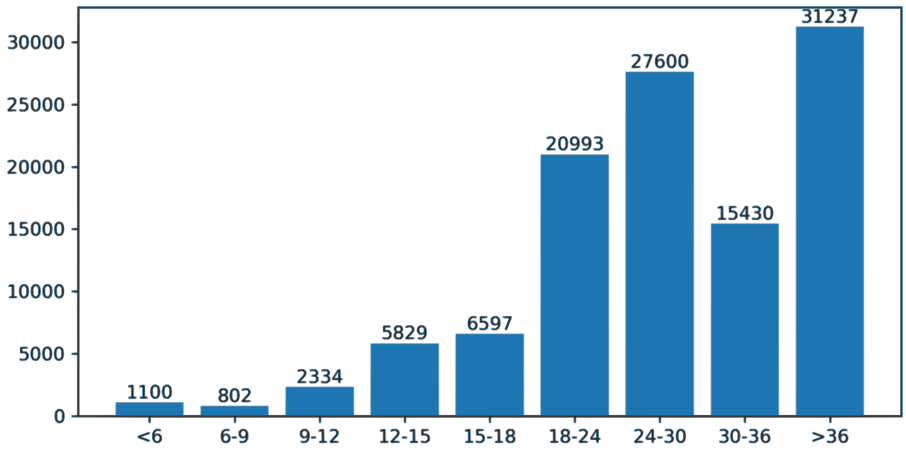


Fig. 1. Mean frequency in months of patients visits to the ophthalmologists.

Table 1
Number of visits per patient.

Number of visits	Number of patients	Percentage (%)
1	119 142	51.56
2	61 465	26.6
3	30 257	13.09
4	12 961	5.61
5	4893	2.12
6	1564	0.68
7	551	0.24
8	171	0.074
9	47	0.02
10	11	0.0048
11	1	0.00043
12	1	0.00043

Next, we define $\Omega_{v,b} = \{p_{v,i} | year(p_{v,i}) \in b\}$ as a subset of all data obtained in the same year b for the variable v, or an empty set $\varnothing$ if no data was collected that year.

Then, we can define the aggregation function h for all the values of a variable v in the subset of data of a certain bin b. No aggregation is performed if no data was obtained in that bin.

The aggregation function should use domain knowledge, as Eq. (1). In the DR case, categorical variables (TTM, HTAR, DR) take the maximum value, keeping track of the worst health status. Numerical variables have different aggregations: Age and EVOL use the most up-to-date value for that year, MA uses the maximum (worst) value in the period, while HbA1c, CKDEPI, BMI are aggregated using the mean.

$$h(\Omega_{vb}) = \begin{cases} \text{mean}(\Omega_{vb}) & \text{if } v \in \{\text{HbA1c, CKDEPI, BMI}\} \\ \text{last\_value}(\Omega_{vb}) & \text{if } v \in \{\text{Age, EVOL}\} \\ \max(\Omega_{vb}) & \text{otherwise} \end{cases} \tag{1}$$

The computational complexity of our proposed method is $O(n)$, where n is the number of patients. Because the number of bins is small in this kind of medical series data, and is also constant ($k = 12$), it does not contribute to the overall complexity calculation. Moreover, since each patient's vector of visits can be processed independently, the binning can be easily parallelized. As a result, the method can be easily scaled to handle large datasets without significant increases in computation time.

In our dataset, after binning the data, we have 228,956 patients with sequences of less than 6 entries (i.e. data from at least 6 different years). A total of 2108 patients have sequences between 6 and 12 years, whose frequency is shown in Fig. 2. In the next subsection, the transformation process to obtain equal length time series is explained.

3.3. Time series transformation

As indicated before, most TSC require that all the time series of a dataset are of the same length. Observing the DR binned data distribution, Fig. 2, we selected the length of the time series to be of 10 years. According to the experts, this is a reasonable amount of historic data to be used to perform the DR risk assessment. In this subsection, we propose a procedure to obtain series of 10 years from the DR binned data.

We denote each patient time series obtained after binning the EHR data as $T_p = \left(t_1, t_2, \ldots, t_{l_p}\right)$, where l_p is the length of the time series. Patients with $l_p < l_{short}$ (i.e., data less than threshold l_{short}) are excluded from the dataset. This application-dependent threshold has to be considered as the minimum amount of past information needed to perform the transformation without inferring too much data, which might lead to incorrect data.

The transformation procedure aims to generate sequences with length equal to l_c, called complete series. We may consider three cases based on the length l_p of the time series (Eq. (2)).

$$T_p = \begin{cases} \text{truncate}(T_p, l_c) & \text{if } l_p > l_c \\ T_p & \text{if } l_p = l_c \\ \text{interpolate}(T_p, l_c) & \text{if } l_p < l_c \end{cases} \tag{2}$$

For the DR problem, we set $l_{short} = 6$ and $l_c = 10$, as recommended by ophthalmologists. In the third case of Eq. (2), as we have sequences of $6 \leq l_p < 10$, the maximum number of fictive entries is 4.

To generate new entries in the series, one option is to use a linear interpolation, where existing values are used to approximate some function $y = f(x)$, which is then used to find the values of the missing points between the extreme values. However, for this medical problem, this method is too simple. Instead, we propose to use a double interpolation approach, which takes into account the specific characteristics of the patients' variables. It is composed of the following steps:

1. **Data initialization:** each binned data entry T_i is assigned to a year. The years without available data are considered missing data. As example, the first column on Fig. 3 shows the 6 binned data values for a patient visited between 2012 and 2019 (8 years).
2. **First interpolation:** missing time-points on the initial data are filled according to each variable. Numerical variables are interpolated, taking into account the length of intervals to be filled. The meaning and mathematical properties of the variables are taken into account, for example, age and EVOL must be monotonic non-decreasing, as they cannot decrease from one year to the next. For categorical variables, a backfill interpolation is

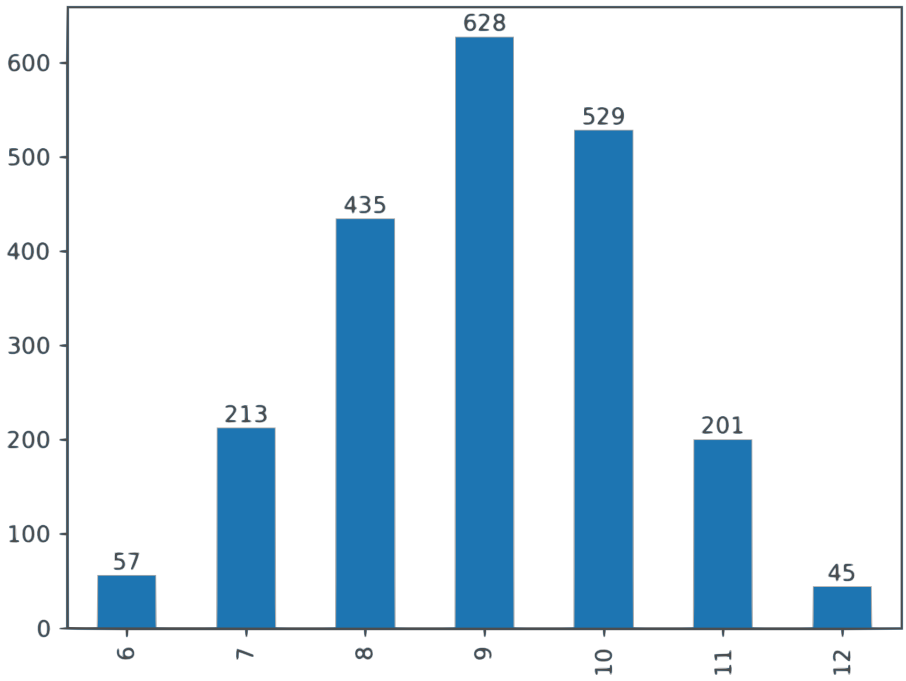


Fig. 2. Frequency of the length of binned time series.

performed. It uses the next existing value, so we use the latest available category. An example is shown in the second column of Fig. 3, where two entries have been added in 2013 and 2017 for all variables.

3. **Second interpolation:** in the previous step, a complete time series has been obtained, but it must still not have the minimum required length. In this step, it will be linearly interpolated to the desired length, which is 10 for the DR case study. Additionally, for categorical variables, decimal values obtained at interpolation are rounded to avoid nonexistent categories. Because of the quality of first interpolation step, we are not expecting the rounding to result in a significant change in the category assigned. Finally, time stamps are also replaced with ordered integer values $[0, 10)$, since it is not important the specific year, but their order. The third column in Fig. 3 shows the final output of this step for this patient example.

The computational complexity of our proposed method is $O(n \cdot l_n)$, where n is the number of patients and l_n the length of each patient's time series. The average length of these time series ($\approx$10) can be considered a small constant value. As such, its effect on the overall complexity is negligible. Similar to the binning method, our approach allows for efficient parallelization by processing patients' vectors independently. Even with large datasets, this complexity remains comparable to that of a basic linear interpolation. The proposed time series interpolation method has been compared to a basic linear interpolation applied the initial sequence of data. The comparison has been evaluated with the widely-used Dynamic Time Warping (DTW) as distance measure between the sequences obtained with the two methods. The dependent version of DTW, DTW_D, has been selected based on the study made by Pasos et al. [17]. Results are shown as histograms in Fig. 4.

It is clear from the distribution represented in the two histograms that the proposed interpolation method leads to equal-length time series that are more similar to the original short sequence than using a single linear interpolation. For the double-interpolation method, the distance is below 4 in most of the series, with a few exceptions greater than 6, while for the linear interpolation, most of the sequences have a distance between 4 and 6, with almost no cases of very similar sequences (those at distance below 3).

After the interpolation stage, the last transformation applied to the series consists on a process of data standardization and encoding. Categorical variables have been encoded using one-hot encoding (OHE), whereas the numerical ones have been standardized by subtracting the mean and scaling to unit variance. This encoding is a requirement for the classifiers that we have used in this study.

4. Time series generation

After applying the previous data preparation steps, a multivariate multiclass time series dataset with 2108 patients is obtained. This quantity is quite reduced for training a classifier. It is due to the fact that many patients data were discarded after the binning stage, specifically, the ones with sequences of short length, as the interpolation method could not find appropriate values. In this section, we propose to take advantage of this patient's data to generate partially-synthetic instances for the minority classes.

The proposed method for generation of examples distinguishes two types of data sets, defined as follows:

- C_p are sets of complete time series for the minority classes, with length l_c. In the case of DR, it consists of 3 sets, one for each of the DR positive categories, $p \in \{1, 2, 3\}$.
- I is the set with incomplete time series, i_j, each one with a short length l_j. The length must be in the range $l_{min} \leq l_j < l_c$, where l_{min} must be determined by the characteristics of the data, in case of DR, $l_{min} = 5$.

The generation of new examples of length l_c will be done by means of extending (i.e. boosting) the information available in existing short series in I. For each minority class p, the additional entries added at the end of the existing sequence will take into account the information available in the set C_p. In that way, we introduce data values they are feasible, as some other patient has had similar values. In the following subsections, the method to boost the incomplete set I using the complete set C_p is explained. A different treatment is done depending on the nature of the variables. The method is applied to all examples of the incomplete set, $i_j \in I$, for each of the minority classes $p \in \{1, 2, 3\}$. Examples in I do not have a ground truth value about the DR diagnostic, hence, they can be completed using examples from different classes, generating different sequences for each class.

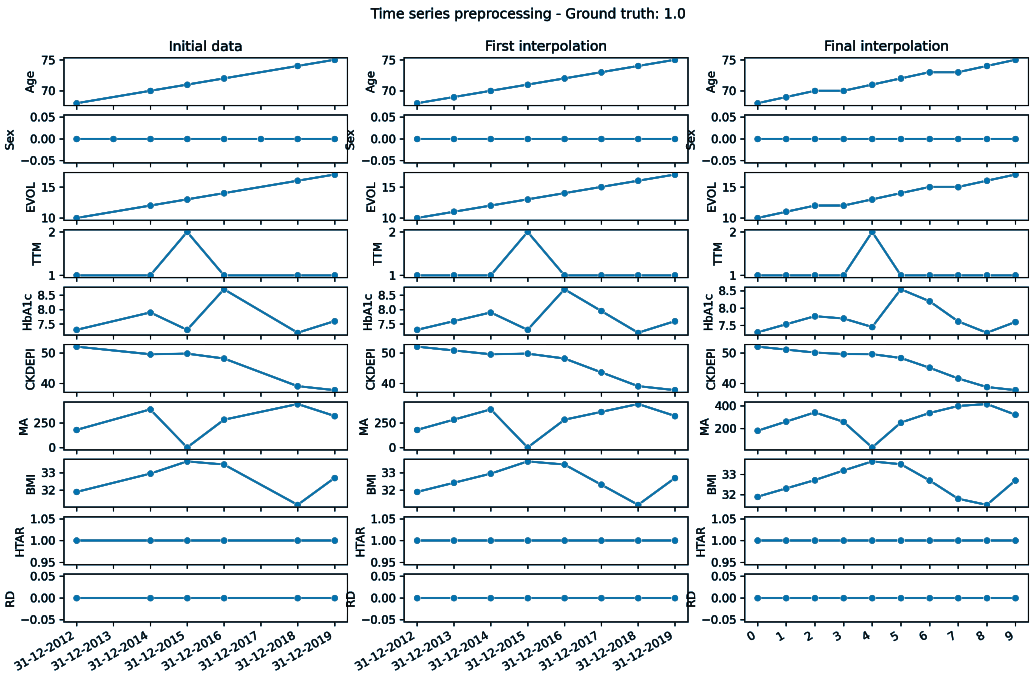


Fig. 3. Double interpolation example for one patient.

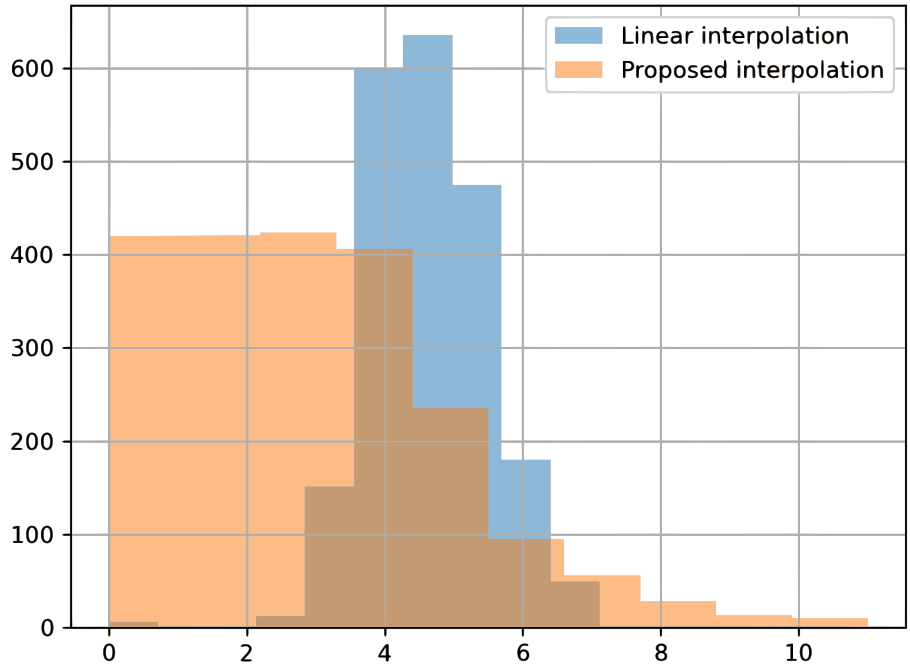


Fig. 4. Histogram of DTW distances, comparing a linear interpolation with the proposed double interpolation.

4.1. Demographic variables

First, we consider the demographic variables, whose progression is known in advance. In the DR case study, they are age, gender and EVOL (duration of diabetes). Age and EVOL are numerical, and they are measured in years, so at each time point in the series (yearly intervals), they increase in 1 unit. Gender is a categorical variable with a value fixed along the time, so the same category (woman/man) is maintained equal in all the new entries for each given time series i_j.

4.2. Medical variables

These are variables that store clinical and analytical information related to health. For the medical variables, we calculate the distance between a given incomplete series $i_j \in I$ (with length l_j) and a complete series $c_{p,k} \in C_p$ (with length l_c). As $l_j < l_c$, for the complete series we only consider the first l_j entries for the distance calculation. Dynamic Time Warping (DTW) has been used as the distance measure for comparing the sequences.

This comparison is performed for each of the minority classes p. For each class, we find the example from the complete set with the minimum distance to i_j, i.e. the most similar in class p, denoted c_{p,sim_j}.

$$c_{p,sim_j} = argmin_k(DTW(i_j, c_{p,k}))\ \forall c_{p,k} \in C_p \quad (3)$$

Once we know the most similar complete series to an incomplete one, the procedure for assigning the following missing values of the sequence depends on being a numerical or categorical variable.

4.2.1. Categorical variables

We have three categorical variables: TTM, HTAR, and the class label DR. Their values in the incomplete entries of i_j are completed using the categorical values of the most similar series, c_{p,sim_j}. This corresponds to the missing time points $t \in (l_j, l_c]$. Regarding the class variable DR, which must be monotonic non-decreasing according to the medical specialists, a forward fill is applied in the time points where copying the value would produce a decrease from the previous DR level. This process mainly affects the first generated time points, where some discrepancies between the incomplete and complete sequences could be found.

4.2.2. Numerical variables

For the management of numerical values, we propose a procedure based on fuzzy sets. Doctors usually work with ranges of values with fuzzy boundaries rather than with precise numerical values. We consider that for each numerical variable $a \in A$, we can define a linguistic fuzzy variable f_a with a fixed set of ordered labels. Each label has a fuzzy set, with its corresponding membership function $\mu_{x \in f_a}$. In our case study, ophthalmologists provided appropriate linguistic labels and fuzzy sets for the numerical variables $A = \{CKDEPI, HbA1c, MA, BMI\}$. For each variable ($a \in A$) and for all missing time points $t \in (l_j, l_c]$, the following procedure is proposed:

1. A forecasting method is used to predict the next numerical value for the incomplete time series, $i_j(a,t)$. Drift forecasting has been chosen because of its simplicity. Moreover, the amount of available past data is limited, so more complex forecasting techniques were not needed. It fits a line between the first and last points of the series, and extrapolates them to the future.
2. The value of the same time point is obtained from the nearest complete time series, $c_{p,sim_j}(a,t)$.
3. The fuzzy sets of the variable f_a are then used to obtain the label with maximum activation for both the incomplete and complete time series values, x and y, respectively. If $x = y$, the forecasted value is stored in $i_j(a,t)$. Otherwise, a random value with maximum activation on the fuzzy term y is the one assigned to $i_j(a,t)$.

By forcing the forecasted value to be similar to one in the complete sequence, we can generate new examples that, although not having the same values, are similar. The use of fuzzy sets permits to assign values that are fuzzified with the same label, which means that they are falling in the same category according to the vocabulary given by the ophthalmologists.

The overall computational complexity of our fuzzy-based sample generation method is $O(n \cdot m)$, where $n = |C_p|$ (i.e., the number of complete series of minority classes) and $m = |I|$ (i.e., the number of incomplete series that need to be extended). While the upfront computation time may seem significant due to this complexity, it can be effectively mitigated by parallelizing the process.

In Section 6.3 random oversampling is compared to this new method for the DR dataset.

5. Multivariate multiclass time series classifiers

In this section, we introduce the four multivariate multiclass time series classifiers we have tested. They are the best performing classifiers according to the review of Pasos et al. [17] Moreover, they are representatives of different kinds of approaches to classification.

5.1. K-nearest neighbours

K-nearest neighbours (KNN) is one of the most simple, yet good performing methods in classification problems. In MTSC problems, KNN is usually taken as the baseline results by using 1-nearest neighbour (i.e. $K = 1$). As it is a distance-based classifier, the selection of the distance measure is crucial. In this work, for time series comparison, Dynamic Time Warping is used. Several strategies can be used to compute a DTW distance on the multivariate case:

1. Independent warping (DTW_I): the distance between time series is computed independently for each variable. The resulting DTW distance is the sum of the independent distances.
2. Dependent warping (DTW_D): a warping is assumed to be correct across all time series. The Euclidean distance is computed between the two vectors representing all time series. Then DTW is applied over all time series simultaneously.
3. Adaptive warping (DTW_A): instead of selecting one of the previous two approaches, the choice of using DTW_I or DTW_D is made based on the characteristics of the data.

5.2. ROCKET

ROCKET (Random Convolutional Kernel Transform) combines numerous convolutional kernel transforms with a linear classifier [18]. Kernels are randomly chosen with a variety of lengths, dilations, paddings, weights and biases. For the multivariate case, they are also randomly assigned a time series. A feature map is created by convolving a kernel, and it is aggregated to produce two features per kernel, which are the maximum value and the proportion of positive values (ppv). A linear classifier, such as ridge regression classifier, is then trained on the extracted features.

5.3. TapNet

TapNet (Time series attentional prototype network) [31] defines an architecture to combine the strengths of both traditional and deep learning approaches to MTSC. It consists of three main components. First, a dimension permutation, which randomly combines the time series, in order to model the interactions between the different variables. Second, embeddings are learned using a Long Short-Term Memory structure and 1-dimensional Convolutional Neural Network, in order to model the sequential information of the time series. Finally, Attentional Prototype Learning is applied using the learnt embeddings. It generates an embedding prototype of each class, and the classification is then based on the distance to these prototypes.

5.4. Convolutional neural networks

Convolutional Neural Networks (CNN) are well known neural networks in deep learning. They have been mainly used in image analysis problems, although Zhao et al. proposed a CNN for time series classification [32]. They alternate the usage of convolutional and pooling operations to obtain deep features of the data. The final representation of the data is in a feature layer, which is created by connecting all the obtained feature maps. Finally, the feature layer is used to perform the classification using a multi-layer perceptron model.

Table 2
Diabetic retinopathy time series data.

Class/Dataset	Training (70%)	Testing (30%)	Total
DR = 0	1212 (82.2%)	518 (81.8%)	1730 (82.1%)
DR = 1	148 (10%)	61 (9.64%)	209 (9.9%)
DR = 2	92 (6.2%)	41 (6.5%)	133 (6.3%)
DR = 3	23 (1.6%)	13 (2.1%)	36 (1.7%)
Total	1475	633	2108

6. Experimental results

This section presents the obtained experimental results. In Section 6.1 the diabetic retinopathy dataset is presented. Section 6.2 compares the classification performance of non-temporal classifiers on long-term diabetic patients. Finally, Section 6.3 discusses the obtained results on the temporal datasets, and a comparison between the tested classifiers is performed.

6.1. Dataset

The data used in this study comes from diabetic population on our region, Catalonia, from period 2010 to 2021. It is a private dataset provided to us in the framework of a national research project. Even it is an anonymized dataset, it contains sensible healthcare data from patients. Therefore, it is protected with the national privacy laws, and it cannot be disclosed. After performing the pre-processing steps explained in Section 3, we obtained a set with 2108 sequences of 10 entries. This dataset has been divided into two: training and testing, with 70% and 30% of the data, respectively (Table 2). It can be clearly seen the high imbalance towards the negative ($DR = 0$) class, which makes it more difficult to predict the classes with higher DR risk.

When balancing is needed, we either applied oversampling or the method proposed in Section 4 for completing short sequences for the three minority classes. In the latter, we use the incomplete set I. It is composed by previously discarded time series because of their short length. In this work, we just considered the incomplete series of the maximum length available, $l_j = 5$, which correspond to 4547 patients. This is because the shorter the incomplete time series are, more fictive data has to be introduced, increasing the probabilities of introducing erroneous data.

6.2. Long-term DR patients classification

The first study consists in comparing the performance of the Retiprogram model [11,12,14] on patients suffering from long-term diabetic retinopathy. Retiprogram does not take into account the history and evolution of the patient since the first diagnosis of DR. On the contrary, it uses a single point state of the patient to diagnose the risk of DR. We want to prove that such model is good for the initial diagnosis of patients who are starting DR (called new DR patients), but not for patients with an advanced progression of DR (called long-term DR patients).

The tests consists in comparing how Retiprogram classifies patients the first time they are diagnosed with diabetes, and how it classifies them some years later. To perform this comparison, we have created two non-temporal DR datasets. One for new patients, which contains the first entry of the temporal series; another for long-term patients, which contains the latest entry, instead.

All the available data of the 2108 patients with complete series has been used, and Retiprogram was tested on both datasets. As the DR class in the ground truth is not the same in both datasets, the number of patients in each class is different in each dataset. The obtained confusion matrices for both tests are depicted in Fig. 5, and their corresponding metrics in Table 3.

Table 3
Performance indicators of new patients and long-term patients.

Classifier/Metric (%)	New patients	Long-term patients
Accuracy	75.6	35
Kappa	0.095	0.068
Precision	44.9	30.6
Weighted precision	69.4	35
Macro recall	28.2	27.8
Weighted recall	75.6	35
Macro F1	27	20.4
Weighted F1	68.6	24

Table 4
Classifiers parameters.

Classifier	Parameters
KNN oversample	k: 9, distance: DTW_D
KNN Fuzzy Gen.	k: 3, distance: DTW_D
ROCKET	Number of kernels: 200
TapNet	Epochs: 20, batch: 16, activation: softmax, loss: categorical crossentropy, filter: (32, 32, 16), kernel: (8, 5, 3), layers: (50, 30)
CNN	Epochs: 20, batch: 16, activation: softmax, loss: categorical crossentropy, kernel: 5, avg pool size: 2

All metrics about the classification of new patients are better than the ones for long-term patients. Having in mind that the ground truth is different in each dataset, it is still clear that the model has more difficulties to correctly classify the long-term patients. Most errors on long-term patients are due to mis-classifications for patients that belong to $DR = 0$, as shown on its confusion matrix in Fig. 5. Once a diabetic patient has been diagnosed with this disease, he/she starts taking some medications in order to have under control some dangerous values. These changes generates a confusion to Retiprogram when evaluating the current patient's state, making it really hard for the model to predict the current status of DR. These results confirm the need of creating a new model that uses all the temporal information available for each patient, thus, taking all states into account to perform a good prediction of the DR degree for those long-term patients.

6.3. Results and discussion

We tested the performance of the multivariate time series preparation and classification method explained in Section 5 on the DR dataset presented in Section 6.1. To perform the tests, we have used the *sktime* toolkit [33]. Tests have been performed both oversampling the data, and using the proposed time series fuzzy generation method. Some of the default parameters of the classifiers have been used, but others have been adjusted, such as the ones related to length of the series. Moreover, the epochs, activation and loss functions of the tested neural networks have also been experimentally tunned to fit with our ordinal multiclass problem. In the case of KNN, we found that oversampling requires a higher k value than the fuzzy generation method to avoid overfitting. The final configuration values for these parameters are given in Table 4.

A 10-fold cross-validation has been chosen to validate the performance of the parameters, choosing the values with the best performance. The whole training set shown in Table 2 has been used to perform the validation. First, the data is split among 10 folds. One of them is selected as the test set. The rest is combined to form a training set for that fold. The training data is then balanced either by means of oversampling or by employing our proposed fuzzy sample generation method. Finally, the model is trained using the balanced training data and then tested using the test fold. The process is repeated using all folds as test data, and the obtained performance metrics in all folds are averaged. Those results are shown in Table 5. Columns contain each of

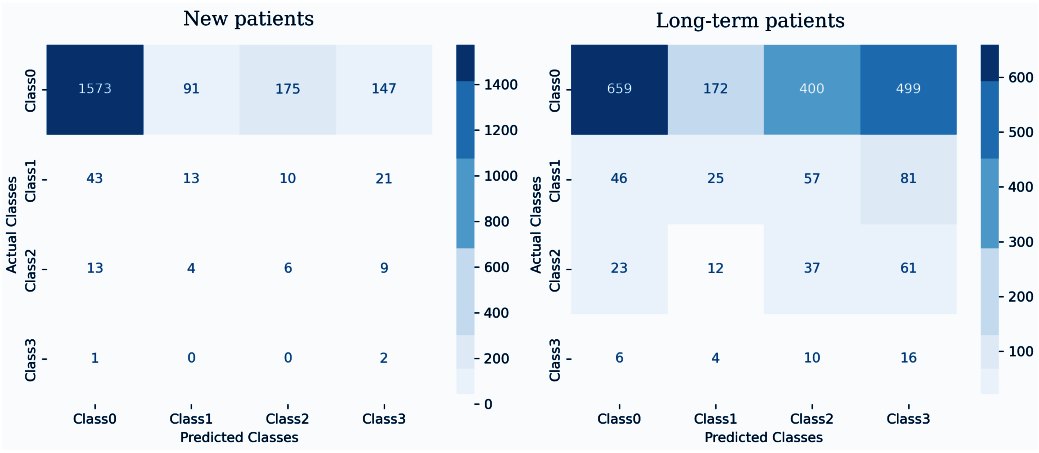


Fig. 5. Comparison between new patients (left) and long-term patients (right).

Table 5
10-fold cross-validation performance indicators of different DR series classifiers.

	CNN		KNN		ROCKET		TapNet	
	FG	O	FG	O	FG	O	FG	O
Accuracy	94	90.6	91.2	89.6	91.1	93.2	**94.1**	86.4
M. Precision	80.4	74.7	71.3	68.8	86.7	**91.1**	82.8	76.7
M. Recall	75.9	83.1	60.5	61.1	64.9	75.4	75.4	**90**
M. F1	77.1	77.4	63.8	63	71.3	**80.3**	77.1	79.9
W. Precision	93.7	92.4	90.9	89.6	90.7	93.3	**93.9**	93.3
W. Recall	94	90.6	91.2	89.6	91.1	93.2	**94.1**	86.4
W. F1	**93.6**	91.3	90.6	89.1	89.7	92.7	**93.6**	88.1
Kappa	0.856	0.764	0.640	0.670	0.736	0.805	**0.868**	0.738

Table 6
Performance indicators of different DR series classifiers in testing stage.

	CNN		KNN		ROCKET		TapNet	
	FG	O	FG	O	FG	O	FG	O
Accuracy	91.63	81.36	85.15	84.04	90.68	89.26	**93.68**	91.94
M. Precision	78.83	52.17	55.81	70.63	**89.71**	84.96	84.26	77.41
M. Recall	76.49	63.87	50.15	45.97	64.42	66.22	**79.39**	70.25
M. F1	76.89	56.11	51.91	48.32	73.47	73.36	**81.17**	72.58
W. Precision	91.78	86.14	84.28	83.53	90.20	88.60	**93.41**	91.55
W. Recall	91.63	81.36	85.15	84.04	90.68	89.26	**93.68**	91.94
W. F1	91.52	83.21	84.42	82.92	89.43	88.38	**93.36**	91.25
Kappa	0.840	0.650	0.551	0.623	0.734	0.729	**0.868**	0.856

the tested classifiers, either balancing the data using oversampling (O), or the proposed fuzzy sample generation (FG). Several standard multi-classification metrics have been used to measure the performance of the different classifiers. In particular, we consider accuracy, precision, recall, F1-Score and the quadratic weighted kappa. For precision, recall and F1-Score we have taken both macro and weighted average. Because of the class imbalance, we are expecting the macro average to be lower than the weighted average. Quadratic weighted kappa is a relevant metric for this ordinal multiclass problem, because it penalizes the mistakes according to the distance between the ground truth and the predicted class. In medical decision support, a short difference between the correct class and the predicted one is crucial in order to not affect the health of the patient. Hence, we aim to minimize it as much as possible. Following the kappa interpretation of Landis and Koch [34], a kappa in the interval $[0.61 - 0.8]$ describes a substantial strength of agreement between the predictions and the ground truth. Values greater than 0.8 indicate an almost perfect agreement, which would be a strong indicator of a good medical decision support system.

Comparing the results obtained by using oversampling to fuzzy sample generation, overall, fuzzy sample generation (FG) obtains better accuracy, kappa, and weighted metrics. In contrast, oversampling has better or similar results on the macro recall and macro F1-Score. The only exception is the ROCKET classifier, where oversampling obtains better results in all cases. The best result for each of the metrics is marked in bold. TapNet gets the best overall results, with the best accuracy, weighted metrics and quadratic weighted kappa. CNN and ROCKET also have good overall results. CNN draws with TapNet on the best weighted F1-Score, and ROCKET has the best macro precision and macro F1-Score. KNN, in the other hand, has the worse results among the tested classifiers.

A second evaluation has been done with the traditional training/testing division using a percentage split evaluation. In that way, testing data has not seen before by the trained classifier. Thus, using the best configuration parameters, the models were trained again on the whole training set (70% of the data), and tested on the test set (30% of the data). The training data was also balanced using O and FG techniques. The obtained performance indicators for the different tested classifiers and balancing techniques is shown in Table 6.

In this evaluation, it can also be observed how fuzzy sample generation again obtains better overall results than oversampling. The classifiers that make greatest improvements when using the proposed balancing technique are deep learning classifiers (i.e., CNN and TapNet). The classifier performing worst among the tested ones is KNN, for which the obtained metrics are lower than the rest of the classifiers.

In contrast, TapNet using fuzzy sample generation has achieved the best results for all metrics except the precision. It specially stands out on the quadratic weighted kappa, with a value of 0.868, which is a remarkably high score for this measure. These results confirm the ones obtained with cross-validation. We also see that testing results are quite close to the ones of 10-fold cross-validation, confirming the lack of overfitting.

The other two classifiers, ROCKET and CNN, are similar in terms of the metrics that take into account the class imbalance, such as accuracy, weighted recall, and weighted F1-Score. That is because they are able to properly classify most of the negative examples.

Metrics that do not compensate for class imbalance, macro recall and macro F1, are more interesting for our use case because they will show which classifiers can better identify the positive patients. Here, KNN is again the worst. ROCKET and CNN have similar results, with CNN performing slightly better. Tapnet shows the best macro recall and the second best in macro precision. To further analyse the results obtained by TapNet, the confusion matrix for this test is depicted in Fig. 6.

Tapnet is excellent in classifying the minority classes (categories 2 and 3), which is important in the medical domain. Consequently, the study concludes that TapNet is the classifier that exhibits superior performance among the tested classifiers for short time series in

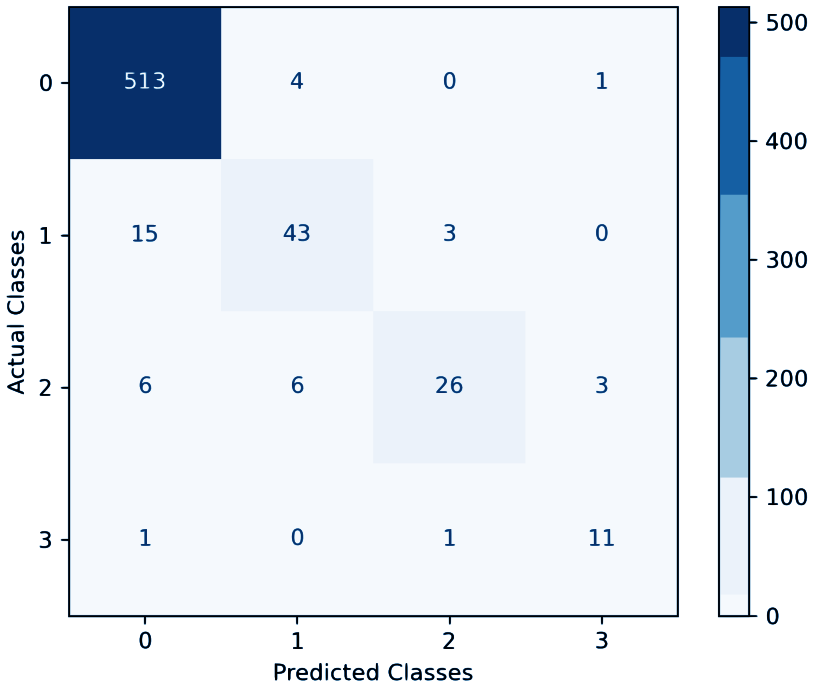


Fig. 6. Confusion matrix in testing with TapNet and fuzzy sample generation balancing.

the diabetic retinopathy problem. Its high quadratic weighted kappa demonstrates its ability to properly classify positive examples in the ordinal case. Additionally, when an error is made, the class assigned is close to the correct one.

6.4. Ablation study

Ablation studies are common for evaluating the importance of individual components within a proposed model. However, the pipeline of techniques of our preprocessing method cannot be tested separately, as the classifiers need to receive a series with equal length and equal distributed data. Therefore, each step in our processing pipeline serves a specific purpose, and it is necessary for producing the final required time series data.

One exception is the data balancing step, which can be removed from the pipeline without affecting its functionality. Fig. 7 depicts the confusion matrix for TapNet without applying any data balancing technique. This result is included to provide a baseline comparison to applying our proposed data balancing technique. The significant imbalance in the original dataset (as evident from Table 2) leads to poor performance on minority classes, as seen in this figure. Without balancing, we obtain an accuracy of 84.4% and a Kappa of 0.646, whereas with the proposed balancing method, accuracy increases to 93.7% and Kappa to 0.87. This result shows that the impact of balancing the minority classes is clear on improving classifier performance, particularly for the more severe classes.

6.5. Comparison with retiprogram

As explained in Section 6.2, the system Retiprogram only uses the last data values obtained from a patient. In that section, we have seen a comparison of the performance of Retiprogram with data from a new patient (in his/her first visit) and with data of the current visit in a chronic patient (last visit in the series). A decrease from 75% to 35% in accuracy has been observed.

Now, in order to know the improvement in the classification performance of using a sequence of retrospective data, we have used the same test set of Section 6.1 to evaluate the performance of Retiprogram. We have used only the last time point of the DR series. An accuracy of 36.7% and a Kappa of 0.076 are obtained. The results obtained are shown in Fig. 8. We observe much confusion in the class assignments made for patients in all rows, and large errors in case of patients with $DR = 0$.

From the obtained results, it can be concluded that using the temporal information on long-term diabetic patients is beneficial in order to correctly classify them into their current DR degree.

7. Conclusions and future work

In this paper, we presented a novel approach for estimating the different levels of diabetic retinopathy using only retrospective data from the EHR of the patient. Motivated by the discussed special characteristics of such medical series data, a technique for pre-processing EHR data has been presented. As first contribution, to construct multivariate time series of the same length for all the patients, missing entries have been completed with a double interpolation technique. Results show a much closer similarity with the original sequence when using the proposed method, compared with a classic linear interpolation. As second contribution, the paper presented a new fuzzy-based approach to boost short time series to generate new examples that may alleviate the problem of class imbalance in health care data. It consists in completing short sequences by using information from similar completed ones. Three types of variables have been distinguished when completing their values. For medical numerical values, a method based on the use of linguistic fuzzy variables has been proposed. By using the membership functions, we can find new input values that generate series similar enough to real examples.

Thirdly, several multivariate time series classifiers have been compared using the DR EHR data time series prepared with the previous techniques. From the results, we conclude that TapNet is the best classifier for this problem. TapNet is able to appropriately learn the underlying patterns of the minority DR examples, as metrics depict. It has achieved a kappa value of 0.868 and accuracy of 93.7%, which are outstanding results for such an imbalanced health-care problem.

Future work includes addressing some limitations and exploring new possibilities to further enhance our series processing method.

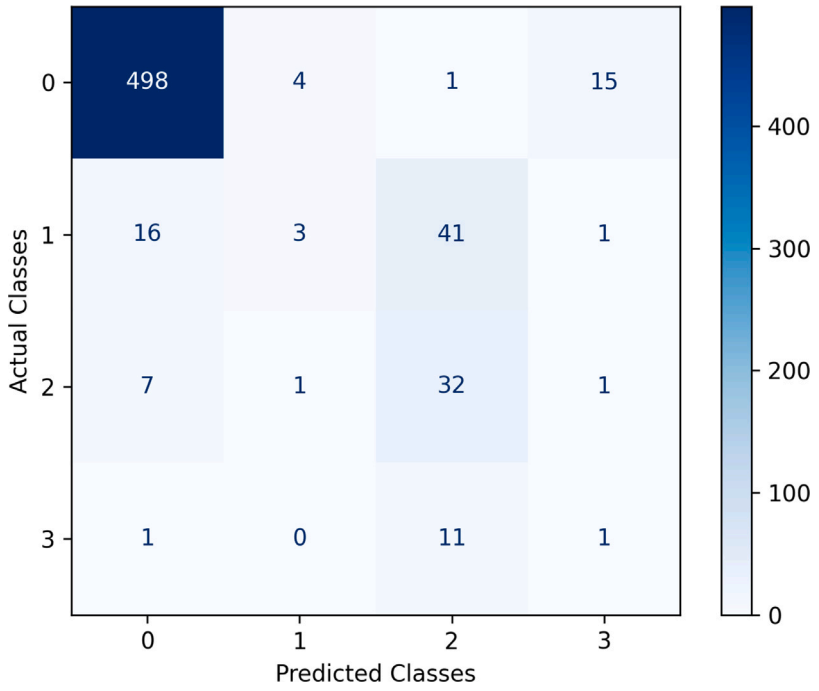


Fig. 7. Confusion matrix in testing with TapNet and no data balancing.

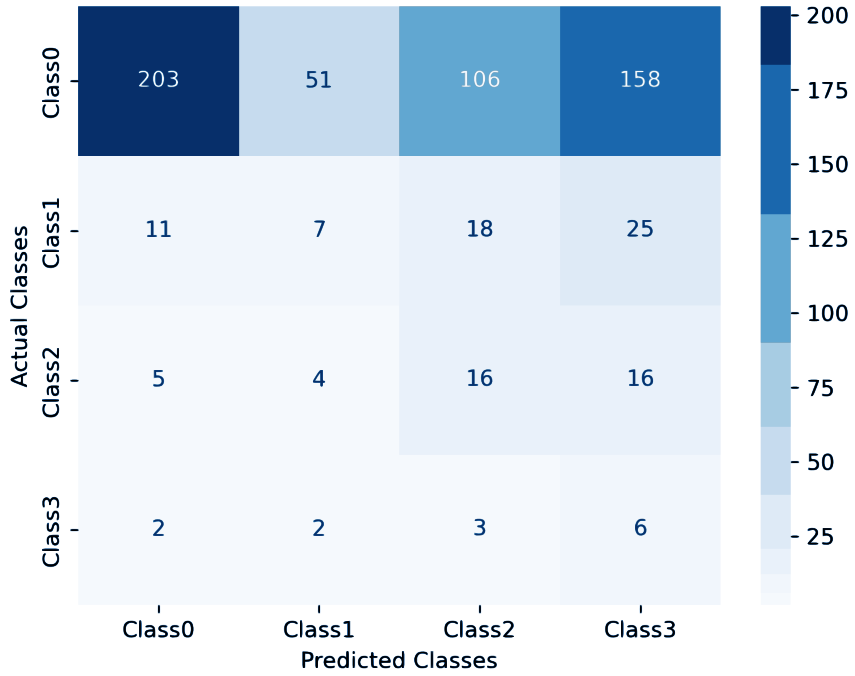


Fig. 8. Retiprogram results on the test set.

One major challenge is that our approach requires at least 6 years of historical data, which might not be feasible or practical in all medical contexts where patient records have limited duration. To overcome this issue, we propose exploring the possibility of handling shorter time series as well as applying transfer learning as future work. Additionally, while our proposed method has shown promising results on the DR dataset, it is essential to evaluate its generalizability across different datasets and clinical settings. Future work should focus on testing our approach with additional EHR datasets and exploring its applicability in other situations.

It would be valuable to investigate how recent classification techniques such as Densely Knowledge-aware Network for Multivariate Time Series Classification [35], DTCM (Deep Transformer Capsule Mutual distillation method) [36] or CapMatch (Contrastive transformer capsule method with feature-based knowledge distillation) [37] perform when applied to series of EHR data preprocessed with the techniques proposed in this paper. By combining the strengths of these classification methods with this novel preprocessing approach, we may improve the accuracy of CDSSs for disease diagnosis in chronic patients. The possibility of using or adapting the proposed method for shorter incomplete time series should also be studied. We also plan to make

a comparison with other balancing techniques (T-SMOTE). Finally, a study of different ways of encoding the numerical variables would also be interesting, since they are quite common and relevant in clinical decision support systems.

CRediT authorship contribution statement

Jordi Pascual-Fontanilles: Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Aida Valls:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Pedro Romero-Aroca:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgements

This work was supported by Instituto de Salud Carlos III, Spain (ISCIII) [project PI21/00064] and co-funded by the European Union; Universitat Rovira i Virgili (URV) [projects number 2023PFR-URV-114, 2022PFR-URV-41]; ITAKA funding from AGAUR [2021-SGR-00114]; and the first author had a pre-doctoral FI grant from Generalitat de Catalunya and Fons Social Europeu [grant number 2022 FI_B1 00036].

References

[1] R.A. Goodman, S.F. Posner, E.S. Huang, A.K. Parekh, H.K. Koh, Peer reviewed: Defining and measuring chronic conditions: Imperatives for research, policy, program, and practice, Prev. Chronic Dis. 10 (2013) 1–16, http://dx.doi.org/10.5888/PCD10.120239.

[2] J. Pascual-Fontanilles, A. Valls, A. Moreno, P. Romero-Aroca, Challenges in the Exploitation of Historical Clinical Data for the Classification of Diabetic Retinopathy Patients, IOS Press, 2023, pp. 204–207, http://dx.doi.org/10.3233/FAIA230683, URL https://ebooks.iospress.nl/doi/10.3233/FAIA230683.

[3] L.A. Zadeh, Knowledge representation in fuzzy logic, in: An Introduction To Fuzzy Logic Applications in Intelligent Systems, Springer, 1992, pp. 1–25.

[4] H. Ahmadi, M. Gholamzadeh, L. Shahmoradi, M. Nilashi, P. Rashvand, Diseases diagnosis using fuzzy logic methods: A systematic and meta-analysis review, Comput. Methods Programs Biomed. 161 (2018) 145–172, http://dx.doi.org/10.1016/J.CMPB.2018.04.013.

[5] R.R. Bourne, J.D. Steinmetz, M. Saylan, A.M. Mersha, A.H. Weldemariam, T.G. Wondmeneh, C.T. Sreeramareddy, M. Pinheiro, M. Yaseri, C. Yu, M.S. Zastrozhin, A. Zastrozhina, Z.J. Zhang, S.R. Zimsen, N. Yonemoto, G.W. Tsegaye, G.T. Vu, A. Vongpradith, A.M. Renzaho, M.B. Sorrie, A.A. Shaheen, W.S. Shiferaw, V.Y. Skryabin, A.A. Skryabina, G.K. Saya, V. Rahimi-Movaghar, M. Shigematsu, M.A. Sahraian, H. Naderifar, S. Sabour, P. Rathi, B. Sathian, T.R. Miller, A. Rezapour, L. Rawal, H.Q. Pham, U. Parekh, V. Podder, O.E. Onwujekwe, M. Pasovic, N. Otstavnov, H. Negash, S. Pawar, M.D. Naimzada, A.A. Montasir, F.A. Ogbo, M.O. Owolabi, K. Pakshir, Y. Mohammad, M.A. Moni, V. Nunez-Samudio, G.F. Mulaw, M. Naveed, S. Maleki, I.M. Michalek, S. Misra, S.N. Swamy, J.A. Mohammed, S. Flaxman, E.C. Park, P.S. Briant, G.G. Meles, K. Hayat, I. Landires, G.R. Kim, X. Liu, K.E. LeGrand, H.R. Taylor, S.M. Kunjathur, T.A.M. Khoja, B.K. Bicer, R. Khalilov, A. Hashi, G.A. Kayode, V.L. Carneiro, T. Kavetskyy, S. Kosen, V. Kulkarni, R. Holla, R. Kalhor, S. Jayaram, S.M.S. Islam, S.A. Gilani, S. Eskandarieh, M.D. Molla, R. Itumalla, F. Farzadfar, N.G. Congdon, H.R. Elhabashy, R. Elayedath, R.A. Couto, N. Dervenis, E.A. Cromwell, S.M. Dahlawi, S. Resnikoff, R.J. Casson, A. Abdoli, J.Y.J. Choi, F.L.C.D. Santos, W.A. Abrha, S.B. Nagaraja, A. Abualhasan, T.G. Adal, B.B. Aregawi, M. Beheshti, E. Abu-Gharbieh, A. Afshin, H. Ahmadieh, S.A. Alemzadeh, A. Arrigo, D.D. Atnafu, C. Ashbaugh, E. Ashrafi, W. Alemayehu, A.S. Alfaar, V. Alipour, E.W. Anbesu, S. Androudi, J. Arabloo, A. Arditi, E. Bagli, A.A. Baig, T.W. Bärnighausen, M.B. Parodi, A.S. Bhagavathula, N. Bhardwaj, P. Bhardwaj, K. Bhattacharyya, A. Bijani, M. Bikbov, M. Bottone, T. Braithwaite, A.M. Bron, Z.A. Butt, C.Y. Cheng, D.T. Chu, M.V. Cicinelli, J.M. Coelho, X. Dai, R. Dana, L. Dandona, R. Dandona, M.A.D. Monte, J.P. Deva, D. Diaz, S. Djalalinia, L.E. Dreer, J.R. Ehrlich, L.B. Ellwein, M.H. Emamian, A.G. Fernandes, F. Fischer, D.S. Friedman, J.M. Furtado, S. Gaidhane, G. Gazzard, B. Gebremichael, R. George, A. Ghashghaee, M. Golechha, S. Hamidi, B.R. Hammond, M.E.R. Hartnett, R.K. Hartono, S.I. Hay, G. Heidari, H.C. Ho, M. Househ, S.E. Ibitoye, I.M. Ilic, J.J. Huang, M.D. Ilic, A.D. Ingram, S.S.N. Irvani, R.P. Jha, R. Kahloun, H. Kandel, A.S. Kasa, J.H. Kempen, M. Khairallah, E.A. Khan, R.C. Khanna, M.N. Khatib, J.E. Kim, Y.J. Kim, A. Kisa, S. Kisa, A. Koyanagi, O.P. Kurmi, V.C. Lansingh, J.L. Leasher, N. Leveziel, H. Limburg, N. Manafi, K. Mansouri, C. McAlinden, S.F. Mohammadi, A.H. Mokdad, A.R. Morse, M. Naderi, K.S. Naidoo, V. Nangia, H.L.T. Nguyen, K. Ogundimu, A.T. Olagunju, S. Panda-Jonas, K. Pesudovs, T. Peto, M.H.U. Rahman, P.Y. Ramulu, D.L. Rawaf, S. Rawaf, N. Reinig, A.L. Robin, L. Rossetti, S. Safi, A. Sahebkar, A.M. Samy, J.B. Serle, M.A. Shaikh, T.T. Shen, K. Shibuya, J.I. Shin, J.C. Silva, A. Silvester, J.A. Singh, D. Singhal, R.S. Sitorus, E. Skiadaresi, A. Soheili, R.A. Sousa, D. Stambolian, E.G. Tadesse, N. Tahhan, M.I. Tareque, F. Topouzis, B.X. Tran, M.K. Tsilimbaris, R. Varma, G. Virgili, N. Wang, Y.X. Wang, S.K. West, T.Y. Wong, J.B. Jonas, T. Vos, Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: The right to sight: An analysis for the global burden of disease study, Lancet. Glob. health 9 (2021) e144–e160, http://dx.doi.org/10.1016/S2214-109X(20)30489-7.

[6] P. Romero-Aroca, M. López-Galvez, M.A. Martinez-Brocca, A. Pareja-Ríos, S. Artola, J. Franch-Nadal, J. Fernandez-Ballart, J. Andonegui, M. Baget-Bernaldiz, Changes in the epidemiology of diabetic retinopathy in Spain: A systematic review and meta-analysis, Healthcare (Switzerland) 10 (2022) 1318, http://dx.doi.org/10.3390/HEALTHCARE10071318.

[7] L. Prothero, M. Cartwright, F. Lorencatto, J.M. Burr, J. Anderson, P. Gardner, J. Presseau, N. Ivers, J.M. Grimshaw, J.G. Lawrenson, Barriers and enablers to diabetic retinopathy screening: A cross-sectional survey of young adults with type 1 and type 2 diabetes in the UK, BMJ Open Diabetes Res. Care 10 (2022) 2971, http://dx.doi.org/10.1136/BMJDRC-2022-002971.

[8] M.Z. Atwany, A.H. Sahyoun, M. Yaqub, Deep learning techniques for diabetic retinopathy classification: A survey, IEEE Access 10 (2022) 28642–28655.

[9] S. Dubey, M. Dixit, Recent developments on computer aided systems for diagnosis of diabetic retinopathy: A review, in: Multimedia Tools and Applications 2022 82:10, Vol. 82, Springer, 2022, pp. 14471–14525, http://dx.doi.org/10.1007/S11042-022-13841-9.

[10] J. Escorcia-Gutierrez, J. Cuello, M. Gamarra, P. Romero-Aroca, E. Caicedo, A. Valls, D. Puig, Grading diabetic retinopathy using transfer learning-based convolutional neural networks, in: Computer Information Systems and Industrial Management, Springer, Cham, 2023, pp. 240–252.

[11] P. Romero-Aroca, A. Valls, A. Moreno, R. Sagarra-Alamo, J. Basora-Gallisa, E. Saleh, M. Baget-Bernaldiz, D. Puig, A clinical decision support system for diabetic retinopathy screening: Creating a clinical support application, Telemed. e-Health 25 (2019) 31–40, http://dx.doi.org/10.1089/tmj.2017.0282.

[12] E. Saleh, J. Błaszczyński, A. Moreno, A. Valls, P. Romero-Aroca, S. de la Riva-Fernández, R. Słowiński, Learning ensemble classifiers for diabetic retinopathy assessment, Artif. Intell. Med. 85 (2018) 50–63, http://dx.doi.org/10.1016/j.artmed.2017.09.006.

[13] A. Valls, A. Moreno, J. Pascual-Fontanilles, J. Cristiano, D. Puig, P. Romero-Aroca, RETIPROGRAM and MIRA software, in: Inteligencia Artificial Y Oftalmología: Estado Actual En Cataluña, Vol. 31, Òrgan de la Societat Catalana d'Oftalmologia, 2023, pp. 206–213.

[14] J. Pascual-Fontanilles, L. Lhotska, A. Moreno, A. Valls, Adapting a fuzzy random forest for ordinal multi-class classification, Frontiers Artificial Intelligence Appl. 356 (2022) 181–190, http://dx.doi.org/10.3233/FAIA220336.

[15] C.P. Wilkinson, F.L. Ferris, R.E. Klein, P.P. Lee, C.D. Agardh, M. Davis, D. Dills, A. Kampik, R. Pararajasegaram, J.T. Verdaguer, Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales, Ophthalmology 110 (2003) 1677–1682, http://dx.doi.org/10.1016/S0161-6420(03)00475-5.

[16] W.K. Wang, I. Chen, L. Hershkovich, J. Yang, A. Shetty, G. Singh, Y. Jiang, A. Kotla, J.Z. Shang, R. Yerrabelli, A.R. Roghanizad, M.M.H. Shandhi, J. Dunn, A systematic review of time series classification techniques used in biomedical applications, Sensors 22 (2022) 8016, http://dx.doi.org/10.3390/S22208016.

[17] A.P. Ruiz, M. Flynn, J. Large, M. Middlehurst, A. Bagnall, The great multivariate time series classification bake off: A review and experimental evaluation of recent algorithmic advances, Data Min. Knowl. Discov. 35 (2021) 401–449, http://dx.doi.org/10.1007/s10618-020-00727-3.

[18] A. Dempster, F. Petitjean, G.I. Webb, ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels, Data Min. Knowl. Discov. 34 (2020) 1454–1495, http://dx.doi.org/10.1007/s10618-020-00701-z.

[19] Y. Sun, D. Zhang, Diagnosis and analysis of diabetic retinopathy based on electronic health records, IEEE Access 7 (2019) 86115–86120, http://dx.doi.org/10.1109/ACCESS.2019.2918625.

[20] Y. Zhao, X. Li, S. Li, M. Dong, H. Yu, M. Zhang, W. Chen, P. Li, Q. Yu, X. Liu, Z. Gao, 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, Front. Endocrinol. 13 (2022) 885, http://dx.doi.org/10.3389/FENDO.2022.876559.

[21] H.Y. Tsao, P.Y. Chan, E.C.Y. Su, Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms, BMC Bioinform. 19 (2018) 111–121, http://dx.doi.org/10.1186/S12859-018-2277-0.

[22] O.I. Ogunyemi, M. Gandhi, C. Tayek, Predictive models for diabetic retinopathy from non-image teleretinal screening data, AMIA Summits Transl. Sci. Proc. 2019 (2019) 472.

[23] O. Ogunyemi, D. Kermah, Machine learning approaches for detecting diabetic retinopathy from clinical and public health records, AMIA Annu. Symp. Proc. 2015 (2015) 983.

[24] N. Itzhak, I.M. Pessach, R. Moskovitch, Prediction of acute hypertensive episodes in critically ill patients, Artif. Intell. Med. 139 (2023) 102525, http://dx.doi.org/10.1016/J.ARTMED.2023.102525.

[25] S. Sheikhalishahi, A. Bhattacharyya, L.A. Celi, V. Osmani, An interpretable deep learning model for time-series electronic health records: Case study of delirium prediction in critical care, Artif. Intell. Med. 144 (2023) 102659, http://dx.doi.org/10.1016/J.ARTMED.2023.102659.

[26] S. Rabhi, F. Blanchard, A.M. Diallo, D. Zeghlache, C. Lukas, A. Berot, B. Delemer, S. Barraud, Temporal deep learning framework for retinopathy prediction in patients with type 1 diabetes, Artif. Intell. Med. 133 (2022) 102408, http://dx.doi.org/10.1016/J.ARTMED.2022.102408.

[27] Z.L. Teo, Y.C. Tham, M. Yu, M.L. Chee, T.H. Rim, N. Cheung, M.M. Bikbov, Y.X. Wang, Y. Tang, Y. Lu, I.Y. Wong, D.S.W. Ting, G.S.W. Tan, J.B. Jonas, C. Sabanayagam, T.Y. Wong, C.Y. Cheng, Global prevalence of diabetic retinopathy and projection of burden through 2045: Systematic review and meta-analysis, Ophthalmology 128 (2021) 1580–1591, http://dx.doi.org/10.1016/j.ophtha.2021.04.027.

[28] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, 2002, http://dx.doi.org/10.1613/jair.953.

[29] P. Zhao, C. Luo, B. Qiao, L. Wang, S. Rajmohan, Q. Lin, D. Zhang, T-SMOTE: Temporal-oriented synthetic minority oversampling technique for imbalanced time series classification, in: Proceedings of IJCAI, 2022.

[30] B.K. Iwana, S. Uchida, An empirical survey of data augmentation for time series classification with neural networks, PLoS One 16 (2021) e0254841, http://dx.doi.org/10.1371/JOURNAL.PONE.0254841.

[31] X. Zhang, Y. Gao, J. Lin, C.T. Lu, TapNet: Multivariate time series classification with attentional prototypical network, Proc. AAAI Conf. Artif. Intell. 34 (2020) 6845–6852, http://dx.doi.org/10.1609/AAAI.V34I04.6165.

[32] B. Zhao, H. Lu, S. Chen, J. Liu, D. Wu, Convolutional neural networks for time series classification, J. Syst. Eng. Electron. 28 (2017) 162–169, http://dx.doi.org/10.21629/JSEE.2017.01.18.

[33] M. Löning, F. Király, T. Bagnall, M. Middlehurst, S. Ganesh, G. Oastler, J. Lines, M. Walter, ViktorKaz, L. Mentel, chrisholder, L. Tsaprounis, RNKuhns, M. Parker, T. Owoseni, P. Rockenschaub, danbartl, jesellier, eenticott-shell, C. Gilbert, G. Bulatova, Lovkush, P. Schäfer, S. Khrapov, K. Buchhorn, K. Take, S. Subramanian, S.M. Meyer, AidenRushbrooke, B. rice, Sktime/sktime: v0.13.4, 2022, http://dx.doi.org/10.5281/zenodo.7117735.

[34] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174, http://dx.doi.org/10.2307/2529310.

[35] Z. Xiao, H. Xing, R. Qu, L. Feng, S. Luo, P. Dai, B. Zhao, Y. Dai, Densely knowledge-aware network for multivariate time series classification, IEEE Trans. Syst. Man Cybern.: Syst. 54 (2024) 2192–2204, http://dx.doi.org/10.1109/TSMC.2023.3342640.

[36] Z. Xiao, X. Xu, H. Xing, B. Zhao, X. Wang, F. Song, R. Qu, L. Feng, DTCM: Deep transformer capsule mutual distillation for multivariate time series classification, IEEE Trans. Cogn. Dev. Syst. (2024) http://dx.doi.org/10.1109/TCDS.2024.3370219.

[37] Z. Xiao, H. Tong, R. Qu, H. Xing, S. Luo, Z. Zhu, F. Song, L. Feng, CapMatch: Semi-supervised contrastive transformer capsule with feature-based knowledge distillation for human activity recognition, IEEE Trans. Neural Netw. Learn. Syst. (2023) http://dx.doi.org/10.1109/TNNLS.2023.3344294.