Learning Ensemble Classifiers for Diabetic Retinopathy Assessment

Emran Saleh[a], Jerzy Błaszczyński[c], Antonio Moreno[a], Aida Valls[a,*], Pedro Romero-Aroca[b], Sofia de la Riva - Fernández[b], Roman Słowiński[c,d]

[a] *Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Tarragona, Spain*
[b] *Ophthalmic Service, University Hospital Sant Joan de Reus, Institut d'Investigació Sanitària Pere Virgili (IISPV), Universitat Rovira i Virgili, Reus (Tarragona), Spain*
[c] *Institute of Computing Sciences, Poznań University of Technology, 60-965 Poznań, Poland*
[d] *Systems Research Institute, Polish Academy of Sciences,01-447 Warsaw, Poland*

Abstract

Diabetic Retinopathy is one of the most common comorbidities of diabetes. Unfortunately, the recommended annual screening of the eye fundus of diabetic patients is too resource-consuming. Therefore, it is necessary to develop tools that may help doctors to determine the risk of each patient to attain this condition, so that patients with a low risk may be screened less frequently and the use of resources can be improved. This paper explores the use of two kinds of ensemble classifiers learned from data: Fuzzy Random Forest and Dominance-Based Rough Set Balanced Rule Ensemble. These classifiers use a small set of attributes which represent main risk factors to determine whether a patient is in risk of developing Diabetic Retinopathy. The levels of specificity and sensitivity obtained in the presented study are over 80%. This study is thus a first successful step towards the construction of a personalized decision support system that could help physicians in daily clinical practice.

Keywords: `Diabetic Retinopathy, Decision Support Systems, Rule-based Models, Fuzzy Decision Trees, Random Forest, Ensemble Classifiers, Dominance-based Rough Set Approach, Class Imbalance.`

1. Introduction

Diabetes Mellitus (DM) is a disease currently attaining over 400 million people around the world (IDF, 2016). Its incidence is constantly growing, and it is expected to affect 10% of the world's population by 2040. It is considered one of the main global causes of death, overcoming other diseases like HIV/AIDS, tuberculosis and malaria (IDF, 2015; Alghadyan,

*Corresponding author

Email addresses: `emran.saleh@urv.cat` (Emran Saleh), `jerzy.blaszczynski@cs.put.poznan.pl` (Jerzy Błaszczyński), `antonio.moreno@urv.cat` (Antonio Moreno), `aida.valls@urv.cat` (Aida Valls), `pedro.romero@urv.cat` (Pedro Romero-Aroca), `delariva.sofia@gmail.com` (Sofia de la Riva - Fernández), `roman.slowinski@cs.put.poznan.pl` (Roman Słowiński)

August 29, 2017

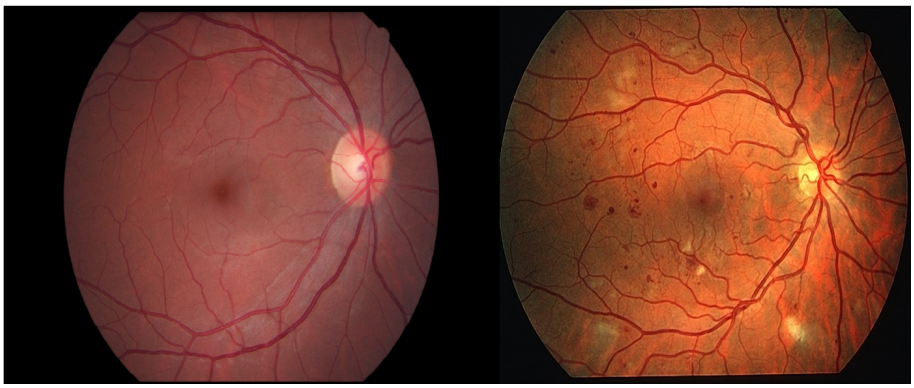

Figure 1: Normal eye (left), versus eye with signs of diabetic retinopathy (right)

2011). It has been estimated that 46% of the people attained by this illness are never properly diagnosed and treated (IDF, 2013). Furthermore, DM may lead to many health complications such as kidney failure, amputation and blindness.

Diabetic Retinopathy (DR) is a disease characterized by a progressive damage of the retina. Many factors are related to the development of DR, like the period of diabetes evolution, genetic factors and metabolic control density (Control et al., 1993). Diabetic retinopathy appears when diabetes harms the blood vessels. In a first stage, the arteries in the retina weaken and begin to leak, forming small, dot-like hemorrhages. These leaking vessels often from deposits of lipoproteins (exudates) in the retina, which produces a blur vision. A second complication is the growth of new weak blood vessels that break and leak blood into the eye, harming the macula so the retina cannot project images to the brain. The result is a loss of sight. Consequently, DR is the most serious ophthalmic condition induced by DM if it is not detected early and properly treated (Bourne et al., 2013; Zhang et al., 2011).

As DR is a main cause of vision loss among people suffering from diabetes, all these patients must be periodically screened in order to detect signs of retinopathy development in an early stage. The screening process consists in taking an image of the eye fundus, which is analyzed by expert ophthalmologists for signs of micro-aneurisms or exudates. Figure 1 shows an eye of a healthy person on the left, and an eye with diabetic retinopathy on the right. The central dark circle of the eye is the macula. Small red or yellow spots can be seen near the macula on the eye in the right, corresponding to microaneurisms and hemorrages. Yellow bigger spots in the peripheral retina are exudates.

Early and frequent screening of diabetic patients reduces the chance of blindness and decreases the overall load on the health care centres (Romero Aroca et al., 2012). However, as the number of diabetic patients is very large and it is continually increasing, it is already very resource-consuming to perform a yearly screening to all of them (which is the frequency recommended by international medical associations such as the American Diabetes Association (ADA, 2015), the American Academy of Ophthalmology and the Royal College of Ophthalmologists (RCO, 2012)).

Some studies ((Olafsdottir & Stefansson, 2007), (Chalk et al., 2012)) have argued that

a high percentage of diabetic patients could be safely screened every 2 or 3 years, because their medical condition indicates that they are very unlikely to develop DR. These results coincide with those of a local study developed at *Sant Joan de Reus University Hospital* (SJRUH), which observed that just a 9% of diabetic patients develop DR. Detecting which patients do not need a screening, could permit more frequent screening of the patients that have a higher risk of DR. This may result in an improvement of the usage of Ophthalmology service resources without reducing the quality of health care.

The Ophthalmology unit of SJRUH has been collecting detailed information on the screening of thousands of diabetic patients in the area of Reus (Tarragona, Spain) for several years. This database is highly valuable, as it is very uncommon for a hospital to collect and store systematically such accurate data about the patients' condition on each visit to the medical centre. Expert ophthalmologists analyzed these data to identify the components of the *Electronic Health Record* (EHR) that are more relevant in the assessment of the risk of developing DR (Romero-Aroca et al., 2016). As a result of this study, nine numerical and categorical attributes were selected as the key factors to be taken into account when determining whether a diabetic patient is likely to have DR.

At this point, the main aim of our study is to explore the feasibility of using different *Machine Learning* (ML) techniques to build computational models based on the values of the attributes available in the EHR (without the need of an image). Such a classification model, given the data of a patient, can assess the personalised risk level of developing DR and, therefore, can help the physician to decide the best screening time. Those patients with a high risk should be screened more frequently, whereas those with a low risk could be safely screened again in 2-3 years.

In this paper, we study whether the use of ensemble classification techniques based on uncertainty models may lead to a good classification accuracy for the problem of risk assessment of DR based on EHR data. Specifically, *Fuzzy Random Forests* (FRF) and *Dominance-based Rough Set Balanced Rule Ensemble* (DRSA-BRE) are applied to the data of SRJUH. The numerical attributes have been discretized before the analysis. FRF and DRSA-BRE are particularly appropriate for dealing with uncertain data, providing qualitative classification rules which are easier to understand by medical physicians than rules referring to particular numerical values. The classification models constructed by these techniques assign a relevance measure to each of the classification rules, which may also be a highly valuable information for the doctors. Both techniques are examples of *ensemble* methods, in which a group of classifiers is built and the final prediction for a particular patient is made taking into account the opinions of all the classifiers.

The rest of the paper is organized as follows. Section 2 presents the background and the related work on risk-prediction models for DR. It also includes a brief introduction to FRF and DRSA-BRE followed by reviews of their medical application. Section 3 explains how the dataset was constructed, which are the attributes included in this study and how the numerical ones were discretized. It also explains how the FRF and DRSA-BRE models have been applied for DR detection. Section 4 presents and discusses the experimental results on the use of the two classification learning techniques on these data. Conclusions and an outline of the future lines of research are presented in Section 5. The article ends with

Appendix A and B with a detailed definition of FRF and DRSA-BRE, respectively.

2. Background and related works

This section begins with related works on diabetic retinopathy detection, which are mainly focused on the computational analysis of the images obtained from the eye fundus (and not on the physical and clinical features of the patient, as in this work). After that, the two studied rule induction methods are presented: Fuzzy Random Forests and Dominance-Based Rough Set Balanced Rule Ensembles. A brief review of medical applications of these methods is also included.

2.1. Diabetic Retinopathy detection

Research on DR detection has mainly focused on the automatic analysis of the eye images. Currently researchers are studying how to extract signs of DR from images taken from the eye fundus. Several computer vision techniques have been used to build models for the detection of these signs. There exists a medical definition of 5 levels of severity of DR. Some works consider only the distinction between normal eyes and eyes with lesions, while other works try to differentiate the 5 categories according to the number, size and shape of different types of lesions (microaneurysms, hemorrhages and exudates). There are mainly two kinds of computer vision approaches: those based on the classical procedure of feature extraction and classification, and those based on the use of advanced neural networks.

As a recent example of the feature extraction approach, (Saleh & Eswaran, 2012) presented a system for the detection of microaneurysms and hemorrhages. In this work, first a morphological pre-processing stage extracted and removed the optic disc from the image. Then, depending on the illumination disparity, the background of the image was removed so that the resulting image contained only the blood vessels and the microaneurysms and hemorrhages. Finally, the image was classified using the geometry of these two types of lesions (Early Treatment Diabetic Retinopathy Study Research Group, 1991). A key point in this kind of approach is the segmentation of the optic disc. Several techniques have been applied in this task, such as the Active Contour Model (Popescu & Ichim, 2015) or Grabcut (Escorcia-Gutiérrez et al., 2016).

The second approach consists in training a neural network using a set of images already tagged with the DR category. This kind of supervised Machine Learning approaches have obtained quite good results since the advent of *Convolutional Neural Networks* (CNN). Deep neural networks have been heavily used in the last years to classify images because they have demonstrated a significant increase in classification accuracy compared with the traditional simple neural network models. For example, Pratt et al. (2016) have recently proposed a CNN model that has 10 convolutional layers and a final classifying layer. They are able to classify images in the 5 severity classes, although the system suffers from a low sensitivity. As microaneurysms are among the most important visual features to detect diabetic retinopathy, (Haloi, 2015) has proposed a deep learning approach to classify every pixel as being a microaneurysm or not. The model, based on three convolutional layers, achieved a good classification accuracy on a well-known data set. Some other examples of

recent works on the use of deep CNN for DR detection are (Chandrakumar & Kathirvel, 2016) and (La Torre et al., 2016).

CNNs have also been combined with other techniques such as Random Forests (Wang et al., 2015). This hybrid method focused on segmenting the retinal blood vessels. The CNN extracted the features from the images while the Random Forest was used as an ensemble classifier.

All these DR detection methods require special equipment and trained personnel to acquire the eye fundus images. Another line of work, which requires less material and human resources, takes advantage of all the data stored in the Electronic Health Records (EHRs) of the patients. Nowadays EHRs include many variables with information about the patient's conditions, which are updated during regular visits to the family physician. However, few works have used this information to help in the diagnosis of DR. In a previous work (Sanromà et al., 2016), we used some categorical machine learning methods (regression, k-nearest neighbours, decision trees, and random forests) to solve this classification problem. The best results were obtained with the Random Forests technique (using crisp decision trees). A classifier based on a combination of neural networks has been proposed in (Skevofilakas et al., 2010). However, such kind of technique is a black box model for the medical personnel that will use the system. They need to know how the decision is made by the decision support system in order to make a conscious decision upon a given patient.

2.2. Brief overview of FRF and a review of its medical applications

Ensemble methods have been extensively applied to complex problems in the last twenty years. Their basic idea is that the analysis of different aspects of the problem, through the use of a diverse set of classifiers, can improve the performance of any single individual classification system. A *Random Forest* (RF, introduced by Breiman (2001)) is a collection of decision trees. Two main random elements are introduced in the construction of each tree. On the one hand, the elements in the training set are only a subset of the full training set. On the other hand, in each node of the tree only a random subset of the available attributes is considered. This double randomization in the construction of the classifiers alleviates some of the well-known drawbacks of using a single decision tree, like overfitting and bias towards attributes with more values. Moreover, it increases the robustness and predictive power when compared with classification techniques based on a single tree (Breiman, 2001).

More complex variants of Random Forests have also been proposed (see a recent survey in (Kulkarni & Sinha, 2013)). One extension that is particularly interesting is the use of fuzzy logic, which allows to work with uncertainty. *Fuzzy Decision Trees (FDT)* permit to manage uncertain or imprecise data, represented with linguistic labels whose meaning is defined with a fuzzy set. *A Fuzzy Random Forest* (FRF) is composed of a set of fuzzy decision trees. Given an object to be classified, each branch of each tree assigns a particular confidence to the object belonging to a certain class. A fuzzy aggregation procedure is used to combine the output of the rules and make the final assignment.

There are several learning methods that may be used to construct an FDT from a set of labelled examples. They may be roughly classified in two categories, depending on whether they use information theory measures to optimize the *fuzzy entropy* (Kosko, 1986) or if they

are based on the notion of *classification ambiguity* (Wang et al., 2001). In this work we have considered Yuan and Shaw's FDT induction algorithm, which falls into the latter category (Yuan & Shaw, 1995). Yuan and Shaw's induction algorithm is an extension of the classic ID3 method for crisp data. It incorporates two parameters to manage the uncertainty:

- The *significance level* (α) is used to filter evidence that is not relevant enough. If the membership degree of a fuzzy evidence E is lower than α, it is not used in the rule induction process.
- The *truth level threshold* (β) fixes the minimum truth of the conclusion given by a rule. Thus, it controls the growth of the decision tree. Lower values of β may lead to smaller trees but with a lower classification accuracy.

The main steps of the induction process of an FDT are the following (for more details see Appendix A):

1. Select the attribute with the smallest ambiguity (i.e. the attribute whose values help to discrimante better the class of each example) as the root node of the tree.
2. Create a new branch for each of the values of the attribute for which we have examples with support at least α.
3. Calculate the *truth level of classification* of the objects within a branch into each class.
4. If the truth level of classification is above β for at least one of the classes X_i, terminate the branch with a leaf with label X_i, corresponding to the class with the highest truth level.
5. If the truth level is smaller than β for all classes, check if an additional attribute will further reduce the *classification ambiguity*.
6. If it does, select the attribute with the smallest classification ambiguity with the accumulated evidence as a new decision node for the branch. Repeat from step 2 until no further growth is possible.
7. If it doesn't, terminate the branch as a leaf with a label corresponding to the class with the highest truth level.

This type of ensemble methods have been recently applied to different medical problems. Crisp RF have been used for diagnosis of heart arrhythmia (Alickovic & Subasi, 2016) and brain tumours (Koley et al., 2016), or for the analysis of medical images to differentiate between fat, muscle and edema tissues in MRIs (Kovacs et al., 2016).

Concerning the fuzzy approach, Fuzzy Random Forests have been applied to different kinds of problems like feature selection (Cadenas et al., 2013), facial biometric identification (Jiang et al., 2016) or terrain classification (Zhang et al., 2012). In the medical domain, FRFs and classifiers based on fuzzy rules have been used for indoor localization of elderly people (Trawiński et al., 2013), characterization of medical data (Marsala, 2009) and gene prioritization for cancer diagnosis (Cadenas et al., 2016).

In the literature, fuzzy rule-based approaches have been extensively utilized in detection, classification and prediction related to diabetes. Lukmanto & Irwansyah (2015) proposed

a fuzzy hierarchical model for early detection of Diabetes Mellitus. To optimize the fuzzy rules, a modification of the bee colony algorithm has been integrated with a fuzzy rule-based classifier for diabetic diagnosis in (Beloufa & Chikh, 2013). A fuzzy rule-based model has been proposed in (Meza-Palacios et al., 2017) to support the physicians in nephropathy control on Type-2 diabetic patients.

2.3. Brief overview of DRSA and a review of its medical applications

In this study, we are using an ensemble classifier adapted to class imbalanced data which are partially inconsistent, called DRSA-BRE. It combines the rough set methodology, called *Dominance-based Rough Set Approach* (DRSA), and a special bagging extension designed to construct a balanced ensemble of rule classifiers from data structured using DRSA, called *Dominance-based Rough Set Balanced Rule Ensemble* (DRSA-BRE).

For a complete presentation of the DRSA methodology see (Greco et al., 2001; Słowiński et al., 2014, 2015). In DRSA, information about objects (classification examples) is represented in the form of an *information table*. The rows of the table are labeled by objects, whereas columns are labeled by attributes and entries of the table are attribute values. The set of attributes is, in general, divided into a set C of condition attributes and set D of decision attributes (in most of the cases, a singleton decision attribute d designating class labels). DRSA approach is particularly interesting for decisions where the condition attributes and decision attributes are ordinal. Then, the rules constructed represent different kinds of relationships between C and D. A positive relationship means that the greater the value of the condition attribute, the higher the class label (i.e. the value of the decision attribute), and a negative relationship means that the greater the value of the condition attribute, the lower the class label. Using DRSA, we get rough approximations of each decision class X_k and its complement $\neg X_k$. These approximations serve to induce "*if..., then...*" decision rules recommending assignment to class X_k (argument pros) or to its complement $\neg X_k$ (argument cons). Rules are constructed using elementary building blocks, known as *dominance cones*, with origins in each example in the attribute space. Based on the rough set concept, introduced by Pawlak (1991), rules for lower and upper approximation of each decision class are obtained from the training observations. More details about DRSA and DRSA-BRE can be found in Appendix B.

The choice of DRSA is motivated by the aim of discovering synthetic rules that exhibit monotonic relationships between values of attributes describing the objects from the universe of discourse and their classification. DRSA is able to deal with possible inconsistencies in data prior to the induction of rules. Rules represent knowledge discovered from data. They are presented to the user without irrelevant facts, which could obfuscate cause-effect relationships. Moreover, they are helpful to predict the classification of new objects. Finally, DRSA permits to assess the relevance of particular attributes, using a Bayesian confirmation measure on the responses of the rules applied on testing examples (see Appendix B.3).

Although the DRSA methodology is based on the rough set concept, it has been adapted to data with the above mentioned monotonic relationships (Greco et al., 2001). DRSA appears to be more suitable for the analysis of this kind of qualitative and quantitative

data than statistical methods that are well suited for quantitative data without monotonic relationships. The state-of-the-art articles about rough sets and DRSA can be found in (Słowiński et al., 2014; Yao et al., 2015; Słowiński et al., 2015). The DRSA methodology has been implemented in the jRS library and jMAF desktop application[1]

DRSA-BRE is an ensemble of rule classifiers induced from bootstrap samples of objects derived from data structured by DRSA. It has been noticed that, when learning from class imbalanced data (as it is the case for Diabetic Retinopathy), the global imbalance ratio (i.e., ratio of the number of objects in the minority class to the number of objects in other classes) is not the only or even not the most important factor which makes learning difficult. Other data difficulty factors such as class overlapping, small disjunct or lack of representativeness significantly deteriorate the quality of the induced model even on exactly balanced data (Napierala & Stefanowski, 2016). The method that we use to construct an ensemble of rule classifiers from balanced bootstrap samples of objects is called *Neighbourhood Balanced Bagging* (NBBag) (Błaszczyński & Stefanowski, 2015). It extends the standard *bagging* scheme proposed by Breiman (1996). The samples of objects generated by NBBag are controlled by a balancing factor, which allows to handle difficulty factors typical for imbalanced data by changing the distribution of objects in the constructed samples (see Appendix B.2).

The record of applications of the DRSA methodology in medicine and biochemistry is quite long. A large application area concerns the analysis of relationships between antimicrobial activity and the chemical structure of new compounds. Strong rules discovered by DRSA enable creating prognostic models of new compounds with favorable antimicrobial properties. Moreover, the relevance of the attributes estimated from the discovered rules allows to distinguish which of the compound features have the strongest and the weakest influence on the antimicrobial properties. In (Pałkowski et al., 2014b), relationships between chemical structure, surface active properties and antibacterial activity of bis-quaternary imidazolium chlorides were analyzed. In (Pałkowski et al., 2014a), a SAR (structure-activity-relationship) study was performed on another set of imidazolium-based chlorides, using DRSA. In (Pałkowski et al., 2015), a series of bis-quaternary imidazolium chlorides was analyzed with respect to their biological activity against Candida albicans as one of the major opportunistic pathogens causing a wide spectrum of diseases in human beings; the DRSA results show that the antifungal activity is dependent on the type of substituents and their position at the chloride moiety, as well as on the surface active properties of the compounds. The rough set approach has been recently counted into prospective tools of chemoinformatics in a survey article on knowledge discovery from chemical data (Gardiner & Gillet, 2015).

Another field of applications of DRSA concerns analysis of biomedical data. In (Blasco et al., 2015), DRSA has been applied to the analysis of metabolomic data in view of discovering diagnostic biomarkers for amyotrophic lateral sclerosis. Yet another application of DRSA was reported in (Cinelli et al., 2015), where it served to induce rules for a greener synthesis of silver nanoparticles from a data set describing synthesis protocols.

[1] http://www.cs.put.poznan.pl/jblaszczynski/Site/jRS.html

Classic rough set based models have been used in diabetes detection. They have been tested on the Pima Indian Diabetes Dataset (Khan & Revett, 2004). Moreover, they were used to evaluate the importance of different attributes for children with Diabetes Mellitus Type-1 (Stepaniuk, 1999) and for classification of Type-2 diabetes from three-dimensional body surface anthropometrical data (Su et al., 2006). Some works address diseases derived from diabetes, such as the detection of macro-angiopathy diagnoses for diabetic patients (Nakayama et al., 1999). Nevertheless, no prior study with the DRSA approach for diabetes has been found.

3. Data and Methods

This section explains how the two ensemble methods explained in the previous section, *Fuzzy Random Forests* and *Dominance-Based Rough Sets-Balanced Rule Ensemble*, have been used for diabetic retinopathy detection. First, the dataset used for this study is presented. The data has been provided by *Sant Joan de Reus University Hospital*, a hospital located in the city of Reus (Tarragona) in the region of Catalonia (north-est of Spain).

3.1. Data from the Electronic Health Record

Sant Joan de Reus University Hospital (SJRUH) serves an area with a population of 247,174 inhabitants, in which there are 17,792 patients with Diabetes Mellitus (Romero-Aroca et al., 2016).

The Ophthalmologic unit of SJRUH has been collecting detailed information about thousands of diabetic patients for several years. According to the current medical protocols, these patients are regularly screened using non-mydriatic cameras. From the eye fundus images ophthalmologists are able to determine the presence or absence of Diabetic Retinopathy.

Since 2007 several analytical, metabolic and demographic data have been systematically collected and stored in the Electronic Health Records of the diabetic patients. The dataset used in this study contains 2323 diabetic patients, who were already labeled regarding DR: 579 patients presented DR (class 1) and 1744 patients were not suffering from DR (class 0). The dataset was divided in two parts, one for training with 1212 examples (871 from class 0 and 341 from class 1) and another for testing with 1111 patients (873 from class 0 and 238 from class 1). The parameters of the models are optimized using a 10-fold cross-validation on the training set. The best classification models are then applied to the test set for validation.

Previously, a statistical analysis on the data for the 8-year period from 2007 to 2014 was made in order to determine the changes in the incidence of DR (Romero-Aroca et al., 2016). It was observed that incidence was stable between 2007 and 2011 (around 8.1%), but since 2011 it has continuously increased until almost 9%. The same study also analyzed which are the main risk factors for developing DR. As a result of this study, nine attributes were identified as the most relevant ones in the appearance of DR. These attributes are the ones that have been considered in the construction of the classifiers reported in this paper.

There are two types of attributes: *numerical* (age, evolution time of diabetes, body mass index, HbA1, microalbuminuria and creatinine) and *categorical* (sex, HTAR and medical treatment). In the FRF and DBR-BRE classifiers, information is managed at a linguistic level; therefore, numerical attributes had to be discretized. The meaning of the intervals is of great importance to understand the classification rules. Therefore, the numerical attributes' values and their boundaries were defined with the aid of a team of expert ophtalmologists of SJRHU. In some of the indicators (*e.g.* Body Mass Index-BMI) standard intervals were taken; in most cases the medical doctors defined the most appropriate intervals from their expert knowledge and the findings of the statistical analysis (Romero-Aroca et al., 2016).An appropriate label was assigned to each interval to define an easily interpretable category.

Furthermore, in the case of the FRF method, a second step which smoothed the boundaries between consecutive categories was applied using membership functions. Appropriate membership functions were defined, fulfilling always the definition of *fuzzy partition* (*i.e.* each point in the numerical range belongs to a maximum of two categories and the sum of its membership values is 1). As an example, Fig. 2 shows the fuzzy sets associated to the linguistic labels considered for the age and BMI attributes.

It is possible to distinguish 5 levels of severity of DR (Wilkinson et al., 2003). However, in this work the doctors were interested in distinguishing only patients that need a screening of the eye from the ones that do not have any sign of DR and, hence, do not need the screening test. For this reason, a binary classification into two classes was made: X_1 patients with risk of DR (positive class) and X_0 for non-risky patients (negative class).

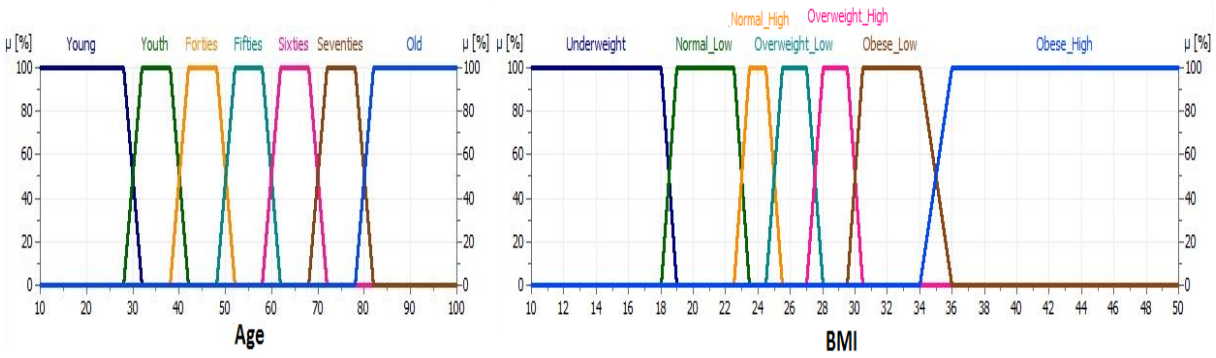

Figure 2: Definition of linguistic labels for Age and Body Mass Index.

This dataset has two characteristics that must be properly handled when applying the machine learning method. First, the set is highly imbalanced because only 25% of the patients belong to class 1 (i.e. have DR). Second, once the data is discretized, we have some contradictory examples, which correspond to two patients with the same values on all the criteria but one of the patients belongs to class 0 and the other to class 1. These two issues have been properly addressed in each of the two methods applied, as will be explained in the next subsections.

3.2. Fuzzy Random Forests in DR detection

In order to construct a fuzzy random forest for DR using the (Yuan & Shaw, 1995) approach presented in Appendix A , some parameters have to be fixed:

1. Choose a random subset of the training examples for training (*bootstrap*). It is important to keep a balanced distribution of the classes in each bootstrap. The repetition of the examples in a bootstrap is acceptable. It is generally recommended that the size of each bootstrap should be around two thirds of the training dataset. Each FDT in the forest trains with 1/3 of the total examples which are labeled with class 0 (the patients who do not suffering from DR) and 1/3 of the total examples which are labeled with class 1 (the patients suffer from DR)
2. During the tree construction a random subset of the attributes of size γ will be taken when deciding a new splitting of a tree node. In this study, several tests were done with $\gamma = \{1, 2, 3, 4\}$.
3. The number of fuzzy decision trees (n) has to be large enough. Tests have been done with $n = \{100, 200, 300\}$.

Fixing these parameters, the Yuan and Shaw induction procedure is used for training and a set of fuzzy decision trees is constructed. Notice that the use of random balanced subsets of examples enables to minimize the impact of the imbalance of the original whole dataset.

Once the trees are created, the 1111 patients of the testing set are classified using the fuzzy rules. Each tree has a set of decision rules that can be activated for a patient, giving different predicted classes. Many techniques can be used to establish the final decision of the fuzzy tree. The method used in this work to classify a patient is the well known Mamdani inference procedure using the t-norm minimum (on satisfaction level of a rule) and the t-conorm maximum (on aggregation of outputs). When a new patient is fed into the classifier system, each rule is activated with a certain degree of satisfaction. Then, as each rule has a certain degree of support (obtained in the construction procedure) the level of activation produced by the patient is multiplied by the degree of support of the rule, obtaining the membership degree μ_{X_k} to the conclusion class X_k .

Using this inference procedure, every activated rule of the tree leads to one of the two classes with degree μ_{X_k}.

This procedure is applied to each rule of each of the n different trees, constructed with the different configurations of the parameters. The aggregation of the classification output of all the rules may be done in different ways. There are basically two approaches: merge the information of all the branches to make a prediction for every tree and aggregate these predictions, or take into account directly all the scores of all the branches of all the trees to make a single global prediction (Cadenas et al. (2012), Breiman (2001)).

Figures 3 and 4 show the two possibilities (in those figures $Conf_{c,j,i}$ refers to the confidence on a prediction on class c, according to the i-th rule of the j-th classifier). The one in Fig.3 considers a unique aggregation step that assigns a final class for a certain individual based on all the conclusions reached by all the rules. The procedure shown in Fig. 4 has

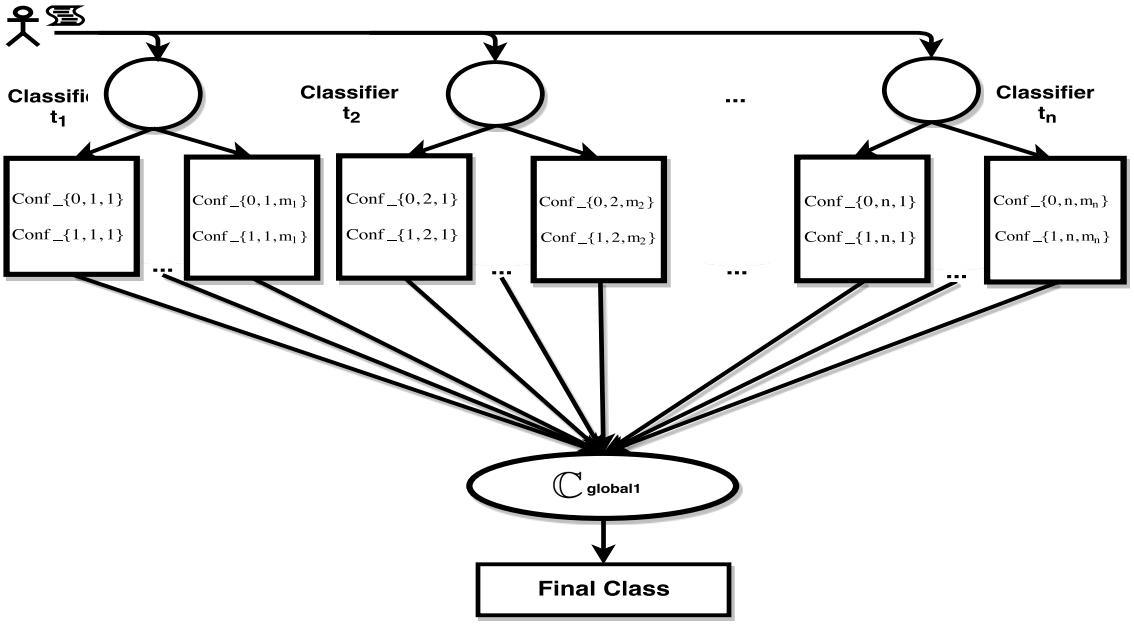

Figure 3: 1-step classification process

two steps. First, a class is assigned to the individual on each classifier by aggregating only the outcomes of the rules of that classifier. Afterwards, the classes proposed by the different classifiers are aggregated to decide the final class.

These two approaches have been used to validate the Fuzzy Random Forest classifiers with different values for its parameters. The aggregation operators applied are based on the Mamdami inference procedure for fuzzy rules using min as t-norm, and max as t-conorm. The confidence on each label is, in this case, the membership degree obtained from the rule.

In the case of the 1-step aggregation procedure, the consensus of the different classifiers is done with the operator shown in Eq. 1, where the output classes are labeled X_0 and X_1, the number of FDTs is n, the j-th tree has m_j branches and $\mu_{X_k,i,j}$ is the confidence on class X_k according to the i-th branch (*i.e.* fuzzy rule) of the j-th tree.

Given all the rules of all the trees, the decision on the final class assignment is done with the aggregation operator:

$$\mathbb{C}_{global1} = max_k(\mu_{X_k,j,i}), \text{ for all } k = \{0,1\},\ j = 1,\ldots,n,\ i = 1,\ldots,m_j. \tag{1}$$

In order to deal with data inconsistency as well as with the possibility of error in the class assignment (due to having rules with low confidence), we have introduced a threshold in order to compare the degree of membership to the two possible classes. If the difference on the membership degree is smaller than the threshold δ_1, the patient is assigned to a class denoted as "Unknown".

In the case of the 2-steps aggregation procedure, a consensus is found for each j-th classifier ($j = 1..n$) with Eq. 2, which may also lead to an "Unknown" assignment. Afterward, a

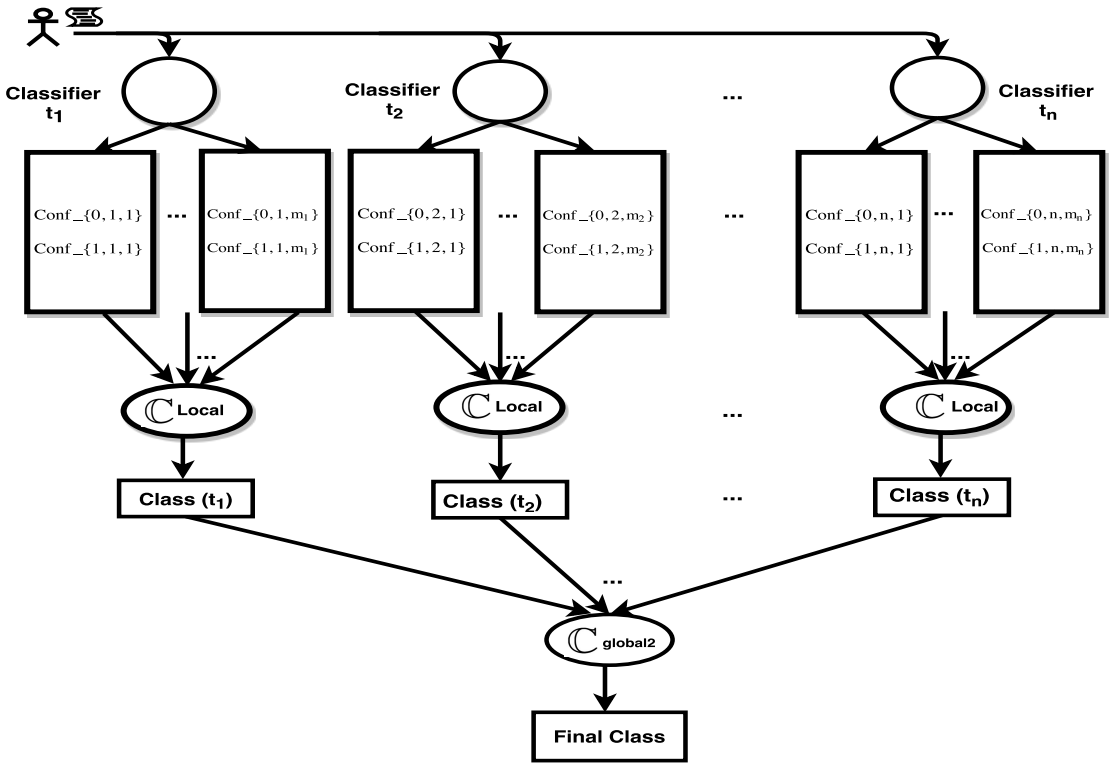

Figure 4: 2-steps classification process

second aggregation operator $\mathbb{C}_{global2}$ is used to merge the output given by each tree by means of a majority voting among the three possible assignments: X_0, X_1 or "Unknown". After voting, if the difference between the two majority classes is lower than a certain threshold δ_2, the class is labeled as *"Unknown"*; otherwise, the final decision is the class with the majority of the votes of the trees.

$$\mathbb{C}_{local}(j) = max_k(\mu_{X_k,j,i}), \text{ for all } k = \{0,1\},\ i = 1,\ldots,m_j. \tag{2}$$

It is worth nothing that the system avoids to make a decision when there is not enough support for one of the two classes in order to deal with the uncertainty and possible inconsistencies that naturally appear in DR assessment. For this reason, two parameters (δ_1 and δ_2) have been introduced in order to detect cases where a patient is not clearly classified with the current rules. In this particular medical diagnosis problem, it is preferred to avoid mistakes (specially false negatives). In this way, the method only classifies objects for which we can find a certain degree of support from the rules of the different trees. Otherwise, the output is *"Unknown"* and the decision is left to the physician, who can observe other evidences from the patient's health record.

3.3. DRSA-BRE in DR detection

In this section the application of the DRSA-BRE method to the case of DR risk assessment is explained. First issue to consider is the nature of the data. In the dataset of DR we

have a set of non-ordinal condition attributes. Therefore, in a pre-processing stage, the value sets of all condition attributes were number-coded. While this is natural for quantitative attributes, nominal attributes must be binarized and get 0-1 codes for absence or presence of a given nominal value. In this way, the value sets of all non-ordinal attributes get ordered (as all sets of numbers are ordered).

In the analysis of DR data, we are considering only two classes, coded 0-1, where 1 corresponds to the high risk of DR, while 0 corresponds to the absence of such risk. In this application where the objects are patients, the syntax of the rules can be written as:

"*if* $q_l(patient) \geq val_l$ *and* $q_j(patient) \geq val_j$ *and* ... *and* $q_r(patient) \geq val_r$, *then class* 1",
"*if* $q_h(patient) \leq val_h$ *and* $q_s(patient) \leq val_s$ *and* ... *and* $q_t(patient) \leq val_t$, *then class* 0",

where q_j is the j-th condition attribute and val_j is a threshold value of this attribute which makes an elementary condition $q_j(patient) \geq val_j$ or $q_j(patient) \leq val_j$ present in the condition part of a rule indicating the assignment of a patient to either class 1 or class 0, respectively. In the above syntax, it is assumed that all condition attributes are number-coded and their value sets are ordered such that the greater the value, the more likely it is that the patient is assigned to class 1; analogously, it is assumed that the smaller the value, the more likely it is that the patient is assigned to class 0. Attributes ordered in this way are called gain-type. Cost-type attributes have value sets ordered in the opposite direction, such that elementary conditions on these attributes have opposite relation signs, i.e., the $\geq$ and $\leq$ signs in the above rules should be inverted for cost-type attributes. In all cases, the threshold values are discovered from data in the course of rule induction.

In case of DR data, it is not possible to know a priori if the considered attributes are gain or cost attributes, and therefore we need to proceed as described in (Błaszczyński et al., 2012): each original attribute is considered in two copies, with the first copy assumed to be gain-type, and the second cost-type. The applied transformation of data is non-invasive, i.e., it does not bias the relationships identified between condition attributes and the class code. Then, the induction algorithm constructs decision rules involving elementary conditions on one or both copies of particular attributes. For example, in a rule indicating the assignment of a patient to class 1, the following elementary conditions concerning attribute q_j may appear:

- $\uparrow q_j(patient) \geq val_j^{\uparrow}$,
- $\downarrow q_j(patient) \leq val_j^{\downarrow}$,
- $\uparrow q_j(patient) \geq val_j^{\uparrow}$ *and* $\downarrow q_j(patient) \leq val_j^{\downarrow}$,
  which boils down to $q_j(patient) \in \left[val_j^{\uparrow}, val_j^{\downarrow}\right]$, when $val_j^{\uparrow} \leq val_j^{\downarrow}$,

where $\uparrow q_j$ and $\downarrow q_j$ are gain-type and cost-type copies of attribute q_j, respectively. Note that this transformation of attributes allows global and local monotonic relationships between attribute values and class assignment to be discovered. The monotonic relationship is global when it can be expressed by a single elementary condition concerning gain-type or cost-type attribute. Local monotonicity relationship requires conjunction of two elementary conditions

of different type, e.g., in case of attribute Age, and assignment of a patient to class 1, a local monotonicity holds in the range of values $[Sixties, Seventies]$ if in the condition part of the rule there are conditions: $\uparrow Age \geq Sixties$ (the more above *Sixties*, the higher the risk of Diabetic Retinopathy) and $\downarrow Age \leq Seventies$ (the more below *Seventies*, the higher the risk of Diabetic Retinopathy), which boils down to condition $Age \in [Sixties, Seventies]$. Examples of rules discovered from DR data by DRSA-BRE are given in Section 4.2.

Decision rules represent the most important cause-effect relationships between values of condition attributes and the value of the decision attribute. The rules are characterized by various parameters, such as *strength* (i.e., the proportion of objects covered by the rule premise that are also covered by the conclusion), *consistency measure* (e.g., measure defined as the ratio of the number of objects covered by the rule premise that belong to the lower approximation of the conclusion class, to the number of objects covered by the rule premise (Błaszczyński et al., 2011b)), or *rule relevance*.

In this study, to assess the relevance of the rules for diabetic retinopathy classification, we use a *Bayesian confirmation measure* that is quantifying the degree to which the rule premise E provides evidence for the conclusion H (Greco et al., 2016). Many Bayesian confirmation measures have been described in the literature, of which we used the measure $s(H, E)$. This measure allows a clear interpretation in terms of a difference of conditional probabilities involving H and E, i.e., $s(H, E) = \Pr(H|E) - \Pr(H|\neg E)$, where probability $\Pr(\cdot)$ is estimated from the information table. In addition, attribute relevance is also calculated with a similar approach based also on Bayesian confirmation measures (details can be found in Appendix B).

A characteristic of this particular application field is the presence of inconsistent examples in the training dataset. Due to undetermined external factors (e.g., personal conditions, genetic data, co-morbidities), we can find two patients with similar values on all the condition attributes but with different decision class (one has developed Diabetic Retinopathy and the other patient not). The result of the inconsistency analysis is presented in Section 4.2.

In this situation, it is better to relax to some extent the definition of the lower approximations, and permit some inconsistent objects to enter the lower approximations. Consequently, in this work we have used a relaxed variant of DRSA called *Variable Consistency DRSA* (VC-DRSA) (Błaszczyński et al., 2009). Therefore, the rules obtained for DR classification are characterized by a consistency measure (Błaszczyński et al., 2011b).

Another feature of the dataset is the imbalance, already mentioned before. The NBBag (Neighbourhood Balanced Bagging) strategy has been applied. It consists on focusing the bootstrap sampling toward the minority examples, in this case patients with DR. First, a global balancing factor is calculated as the class imbalance ratio, which in this case is 28.13%. Second, a local balancing factor is calculated for each positive (minority) patients. Weight of a certain patient is calculated from the analysis of class labels among its k nearest neighbours. The value of k that has been used in this study is 5.

Finally, the aggregation of the output of the rules was done with the majority voting technique. In this case, thus, the merging is done with 1-step aggregation.

4. Experimental Results

In this section we study the results obtained with the FRF and DRSA-BRE methods on the data set presented in Section 3.1. A comparison of the performance of both methods is also done. The goal is to achieve enough quality in the classification of diabetic people in two classes: class0 (low risk of DR - negative class), and class1 (high risk of DR - positive class). Different values of the parameters of the algorithms are studied. Sensitivity (also called *recall*) and specificity (Equation 3) are the measures used to evaluate the performance of each classifier, as commonly used in the medical domain. In addition, accuracy has been calculated in two ways: (Acc1) the number of correct classifications divided by the total number of instances to be classified, and (Acc2) the number of correct classifications divided by the total number of classified instances (removing the ones classified as *Unknown*).

$$Sensitivity = \frac{TP}{TP + FN}, \qquad Specificity = \frac{TN}{TN + FP} \tag{3}$$

4.1. Results of the FRF analysis

Several parameters are used in the fuzzy random forest method. The values shown in Table 1 have been considered in the tests.

α	β	γ	δ_1	δ_2	n
0, 0.1, ..., 1	0, 0.1, ..., 1	1,2,3,4	0, 0.1, ..., 0.5	0%,5%,10%	100,200,300

Table 1: Parameters in the FRF models

Each combination of values of the parameters was used to train and validate a model using 10-fold cross-validation on the training set. The best configurations of parameters are shown in the tables of this section (both for the 1-and 2-steps aggregation procedures). In each case four models were selected, optimizing 4 different indexes: Acc1, Acc2, Specificity and Sensitivity. Averaged values of the 10-folds are given on the corresponding tables. After that, those models were constructed again using the entire training data set and applied to the testing set. Due to the randomization of the FRF algorithm, each model was trained and tested 5 times and the results of the best one are shown in the following tables.

Tables 2 (10-fold cross-validation on the training set) and 3 (Testing) show the results for the 2-steps aggregation. The best models were found with 100 trees, $\gamma = \{1, 2\}$, δ_1= 0.2, a high $\beta = \{0.8, 1\}$ and 5-10% in δ_2. The performance is much better with the testing set than with the training set, probably due to the fact that the size of the validation fold is relatively small (121 patients). These tables also show the number of the true positives (TP), false negatives (FN), false positives (FP), true negatives (TN) and unclassified patients (Unk).

When using the one-step aggregation of the outputs of each rule, the best results are found with smaller values of δ_1 (0-0.2), but the rest of parameter values are similar (see Table 4). The cross-validation results are better than those of the 2-steps aggregation procedure in terms of Acc2, sensitivity and specificity; however, the results on the testing set are worse in terms of specificity, Acc1 and number of unclassified objects. Therefore, the two-steps aggregation procedure seems to be the best option in this case.

n	δ_2	δ_1	γ	α	β	Acc2	Sens	Spec	Acc1	TP	FN	FP	TN	Unk	Ind
100	10%	0.2	1	0.5	1.0	76.76	71.93	78.52	70.01	22	8	17	62	12	Acc2
100	5%	0.2	1	0.4	0.8	74.50	70.10	76.21	71.90	23	10	20	64	4	Acc1
100	10%	0.2	1	0.8	0.8	76.34	68.12	79.49	68.42	21	10	16	62	13	Spec
100	10%	0.2	2	0.4	1	75.01	74.22	75.31	68.26	23	8	20	60	11	Sens

Table 2: Cross-validation results with FRF with two-steps classification technique

n	δ_2	δ_1	γ	α	β	Acc2	Sens	Spec	Acc1	TP	FN	FP	TN	Unk	Ind
100	10%	0.2	1	0.5	1.0	84.23	80.38	85.25	75.97	168	41	117	676	109	Acc2
100	5%	0.2	1	0.4	0.8	80.93	83.93	80.12	76.78	188	36	165	665	90	Acc1
100	10%	0.2	1	0.8	0.8	82.75	82.79	82.74	76.42	178	37	140	671	85	Spec
100	10%	0.2	2	0.4	1	81.26	81.33	81.23	76.87	183	42	155	671	60	Sens

Table 3: Best classification results with FRF in the testing phase with two-steps aggregation

						Cross validation				Testing					
n	δ_2	δ_1	γ	α	β	Acc2	Sens	Spec	Acc1	Acc2	Sens	Spec	Acc1	Unk	Index
100	10%	0.1	1	0.7	0.8	78.48	71.94	80.96	65.45	84.94	86.07	84.64	72.10	168	Acc2
100	0%	0.2	1	0.4	0.8	74.71	70.88	76.21	74.71	80.09	79.83	80.07	74.71	0	Acc1
100	10%	0	1	0.4	1.0	77.76	70.99	80.35	65.37	81.95	87.02	80.52	69.48	169	Spec
100	10%	0.2	2	0.3	1	75.89	75.61	76.06	67.02	82.25	83.56	81.89	75.07	97	Sens

Table 4: Cross-validation results and best classification results with FRF with the one-step classification technique

As the DRSA-BRE method classifies all the objects, in order to make a fair comparison with FRF the same analysis was repeated, but the parameters δ_1 and δ_2 were set to 0 to force the classification of all the patients. Table 5 shows the performance indexes for this case (with 100 trees). The values are lower than in the previous tables because the model makes more mistakes in those patients for which there is not a clear consensus on the predictions made by the different classifiers.

						Cross validation			Testing			
n	δ_2	δ_1	γ	α	β	Acc2	Sens	Spec	Acc2	Sens	Spec	Index
100	0%	0	1	0.7	1.0	74.49	70.71	75.91	80.05	81.78	79.58	Acc
100	0%	0	4	1.0	1.0	70.16	73.89	68.71	75.34	86.97	72.14	Sens
100	0%	0	1	0.5	0.8	73.72	68.24	80.18	80.29	80.67	80.18	Spec

Table 5: Cross-validation results and best classification results with FRF with the two-steps aggregation with δ_1 =0 and δ_2=0%

4.2. Results of the DRSA-BRE analysis

The DRSA-BRE analysis starts by checking the consistency of the training set. The results of this analysis, presented in Table 6, show that there is a relatively high level of inconsistency in the training set, i.e. there are many cases in which it is possible to find patients of both classes that have the same values in the nine attributes.

	class0 (low risk of DR)	class1 (high risk of DR)
Lower approximation	815	295
Upper approximation	917	397
Boundary[a]	102	102
Accuracy of approximation[b]	0.889	0.743

[a] Difference between lower and upper approximation, [b] Ratio of the number of patients in the lower approximation to the number of patients in the upper approximation

Table 6: Number of patients consistent with the assignment to its class and resulting accuracy of the approximation

Taking into account this remarkable degree of inconsistency in the training set, the DRSA-BRE model resulting from applying the *Neighbourhood Balanced Bagging* (NBBag) method with *the VC-DomLEM component classifiers* showed a good classification performance after applying a 10-fold stratified cross validation, in which the cross validation procedure was performed several times to reduce the effect of randomness. On average, the results were a 72.84% level of accuracy (percentage of correctly classified cases), a sensitivity of 73.12% and a specificity of 72.56%. In the validation with the testing set, the best results yielded an accuracy of 77.32%, sensitivity of 76.89% and a specificity of 77.43%

The values of the Bayesian confirmation measure calculated for all condition attributes give more insight into the constructed classification model (see Figure 5).

The attributes with the highest values of the confirmation measure s (see Appendix B for more information about s) are those which are the most relevant from the viewpoint of correct prediction by the DRSA-BRE classifier. These attributes are medical treatment (TTM), hypertension (HTAR) and age. The less relevant attributes are sex, Body Mass Index (BMI) and creatinine (Creat).

This analysis can be followed by an examination of selected decision rules. We show below several rules that classify patients in class 1, distinguished by relatively high support and confirmation value. These rules usually involve the more relevant attributes, which were presented above in Figure 5. Note, however, that a non-relevant attribute may also appear in relevant rules when it is also present in some non-relevant rules (this co-occurrence decreases its relevance). This is the case of Sex. As we are presenting below only some relevant rules, it would be wrong to consider all attributes present in these rules as relevant. Let us also remind that a value of a confirmation measure s for a rule is different from the value of s for an attribute - they are two different applications of s. In both cases, however, $s \in [-1, 1]$, and the closer is the s-value to 1, the more relevant is either the rule or the attribute.

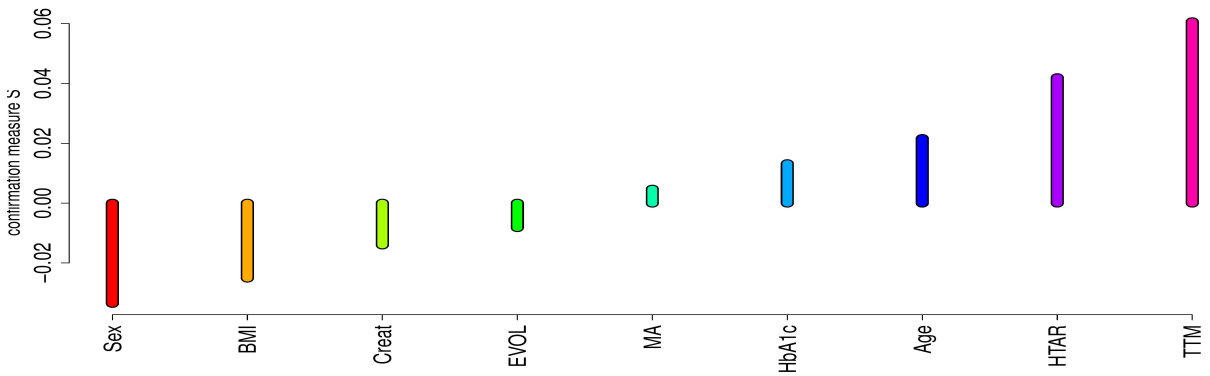

Figure 5: Bayesian confirmation measure computed for each condition attribute. This plot shows how each of the variables used in the DRSA-BRE model confirms the correct classification of DR patients.

#1: "*if* $(Age \in [Sixties, Seventies])$ *and* $(Sex = Man)$ *and* $(TTM \in [Diet, OralAntidiab])$ *and* $(HbA1c \geq 8to9)$ *and* $(MA = Correct)$ *and* $(HTAR = badControl)$,
then the patient has DR (class1)",
Support: 19 patients, s : 0.725

#2: "*if* $(Age \geq Fifties)$ *and* $(Sex = Man)$ *and* $(EVOL \geq 10to15)$ *and* $(HbA1c \geq 8to9)$ *and* $(MA = Correct)$ *and* $(HTAR = badControl)$,
then the patient has DR (class1)",
Support: 32 patients, s : 0.725

#3: "*if* $(EVOL \leq 10to15)$ *and* $(TTM = Insuline)$ *and* $(Creat = Normal)$ *and* $(MA = Correct)$ *and* $(BMI \leq ObeseLow)$ *and* $(HTAR = badControl)$,
then the patient has DR (class1)",
Support: 34 patients, s : 0.731

#4: "*if* $(EVOL \geq 5to10)$ *and* $(TTM = Insuline)$ *and* $(HbA1c \leq 7to8)$ *and* $(Creat \leq Normal)$ *and* $(BMI \geq OverweightHigh)$ *and* $(HTAR = badControl)$,
then the patient has DR (class1)",
Support: 26 patients, s : 0.75

#5: "*if* $(Age \leq Sixties)$ *and* $(Sex = Man)$ *and* $(TTM = Insuline)$ *and* $(HbA1c \leq 8to9)$ *and* $(MA = Correct)$ *and* $(HTAR = badControl)$,
then the patient has DR (class1)",
Support: 25 patients, s : 0.737

#6: "*if* $(TTM = Insuline)$ *and* $(HbA1c \in [7to8, 8to9])$ *and* $(Creat \leq Normal)$ *and* $(MA = Correct)$ *and* $(BMI \geq ObeseLow)$ *and* $(HTAR = badControl)$,
then the patient has DR (class1)",
Support: 23 patients, s : 0.731

4.3. Comparison between FRF and DRSA-BRE

In this section the two ensemble classifiers proposed in this work are compared using the best parameters obtained using the cross validation. DRSA-BRE is an ensemble of 50 classifiers, whereas FRF employs 100 FDTs. In the cross-validation stage DRSA-BRE obtained better sensitivity results, although the accuracy and specificity of FRF were better (see Table 7).

Model	n	Accuracy	Sensitivity	Specificity
FRF	100	74.49	70.71	75.91
DRSA	50	72.84	73.12	72.56

Table 7: Best cross validation results of FRF and DRSA

The same parameters were used to validate the models on the testing data set. FRF gave better accuracy, sensitivity and specificity results with all values slightly over 80% (see Table 8).

Model	n	Accuracy	Sensitivity	Specificity	TP	FN	FP	TN
FRF	100	80.29	80.67	80.18	192	46	173	700
DRSA	50	77.32	76.89	77.43	183	55	197	676

Table 8: Comparison between FRF and DRSA models on the testing data set using the best parameters obtained using cross validation

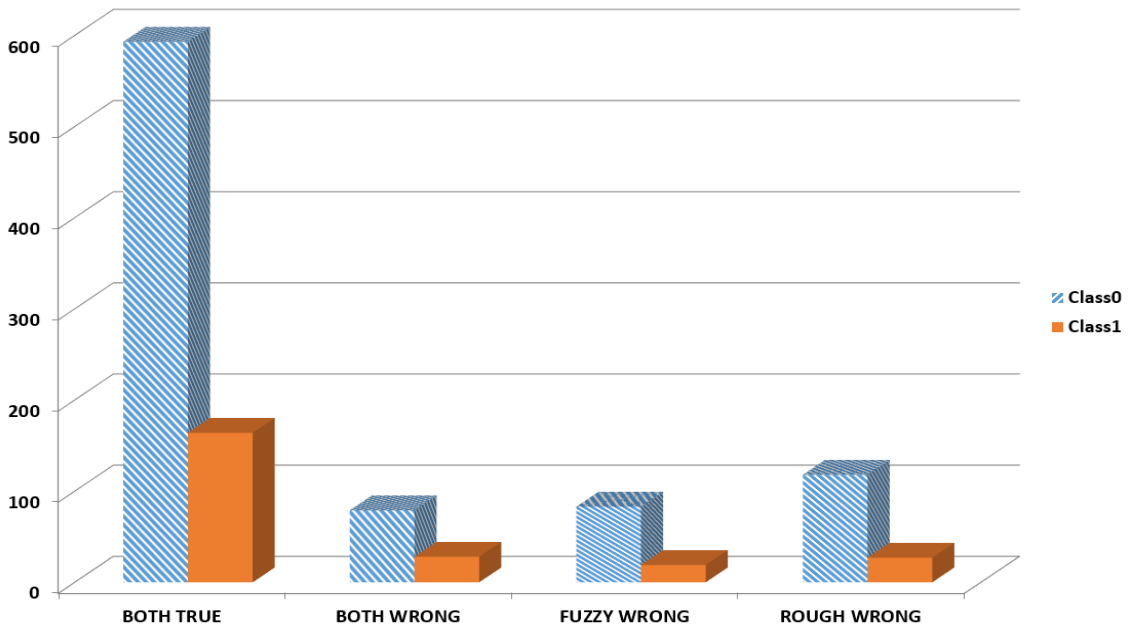

Figure 6: Comparison FRF and DRSA-BRE algorithms

A comparison of the diagnosis made for each patient in the testing set (1111 patients) was done. Figure 6 shows the number of patients that are classified correctly and incorrectly by each method for each class. We divided the results into four groups: patients that are classified correctly by both methods, patients incorrectly classified by both methods and patients who are wrongly classified by only one method (i.e. Fuzzy wrong, Rough wrong). Out of this study, it can be observed that a high percentage of the cases are classified correctly. The figure shows that both models make more mistakes with the examples from class0 than from class1, which corresponds to False Positives. The number of these patients is low (less than 100 in each class). With these patients, it is now challenging to study what motivates the wrong classification in order to improve the models.

5. Conclusions and future work

Ensemble classifiers have demonstrated good performance in many medical applications. In this paper, they have been applied to the case of Diabetic Retinopathy risk assessment. Two learning methods, which take special care of data uncertainty and inconsistency, have been analyzed.

On the one hand, Fuzzy Random Forests have been able to achieve an accuracy of 84% on the testing dataset, with an 80% sensitivity and an 85% specificity (however, almost 10% of the patients could not be classified). These results improve those of our previous works (Sanromà et al., 2016; Saleh et al., 2016), in which 80% levels were achieved using a non-fuzzy Random Forest (although, in that case, the number of unclassified objects was almost null). Training performance is always worse than in testing. This may be due to the imbalance on the data, as the cross-validation folders are more affected than the large testing set. Additionally, two aggregation methods for the fuzzy rules have been applied and compared (one-step and two-steps). The 2-steps aggregation procedure is recommended because it offers a better performance on the testing set, specially in terms of specificity, accuracy1 and the number of unknown predictions.

On the other hand, the DRSA-BRE method also obtains good results, with accuracy, sensitivity and specificity levels around 77% on the testing dataset (contrary to FRF, however, these results were obtained while classifying all the patients). The results obtained with DRSA-BRE are, moreover, affected by the discretization of attribute domains. This discretization is not necessary from the perspective of the method. It was demonstrated in the experiments that the discretization introduced a significant amount of inconsistency into the training set. Future work should include improving the proposed discretization intervals in close collaboration with medical experts. The DRSA-BRE method provides not only an initial analysis of the inconsistency of the data, but also a set of decision rules that are easy to understand by the doctors, and an analysis of the relevance of each attribute in the classification process, which give additional insight into the problem.

Currently, a clinical decision support system (CDSS) for Catalan health care centers is being developed out of this work. This CDSS will be used by family physicians who are the ones that regularly have control visits with diabetic patients. As they are not experts in ophthalmology, this CDSS will help them to evaluate the risk of developing retinopathy.

In a first stage, the CDSS would analyze the patients' data stored in the EHR (using the classifiers presented in this paper). If the risk of developing DR would be null or low, the next visit would be after two or three years. Otherwise, an eye fundus photo using a non-mydriatic camera would be taken by an specialist. In a second stage (research still in progress), this photo would be provided to the CDSS, which would apply computer vision methods to detect if there any harms in the blood vessels or intraocular haemorrhages (Escorcia-Gutiérrez et al., 2016; La Torre et al., 2016). The results would be shown to the physician and the next visit would be scheduled. This tool will avoid the unnecessary screening of the eye fundus to many diabetic patients and will help to improve the use of the human and material resources of the hospital, without reducing the quality of the health care.

To improve the system in its first stage, two main lines of work should be addressed: first, study the combination of the FRF and DRSA-BRE methods to improve the classification accuracy; second, consider the possibility of providing a linguistic explanation of the classification made by the system.

Hybrid approaches that combine rule-based systems with other machine learning techniques have been used successfully for diagnosis and monitoring of diabetic patients (Cvetković et al., 2016; Kafali et al., 2014). In this line, the combination of the two methods studied in this paper seems promising because when both classifiers predict the same class they are usually right (both classifiers are wrong only in 8-9% of the cases). The high degree of inconsistency in the data makes it impossible to achieve a 100% of correct classifications. There is still some space for improvement, however, because there are about 20% of the cases in which the answers differ, so one of the methods is giving the correct answer. It may also be possible to construct a hybrid fuzzy-rough classification method. However, such a hybrid method, although possibly better in terms of the capacity of prediction, would likely lose on understandability/readability. On the contrary, the integration of the outputs of two separate classifiers could be more helpful for a physician because both methods provide easily interpretable rules. For each patient, the rules matching this patient could be presented to the doctor. The number of matching rules might be large, however, thus requiring the selection of only the most relevant ones in order to facilitate the work of the medical personnel. Thus, a procedure for finding commonalities in the rules coming from both methods would be worth studying.

Acknowledgements

The authors from Univ. Rovira i Virgili acknowledge the support given by the Spanish research projects PI15/01150 and PI12/01535 (Instituto de Salud Carlos III and FEDER funds) and the URV grants 2015PFR-URV-B2-60 and 2016PFR-URV-B2-60. The first author would like to acknowledge the Martí-Franquès URV research fellowship programme for the research funding 2016PMF-PIPF-24. The research of the second author was funded by the Polish National Science Center, grant no. DEC-2013/11/B/ST6/00963. Close cooperation with Marcin Szelag at the initial stage of the DRSA analysis is also acknowledged.

References

ADA (2015). American Diabetes Association. Standards of medical care in diabetes. Microvascular complications and foot care. *Diabetes Care*, *38*, 858–866.

Alghadyan, A. A. (2011). Diabetic retinopathy–an update. *Saudi Journal of Ophthalmology*, *25*, 99–111.

Alickovic, E., & Subasi, A. (2016). Medical decision support system for diagnosis of heart arrhythmia using DWT and random forests classifier. *Journal of medical systems*, *40*, 1–12.

Beloufa, F., & Chikh, M. (2013). Design of fuzzy classifier for diabetes disease using modified artificial bee colony algorithm. *Computer Methods and Programs in Biomedicine*, *112*, 92 – 103.

Blasco, H., Błaszczyński, J., Billaut, J., Nadal-Desbarats, L., Pradat, P.-F., Devos, D., Moreau, C., Andres, C. R., Emond, P., Corcia, P. et al. (2015). Comparative analysis of targeted metabolomics: Dominance-based rough set approach versus orthogonal partial least square-discriminant analysis. *Journal of biomedical informatics*, *53*, 291–299.

Błaszczyński, J., Greco, S., & Słowiński, R. (2012). Inductive discovery of laws using monotonic rules. *Engineering Applications of Artificial Intelligence*, *25*, 284–294.

Błaszczyński, J., Greco, S., Słowiński, R., & Szelkag, M. (2009). Monotonic variable consistency rough set approaches. *International journal of approximate reasoning*, *50*, 979–999.

Błaszczyński, J., Słowiński, R., & Susmaga, R. (2011a). Rule-based estimation of attribute relevance. In *International Conference on Rough Sets and Knowledge Technology* (pp. 36–44). Springer.

Błaszczyński, J., Słowiński, R., & Szelkag, M. (2011b). Sequential covering rule induction algorithm for variable consistency rough set approaches. *Information Sciences*, *181*, 987–1002.

Błaszczyński, J., & Stefanowski, J. (2015). Neighbourhood sampling in bagging for imbalanced data. *Neurocomputing*, *150*, 529–542.

Bourne, R. R. A., Stevens, G. A., White, R. A., Smith, J. L., Flaxman, S. R., Price, H., Jonas, J. B., Keeffe, J., Leasher, J., Naidoo, K., Pesudovs, K., Resnikoff, S., & Taylor, H. R. (2013). Causes of vision loss worldwide, 19902010: a systematic analysis. *The Lancet Global Health*, *1*, e339 – e349.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*, 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Cadenas, J. M., Garrido, M. C., & Martínez, R. (2013). Feature subset selection filter–wrapper based on low quality data. *Expert systems with applications*, *40*, 6241–6252.

Cadenas, J. M., Garrido, M. C., Martínez, R., & Bonissone, P. P. (2012). Extending information processing in a fuzzy random forest ensemble. *Soft Computing*, *16*, 845–861.

Cadenas, J. M., Garrido, M. C., Martínez, R., Pelta, D., & Bonissone, P. P. (2016). Gene priorization for tumor classification using an embedded method. In *Computational Intelligence* (pp. 363–380). Springer.

Chalk, D., Pitt, M., Vaidya, B., & Stein, K. (2012). Can the retinal screening interval be safely increased to 2 years for type 2 diabetic patients without retinopathy? *Diabetes care*, *35*, 1663–1668.

Chandrakumar, T., & Kathirvel, R. (2016). Classifying diabetic retinopathy using deep learning architecture. *International Journal of Engineering Research and Technology*, *5*.

Cinelli, M., Coles, S. R., Nadagouda, M. N., Błaszczyński, J., Słowiński, R., Varma, R. S., & Kirwan, K. (2015). A green chemistry-based classification model for the synthesis of silver nanoparticles. *Green Chemistry*, *17*, 2825–2839.

Control, D., Group, C. T. R. et al. (1993). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N Engl j Med*, *1993*, 977–986.

Cvetković, B., Janko, V., Romero, A. E., Kafalı, Ö., Stathis, K., & Luštrek, M. (2016). Activity recognition for diabetic patients using a smartphone. *Journal of Medical Systems*, *40*, 256.

Escorcia-Gutiérrez, J., Torrents-Barrena, J., Romero-Aroca, P., Valls, A., & Puig, D. (2016). Interactive optic disk segmentation ia discrete convexity shape knowledge using high-order functionals. In *Artificial Intelligence Research and Development: Proceedings of the 19th International Conference of the Catalan Association for Artificial Intelligence, Barcelona, Catalonia, Spain, October 19-21, 2016* (p. 39). IOS Press volume 288.

Gardiner, E. J., & Gillet, V. J. (2015). Perspectives on knowledge discovery algorithms recently introduced

in chemoinformatics: Rough set theory, association rule mining, emerging patterns, and formal concept analysis. *Journal of chemical information and modeling*, *55*, 1781–1803.

Greco, S., Matarazzo, B., & Słowiński, R. (2001). Rough sets theory for multicriteria decision analysis. *European journal of operational research*, *129*, 1–47.

Greco, S., Słowiński, R., & Szczkech, I. (2016). Measures of rule interestingness in various perspectives of confirmation. *Information Sciences*, *346*, 216–235.

Early Treatment Diabetic Retinopathy Study Research Group (1991). Grading diabetic retinopathy from stereoscopic color fundus photographsan extension of the modified airlie house classification: {ETDRS} report number 10. *Ophthalmology*, *98*, 786 – 806.

Haloi, M. (2015). Improved microaneurysm detection using deep neural networks. *arXiv preprint arXiv:1505.04424*, .

IDF (2013). International Diabetes Federation. *Diabetes Atlas 2013*, .

IDF (2015). International Diabetes Federation. *Diabetes Atlas 7th edition, 2015*, .

IDF (2016). International Diabetes Foundation. *Diabetes Atlas 2016*, .

Jiang, R., Bouridane, A., Crookes, D., Celebi, M. E., & Wei, H.-L. (2016). Privacy-protected facial biometric verification using fuzzy forest learning. *IEEE Transactions on Fuzzy Systems*, *24*, 779–790.

Kafali, Ö. ., Schaechtle, U., & Stathis, K. (2014). Hydra: A hybrid diagnosis and monitoring architecture for diabetes. In *2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom)* (pp. 531–536). doi:10.1109/HealthCom.2014.7001898.

Khan, A., & Revett, K. (2004). Data mining the pima dataset using rough set theory with a special emphasis on rule reduction. In *Multitopic Conference, 2004. Proceedings of INMIC 2004. 8th International* (pp. 334–339). IEEE.

Koley, S., Sadhu, A. K., Mitra, P., Chakraborty, B., & Chakraborty, C. (2016). Delineation and diagnosis of brain tumors from post contrast T1-weighted MR images using rough granular computing and random forest. *Applied Soft Computing*, *41*, 453–465.

Kosko, B. (1986). Fuzzy entropy and conditioning. *Information sciences*, *40*, 165–174.

Kovacs, W., Liu, C.-Y., Summers, R. M., & Yao, J. (2016). Differentiation of fat, muscle, and edema in thigh MRIs using random forest classification. In *SPIE Medical Imaging* (pp. 978507–978507). International Society for Optics and Photonics.

Kulkarni, V. Y., & Sinha, P. K. (2013). Random forest classifiers: a survey and future research directions. *International Journal of Advanced Computing*, *36*, 1144–1153.

La Torre, J. d., Valls, A., & Puig, D. (2016). Diabetic retinopathy detection hrough image analysis using deep convolutional neural networks. In *Artificial Intelligence Research and Development: Proceedings of the 19th International Conference of the Catalan Association for Artificial Intelligence, Barcelona, Catalonia, Spain, October 19-21, 2016* (p. 58). IOS Press volume 288.

Lukmanto, R. B., & Irwansyah, E. (2015). The early detection of diabetes mellitus (dm) using fuzzy hierarchical model. *Procedia Computer Science*, *59*, 312 – 319.

Marsala, C. (2009). A fuzzy decision tree based approach to characterize medical data. In *Fuzzy Systems, 2009. FUZZ-IEEE 2009. IEEE International Conference on* (pp. 1332–1337). IEEE.

Meza-Palacios, R., Aguilar-Lasserre, A. A., Ureña-Bogarín, E. L., Vázquez-Rodríguez, C. F., Posada-Gómez, R., & Trujillo-Mata, A. (2017). Development of a fuzzy expert system for the nephropathy control assessment in patients with type 2 diabetes mellitus. *Expert Systems with Applications*, *72*, 335–343.

Nakayama, H., Hattori, Y., & Ishii, R. (1999). Rule extraction based on rough set theory and its application to medical data analysis. In *Systems, Man, and Cybernetics, 1999. IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on* (pp. 924–929). IEEE volume 5.

Napierala, K., & Stefanowski, J. (2016). Types of minority class examples and their influence on learning classifiers from imbalanced data. *Journal of Intelligent Information Systems*, *46*, 563–597.

Olafsdottir, E., & Stefansson, E. (2007). Biennial eye screening in patients with diabetes without retinopathy: 10-year experience. *British journal of ophthalmology*, *91*, 1599–1601.

Pałkowski, Ł., Błaszczyński, J., Skrzypczak, A., Błaszczak, J., Kozakowska, K., Wróblewska, J., Kożuszko, S., Gospodarek, E., Krysiński, J., & Słowiński, R. (2014a). Antimicrobial activity and sar study of new

gemini imidazolium-based chlorides. *Chemical biology & drug design*, *83*, 278–288.

Pałkowski, Ł., Błaszczyński, J., Skrzypczak, A., Błaszczak, J., Nowaczyk, A., Wróblewska, J., Kożuszko, S., Gospodarek, E., Słowiński, R., & Krysiński, J. (2015). Prediction of antifungal activity of gemini imidazolium compounds. *BioMed research international*, *2015*.

Pałkowski, Ł., Krysiński, J., Błaszczyński, J., Słowiński, R., Skrzypczak, A., Błaszczak, J., Gospodarek, E., & Wróblewska, J. (2014b). Application of rough set theory to prediction of antimicrobial activity of bis-quaternary imidazolium chlorides. *Fundamenta Informaticae*, *132*, 315–330.

Pawlak, Z. (1991). Rough sets: Theoretical aspects of reasoning about data. , .

Popescu, D., & Ichim, L. (2015). Computeraided localization of the optic disc based on textural features. In *Advanced Topics in Electrical Engineering (ATEE), 2015 9th International Symposium on* (pp. 307–312). IEEE.

Pratt, H., Coenen, F., Broadbent, D. M., Harding, S. P., & Zheng, Y. (2016). Convolutional neural networks for diabetic retinopathy. *Procedia Computer Science*, *90*, 200–205.

RCO (2012). The Royal College of Ophthalmologists. Diabetic retinopathy guidelines. URL: `rcophth.ac.uk`.

Romero Aroca, P., Reyes Torres, J., Sagarra Alamo, R., Basora Gallisa, J., Fernández-Balart, J., Pareja Ríos, A., & Baget-Bernaldiz, M. (2012). Resultados de la implantación de la cámara no midriática sobre la población diabética. *Salud (i) cienc.*, *19*, 214–219.

Romero-Aroca, P., de la Riva-Fernandez, S., Valls-Mateu, A., Sagarra-Alamo, R., Moreno-Ribas, A., & Soler, N. (2016). Changes observed in diabetic retinopathy: eight-year follow-up of a spanish population. *British Journal of Ophthalmology*, *100*, 1366–1371. URL: `http://bjo.bmj.com/content/100/10/1366`. doi:10.1136/bjophthalmol-2015-307689.

Saleh, E., Valls, A., Moreno, A., Romero-Aroca, P., de la Riva-Fernandez, S., & Sagarra-Alamo, R. (2016). Diabetic retinopathy risk estimation using fuzzy rules on electronic health record data. In *Modeling Decisions for Artificial Intelligence* (pp. 263–274). Springer.

Saleh, M. D., & Eswaran, C. (2012). An automated decision-support system for non-proliferative diabetic retinopathy disease based on mas and has detection. *Computer methods and programs in biomedicine*, *108*, 186–196.

Sanromà, S., Moreno, A., Valls, A., Romero, P., de la Riva, S., & Sagarra, R. (2016). Assessment of diabetic retinopathy risk with random forests. In *24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2016)* (pp. 313–318).

Skevofilakas, M., Zarkogianni, K., Karamanos, B. G., & Nikita, K. S. (2010). A hybrid decision support system for the risk assessment of retinopathy development as a long term complication of type 1 diabetes mellitus. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE* (pp. 6713–6716). IEEE.

Słowiński, R., Greco, S., & Matarazzo, B. (2014). Rough-set-based decision support. In *Search Methodologies* (pp. 557–609). Springer.

Słowiński, R., Greco, S., & Matarazzo, B. (2015). Rough set methodology for decision aiding. In *Springer Handbook of Computational Intelligence* (pp. 349–370). Springer.

Stepaniuk, J. (1999). Rough set data mining of diabetes data. *Foundations of Intelligent Systems*, (pp. 457–465).

Su, C.-T., Yang, C.-H., Hsu, K.-H., & Chiu, W.-K. (2006). Data mining for the diagnosis of type ii diabetes from three-dimensional body surface anthropometrical scanning data. *Computers & Mathematics with Applications*, *51*, 1075–1092.

Trawiński, K., Alonso, J. M., & Hernández, N. (2013). A multiclassifier approach for topology-based wifi indoor localization. *Soft Computing*, *17*, 1817–1831.

Wang, S., Yin, Y., Cao, G., Wei, B., Zheng, Y., & Yang, G. (2015). Hierarchical retinal blood vessel segmentation based on feature and ensemble learning. *Neurocomputing*, *149*, 708–717.

Wang, X., Yeung, D. S., & Tsang, E. C. C. (2001). A comparative study on heuristic algorithms for generating fuzzy decision trees. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *31*, 215–226.

Wilkinson, C., Ferris, F. L., Klein, R. E., Lee, P. P., Agardh, C. D., Davis, M., Dills, D., Kampik, A.,

Pararajasegaram, R., Verdaguer, J. T. et al. (2003). Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, *110*, 1677–1682.

Yao, Y., Greco, S., & Słowiński, R. (2015). Probabilistic rough sets. In *Springer Handbook of Computational Intelligence* (pp. 387–411). Springer.

Yuan, Y., & Shaw, M. J. (1995). Induction of fuzzy decision trees. *Fuzzy Sets and systems*, *69*, 125–139.

Zhang, L., Wei, H., Zhu, J., de la Cruz, J., Gonzalez, H. J., & Yadegar, J. (2012). An ontology-based multi-class terrain surface classification system for aerial imagery. In *Emerging Signal Processing Applications (ESPA), 2012 IEEE International Conference on* (pp. 95–98). IEEE.

Zhang, W., Liu, H., Al-Shabrawey, M., Caldwell, R. W., & Caldwell, R. B. (2011). Inflammation and diabetic retinal microvascular complications. *Journal of cardiovascular disease research*, *2*, 96–103.

Appendix A. Yuan and Shaw's Fuzzy Decision Tree Induction

For the sake of completeness, the notation and the formulation of the different measures used in the algorithm for fuzzy decision tree induction defined by (Yuan & Shaw, 1995) induction are given here.

Let us consider the universe of discourse $U = \{x_1, x_2, ..., x_m\}$, where x_i is an object described by a collection of attributes $C = \{q_1, ..., q_r\}$. Each attribute $q_j \in C$ takes values on a linguistic fuzzy partition $T = \{t_1, ..., t_s\}$ with membership functions $\mu_{t_i} \in \mu_T$. These membership functions can be understood as a possibility distribution. The *U-uncertainty* (or non-specificity measure) of a possibility distribution π on a set $Y = \{y_1.y_2, ..., y_d\}$ is defined as:

$$g(\pi) = \sum_{i=1}^{d} (\pi_i^* - \pi_{i+1}^*) \ln i \tag{A.1}$$

where $\pi^* = \{\pi_1^*, \pi_2^*, ..., \pi_d^*\}$ is a permutation of $\pi = \{\pi(y_1), \pi(y_2), ..., \pi(y_d)\}$ such that $\pi_i^* \geq \pi_{i+1}^*$, for $i = 1, ..., d$, and $\pi_{d+1}^* = 0$. The *ambiguity* of an attribute q_j $(j = 1, \ldots, r)$ is calculated as

$$Ambiguity(q_j) = \frac{1}{m} \sum_{i=1}^{m} g(\pi_T(x_i)) \tag{A.2}$$

where π_T is the normalized possibility distribution of μ_T on U:

$$\pi_{t_r}(x_i) = \mu_{t_r}(x_i) / max_{1 \leq j \leq s}\{\mu_{t_j}(x_i)\} \tag{A.3}$$

Having a set of classes $\mathbf{X} = \{X_1, ..., X_p\}$, the *truth level of classification* indicates the possibility of classifying an object u_i into a class $X_k \in \mathbf{X}$ given the fuzzy evidence E, which is a fuzzy set defined on the linguistic values taken by one or more attributes (*i.e.* a condition given by one branch of the decision tree).

$$Truth(X_k|E) = S(E, X_k) / max_{1 \leq j \leq p}\{S(E, X_j)\} \tag{A.4}$$

where S is the subsethood of the fuzzy set A on the fuzzy set B

$$S(A, B) = \frac{M(A \cap B)}{M(A)} = \frac{\sum_{i=1}^{m} min(\mu_A(x_i), \mu_B(x_i))}{\sum_{i=1}^{m} \mu_A(x_i)} \tag{A.5}$$

and $M(A)$ is the cardinality or sigma count of the fuzzy set A.

The truth level can be understood as a possibility distribution on the set U, and $\pi(\mathbf{X}|E)$ is the corresponding normalisation.

Having a fuzzy partition $P = \{E_1, ..., E_k\}$ on fuzzy evidence F, the *classification ambiguity*, denoted by $G(P|F)$, is calculated as

$$G(P|F) = \sum_{i=1}^{k} W(E_i|F) g(\pi(\mathbf{X}|E_i \cap F)) \tag{A.6}$$

where $W(E_i|F)$ is the weight which represents the relative size of the subset $(E_i \cap F)$ with respect to F (*i.e.* $W(E_i|F) = M(E_i \cap F)/\sum_{i=1}^{k} M(E_i \cap F)$).

Appendix B. Description of Dominance-based Rough Set Balanced Rule Ensemble (DRSA-BRE)

The description of DRSA-BRE starts with the *Dominance-based Rough Set Approach* (DRSA), and proceeds with the explanation of a bagging method developed for class imbalanced data and used to construct an ensemble classifier called *Balanced Rule Ensemble* (BRE).

The set of attributes is divided into a set $C = \{q_1, q_2, \ldots, q_r\}$ of condition attributes and a set $D = \{d\}$ with the decision attribute designating class labels. Condition attributes whose value sets are ordered are called *ordinal attributes*. Without loss of generality, given an ordinal attribute $q_j \in C$, $\phi : U \to \mathbb{R}$, for all objects $x_i, x_h \in U, \phi(x_i) \geq \phi(x_h)$ means "x_i is evaluated at least as high as x_h on ordinal attribute q_j", which is denoted $x_i \succeq_{q_j} x_h$. Therefore, it is supposed that $\succeq_{q_j}$ is a complete preorder, i.e., a strongly complete and transitive binary relation, defined on $U = \{x_1, x_2, \ldots, x_m\}$ on the basis of evaluations $\phi(\cdot)$. An ordinal attribute q_j may have a positive or negative monotonic relationship with the decision attribute d (which is also ordinal).

Furthermore, the values of decision attribute d make a partition of U into a finite number of decision classes, $\mathbf{X} = \{X_k, k = 1, \ldots, p\}$, such that each object $x_i \in U$ belongs to one and only one class $X_k \in \mathbf{X}$. It is supposed that the classes are ordered, i.e., for all r,$s \in \{1, \ldots, p\}$, such that $r > s$, the objects from X_r are in a higher class than the ones from X_s. More formally, if $\succeq$ is a *comprehensive weak order relation* on U, i.e., if for all x_i,$x_h \in U$, $x_i \succeq x_h$ means "x_i is ranked at least as high as x_h", it is supposed: $[x_i \in X_r, x_h \in X_s, r > s] \Rightarrow [x_i \succeq x_h$ and *not* $x_h \succeq x_i]$. If it is not so, then we observe an *inconsistency* between x_i and x_h. The above assumptions are typical for consideration of *ordinal classification problems with monotonicity constraints*, also called *multiple criteria sorting problems*.

As it was shown in (Błaszczyński et al., 2012), *non-ordinal classification problems* can be analyzed by DRSA. Such problems need a proper transformation of the information table, that does not bias the matter of discovered relationships. The intuition which stands behind this transformation is the following. In case of ordinal condition attributes, for which the presence and the sign of the monotonicity relationship between values of condition and decision attributes is known a priori, no transformation is required and DRSA can be applied

directly. Each non-ordinal condition attribute, for which the presence or absence and the possible sign of the monotonicity relationship is not known a priori, is doubled; for the first attribute in the pair it is supposed that the monotonicity relationship is potentially positive, while for the second attribute, it is taken to be potentially negative. Due to this transformation, using DRSA one will be able to find out if the actual monotonicity is global or local, and if it is positive or negative. The decision attributes are transformed such that:

- In case of a non-ordinal decision attribute, each value of this attribute representing a given feature is replaced by a new decision attribute with two values corresponding to the presence and absence of this feature, respectively.
- In case of an ordinal decision attribute, each value of interest k is replaced by a new decision attribute with two values corresponding to original values under and over k, respectively.

More precisely, given a finite set of objects (universe) U described by condition and decision attributes, we assume that the decision attribute makes a partition of U into a finite set of classes $X_1, \ldots, X_k, \ldots, X_p$. To discover rules relating values of condition attributes with class assignment, in case of non-ordinal classification problems, we have to consider p ordinal binary classification problems with two sets of objects: class X_k and its complement $\neg X_k,\ k = 1, \ldots, p$, which are number-coded by 1 and 0, respectively.

Appendix B.1. Rule induction

A *decision rule* is a consequence relation of the form: $E \rightarrow H$. Rule induction is preceded by data structuring using the dominance-based rough set concept. Each of the classes (class 1 or class 0) is approximated using elementary building blocks which in DRSA are positive or negative dominance cones with origins in each object from U in the r-dimensional condition attribute space. As a result of this approximation, one gets two classic sets approximating each class: a *lower approximation* of class 1, composed of all objects from U whose positive dominance cones include only objects from class 1, and an *upper approximation* of class 1, composed of all objects from U included in the positive cones with the origins in objects from class 1. Analogously for class 0: there is a *lower approximation* of class 0, composed of all objects from U whose negative dominance cones include only objects from class 0, and an *upper approximation* of class 0, composed of all objects from U included in the negative cones with the origins in objects from class 0. Thus, the lower approximations include only those objects from U which certainly belong to a given class, because they are consistent with the *dominance principle* (which says that any object x_i with an evaluation on all attributes from C being at least as good as evaluations of some object x_h should not be classified to a worse class than x_h), while the upper approximations include objects from U which possibly belong to a given class. The lower approximation of a class is included in its upper approximation, and their difference is a boundary set composed of inconsistent objects for which one cannot decide on the base of attributes from C if they belong to class 1 or class 0.

When the decision rules are induced from lower approximations of decision classes (i.e., when only lower approximations provide positive examples for the induction), then they are certain in the sense of having confidence ratio equal to 1 (this is the ratio of the number of objects covered by the rule premise that belong to the conclusion class, to the number of objects covered by the rule premise). In the course of some practical applications of DRSA, it appeared, however, that it is better to relax to some extent the definition of the lower approximations, and permit some inconsistent objects to enter the lower approximations. Such a relaxed approach has been called *Variable Consistency DRSA* (VC-DRSA) (Błaszczyński et al., 2009). As a result of applying VC-DRSA, the induced rules are no longer certain, and they are characterized by a chosen consistency measure (Błaszczyński et al., 2011b).

Appendix B.2. Balanced rule ensemble classifier

The classifier that we consider in this work is the *Dominance-based Rough Set Balanced Rule Ensemble* (DRSA-BRE), which is an ensemble of so-called VC-DomLEM rule classifiers described in (Błaszczyński et al., 2011b). The DRSA-BRE is composed of rule classifiers induced on bootstrap samples of objects from the information table. It has been noticed that, when learning from imbalanced data, the global imbalance ratio (i.e., ratio of the number of objects in the minority class to the number of objects in other class) is not the only or even not the most important factor which makes learning difficult. Other data difficulty factors such as class overlapping, small disjunct or lack of representativeness significantly deteriorate the quality of the induced model even on exactly balanced data (Napierala & Stefanowski, 2016).

The samples of objects used in the induction process are controlled by a balancing factor. The approach applied to this end, described in (Błaszczyński & Stefanowski, 2015), is called *Neighbourhood Balanced Bagging* (NBBag). It extends the standard *bagging* scheme proposed by Breiman (1996). Let us remark that in the standard bagging, several classifiers, called component or base classifiers, are induced using the same learning algorithm over different distributions of input objects, which are bootstrap samples obtained by uniform sampling with replacement. NBBag focuses bootstrap sampling toward difficult minority examples by using certain type of weights. The weight of an object from the minority class depends on the analysis of class labels among its k nearest neighbours. Such object is considered the more unsafe, the more it has examples from other classes in its neighbourhood. Thus, this part of the weight reflects a local balancing factor. Moreover, the local part of the weight is also aggregated with a global balancing factor, which takes into account the imbalance ratio between classes. Objects from other classes are assigned weights which reflect only the global balancing factor.

Appendix B.3. Assessing attribute relevance

Assessing attribute relevance is a part of the DRSA methodology. It involves measures that satisfy the property of Bayesian confirmation (Błaszczyński et al., 2011a). These measures take into account the interactions between attributes represented by decision rules. In this case, the property of confirmation is related to a quantification of the degree to which the presence of an attribute in the premise of a rule provides evidence for or against the

conclusion of the rule. The measure increases when more rules involving an attribute suggest a correct decision, or when more rules that do not involve the attribute suggest an incorrect decision, otherwise it decreases.

Before defining a relevance measure, let us give some necessary definitions. Considering a decision rule and a finite set of condition attributes $C = \{q_1, \ldots q_r\}$, we can define the condition part of the rule as a conjunction of elementary conditions on a particular subset of attributes:

$$E = e_{j_1} \wedge e_{j_2} \wedge \ldots \wedge e_{j_v}, \tag{B.1}$$

where $\{j_1, j_2, \ldots j_v\} \subseteq \{1, \ldots, r\}$, $v \leq r$, and e_{j_h} is an elementary condition defined on the value set of attribute q_{j_h}, e.g., $e_{j_h} \equiv q_{j_h} \geq 0.5$, $j_h \in \{j_1, \ldots, j_v\}$.

The set of rules R induced from data set L can be applied to objects from L or to objects from a testing set T. A rule $r \equiv E \rightarrow H$, $r \in R$, covers object x ($x \in L$ or $x \in T$) if x satisfies the condition part E. We say that the rule is correctly classifying x if it both covers x, and x satisfies the decision part H. If the rule covers x, but x does not satisfy the decision part H, then we say that the rule classifies x incorrectly. In other words, we say that rule r is true for object x if it classifies this object correctly, and it is not true otherwise. We denote the fact that E includes an elementary condition e_j involving attribute q_j by $q_i \triangleright E$, $j \in \{1, \ldots, r\}$. An opposite fact will be denoted by $q_j \not\triangleright E$. Formally, a relevance measure $f(H, (q_j \triangleright E))$ has the property of Bayesian confirmation if and only if it satisfies the following conditions:

$$f(H, (q_j \triangleright E)) = \begin{cases} > 0 & \text{if} \quad \Pr(H|(q_j \triangleright E)) > \Pr(H|(q_j \not\triangleright E)), \\ = 0 & \text{if} \quad \Pr(H|(q_j \triangleright E)) = \Pr(H|(q_j \not\triangleright E)), \\ < 0 & \text{if} \quad \Pr(H|(q_j \triangleright E)) < \Pr(H|(q_j \not\triangleright E)). \end{cases} \tag{B.2}$$

The conditions of definition (B.2) thus equate the confirmation with an increase of the probability of the hypothesis caused by the evidence, and disconfirmation with a decrease of the probability of the hypothesis caused by the evidence. Finally, neutrality is identified in case of lack of influence of the evidence on the hypothesis.

From among many Bayesian confirmation measures proposed in the literature, we used as a relevance measure the measure $s(H, (q_j \triangleright E))$ for its good properties and a clear interpretation in terms of a difference of conditional probabilities (see (Greco et al., 2016)): $s(H, (q_j \triangleright E)) = \Pr(H|(q_j \triangleright E)) - \Pr(H|(q_j \not\triangleright E))$, where probability $\Pr(\cdot)$ is estimated on the testing samples of objects. In this way, attributes present in the premise of a rule that assigns objects correctly, or attributes absent in the condition part of a rule that assigns objects incorrectly, get more relevant.