Development of Machine Learning Techniques for Diabetic Retinopathy Risk Estimation

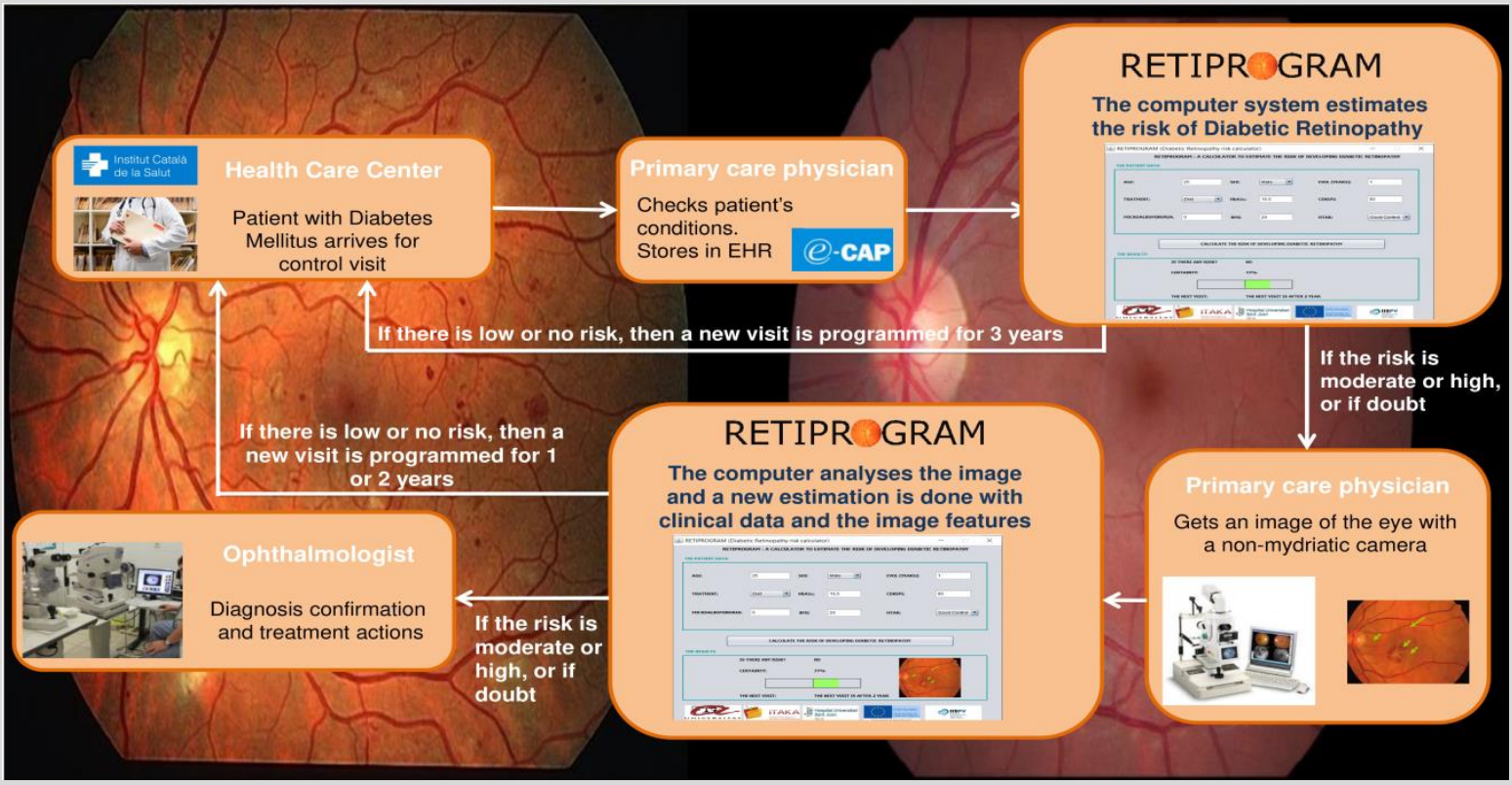
Emran Saleh Ali Ali

UNIVERSITAT ROVIRA i VIRGILI

Development of Machine Learning Techniques for

Diabetic Retinopathy Risk Estimation

EMRAN SALEH ALI ALI



DOCTORAL THESIS
2020

Development of Machine Learning Techniques for Diabetic Retinopathy Risk Estimation

DOCTORAL THESIS

Author:
Emran Saleh Ali Ali

Advisors:
Dr. Aida Valls
Prof. Dr. Antonio Moreno
Dr. Pedro Romero Aroca

Departament d'Enginyeria Informàtica i Matemàtiques

Tarragona
2020

UNIVERSITAT ROVIRA I VIRGILI
Departament d'Enginyeria Informàtica i Matemàtiques

Av. Paisos Catalans, 26
43007 Tarragona
Tel. +34 977 55 95 95
Fax. +34 977 55 95 97

We STATE that the present study, entitled "Development of Machine Learning Techniques for Diabetic Retinopathy Risk Estimation", presented by Emran Saleh Ali Ali for the award of the degree of Doctor, has been carried out under our supervision at the Department of Computer Engineering and Mathematics of this university.

Tarragona, July 10th, 2020

Doctoral Thesis Supervisors

Dr. Aida Valls

Prof. Dr. Antonio Moreno

Dr. Pedro Romero Aroca

To my mother's soul, my father, my wife, my son and my family.

Acknowledgements

The author was supported by a predoctoral FI grant provided by Generalitat de Catalunya.

This thesis' studies were funded by the Spanish research projects PI12/01535, PI15- /01150 and PI18/00169 (Instituto de Salud Carlos III and FEDER) and the URV grants 2014PFR-URV-B2-60, 2015PFR-URV-B2-60, 2016PFR-URV-B2-60 and 2017PFR-URV-B2-60.

I would like to express my gratitude to my supervisors Dr. Aida Valls, Prof. Dr. Antonio Moreno and Dr. Pedro Romero Aroca for their useful guidance, insightful comments, and considerable encouragement to complete this thesis. This work would not have seen the light without their continuous guidance and professionalism. I would also like to thank Prof. Vicenç Torra and Prof. Humberto Bustince for their valuable comments and helpful suggestions.

List of Abbreviations

ACC	Accuracy
BMI	Body Mass Index
BRIEF	Binary Robust Independent Elementary Feature
CA	Classification Ambiguity
CBR	Case Based Reasoning
CDSS	Clinical Decision Support System
CKD-EPI	Chronic kidney disease epidemiology collaboration
CM	Confusion Matrix
CNN	Convolutional Neural Network
Creat	Creatinine
DCT	Discrete Cosine Transform
DFTA	Discrete Fourier Transform Attributes
DM	Diabetes Mellitus
DR	Diabetic Retinopathy
EHR	Electronic Health Record
EVOL	Diabetes evolution time
EX	Hard Exudates
FDT	Fuzzy decision trees
FN	False Negative
FP	False Positive
FPE	Fuzzy Partition Entropy
FRBS	Fuzzy rule-based system
FRF	Fuzzy Random Forest
GA	Gabor Attributes
GMM	Gaussian Mixture Model
HbA1c	High hemoglobin A1c levels
HE	Hemorrhage
HM	Harmonic mean
ICT	Information and Communication Technologies
ID3	Iterative Dichotomiser 3
IDRID	Indian Diabetic Retinopathy Image Dataset
KNN	K Nearest Neighbor
LDA	Linear Discriminant Analysis
LR	Logistic Regression
MA	Microalbuminuria
MAs	Microaneurysms
ML	Machine Learning
NB	Naive Bayes
NHS	National Health Surveys
NPV	Negative predictive value
PCA	Principle Component Analysis

List of Abbreviations

RAM	Regression Activation Map
RC	Rule Confidence
RF	Random Forest
RMCC	Rule membership degree to the conclusion class
SE	Soft Exudates
Sensit	Sensitivity
SJRUH	Sant Joan de Reus University Hospital
Spec	Specificity
SRDA	Spectral Regression Discriminant Analysis
SURF	Speeded up Robust Features
SVM	Support Vector Machines
TP	True Positive
TN	True Negative
TTM	Treatment
WR	Winner Rule

Abstract

Diabetic retinopathy (DR) is a chronic illness. It is one of the main complications of diabetes, and an essential cause of vision loss among people suffering from diabetes. Diabetic patients must be periodically screened in order to detect signs of diabetic retinopathy development in an early stage. Early and frequent screening decreases the risk of vision loss and minimizes the load on the health care centres. The number of the diabetic patients is huge and rapidly increasing, so it is hard and resource-consuming to perform a yearly screening to all of them.

The main goal of this Ph.D. thesis is to build a *clinical decision support system (CDSS)* based on electronic health record (EHR) data. This CDSS will be utilised to estimate the risk of developing RD.

In this Ph.D. thesis, I focus on developing novel Machine Learning systems based on Fuzzy Logic. The output of such systems makes the physician know what combinations of the features can cause the risk of developing DR.

In this work, I propose a method to reduce the uncertainty in classifying diabetic patients using fuzzy decision trees. A Fuzzy Random Forest (FRF) approach is proposed as well to estimate the risk for developing DR.

Several policies, ranging from conservative to permissive, are proposed to merge the classification results achieved by different Fuzzy Decision Trees (FDT) and obtain the final decision.

To improve the final decision of our models, I propose three fuzzy measures that are used with Choquet and Sugeno integrals to aggregate the output of the FDTs rules. The definition of these fuzzy measures is based on the confidence values of the rules. In particular, one of them is a decomposable fuzzy measure in which the hierarchical structure of the FDT is exploited to find the values of the fuzzy measure.

On the practical side, the FRF-based system has been implemented and is currently in use in the University Hospital Sant Joan in Reus (Tarragona).

Keywords: Fuzzy Rule-Based Systems, Fuzzy Measures, Diabetic Retinopathy, Induction of Fuzzy Decision Trees, Random Forests, Ensemble Classifiers.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Motivation

Diabetes mellitus (DM) is a chronic disease, suffered by 382 million adults worldwide. The number of DM patients is estimated to increase up to 592 million adults by 2035 (IDF, 2013). Moreover, specialists estimate that around 46% of diabetic patients are not been diagnosed (IDF, 2013). DM is growing steadily in the last years. In Spain, the National Health Surveys (NHS) detected that diabetes increased from 4.1% of the population in 1993 to 6.4% in 2009. Specialists predict an incidence of more than 3 million DM patients in Spain by 2030 (Shaw et al., 2010).

Some complications of DM make it one of the main reasons of vision loss for young adults around the world (Bourne et al., 2014). The most dangerous ocular

side effect of DM is diabetic retinopathy.

Diabetic retinopathy (DR) is a serious ocular disease related to diabetes. It is a well-known reason of blindness among patients who suffer from DM. Due to the continuous increase on diabetes prevalence, the number of patients who suffer from DR is also growing at a very high rate. That causes a serious concern on health care centers.

Early diagnosis of DR requires an efficient screening procedure. Regular screening of diabetic patients can minimize the risk of blindness development (Romero Aroca et al., 2012). Ophthalmologists currently use non-mydriatic cameras to take images of the fundus of the eye, which is a costly procedure. Because of the huge number of diabetic patients, it is too resource consuming to do this kind of preventive screening by taking images to all the patients. As the incidence of DR is about 10%, in fact, it is not needed to make this kind of image-based screening to all patients. There are some medical and personal conditions that make a person more likely to develop DR. The detection and evaluation of these signs can be made by expert ophthalmologists with a large training. However, diabetic people is usually visited by family doctors, who cannot decide if the eye-fundus screening is needed or not, as they lack of appropriate expertise. For this reason, in this thesis, we aim at developing a decision support system for family physician that helps to determine the risk of developing DR of a given patient. In that way, the patients with risk will be the ones that are screened with an appropriate frequency. This tool may decrease the economic cost of the therapy, reduce the time spent by patients in unnecessary tests and minimize the development of blindness.

Clinical decision support systems (CDSSs) are becoming important tools in the healthcare sector, due to the continuous increase of the available electronic medical data. They may help general practitioners to make diagnoses that would normally be made by specialist physicians. CDSSs may analyse different types of medical information to detect patterns on the data, highlight potential illnesses, suggest the

next steps of treatment or predict the evolution of a patient.

Rule-based systems are a form of Machine Learning (ML) models (Ligeza, 2006), in which knowledge representation and reasoning are accomplished through the use of rules (Durkin, 1993). Rule-based models mimic the human expert reasoning in solving knowledge-based problems. Rule-based systems represent the knowledge into a set of rules of the form (IF A THEN B), in which the antecedent A of each rule is a combination of conditions on the values of some attributes which may lead to conclude a certain output B (Giarratano and Riley, 1998), (Nikolopoulos, 1997).

Fuzzy rule-based systems *(FRBS)* are interpretable models that use linguistic variables, rather than numerical ones. FRBS are a good choice for dealing with medical data for several reasons: (1) they represent the domain knowledge with a linguistic model that can be easily interpreted and understood by doctors, (2) they naturally deal with uncertainty and imprecision, (3) FRBS give also the degree of fulfillment of the classification output, which is an interesting value for doctors, and (4) FRBS usually give a higher classification accuracy than crisp decision trees. In classification problems for diagnosis, doctors are interested not only on the classification result but also on how the system derived the answer (Fan et al., 2011). In contrast, neural networks and linear programming models (Mangasarian, 1992) usually get a high classification accuracy but the decision process is a black box.

The main aim of the work presented in this PhD dissertation is the development of different fuzzy rule-based systems using Machine Learning models to predict the risk of developing DR of individual diabetic patients, considering the values of certain patient's data that is available in the Electronic Health Record (EHR). This work is included in the general framework of several Spanish funded research projects led by the ITAKA research group of URV and the Institute of Health Care Research Pere Virgili, described in the next section.

1.2 Research framework

One of the main goals of the research projects carried out by the ITAKA research group and the Ophthalmology Unit of the University Hospital Sant Joan (Reus, Tarragona) is the development of a CDSS, called RETIPROGRAM, to help clinicians to estimate the personalised risk of developing DR as early as possible. Such system will improve the referral process in Primary Health Care centers, reducing the workload of ophthalmologist services.

Nowadays, all diabetic patients in Catalonia are screened every 2.5 years by taking images of the eyes' fundus (Romero Aroca et al., 2012). Lesions in the eye can be detected in those images, being the main source for DR diagnosis. However, using a fixed-time screening for all patients is not an appropriate model because there are cases in which retinopathy will be detected too late, while most diabetic patients do not require this frequent screening because they will never have this disease.

In a study made on the patients of South Catalonia (from 2007 to 2014) (Romero-Aroca et al., 2016a), it was observed that incidence was stable between 2007 and 2011 (around 8.1%) but since 2011 it has been continuously increasing until almost 9%. If the screening focused on the small subset of diabetic people that have a high risk of developing DR, human and material resources would be used in a more cost-effective way.

In fact, in (Romero-Aroca et al., 2016b) a study of the cost of screening shows that it can be reduced by a personalized screening timing based on each patient's risk factors. This information can be systematically gathered from the EHR. Thus, the goal is that the CDSS can use data from the EHR to decide when to make retinographies, focusing the use of resources on the patients that really need them. In that way, we can avoid unnecessary screenings on patients, saving the time of doctors and patients and the use of the resources on medical centres.

1.2.1 The disease: Diabetic Retinopathy

As already introduced, Diabetic retinopathy (DR) is an ocular illness that can lead up to blindness as consequence of Diabetes Mellitus.

DR does not show any symptom in the early stages but later the patients start noticing a decrease in their vision quality, like problems in seeing distant objects and difficulty of seeing well at night. The patients start having problems in reading and get blurred vision. Another symptom of DR is seeing dark or empty spots. Figure 1.1 shows the difference between the vision of a normal eye and an eye suffering from RD.

Figure 1.1: Left-normal vision. Right-vision with eye with DR lesions.

At the early stages of RD, the eyevessels get damaged and start leaking bloods and other fluids into the retinal tissue (macula) in the middle back of the eye. The macula is the part of the eye responsible to clear the central vision allowing us to see the colors and the fine details of the objects. As a result of that, the macula starts to swell. When this happens, a dark floating spot starts appearing in the retina. Without an early treatment, it may cause scarring.

Ophthalmologists make comprehensive eye examinations by means of eye-fundus

images. These examinations check the retina with a focus on detecting the degree of the damage caused to the central vision as well as observing abnormalities in the growth of the eye vessels.

The retina is the inner and the light-sensitive part of the eye. It contains several layers, that include the photoreceptors. Figure 1.2 shows the components of the retina and the usual DR lesions. The main components *(Anatomical Structures)* of the retina are the following:

- Optic Disc: it is also called optic nerve head. The optic disc represents the beginning of the optic nerve. It is the circular area in the back of the eye where the axons of retinal ganglion cells come together.
- Fovea: it is a small pit in the macula area of the retina. It only contains cones that provide the clearest vision. The fovea is the only part that spreads aside to let the light fell directly into the cells which obtain the most accurate image.
- Macula: it is located in the back of the eye. The macula is responsible of the color vision, central vision and the great details of the view we see. It contains the cells that detect the light and send it to the brain.
- Blood vessel: it is a thin elongated piecewise net in the retina. It contains blood and fluids. The eye vessels have small curvature.

DR produces damages (*Lesions*) in the eyes. These are the main types of lesions caused by DR:

- Exudates: because of the damages happened to the blood vessels in the retina, fats get leaked out of the vessels making exudates. These exudates are the main indicator of having DR.
- Haemorrhages: they show abnormal types of bleeding. They are tiny spots of blood that flow out the eye vessels into the retina.
- Microaneurysms (MAs): are small areas of blood. They are enlarged parts and protrusions in the retina vessels. These protrusions leak blood and fluids into the retinal tissue.

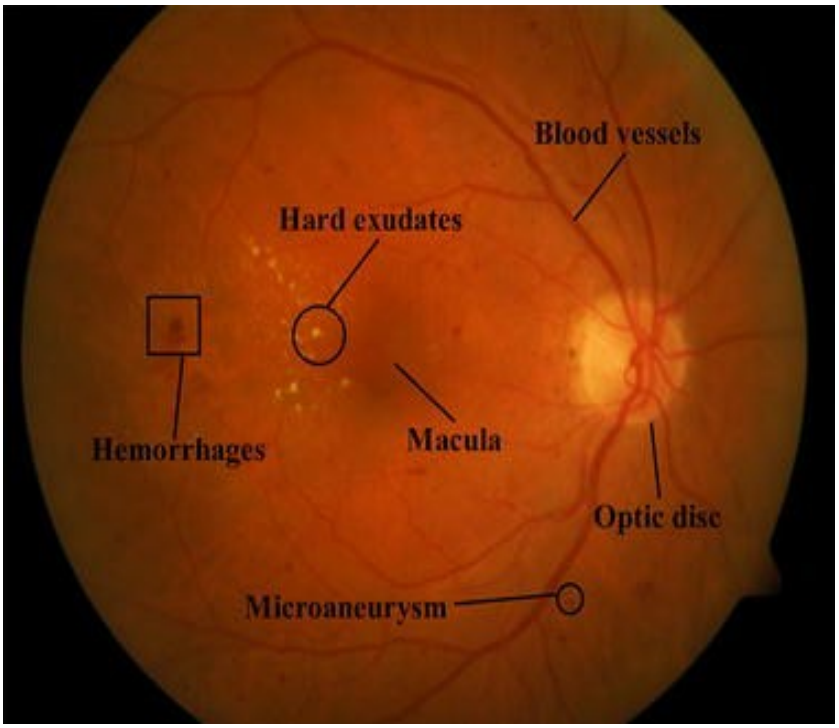


Figure 1.2: Retina image with spots appear because of RD.

In short, DR causes blindness due to the dilatation of small blood vessels (microaneurysms) which produces intra-retinal haemorrhages and fluid leaking composed by lipoproteins and lipids (exudates).

The type of the treatment for DR depends on the stage of the illness. The goal of the treatment is to stop or reduce the progression of the illness. In the early stages of DR, making regular monitoring by ophthalmologists could be the only treatment needed. The DR patients also need to follow their physician's advice for the diet, controlling blood sugar levels and body exercise. That can help to control the progress of the illness.

One possible treatment is the injection of medication in the eye to slow or stop the damages caused by DR. In the advanced levels of RD, the vessels leak blood and fluids into the macula area. That leads up to macular edema. Ophthalmologist make a laser treatment (photocoagulation) to stop the leakage into the retina. A laser beam is aimed to the abnormal vessels area to make burns to seal the blood leakages.

In some advanced situations which happen with proliferative DR, widespread vessels grow in the eye. To treat that, ophthalmologists make scattered laser burns across the eye. As a result, the abnormal vessels shrink and disappear. The patients may lose some side vision with this procedure to safe the central vision.

1.2.2 Using RETIPROGRAM to monitor and screen diabetic patients

The system developed in this dissertation is now integrated in the software tools managed in Sant Joan de Reus University Hospital (SJRUH) and the plan is to integrate it in all the health care centers of Catalonia (Spain). The system gets information directly from the hospital databases and uses the patient's data to make an estimation of the risk of developing DR.

Figure 1.3 shows the workflow for the personalised screening of DR among diabetic patients. It consists of six steps. First, the patient is visited in the health care centre by a family doctor. If the patient is diabetic, the doctor will use RETIPROGRAM to get the values of some relevant attributes from the patient's EHR. Third, a ML model is applied to calculate the risk of DR (different models have been developed in this dissertation, as will be shown in the following chapters). If the risk is low, the next visit of the patient is scheduled in three years.

If the risk is moderate or high, a retinography is made using a non-mydriatic camera. This test is neither invasive nor painful, but it needs to be done by a specialist because the light and calibration conditions are very important to get a good image.

This image is then fed into another module of RETIPROGRAM (not covered in this dissertation), that applies computer vision techniques to automatically detect any kind of lesion in the eye, especially those near the macula. This information is displayed to the doctor and the next visit is scheduled according to the DR risk level. In fact, if lesions are found, the risk will be set to a high level and an urgent

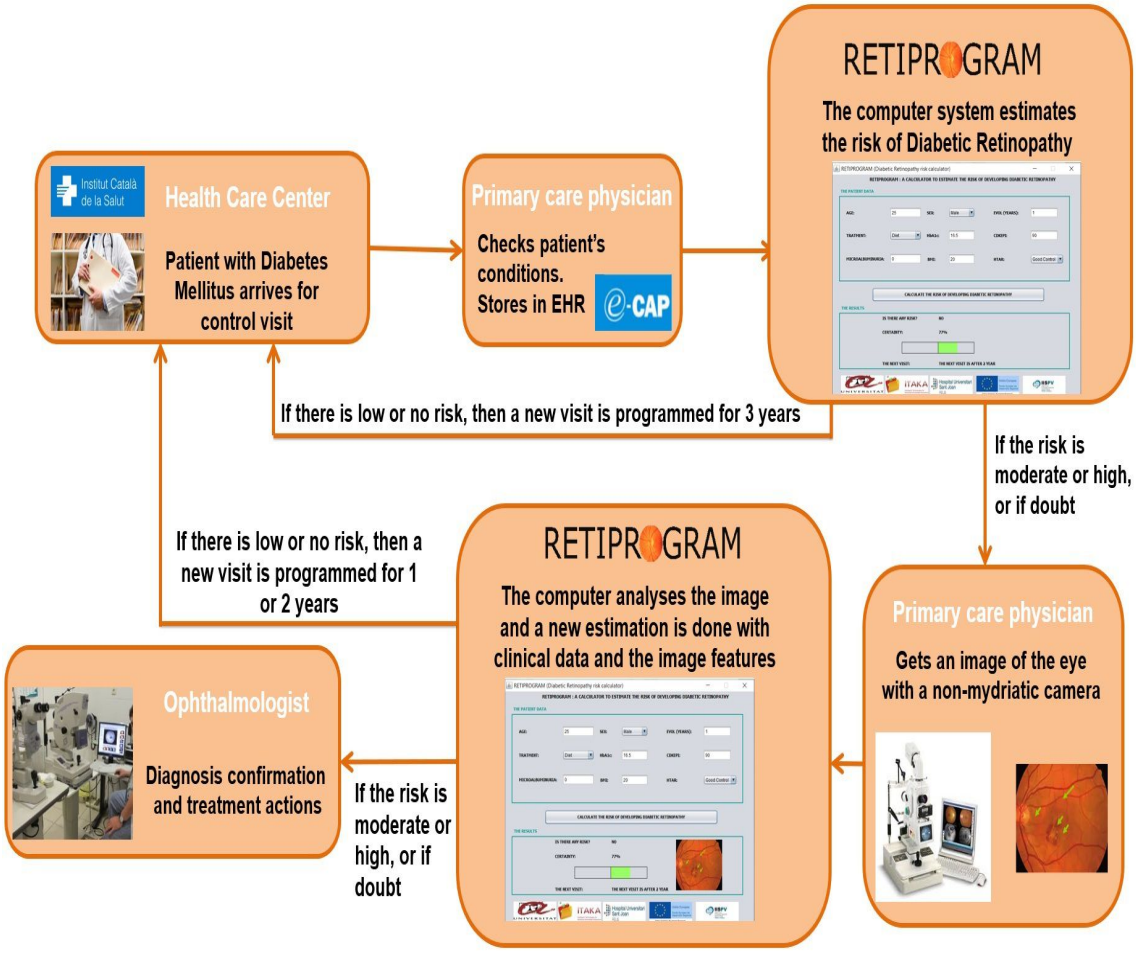


Figure 1.3: Personalised screening of diabetic patients using RETIPROGRAM

appointment with an ophthalmologist will be programmed.

Up to now, there is not any CDSS with these characteristics. Although there exist some works on the assistance of RD detection using computer vision techniques, there are not any tools that also integrate the analysis of the information in the EHR due to the lack of a systematic collection of all data from patients.

The main goal of the research projects of the ITAKA research group in the last years has been to design, develop and test the RETIPROGRAM system using different ML models. RETIPROGRAM includes two main components:

1.2.2.1 PART 1. DR risk estimation from EHR data

The aim of this module is to estimate the personalised risk of developing DR of a particular patient, taking into account his/her medical data, which is stored in the 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).

Very few related works exist based on EHR data. 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. In (Balakrishnan et al., 2013), a predictive system for DR was developed. using decision trees and case based reasoning (CBR). The C5.0 method was applied to produce decision trees. The three most similar cases for the new case were chosen by

KNN and Hamming distance. For the final decision, a voting mechanism is applied.

In our approach we use fuzzy logic to handle the uncertainty and imprecision that may be found in the patient's data. First, nine clinical attributes relevant for the diagnosis of DR were identified by the ophthalmologists (Romero-Aroca et al., 2016a). Using data sets extracted by the Ophthalmology Unit in SJRUH, we have trained and tested different models, which will be explained later in this dissertation. The construction of a single Fuzzy Decision Tree (FDT) was initially proposed (Saleh et al., 2016b). To improve the classification accuracy, an ensemble of classifiers was later implemented using a FRF (Saleh et al., 2016a). The rest of this thesis chapters are devoted to explain this part in detail.

1.2.2.2 PART 2. Analysis of the eye fundus images

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 is 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 and deep learning.

In the classical computer vision approaches, feature descriptors are used to extract the main features in the fundus image. These features are fed into machine learning models to learn and predict the type of DR or segment the area which presents the illness (Quellec et al., 2008; Saleh and Eswaran, 2012; Akyol et al., 2017).

In the classical computer vision techniques line in our research project, we tackled

first the removal of the normal structures of the eye (blood vessels, optic disc). The researchers in our group have explored two algorithms (Convexity Shape Prior and GrabCut) for the segmentation (Escorcia-Gutiérrez et al., 2016, 2018). This cleaning of the image should allow us to leave in the image only the pathological elements and the image background color.

In an ongoing study, the Convexity Shape Prior method is also being used to locate the macula, because lesions near the macula are the ones that cause blindness. Then, some techniques for identification and counting of the types of lesions will be needed. Finally, the features extracted with the image processing should be later analyzed together with the EHR data, to improve the accuracy of the system.

When approaching the problem of detecting the level of DR from an image (not finding the lesions) a usual methodolody 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.

Several deep learning based models for DR classification and segmentation have been proposed since 2015. In 2015, Kaggle lunched their competition for DR detection and open a dataset. Some recent works are (Pratt et al., 2016; Chandrakumar and Kathirvel, 2016; La Torre et al., 2016; Kori et al., 2018).

In the team of our research project, a deep learning based model for DR classification into the 5 possible grades was proposed, (de La Torre et al., 2018, 2020). The proposed model used EyePACS dataset for training and validation the approach. The main feature is that the quadratic weighted kappa (QWK) is used during the learning stage as loss function.

1.3 Thesis objectives

The main objectives of this thesis can be summarized as follows:

1. To analyse the main personalized risk factors to develop DR from the Electric Health Record data of the patient, and build a classifier of each patient into the categories high risk, no risk or unknown. To carry out the study we compare the performance of different ML models with other CDSSs which are used to solve similar problems.
2. To improve the construction of FDTs, proposing a way of selecting the next attribute in the tree branch when there are many attributes that have the same best value, and introducing the possibility of not classifying a patient in some cases.
3. To enhance the performance of rule-based and fuzzy rule-based models by building ensemble systems like Fuzzy Random Forests (FRFs).
4. To improve the final decision of our rule-based models by enhancing the aggregation process of the rules activated for each class.
5. To build and deploy a CDSS using the EHR data of SJRUH. This system will eventually be used in all the Catalan health care centres. The goal is to help the family doctors to assess the risk of developing DR. This system will reduce the load on the health care centres and will allow ophthalmologists to focus on the care of the patients who have a higher risk of developing DR.

1.4 Contributions

The main contributions of this Ph.D. thesis are the following:

1. In order to achieve the first goal of the thesis, a FDT-based model was built. An aggregation method was used to get the certainty degree for each class. To avoid the uncertainty of classifying the patient in different classes that have a similar certainty degree, we introduced a parameter to make sure that the

difference of the certainty degree between the classes is higher than a certain threshold.

This contribution is described in the third chapter of the dissertation. It was presented in the following paper:

- Saleh, E., Valls, A., Moreno, A., Romero-Aroca, P., de la Riva-Fernández, S., & Sagarra-Alamo, R. (2016). Diabetic retinopathy risk estimation using fuzzy rules on electronic health record data. In the 13th International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2016). Lecture Notes in Computer Science, vol 9880 (pp. 263-274). Springer.

2. The problem of decision tree induction when several attributes are the best to expand the tree branch was also tackled. We proposed additional steps which guarantee selecting the best attribute. In this research point we also proposed different ways of merging the outputs from two FDTs which consider different measures in the tree induction procedure.

   The details of this work are described in the third chapter. This work was published in the following paper:

   - Saleh, E., Valls, A., Moreno, A., Romero-Aroca, P., & Virgili, S. P. (2017). Integration of Different Fuzzy Rule-Induction Methods to Improve the Classification of Patients with Diabetic Retinopathy. In the 20th International Conference of the Catalan Association for Artificial Intelligence (CCIA 2017) Frontiers in Artificial Intelligence and Applications, vol 300 (pp. 6-15). Artificial Intelligence Research and Development. IOS Press.

3. Taking into account the third objective of this thesis, the FDT model was extended by building an ensemble model, concretely a FRF. In this approach, different ways of aggregating the rules from the FDTs composing the forest were proposed. A new parameter was added to ensure a minimum degree of

certainty when the forest is classifying the patient into a class.

This approach, described in chapter 4 of this Ph.D. dissertation, was published in the following papers:

- Saleh, E., Błaszczyński, J., Moreno, A., Valls, A., Romero-Aroca, P., de la Riva-Fernández, S., & Słowiński, R. (2018). Learning ensemble classifiers for diabetic retinopathy assessment. Artificial Intelligence in Medicine, 85, (pp. 50-63). Elsevier.
- Saleh, E., Moreno, A., Valls, A., Romero-Aroca, P., & de la Riva-Fernandez, S. (2016). A Fuzzy Random Forest Approach for the Detection of Diabetic Retinopathy on Electronic Health Record Data. In the 19th International Conference of the Catalan Association for Artificial Intelligence (CCIA 2016) Frontiers in Artificial Intelligence and Applications, vol 288 (pp. 169-174). Artificial Intelligence Research and Development. IOS Press.

4. In this thesis we propose three different aggregation operators to merge the output of the FDTs rules. They are based on fuzzy measures and fuzzy integrals. The first two aggregation operators are fuzzy measures based on Sugeno λ-measure and distorted probability. They are used with Choquet and Sugeno integrals and they employ rule confidence values in their calculation. The third aggregation operator is a hierarchically $\perp$-Decomposable fuzzy measure in which we used the hierarchical structure of the FDTs and the values at each node of the FDTs to build the fuzzy measure.

   These works, described in chapter 5, were published in the following papers:

   - Saleh, E., Valls, A., Moreno, A., Romero-Aroca, P., Torra, V., & Bustince, H. (2019). A Hierarchically $\perp$-Decomposable Fuzzy Measure-Based Approach for Fuzzy Rules Aggregation. Vol 27. No Suppl 1. (pp. 59-76). International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems.

- Saleh, E., Valls, A., Moreno, A., Romero-Aroca, P., Torra, V., & Bustince, H. (2018). Learning Fuzzy Measures for Aggregation in Fuzzy Rule-Based Models. In the 15th International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2018). Lecture Notes in Computer Science, (pp. 114-127). Springer.

5. Out of this Ph.D. work, we have built a CDSS software (desktop application with GUI and a web service API) that may be installed in the health care centres. The description of the CDSS has been published in the following papers:

   - Romero-Aroca, P., Valls, A., Moreno, A., Sagarra-Alamo, R., Basora-Gallisa, J., Saleh, E., Baget-Bernaldiz, M. & Puig, D. (2019). A clinical decision support system for diabetic retinopathy screening: creating a clinical support application. Vol 25. No 1. (pp. 31-40) Telemedicine and e-Health, Mary Ann Liebert, Inc.
   - Valls, A., Moreno, A., Puig, D., Saleh, E., Escorcia-Gutierrez, J., & Torrents-Barrena, J. (2017). A clinical decision support system for the diagnosis of Diabetic Retinopathy. Proceedings of the 2nd International Workshop on Artificial Intelligence for Diabetes (AID 2017), pp. 31-34.

6. We built a web service API for the CDSS. This software is already installed in the servers of Sant Joan de Reus University Hospital and it is used for DR risk estimation. By having this web service, it is easy to build websites or mobile applications to access the CDSS and run the system. The deployment of the CDSS is detailed in the following local workshop paper:

   - Saleh, E., Maaroof, N., Jabreel, M., (2020). The deployment of a decision support system for the diagnosis of Diabetic Retinopathy into a Catalan medical center. Proceedings of the 6th URV Doctoral Workshop in Computer Science and Mathematics (DCSM 2020), (pp. 45-48), Universitat Rovira i Virgili.

1.5 Awards

During the development of this thesis we have won two research awards:

- We received the best article award in the 20th International Conference of the Catalan Association for Artificial Intelligence (CCIA 2017). The award-winning work is entitled "Integration of Different Fuzzy Rule-Induction Methods to Improve the Classification of Patients with Diabetic Retinopathy". The authors are Emran Saleh, Dr. Aida Valls, Dr. Antonio Moreno and Dr. Pedro Romero Aroca.
- The ITAKA research group obtained the URV 2018 Social Impact of Research award. This work was given for the work entitled: "Design, construction and evaluation of a clinical decision support system for the personalised screening and follow-up of diabetic retinopathy patients".

More details about the awards can be found in appendix B.

1.6 Thesis organization

The remainder of this document is organized in the following chapters:

- Chapter 2: Background

  In this chapter, we discuss the DR risk assessment as a binary classification problem. We describe the flowchart of the CDSS construction using the EHR data and the main attributes considered in the classification process. This chapter also explains the statistical measures used to evaluate the performance of the different models.
- Chapter 3: Diabetic Retinopathy Risk Estimation Using Fuzzy Decision Trees

  This chapter introduces a FDT-based system for the estimation of the risk of developing DR. This model consists on a set of binary classification rules. Every rule suggests if the patient is going to suffer from DR or not. Preprocessing mechanisms like the fuzzyfication of data and the study of the imbalance of

the data set are detailed in this chapter.

In this chapter we also study the integration of different fuzzy rule-induction methods to improve the classification of patients with DR. Additional steps are proposed in the calculation of the best attribute to expand the tree branch. A new way to merge two FDTs is presented.

- Chapter 4: Learning Ensemble Classifiers for Diabetic Retinopathy Assessment

  In this chapter, a FRF approach is proposed. This FRF classifier is composed of a set of FDTs. It uses a small set of attributes which represent the main risk factors to determine whether a patient is in risk of developing DR. Two aggregation procedures are discussed in this work.

- Chapter 5: Learning fuzzy measures for aggregation in fuzzy rule-based models

  Two fuzzy measures are presented in this chapter. These fuzzy measures are utilised with fuzzy integrals to calculate the confidence of each class and obtain the final class of the patient.

  To improve the performance of the previous fuzzy measures, a new -third-Hierarchically $\perp$-Decomposable fuzzy measure based on the FDT structure is proposed. This work was made in collaboration with Dr. Torra and Dr. Bustince.

- Chapter 6: Retiprogram

  This chapter presents the CDSS software *Retiprogram*. It explains how it works and it shows the GUI.

- Chapter 7: Conclusion and future work

  This chapter shows the conclusions of the Ph.D. thesis and the lines of the future research work.

- Appendix A: This appendix describes the Dominance-based Rough Set Approach (DRSA), developed at the Poznan University of Technology, which was compared with our classification methods. It also explains a bagging method for class-unbalanced data.

- Appendix B: The awards obtained by the work in this Ph.D. dissertation are presented in this appendix, along with the intellectual property protection mechanisms.

CHAPTER 2

Background

2.1 Classification problem

Classification is a well-known identified problem in Artificial Intelligence, which aims to assign a class (i.e. label) to an object, where the set of classes is known in advance. Building automatic classifiers is a research field widely studied in Machine Learning. These methods use a set of examples to construct a model of classification. Each example consists on a description of an object together with its correct class. This is known as a labelled training dataset. Using the training dataset, these techniques start learning the general representation of the data and the relations (mapping) between the input data and the output label (decision). This general representation is called the learning function. This approach is known as Supervised Learning.

In the testing phase, the classification algorithms uses the learning function to predict the output label (decision) for a different dataset, called testing dataset. We also know the correct class of each of the objects in the testing set (this is called Ground Truth), so we can later evaluate the correctness of the output given by the classifier, in comparison to the ground truth label.

The classification and the regression are both supervised learning problems, the difference between them depends on the type of the output labels. In the classification, the output labels are discrete values called classes while in the regression tasks, the output labels are continuous values.

In this thesis, we focus on a binary classification problem. Even though DR has 5 severity levels, the goal of our decision support system is to anticipate the conditions that lead to the appearance of DR; for this reason, we focus on distinguishing patients with DR (called Class 1) from patients without DR (called Class 0).

Figure 2.1 shows the flowchart of the CDSS construction. The source of information is the EHR of diabetic people, who have been already screened and doctors have assigned them to DR=0 or DR=1 classes. In an initial data analysis, DR incidence and the personalized risk factors were identified (Romero-Aroca et al., 2016a; Romero Aroca et al., 2012). They are:

- Current age
- Sex
- Body mass index (BMI)
- Duration of diabetes (Evol)
- Treatment of diabetes (Treatment)
- Control of arterial hypertension (HTAR)
- Glycated hemoglobin (HbA1c)
- Estimated glomerular filtration rate, calculated from plasma creatinine using chronic kidney disease epidemiology collaboration equation (CKD-EPI)
- Microalbuminuria (MA)

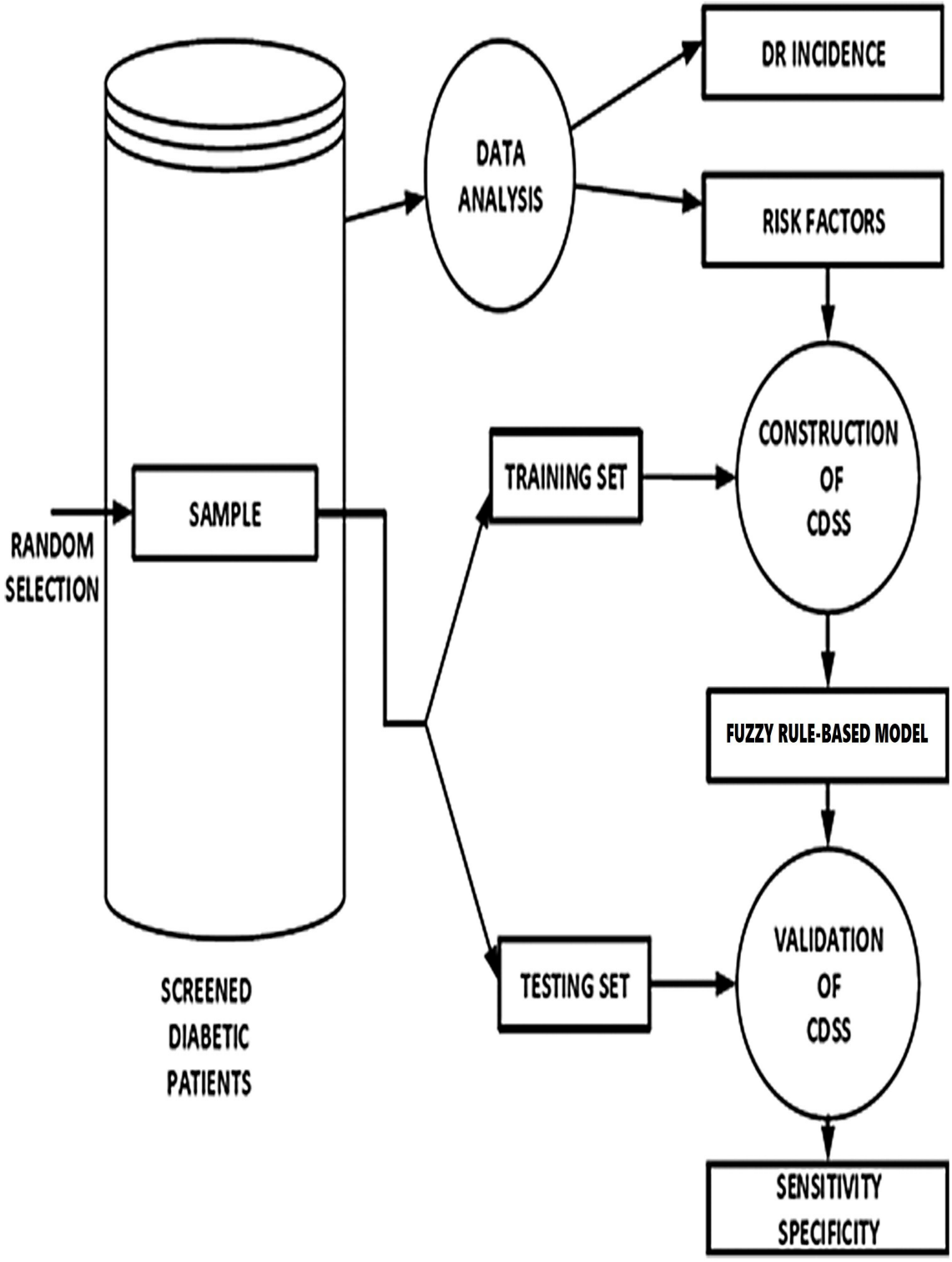


Figure 2.1: The Flowchart of the clinical decision support system (CDSS) construction using the EHR data. Data analysis techniques were used to identify the risk factors.

To build the CDSS, a sample of the diabetic patients data is extracted. We divided this dataset into a training dataset and a testing dataset. The CDSS is built using the training dataset. The CDSS is fuzzy rule-based model which constructs the rules by means of a generalization method. To evaluate the performance of the fuzzy rule-based model, we apply the model on the testing dataset. The model performance evaluation is done using statistical measures, as will be detailed next section.

2.2 Evaluation measures

Statistical measures are required to evaluate the performance of any system. For classification tasks, the following are the base data for all the other performance statistical measures:

- True Positive (TP): The positive examples that the system correctly predicts as positive.
- True Negative (TN): The negative examples that the system correctly predicts as negative.
- False Positive (FP): The negative examples that the system incorrectly predicts as positive.
- False Negative (FN): The positive examples that the system incorrectly predicts as negative.

2.2.1 Confusion matrix

In Machine Learning and specifically in classification problems, the *confusion matrix* is a table that summarizes the performance of the algorithms. The columns in the table represent the number of the examples in the actual class while the rows shows the number of examples in the predicted class. Table 2.1 depicts how the measures are shown in the confusion matrix.

		Actual Class	
		Class 0	Class 1
Predicted Class	Class 0	TP	FP
	Class 1	FN	TN

Table 2.1: Confusion matrix for binary classification.

2.2.2 Accuracy

Accuracy is a measure of how close the predicted values are to the true values. It is calculated using the following equation:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.1}$$

The accuracy is a general statistical measure which is used in wide numbers of tasks in Machine Learning but in some cases it is not the best measure of the algorithm performance, e.g. if the data is imbalanced.

2.2.3 Sensitivity and Specificity

Sensitivity (recall) and specificity are statistical measures. They are used to evaluate the performance of the systems in the classification tasks. In Medicine, sensitivity and specificity are very common.

In many tests, including diagnostic medical tests, sensitivity is the extent to which actual positives are not overlooked (so false negatives are few), and specificity is the extent to which actual negatives are classified as such (so false positives are few).

Sensitivity or recall or the true positive rate is a statistical measurement that calculates the proportion of the examples correctly classified as positive among the examples that are actually positive. The following is the equation to calculate the sensitivity:

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.2}$$

Specificity or (also called the true negative rate) is a statistical measurement that calculatea the proportion of actual negative examples that are correctly classified as negative among the examples that are actually negative. The following is the equation to calculate the specificity:

$$Specificity = \frac{TN}{TN + FP} \tag{2.3}$$

The terms "positive" and "negative" mentioned above don't refer to the values of the condition but they refer to the presence or the absence of the conditions. This condition could be the presence of a type disease or the absence. This presence or absence is the same in all classification tasks.

2.2.4 Precision and Negative Predictive value

Precision, also called the positive predictive value (PPV), is a statistical measurement that refers to the proportion of positive results in classification or diagnostic tests that are really positive. The following is the equation to calculate the precision:

$$Precision = \frac{TP}{TP + FP} \tag{2.4}$$

The negative predictive value (NPV) is a statistical measurement that refers to the proportion of negative results in classification or diagnostic tests that are really negative. The following is the equation to calculate the Negative Predictive Value:

$$NPV = \frac{TN}{TN + FN} \tag{2.5}$$

2.2.5 Harmonic mean

In statistics, the harmonic mean is a type of averaging operator; in particular, the harmonic mean is one of the pythagorean means. It is suitable when the average of rates is desired.

In some parts of this Ph.D. thesis, we need one statistical measure to evaluate and compare the performance of a system with other systems. Thus the harmonic mean (HM) of specificity and sensitivity has been calculated. The following is the equation to calculate the harmonic mean of specificity and sensitivity (HM):

$$HM = \frac{2 * Specificity * Sensitivity}{Specificity + Sensitivity} \tag{2.6}$$

In some classification tasks, a well-known type of the harmonic mean is used to evaluate the systems. Is is called the F1 score (also F-score and F-measure). F1 uses the precision and the recall (sensitivity) in this formula to compute the score of the system. The following is the equation to calculate the F1 score:

$$F1 = \frac{2 * Precision * Sensitivity}{Precision + Sensitivity} \tag{2.7}$$

2.2.6 Cross validation

Cross validation is a statistical technique used for model validation. It is used to understand how the model will generalize its learning function to work well with other independent datasets. It is widely used to evaluate Machine Learning systems. Cross validation is usually done by dividing the dataset into a training dataset and a validation dataset. The following are the most common cross validation techniques:

- K-fold cross validation: In this technique, the dataset is randomly partitioned into K subsamples. One subsample is left to be utilized as a validation dataset and the other K-1 subsamples are the training dataset. In the K-fold cross validation procedure, this process gets repeated K times and each time the subsample used as a validation dataset is different. One of the advantages of this method is that all the examples are used in the validations dataset exactly once. The final performance of the model is calculated by averaging the performance of the model in the K times.
- Leave P out cross validation (LPOCV): In this technique, there are P examples

in the validation dataset and the other are the training dataset. Repeat this all the way for all the examples in the dataset. The other steps are exactly like in the K-fold cross validation technique. Leave One out cross validation (LOOCV) is a particular case of this technique.

- Holdout: In this technique, the examples in the dataset randomly get assigned to a training dataset or a testing dataset so this technique divides the dataset into a training dataset and a testing dataset. This division could be by a ratio of the examples or by determining the exact number of examples for each dataset. Usually the testing dataset is smaller than the training dataset.
- Repeated random sub-sampling validation: this technique is also called (Monte Carlo cross validation). The dataset gets divided by multiple random splits into training datasets and validation datasets. The procedure of the training and the validation is like the one in the K-fold cross validation technique. The disadvantage of this technique is that some examples may not be selected in the validation dataset while other examples may be selected multiple times.

2.3 Conclusion

In this chapter, for the sake of clarity for the reader, we have presented the basic procedure for building a classifier using supervised techniques. We have given some initial general indications of how we have applied this model to the problem of Diabetic Retinopathy risk assessment. The main statistical evaluation measures used to evaluate the systems have also been defined.

The next chapter focuses on the construction of a Fuzzy Decision Tree classification model. We first argue about the utility of data in the EHR. It is very rare to find a classification of DR based solely on this kind of data (usually it considers image analysis). The improvements of existing algorithms will be then presented. In the second part of the next chapter, we analyse the possibility of merging the class assignments made by different FDT models to improve the classification accuracy.

CHAPTER 3

Diabetic Retinopathy Risk Estimation Using Fuzzy Decision Trees

3.1 Introduction

The use of Information and Communication Technologies (ICT) in the health care centres of developed countries has experienced an enormous increase, specially in the last decade. ICTs are crucial in e-Health to improve the quality of medical services and the coordination of providers, to facilitate the access of patients to health professionals and to reduce costs (Quaglio et al., 2016).

Among these digital technologies, the Electronic Health Record (EHR) is a key element, since it collects historical patients' data. The EHR is a tool of paramount

importance for medical professionals because it provides easy access to the patient information and it also facilitates the exchange of information among different units. EHRs store a wide range of patient information, including demographic data, allergies, medications, test results, diagnosis codes and procedures, among others. Two features of EHRs should be highlighted: (1) the integration of all this information in a single entity and (2) the life-long recording of the patient's health conditions. Such a complete information source is basic to have a complete view of the medical history of each patient.

In (Goldstein et al., 2016), it is reported that the use of EHRs in the US has increased dramatically. In 2009, 12.2% of US hospitals had a basic EHR system, increasing to 75.5% by 2014. Similarly, many EU member states have invested great amounts of time and money in the development of ICT in healthcare, starting from the implementation of EHR (Quaglio et al., 2016).

Nowadays it is possible to collect rather large amounts of data that allow performing different kinds of data analysis, for example using Data Mining techniques. This fact provides interesting opportunities, such as developing CDSSs. The main aim of CDSSs is to improve efficiency and reduce costs by avoiding wasteful treatments and laboratory tests. It has already been proved that the analysis of EHR data may reduce clinical errors (Singh et al., 2008), improve chronic care (Dorr et al., 2007) and help to make a more complete and accurate case reporting to public health.

One of the main purposes of EHR exploitation is the construction of prediction models. A survey made in (Goldstein et al., 2016) showed that the phenomena that were most commonly studied were hospital mortality, hospital readmission and diagnosis of some diseases (e.g. stroke and cancer). In fact, disease prognosis and risk models are among the current research lines in many different fields. It is recognized that a drawback of some studies made until now is the limited number of patient cohorts. However, this is rapidly changing due to the increasing rates of EHR penetration and development (Ohno-Machado, 2017).

Two approaches are proposed in this chapter to develop CDSSs in the field of DR. These CDSSs consist on a set of binary classification rules, that assess whether a new patient has a high risk of developing DR or not. It is well known that medical diagnosis has to deal with imprecision and uncertainty (Szolovits et al., 1995) because Medicine is not a matter of precise numerical values. Doctors usually work with linguistic assertions based on ranges of values. Although most of the indicators have some established intervals corresponding to good/bad states, the limits of these intervals are imprecise because they may depend on other factors of each patient. As a result, classification algorithms working with crisp data usually do not give a high accuracy. For this reason, this study proposes a *fuzzy rule-based system* (FRBS).

Fuzzy rule-based systems are a good choice for dealing with medical data because they represent the domain knowledge with a linguistic model that can be easily understood by doctors and they can naturally deal with uncertainty and imprecision. In classification problems for diagnosis, doctors are interested not only on the classification result but also on how the system derived the answer (Fan et al., 2011). In contrast, black-box models like neural networks (Skevofilakas et al., 2010) usually get a high classification accuracy but the decision is not transparent.

The *fuzzy rule-based systems* (FRBS) proposed in this chapter are fuzzy decision trees (FDT), which are well-known structures frequently used for solving real life classification problems. The main reason for their success is that they offer an interpretable representation, which can be easily understood by people from outside the field of Intelligent Data Analysis. FDTs are predictive models that analyse the data while building the model and represent it in a tree-like structure.

The FDT proposed in the first part of this chapter is based on Yuan and Shaw's induction algorithm (Yuan and Shaw, 1995). In the second part of this chapter we propose the integration of two fuzzy rule-based classifiers constructed using two different algorithms of induction of FDTs. Both of them are extensions of the well-known crisp ID3 algorithm. They optimize different criteria during the selection

of the best attribute at each step of the construction of the branches of the tree. One is based on *fuzzy entropy* (Umano et al., 1994) and the other one is based on *classification ambiguity* (Yuan and Shaw, 1995). The fusion of different methods for classification is a common approach. In this way, we can avoid the bias of a model with the help of another one. In this chapter we propose and evaluate different ways of aggregating the class predictions of these two fuzzy rule-based classifiers. The confidence on the prediction of each classifier is used when they make different predictions.

The rest of the chapter is organized as follows. Section 2 presents related work on induction of FDTs. In section 3 we present the FDT approach. In this part, first we introduce the algorithm of induction of FDTs used in this paper, and then we explain the dataset, variables and fuzzification procedure of the DR data, along with the experimental results. The second part of the chapter presents the integration of two decision trees. First we present the FDT models and the new fusion procedure proposed in this work. After that, we explain the dataset, pre-processing treatment and experimental results.

3.2 Related works

In the literature there are diverse methods for the induction of *FDTs*. The classic ones, from the 1990s, are an extension of the ID3 method for crisp data. Two main approaches can be found to identify the best attribute at each step of the construction of the tree: the ones based on *information theory* (Umano et al., 1994) and the ones based on *classification ambiguity* measures (Yuan and Shaw, 1995).

There are different ways of calculating the *fuzzy entropy* in the information theory model. Umanol et al. (Umano et al., 1994) initiated this approach with a fuzzy version of the crisp entropy. Later, Levashenko and Zaitseva (Levashenko and Zaitseva, 2002) defined a fuzzy conditional entropy between attributes, which is based on joint information, to choose an attribute for tree expansion. Some other authors

have focused on the Hartley function as a more general measure of uncertainty (which coincides with Shannon's entropy measure in the case of a uniform probability distribution). For example, Li and Jiang (Li and Jiang, 2011) used a weight function (cut-standard) to solve the influence of different fuzzy levels with Hartley measures. In (Jin et al., 2014) the authors applied a generalized Hartley metric model with a fuzzy consciousness function to consider the non-linearity in membership functions.

Yuan and Shaw proposed the minimization of the *classification ambiguity* as a different criterion to build the tree, instead of *entropy* (Yuan and Shaw, 1995). The new children of a node must reduce the parent's ambiguity to continue growing the tree or the process is stopped. The classification ambiguity is a measure of non-specificity of a possibilistic measure. In (Xiao et al., 2014) Xiao proposed another classification ambiguity function based on the probability distribution of each class on the data set. Using a similar approach, Wang and Yeung (Wang et al., 2001) proposed a selection criterion based on the degree of influence of each attribute in a good classification. In this case a set of weighted rules was obtained, and the reasoning method consisted on a weighted average of similarity.

Some works combine different techniques; for example, the construction of a fuzzy rough tree was proposed in (An and Hu, 2012). Fuzzy rough set theory is used for choosing the nodes and partitioning branches when building the tree. Attributes are evaluated with a fuzzy dependency function. Chang, Fan and Dzan presented a hybrid model that merged three techniques (Chang et al., 2010). First, a case-based method built a weighted distance metric that was used in a clustering algorithm. After that, a fuzzy entropy measure was applied for node selection. Finally, a genetic algorithm was used to increase the accuracy of the decision tree by detecting the best number of terms to consider for each attribute.

FDTs have been applied in many fields, including health care. In the medical area, fuzzy rules have been applied to some well-known diseases (Sikchi et al., 2013). For example, in (Gadaras and Mikhailov, 2009) fuzzy trees were used in the diagnosis

of breast cancer, diabetes and liver malfunction. However, as far as we know there is no previous work on the use of decision trees in the diagnosis of DR.

Wang and Yeung (Wang et al., 2001) made a comparative study between the entropy-based induction algorithm of Umanol et al. (Umano et al., 1994) and the classification ambiguity induction algorithm of Yuan and Shaw (Yuan and Shaw, 1995). The comparative study covered the attribute selection criteria, the complexity of the methods, the reasoning accuracy, the reasoning technique and the tree comprehensibility. This study concluded that the performance is quite similar in both cases. The number of rules is slightly larger in the method of Yuan and Shaw, but the accuracy of the prediction both in the training and testing sets is quite similar (slightly worse in Yuan and Shaw for some datasets). After this analysis, we chose the fuzzy tree induction algorithm proposed by Yuan and Shaw (Yuan and Shaw, 1995) because it takes a possibilistic approach (i.e. a patient can belong to a certain degree to both classes) rather than a probabilistic one.

3.3 Fuzzy Decision Tree

3.3.1 Methodology

To construct the FDT model we have a labeled training dataset of patients from the Sant Joan de Reus University Hospital. The data will be explained in Section 3.3.2. The obtained set of rules is used to classify a different test dataset. We propose a modification of the classic Mamdani inference procedure in subsection 3.3.1.3 in order to detect undecidable cases. The algorithm for the induction of a FDT is described in this section.

3.3.1.1 Preliminaries

Let us consider the universe of discourse $U = \{u_1, u_2, ..., u_m\}$, where u_i is an object described by a collection of attributes $A = \{a_1, ..., a_n\}$. In this case U denotes the

set of users (patients).

Each attribute $a \in A$ takes values on a linguistic fuzzy partition (Bodjanova, Slavka, 1993) $T = \{t_1, ..., t_s\}$ with membership functions $\mu_{t_i} \in \mu_T$. These membership functions can be understood as possibility distributions.

The *U-uncertainty* (or non-specificity measure) of a possibility distribution π on the set $X = \{x_1, x_2, ..., x_d\}$ is defined in (Yuan and Shaw, 1995) as:

$$g(\pi) = \sum_{i=1}^{d} (\pi_i^* - \pi_{i+1}^*) \ln i \tag{3.1}$$

where $\pi^* = \{\pi_1^*, \pi_2^*, ..., \pi_d^*\}$ is a permutation of $\pi = \{\pi(x_1), \pi(x_2), ..., \pi(x_d)\}$ such that $\pi_i^* \geq \pi_{i+1}^*$, for $i = 1, ..., d$, and $\pi_{d+1}^* = 0$.

3.3.1.2 Fuzzy Tree Induction

The induction algorithm proposed in (Yuan and Shaw, 1995) 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.

$$\mu_{E_\alpha}(u_i) = \begin{cases} \mu_E(u_i) & if\ \mu_E(u_i) \geq \alpha \\ 0 & if\ \mu_E(u_i) < \alpha \end{cases}$$

- 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 a FDT are the following:

1. Select the best attribute for the root node v: the one with the **smallest ambiguity**.
2. Create a new branch for each of the values of the attribute v 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 C_i, terminate the branch with a leaf with label C_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 from 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.

The three measures shown in bold in the previous algorithm control the construction of the tree. They are explained in the following paragraphs. Some of these measures are based on the concept of Fuzzy Evidence, 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).

Ambiguity of an attribute $a \in A$: Considering that attribute a takes values on a linguistic fuzzy partition $T = \{t_1, ..., t_s\}$ with membership functions $\mu_{t_i} \in \mu_T$, its ambiguity is calculated as

$$Ambiguity(a) = \frac{1}{m} \sum_{i=1}^{m} g(\pi_T(u_i)) \tag{3.2}$$

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

$$\pi_{t_r}(u_i) = \mu_{t_r}(u_i) / max_{1 \leq j \leq s}\{\mu_{t_j}(u_i)\} \tag{3.3}$$

Truth level of classification: Having a set of classes $C = \{C_1, ..., C_p\}$, the truth

level of classification indicates the possibility of classifying an object u_i into a class $C_k \in C$ given the fuzzy evidence E

$$Truth(C_k|E) = S(E, C_k)/max_{1 \leq j \leq p}\{S(E, C_j)\} \tag{3.4}$$

where S is the subsethood of the fuzzy set X on the fuzzy set Y

$$S(X, Y) = \frac{M(X \cap Y)}{M(X)} = \frac{\sum_{i=1}^{m} min(\mu_X(u_i), \mu_Y(u_i))}{\sum_{i=1}^{m} \mu_X(u_i)} \tag{3.5}$$

and $M(X)$ is the cardinality or sigma count of the fuzzy set X. In this work the classes C are crisp, so μ_{C_k} will be just 0 or 1.

The truth level can be understood as a possibility distribution on the set U. As before, $\pi(C|E)$ is the corresponding normalisation, which is used to define the next concept.

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

$$G(P|F) = \sum_{i=1}^{k} W(E_i|F) g(\pi(C|E_i \cap F)) \tag{3.6}$$

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.* $W(E_i|F) = M(E_i \cap F)/\sum_{i=1}^{k} M(E_i \cap F)$).

3.3.1.3 Classification using the fuzzy rules

The Mamdani inference procedure is used for the binary classification in the classes 0 (no DR) and 1 (suffering DR). An additional step is added at the end of the classic procedure:

1. Calculate the satisfaction degree of a rule $\mu_R(u)$ using the t-norm minimum.
2. Calculate the membership to the conclusion class as the product between the

satisfaction degree $\mu_R(u)$ and the degree of support of the rule.

3. Aggregate all the memberships for the same class, given by different rules, using the t-conorm maximum.
4. Compare the membership degrees for class 0 and class 1:

$$Class(u) = \begin{cases} "Unknown" & ,if|\mu_{c0}(u) - \mu_{c1}(u)| < \delta \\ argmax_{C_k \in \{c0,c1\}}(\mu_{C_k}(u)) & ,otherwise \end{cases} \quad (3.7)$$

The difference threshold (δ) is used to check if an object belongs to the two different classes with a similar membership degree. If the degree of membership is not significantly different, the object is classified as "Unknown". With this additional step we can detect cases where conflicting inference is made, because some variables support the classification in class 0 while others support class 1. The identification of these cases is of extreme importance in this medical application, because this corresponds to an atypical patient that needs to be manually assessed by the doctors as the system is not able to determine the correct class.

3.3.2 The data

Doctors in the Ophthalmology Unit of Sant Joan de Reus University Hospital are screening around 15000 patients for DR every year. They confirmed an incidence of about 8% to 9% of patients in 2012-2015 (Romero-Aroca et al., 2016a). Although it is increasing, this low proportion indicates that many patients could be safely screened only every 2 or 3 years, so that the screening resources could be focused on the part of the diabetic population with more risk to develop DR (Chalk et al., 2012),(Olafsdottir and Stefansson, 2007).

The method presented in the previous section has been applied to data stored in the EHR of patients treated at Sant Joan de Reus University Hospital. This hospital serves an area of Catalonia with a population of 247,174 inhabitants, having 17,792 patients with Diabetes Mellitus (Romero-Aroca et al., 2016a). Various units of

non-mydriatic cameras are used to screen these diabetic patients. Since 2007 several analytical, metabolic and demographic data have been systematically collected and stored in the EHR of the different units. An statistical analysis 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., 2016a). It was observed that incidence was stable between 2007 and 2011 (around 8.1%) but since 2011 it has continuously increased until almost 9%. This study also analysed which are the main risk factors for developing DR. Out of the results of this previous work, a set of 9 attributes have been taken for the construction of the fuzzy rule based system. Most of the attributes are numerical (e.g. age, body mass index) but there are also some categorical attributes (e.g. sex, medical treatment).

These data are used to build and test the classification model which helps the doctor to decide whether or not the patient has a high risk of developing DR on the basis of the selected attributes. The rules also associate a membership degree to the conclusion that indicates the confidence on the class assignment.

3.3.2.1 Data fuzzyfication

The first step was the definition of the fuzzy sets of the linguistic terms for each numerical attribute. The meaning of the intervals is of great importance to understand the classification rules that will be obtained. Thus, the membership functions have been defined according to the standard intervals of some indicators (such as Body Mass Index – BMI) or according to the medical knowledge and the findings of the statistical analysis (Romero-Aroca et al., 2016a), such as the age division (Fig. 1).

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

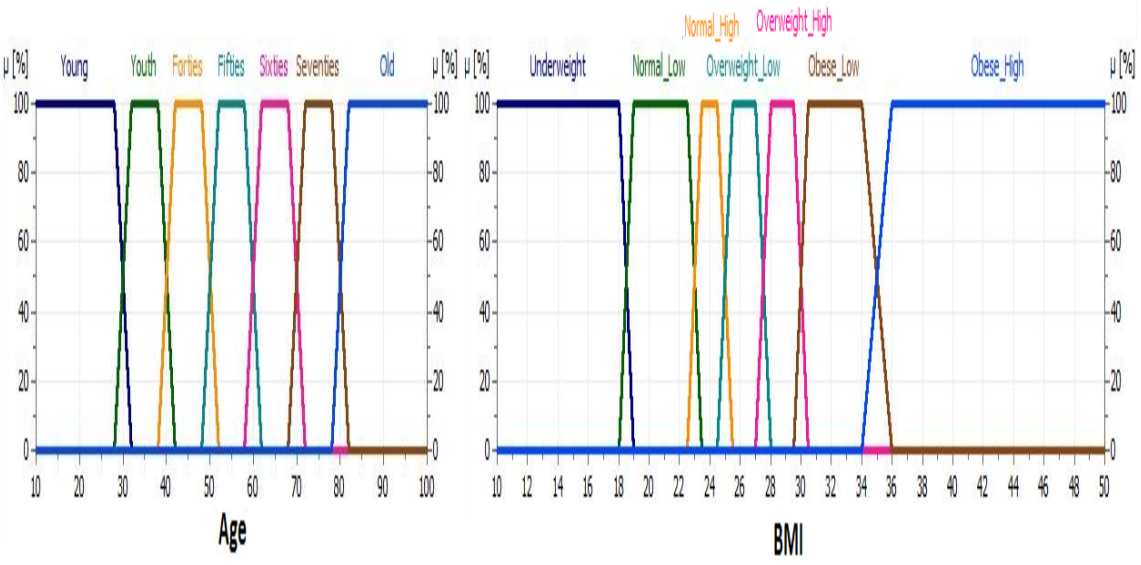


Figure 3.1: Definition of linguistic labels for Age and Body Mass Index.

into two classes was made: X_1 patients with risk of DR (positive class) and X_0 for non-risky patients (negative class).

3.3.2.2 Preprocessing for class imbalance

In medical applications it is common that the low incidence of some diseases generates imbalanced datasets. In this case, the hospital gave us the EHR of 2323 diabetic patients, who were already labeled regarding the DR disease (579 patients with DR (class 1) and 1744 patients 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).

In this dataset only 25% of the patients belong to class 1. Although the incidence of the disease in diabetic patients is much greater than the one in the population (which is below 9%) it still represents a high imbalance in favour of class 0. This situation may cause some problems to the learning algorithm because, as shown previously, it is based on proportions between both classes.

To avoid imbalanced data, a bunch of solutions exist at the algorithmic and

data levels (Kotsiantis et al., 2006). In our case, for the training dataset, we made a random oversampling to balance the class distribution by replicating class 1 examples until they become equal to the number of examples of class 0. Thus, finally the training dataset had 871 patients of each class.

3.3.3 Experimental results

In this section we study the influence of the parameters α, β and δ on the quality of the classification rules. The usual classification measures applied in the medical field are used for the evaluation: specificity and sensitivity.

For values of β below 0.5 the resulting tree was not useful because all the rules predicted class 0 in the all cases (the truth level of classification is too low and the rules cannot find the differential features of the patients with RD). Figure 3.2 shows the results with β=0.6 and β=0.7. Higher values of β generated very large and complex trees without improving either the sensitivity or the specificity.

Figures 3.2 (a),(b) and (c) show the influence of δ on the sensitivity and specificity, for values of α between 0 and 1. With $\alpha \leq 0.5$ the results are quite stable (for β=0.6 they do not change), but when the significance level is increased, the sensitivity improves and the specificity decreases. The figures show that increasing δ improves the sensitivity and specificity for both values of β. The best balance between sensitivity and specificity is found when β=0.7, α=0.2 and δ=0.30, in which sensitivity=82.11 and specificity=82.38 (see black square in Figure 3.2(c)).

In fuzzy classification models one object can be classified in many classes with different membership values. In our model, δ removes the uncertainty of classifying an object into several classes. When the δ parameter is changed, the specificity and sensitivity results behave in the same way depending on α. However, we can see that the increase of the δ value brings an increase of both sensitivity and specificity. The best value for δ is 0.30. Results do not improve with higher values of δ.

Fixing δ=0.3 we can study the number of unclassified patients (Figure 3.2 (d))

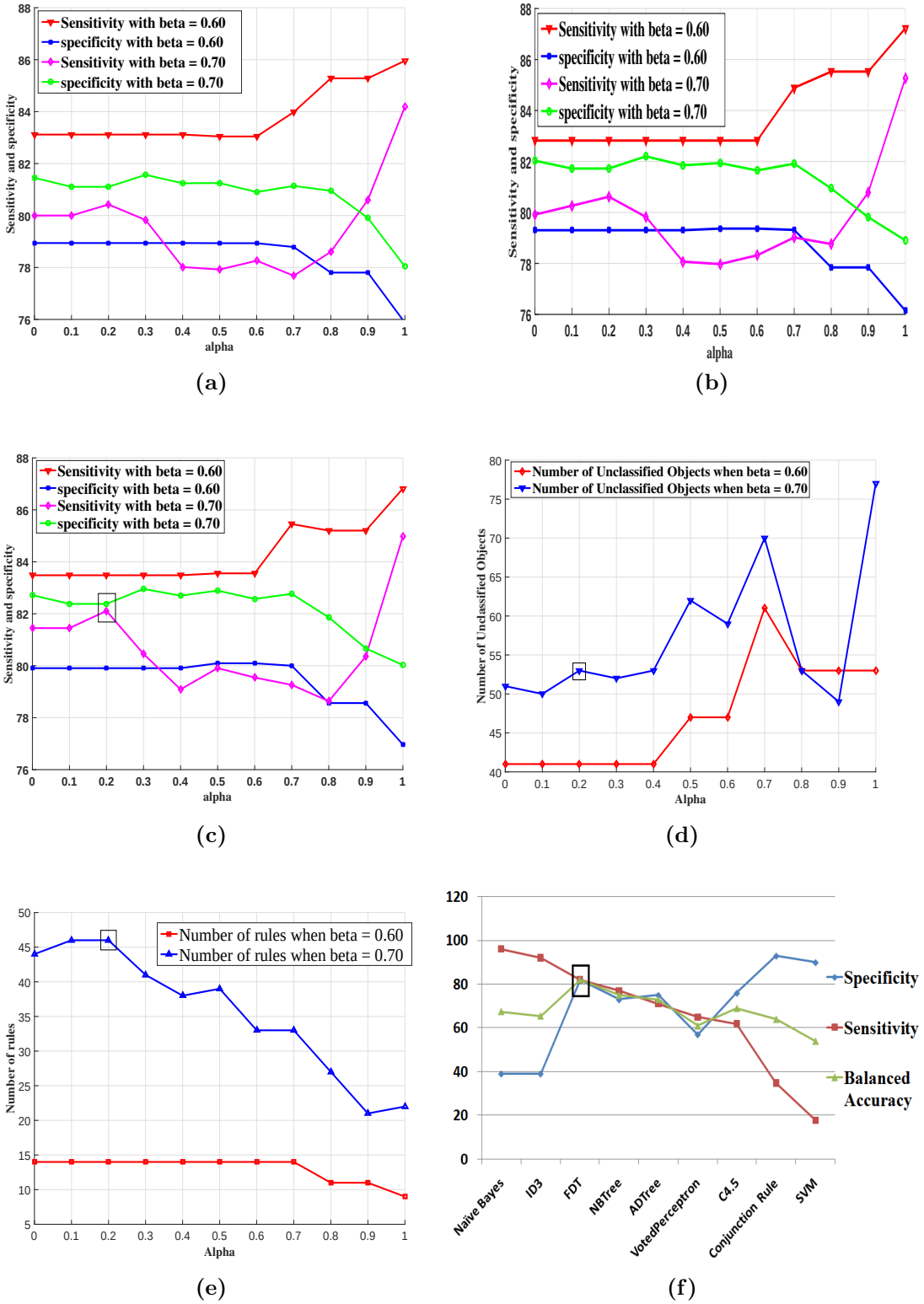


Figure 3.2: Sensitivity and Specificity for $\delta = 0.10$ (a), $\delta = 0.20$ (b), $\delta = 0.30$ (c), Number of unclassified objects (d), Number of rules (e), Comparison with other learning algorithms (f)

and the number of generated rules (Figure 3.2 (e)). Increasing β usually increases the number of unclassified patients. Figure 3.2 (d) shows that the difference between the number of unclassified patients for β=0.6 and β=0.7 is quite stable for values of α lower than 0.8 (between 10 and 15 patients). Increasing α (up to 0.7) increases the number of unclassified patients, but this number decreases when α is 0.8 or 0.9.

β plays an important role in the construction of the FDT, because it is used to decide if the current node is a leaf or not. Only the nodes with a *truth level of classification* higher than β become leaves of the tree. Thus, when β is high, the number and length of the rules increase, but it is still a manageable number for a doctor. Figure 3.2(e) shows the effect of α on the number of rules generated by the algorithm with $\beta = 0.7$ and $\beta = 0.6$. When α increases the system uses less information and the number of rules is smaller. We can see that the number of rules is always around 14 when β=0.6, but it is higher for β=0.7 (between 20 and 46 rules).

Table 3.1 shows some more measures with δ=0.3 and β=0.7 because this is the one with better results, highlighted by the black square in Figure 3.2. The column NPV shows the negative predictive value measure, whereas TP is the number of true positives, FN are the false negatives, FP are the false positives and TN is the number of true negatives. In CDSSs a False Negative is more dangerous than a False Positive, because an ill person is being classified as healthy and is not being properly treated. Thus, it would be desirable to have a low number of FNs. For this reason we want both sensitivity and specificity to be high. In Table 3.1 the combination that gives the best balance between sensitivity and specificity is highlighted.

The method used in this work (FDT) gives better results than some well-known classification algorithms on the current dataset. Figure 3.2(f) shows a comparison between the results of some algorithms based on decision trees (ID3, NBTree, ADTree, C4.5), rule-based algorithms (Conjunction Rule), and other classification algorithms like Support Vector Machines (SVM), Bayes (Naive Bayes) and

α	Precision	Sensitivity	Specificity	NPV	TP	FN	FP	TN
0	55.38	81.45	82.72	94.42	180	41	145	694
0.1	54.88	81.45	82.38	94.41	180	41	148	692
0.2	**54.74**	**82.11**	**82.38**	**94.66**	**179**	**39**	**148**	**692**
0.3	55.31	80.45	82.96	94.18	177	43	143	696
0.4	54.55	79.09	82.70	93.78	174	46	145	693
0.5	55.21	79.91	82.89	93.99	175	44	142	688
0.6	54.69	79.55	82.57	93.85	175	45	145	687
0.7	54.78	79.26	82.77	93.81	172	45	142	682
0.8	53.23	78.64	81.86	93.59	173	47	152	686
0.9	52.63	80.36	80.67	93.89	180	44	162	676
1	52.46	84.98	80.02	95.36	181	32	164	657

Table 3.1: Classification results with $\beta = 0.70$ and $\delta = 0.30$

Perceptrons (VotedPerceptron) (H. Witten and Frank, 2013). Balanced accuracy, sensitivity and specificity are used for the comparison as they are good performance measures for imbalanced datasets (Brodersen et al., 2010). Balanced accuracy is the mean of specificity and sensitivity. Naive Bayes and ID3 give high sensitivity but they have a low specificity and a low balanced accuracy, whereas on the other side Conjunction Rule and SVM give a good specificity but a bad sensitivity. The method in this work gives the best combination of sensitivity and specificity, and the highest balanced accuracy as well.

3.4 Integration of fuzzy rule-induction methods

3.4.1 Methodology

In the previous section we have described how to build a FDT for the binary classification problem of DR risk detection, using an approach based on classification ambiguity. In this section we describe a different way of building a Fuzzy Decision Tree for the same problem, and we propose a novel procedure for integrating the output of the two models. In section 3.4.1.2 several policies to solve the case of having a different classification output in the two models are presented.

3.4.1.1 Induction of fuzzy decision trees

The two fuzzy tree induction methods optimize two different criteria, that will be denoted here as FPE (Fuzzy Partition Entropy) and CA (Classification Ambiguity). Both methods follow the main steps proposed in the classic ID3 algorithm. They propose different extensions for managing the uncertainty of the fuzzy data. The basic algorithm is the following:

1. Assign the attribute that obtains the minimum classification uncertainty to the root node v, that is the one with the *highest information gain*, calculated from the FPE or CA measures, defined below.
2. For each value of the node v, create a branch and check if it could be ended with a class label. Different termination conditions are used in the two methods, but both are based on achieving a minimum threshold of a quality classification indicator, β. In FPE, the condition is based on the fuzzy cardinality of the objects that belong to a class C_i with respect to the whole cardinality of the branch. If the proportion is larger than β, terminate the branch with the class label of the highest class proportion. In CA, if at least at one of the classes C_i, the *truth level of classification* is greater than β, terminate the branch with the class label of the highest truth level of classification.
3. If none of the conditions in step 2 are accomplished, check if it is possible to add one more attribute to the branch. Choose the best attribute in reducing the classification uncertainty. It has been observed that many ties are found in practical applications. Thus, for the case of obtaining the same degree of uncertainty in several attributes, we propose the following selection criteria: a) choose the attribute that classifies more observations; b) if equal, then choose the attribute that uses more terms; c) if equal, then choose the attribute that generates less empty branches (without examples); d) if equal, then choose the attribute that has the term which classifies more patients.
4. If there is not any attribute that reduces the uncertainty, terminate the branch

with a leaf corresponding to the highest class proportion in FPE or highest truth level of classification in CA.

In this algorithm the parameter β is used in step 2. It is a *truth level threshold* used to determine the minimum acceptable confidence on the conclusion of a branch of the tree in order to terminate it with a class label, thus creating a new rule. On the one hand, higher β values may lead to larger trees. On the other hand, very small trees usually have low accuracy, making more prediction mistakes. The difference between the two methods (FPE and CA) relies on the measures used in steps 1 and 2, which are now going to be explained.

The first approach, FPE, is the one proposed in (Umano et al., 1994), based on the *fuzzy entropy* measure for calculating the *information gain* of each attribute. Entropy is a classic measure in Information Theory for calculating the average information content. High entropy indicates more unpredictability, which means a worse classifier. Given the evidence F, corresponding to a set of conditions of a fuzzy rule, and a new attribute $a \in A$ with s terms defining an evidence fuzzy partition $\{E_1, ..., E_s\}$, the information gain is calculated using the fuzzy cardinality measure M.

$$Gain(F, a) = Info(F) - FuzzyEntropy(F, a) \tag{3.8}$$

$$Info(F) = -\sum_{i=1}^{p}(P_i * \ln P_i), \text{ where } P_i = \frac{M(F \cap C_i)}{M(F)} \tag{3.9}$$

$$FuzzyEntropy(F, a) = -\sum_{i=1}^{s}(P_i * Info(F \cap E_i)), \text{ where } P_i = \frac{M(F \cap E_i)}{\sum_{j=1}^{s} M(F, E_j)} \tag{3.10}$$

The second approach, CA is based on *classification ambiguity*, as defined in (Yuan and Shaw, 1995), which refers to the concept of non-specificity or uncertainty of a possibility distribution π on the set $X = \{x_1, x_2, ..., x_d\}$. It indicates the dispersion of the objects into the set of classes. Thus, the goal is to minimize the classification

ambiguity, because it means that the values of the different criteria clearly indicate to which class an object belongs. This approach was explained in detail in section 3.3.1.

To predict the class of a patient from each of the two fuzzy rule-based systems, we use the classical Mamdani's inference procedure, but adding one step to avoid making a prediction when there is not enough difference between the evidences for both classes, as explained in section 3.3.1.3.

3.4.1.2 Fusion of the prediction of the two classifiers

Each of the two approaches for measuring the uncertainty of the attributes and the classification leads to different fuzzy decision rules. As it is usual in Machine Learning, each method has some bias produced by the measures used in the optimization. Wang and Yeung (Wang et al., 2001) made a comparative study between the FE algorithm of Umanol et al. (Umano et al., 1994) and the CA-based induction algorithm of Yuan and Shaw (Yuan and Shaw, 1995), which covered the criterion for attribute selection, the complexity of the methods, the reasoning accuracy, the reasoning technique and the tree comprehensibility. This study concluded that the number of rules is slightly larger in the method of Yuan and Shaw, but the accuracy of the prediction both in the training and testing sets is quite similar (slightly worse in Yuan and Shaw for some datasets).

In order to exploit the benefits of each model while minimizing the mistakes caused by their individual biases, we propose to combine the output of both methods.

In order to be able to compare the confidence of the rules generated in each of the two models, the *truth level of classification* indicator has been selected (equation (3.4)). This measure may be considered as the degree of support of the rule and is going to be used in the fusion procedures that are now proposed.

When observing the two class predictions for a certain patient, different cases can be found. First, if both classifiers predict the same class, then there is full agreement

and this class is assigned to the patient. Otherwise, when the classifiers predict different classes, we need to decide which is the best one. We have studied different fusion policies in order to choose the final class. They are ordered from the most conservative to the most tolerant as follows:

- **Policy 1 (Conservative):** The less risky policy only makes predictions if both classifiers agree on the same class. If they do not agree or if one of the classifiers does not predict a class label (i.e. the decision of classifier is "Unknown"), then the final label will be "Unknown". This approach tries to avoid mistakes, thus it looks for the agreement of the two models on the class assignment.
- **Policy 2:** If one of the decisions is "Unknown" and the other is Class0/Class1 then the final decision is Class0/Class1, respectively. In that way, we rely on the classifier that is able to assign one of the classes when the other cannot. However, if one classifier predicts Class0 and the other one predicats Class1, then when the final decision is "Unknown".
- **Policy 3:** In this policy we rely on the truth level of classification of the rule used by each classifier (i.e. the class assignment support). If one of the decisions is "Unknown" and the other is Class0/Class1, check if the support of Class0/Class1 is above a given threshold. If so, then the final decision is Class0/Class1, respectively; otherwise, the class is "Unknown". If one prediction is Class0 and the other one is Class1, the final class is the one with the highest support. If their support is equal then the final decision is "Unknown".
- **Policy 4 (Risky):** This policy is quite optimistic, as it tries to make all possible class assignments to avoid unclassified objects. If one of the decisions is "Unknown" and the other is Class0/Class1, then the final decision is Class0/Class1, respectively. If one of the decisions is Class0 and the second is Class1, the final decision is the class with the highest degree of support. If the degree of support in both classes is equal then the final decision is "Unknown".

3.4.2 Diabetic Retinopathy data

The method presented in the previous section has been applied to data stored in the EHR of patients treated at Sant Joan de Reus University Hospital (SJRUH). SJRUH gave us the EHR of 3380 diabetic patients with their diagnosis of DR (2527 patients without DR (Class 0) and 853 patients who suffer from DR (Class1)). Note that this dataset is larger than the one used in the first part of this chapter. This dataset was divided into a training dataset and a testing dataset. The training dataset included 2269 EHRs (1654 EHRs from Class0 and 615 EHRs from Class1). The testing dataset contained 1111 EHRs (873 EHRs from Class0 and 238 EHRs from Class1). As commented in the initial FDT algorithm, a balancing strategy was applied, consisting on doing a random oversampling of the EHR of patients of Class1, until their number becomes equal to the number of patients of Class0.

3.4.3 Results

In this section we show the results achieved by the combined model. The values of the parameters for the CA-based induction method were chosen according to our previous study, being $\beta = 0.7$, $\delta = 0.3$. Table 3.2 shows several common quality measures to compare the performance of the different fusion policies: sensitivity, specificity, precision, accuracy and Negative Predictive Value. It also reports the True/False Positive/Negative cases and the number of patients for which the combined system makes an Unknown prediction. For the 3rd policy, the lowest threshold that has been tested corresponds to the β value (which is the minimum truth level required to construct a rule). Few rules achieve a truth level higher than 0.85, so this has been the highest threshold studied.

The two first rows show the results of the two individual rule-induction methods. FRE obtains a higher specificity, but a lower sensitivity. The number of unclassified patients is larger with the CA method (82 vs. 29).

Model	Accuracy	Precision	Sensitivity	Specificity	NPV	TP	FN	FP	TN	Unknown
FPE	81.05	53.37	78.85	81.64	93.57	179	48	157	698	29
CA	79.98	52.57	82.14	79.38	94.11	184	40	166	639	82
Policy1	**84.82**	**61.31**	**84.00**	**85.05**	**94.96**	**168**	**32**	**106**	**603**	**202**
Policy2	**83.13**	**57.52**	**82.63**	**83.27**	**94.59**	**176**	**37**	**130**	**647**	**121**
Policy3, threshold =0.85	84.17	60.99	77.48	86.06	93.14	172	50	110	679	100
Policy3, threshold =0.80	84.19	60.99	77.48	86.08	93.15	172	50	110	680	99
Policy3, threshold =0.75	84.25	60.55	77.78	86.03	93.35	175	50	114	702	70
Policy3, threshold =0.70	82.96	57.42	77.39	84.47	93.25	178	52	132	718	31
Policy4	82.83	57.05	77.39	84.29	93.36	178	52	134	719	28

Table 3.2: Quality indexes of the different classification models on the testing dataset

The fusion policy 1 (conservative) obtains the highest performance (the accuracy achieved by the model is 84.82%, sensitivity is 84.00% and specificity is 85.05%), but the number of unclassified patients is also the largest (202, which means 18% of the testing set). Using the second fusion policy 2, in which the model takes a decision with the opinion of only one classifier, the model also obtains a good classification performance, reducing the unclassified to 121 (11%). In this case, accuracy, sensitivity and specificity are still over 82% (being 80% the usual minimum value required in a medical application of this kind for it to be usable in a real setting).

In policy 3, where the decision depends on the support of the class when the classifiers give different decisions, the performance depends on the threshold used. For a threshold equal to the β parameter (0.7), the results are worse. But for a higher threshold value, the specificity measure is over 86%, which means that the model predicts correctly more examples from Class0. The sensitivity is around 77% with the different tested thresholds. If we decrease the threshold value, the performance of the model slightly decreases but the accuracy is still over 82%. It is also worth noting that, the lower is the threshold, the less unclassified patients we have (only 31 patients in the best case).

Lastly, in the most optimistic policy the model obtains the lowest number of unclassified examples, as expected (28 patients, 2.5 % of the testing set), less than the single classifiers, but at the expense of making more mistakes that the other merging models, specially in False Positives. It means that more people is sent to the retinal screening test when they are healthy, producing unnecessary stress on the patients and the consumption of human and material resources in the health care centers.

However, one of the critical points in medical diagnosis support is avoiding False Negatives, because if people that may develop DR are not detected, they will not receive any early treatment on time. The models with lower FN are highlighted in bold in the table. Policy 2 is the best in this case because it has good quality

indicators and less unclassified patients. The results clearly show that the strategy of fusion plays an important role to increase or decrease the number of unclassified patients. A more conservative policy means more unclassified examples (18 %), whereas a more optimistic policy decreases the number of unclassified examples to only 2.5 % of the testing set.

3.5 Conclusion

The FDT induction method presented at the beginning of this (based on minimizing the classification ambiguity) provides quite good results on the tests done with data coming from the Sant Joan de Reus Hospital. It provides better results than other non-fuzzy approaches, with a specificity and sensitivity above 80% (as required for clinical systems). Therefore, it seems that fuzzy sets are a suitable way of dealing with the ambiguity of the data stored in the EHR. Moreover, obtaining linguistic rules is appreciated by doctors, because they can easily understand the knowledge model and accept it as a valid estimation tool.

Two well-known learning methods for the induction of FDTs have been taken. One of the main contributions of the work is the study of different procedures for merging the class assignments made by these two techniques. We proposed different fusion policies to make the final decision, which range from a very conservative or pessimistic view to a risky or optimistic behaviour. The results show that the number of unclassified patients depends heavily on the chosen policy. In medical systems, it is usually better to avoid mistakes, specially false negatives (i.e. patients with DR that are not detected). However, a too conservative policy increases the number of cases where the system does not give any answer to the physician, which increases the workload and the costs because these patients will be screened with the usual tests to get a confirmation.

The advantage of a fuzzy approach is that the decision is transparent and explicable, which is very important for the physicians. They can know which

combination of evidences has lead to the diagnosis of the disease on each particular patient.

In the next chapter, a more complex classification model, called a Fuzzy Random Forest (FRF), is going to be investigated. The main advantage of building ensemble classifiers is to make the final model overcome the mistakes made by a single classifier.

CHAPTER 4

Learning Ensemble Classifiers for Diabetic Retinopathy Risk Assessment

4.1 Introduction

As described in the previous chapter, the Ophthalmology unit of *Sant Joan de Reus University Hospital* 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 *EHR* that are more relevant in the assessment of the

risk of developing DR (Romero-Aroca et al., 2016a). 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.

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 selected attributes (without the need of an image of the eye fundus). Such a classification model, given the data of a patient, can assess the personalised level of risk 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 (e.g. twice a year), whereas those with a low risk could be safely screened every 3 years.

In the preceding chapter we already described the use of *FDTs* as a classification method. In this chapter, 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, we explore the use of *FRFs* , and compare it with another ensemble method, *Dominance-based Rough Set Balanced Rule Ensemble* (DRSA-BRE), developed by colleagues from the Intelligent Decision Support Systems Laboratory of the Poznan University of Technology (Poznan, Poland) with whom we have been collaborating in this work. A detailed description of this method can be found in Appendix A.

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 chapter is organized as follows. Section 4.2 presents a brief introduction to FRF and DRSA-BRE followed by reviews of their medical applications. Section 4.3 explains how the FRF and DRSA-BRE models have been applied for DR risk assessment. Section 4.4 presents and discusses the experimental results on the use of the two ensemble classification techniques. Conclusions are presented in Section 4.5.

4.2 Background

4.2.1 Fuzzy random forests and their 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 in 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 survey in (Kulkarni and Sinha, 2013)). One extension that is particularly interesting is the use of fuzzy logic, which allows to work with uncertainty. *FDT*, as we have seen in the previous chapter, permit to manage uncertain or imprecise data, represented with linguistic labels whose meaning is defined with a fuzzy set. A *FRF* is composed of a set of FDTs. 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 a 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, as we explained in section 3.3.1.2, we have considered Yuan and Shaw's FDT induction algorithm, which falls into the latter category (Yuan and Shaw, 1995). Yuan and Shaw's induction algorithm is an extension of the classic ID3 method for crisp data.

This type of ensemble methods have been applied to different medical problems. Crisp RF have been used for diagnosis of heart arrhythmia (Alickovic and 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, FRFs 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 and 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 and 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.

4.2.2 DRSA and its medical applications

In this study, we are comparing our approach based on FRF with 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).

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 is particularly interesting for decisions where the condition attributes and decision attributes are ordinal. Then, the constructed rules 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 A. For a complete presentation of the DRSA methodology see (Greco et al., 2001; Słowiński et al., 2014, 2015).

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 A.1.3).

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 are (Słowiński et al., 2014; Yao et al., 2015; Słowiński et al., 2015).

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 DR), 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 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 and 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 and 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 A.1.2).

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 counted into prospective tools of chemoinformatics in a survey article on knowledge discovery from chemical data (Gardiner and Gillet, 2015).

Another field of applications of DRSA concerns the analysis of biomedical data. In (Blasco et al., 2015), DRSA has been applied to the analysis of metabolomic data to try to discover 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.

Classic rough set-based models have been used in diabetes detection. They

have been tested on the Pima Indian Diabetes Dataset (Khan and 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.

4.3 Data and Methods

This section explains how the two ensemble methods explained in the previous section, *FRFs* and *Dominance-Based Rough Sets-Balanced Rule Ensemble*, have been used for DR detection.

The dataset used in this study is the one described in section 3.3.2.2. It 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 were optimized using a 10-fold cross-validation on the training set. The best classification models were then applied to the test set for validation.

This data set 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, there are contradictory examples, which correspond to two patients with the same values on all the criteria but that belong to two different classes. These two issues have been properly addressed in each of the two methods applied, as will

be explained in the next subsections.

4.3.1 Fuzzy Random Forests in DR risk assessment

In order to construct a FRF for DR using the Yuan and Shaw's approach presented in 3.3.1.2 (Yuan and Shaw, 1995), 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 is trained with two thirds of the training dataset. Half of the bootstrap -first third of the dataset- comes from the examples which are labeled with class 0 (the patients who do not suffer from DR) and the second half of the bootstrap -the second third of the dataset- comes from the examples which are labeled with class 1 (the patients that 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 FDTs (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 FRF (*i.e. 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. Half of the training examples for each FDT comes from class 0 and the second half from class 1.

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 the same 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 the satisfaction level of a rule) and the t-conorm maximum (on the aggregation of outputs). When a new patient is fed into the classifier system, each rule is activated with a certain degree of satisfaction (the minimum satisfaction of the conditions of the rule). 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 4.1 and 4.2 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.4.1 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. 4.2 has 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 FRF 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

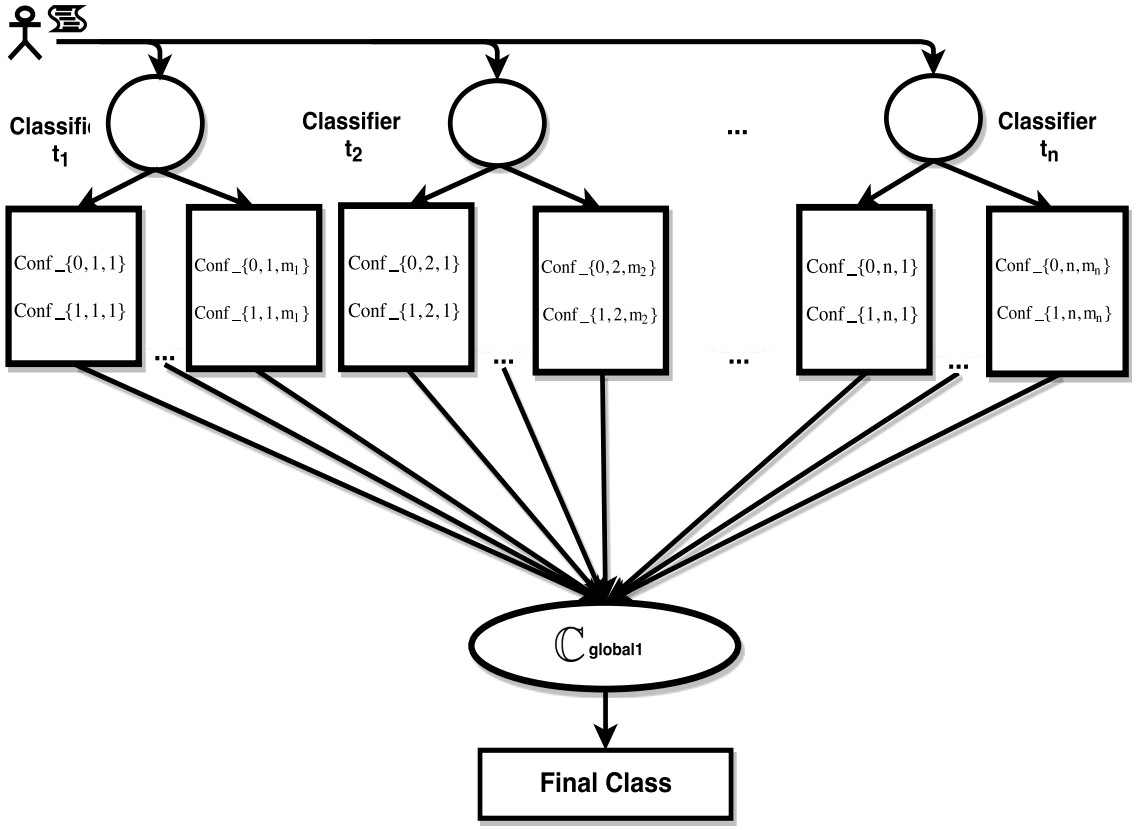


Figure 4.1: 1-step classification process

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. 4.1, 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:

$$\mathbb{C}_{global1} = max_k(\mu_{X_k,j,i}), \text{ for all } k = \{0,1\},\ j = 1,\ldots,n,\ i = 1,\ldots,m_j. \quad (4.1)$$

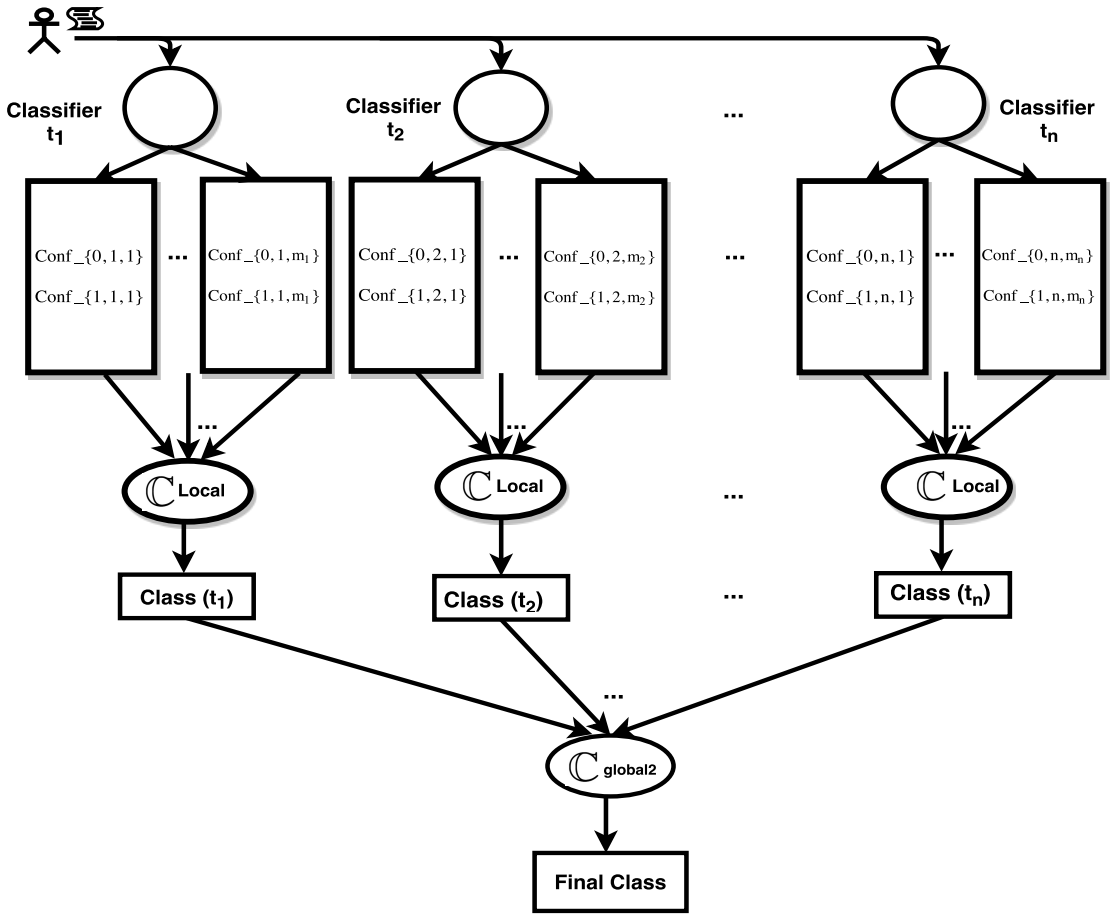


Figure 4.2: 2-steps classification process

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. 4.2, which may also lead to an "Unknown" assignment. Afterwards, a 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.

$$\mathbb{C}_{local}(j) = max_k(\mu_{X_k,j,i}), \text{ for all } k = \{0,1\}, \ i = 1,\ldots,m_j. \tag{4.2}$$

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 risk 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 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.

4.3.2 DRSA-BRE in DR risk assessment

In this section the application of the DRSA-BRE method to the same problem is explained. The first issue to consider is the nature of the data. In the data set 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 the 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 and 1, where 1 corresponds to a 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. 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 the 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. 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 assignments to be discovered. The monotonic relationship is global when it can be expressed by a single elementary condition concerning a gain-type or cost-type attribute. The local monotonicity

relationship requires a conjunction of two elementary conditions of different types, e.g., in the case of the 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 DR) and $\downarrow Age \leq Seventies$ (the more below *Seventies*, the higher the risk of DR), which boils down to condition $Age \in [Sixties, Seventies]$. Examples of rules discovered from DR data by DRSA-BRE are given in Section 4.4.2.

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* (e.g., 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 DR 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 A).

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 DR and the other patient hasn't). The result of the inconsistency

analysis is presented in Section 4.4.2.

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. The weight of a certain patient is calculated from the analysis of the class labels of 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 a 1-step aggregation.

4.4 Experimental Results

In this section we study the results obtained with the FRF and DRSA-BRE methods on the DR data set. 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: class 0 (low risk of DR - negative class), and class 1 (high risk of DR - positive class). Different values of the parameters of the algorithms are studied. Sensitivity and specificity, defined in chapter 2, are the measures used to evaluate the performance of each classifier. In addition, accuracy has been calculated in two ways, to take into account that some patients are not actually classified by the system (they may belong to the *Unknown* additional class): (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 (*i.e.* removing the ones classified as *Unknown*).

4.4.1 Results of the FRF analysis

Several parameters are used in the construction of a FRF. The values shown in Table 4.1 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 4.1: 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-step 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 4.2 (10-fold cross-validation on the training set) and 4.3 (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 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

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

Table 4.2: Cross-validation results with FRF with two-steps classification (100 trees, δ_1=0.2)

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

Table 4.3: Best classification results with FRF in the testing phase with two-steps aggregation (100 trees, δ_1=0.2)

similar (see Table 4.4). 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.

				Cross validation				Testing					
δ_1	γ	α	β	Acc2	Sens	Spec	Acc1	Acc2	Sens	Spec	Acc1	Unk	Index
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
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
0	1	0.4	1.0	77.76	70.99	80.35	65.37	81.95	87.02	80.52	69.48	169	Spec
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 4.4: Cross-validation results and best classification results with FRF with the one-step classification technique (100 trees)

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 4.5 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			
γ	α	β	Acc2	Sens	Spec	Acc2	Sens	Spec	Index
1	0.7	1.0	74.49	70.71	75.91	80.05	81.78	79.58	Acc
4	1.0	1.0	70.16	73.89	68.71	75.34	86.97	72.14	Sens
1	0.5	0.8	73.72	68.24	80.18	80.29	80.67	80.18	Spec

Table 4.5: Cross-validation results and best classification results with FRF with the two-steps aggregation with δ_1 =0, δ_2=0% and 100 trees

4.4.2 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 4.6, 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.

	class 0 (low risk of DR)	class 1 (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 4.6: 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 that results 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 4.3). The attributes with the highest values of the confirmation measure s (see appendix A for more information about this measure) 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 gender (Sex), Body Mass Index (BMI) and creatinine (Creat).

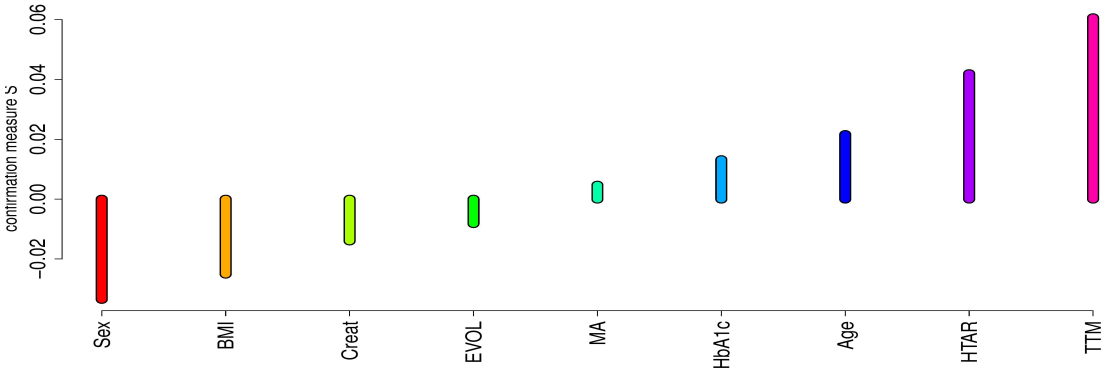


Figure 4.3: Bayesian confirmation measure computed for each attribute. This plot shows how each of the variables used in the DRSA-BRE model confirms the correct classification of DR patients.

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 4.3. 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.

#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 (class 1)",

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 (class 1)",

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 (class 1)",

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 (class 1)",

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 (class 1)",

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 (class 1)",

Support: 23 patients, s : 0.731

4.4.3 Comparison between FRF and DRSA-BRE

In this section the two ensemble classifiers described in this chapter 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 4.7).

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

Table 4.7: 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 4.8).

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

A comparison of the diagnosis made for each patient in the testing set (1111 patients) was done. Figure 4.4 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 class 0 than from class 1, 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.

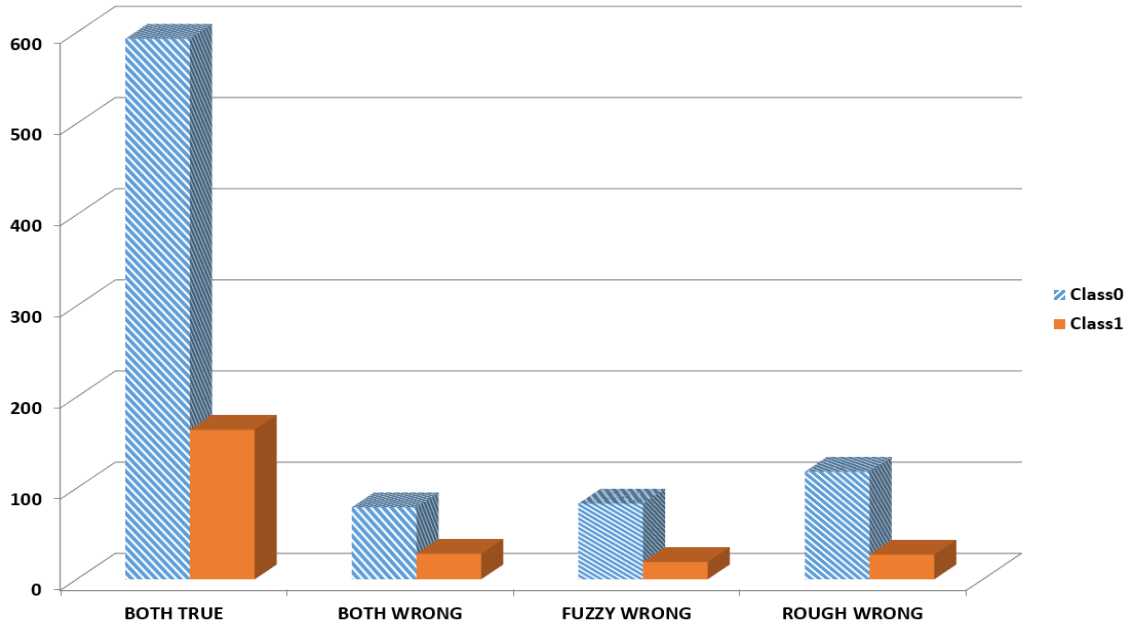


Figure 4.4: Comparison FRF and DRSA-BRE algorithms

4.5 Conclusions

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

On the one hand, FRFs, the core of the work of this dissertation, have been able to achieve an accuracy of 84% on the testing dataset, with a 80% sensitivity and a 85% specificity (however, almost 10% of the patients could not be classified). These results improve those of our initial works (Sanromà et al., 2016; Saleh et al., 2016b), in which 80% levels were achieved using a non-FRF (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, developed at Poznan University of Technology, 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.

FRFs and FDTs have many rules and every rule proposes a decision. Choosing the best way to aggregate the rules' output will certainly improve the overall results. There are many fuzzy-related aggregators that generalize the typical ones. That is what we are going to study in the next chapter.

CHAPTER 5

Learning fuzzy measures for aggregation in fuzzy rule-based models

5.1 Introduction

In the previous chapter it has been described how to use a FRF to assess the personalised risk of developing DR of a particular patient. One of the characteristics of a FRF is that a large number of classification rules are generated using different samples of the data. In our case, we have 100 FDTs with about 100 fuzzy rules in each one. When a new patient has to be classified, his/her data is introduced into the system and all rules are fired at different levels of satisfaction. An aggregation operator is needed to merge the outcome of all these rules. It is usually done with

the Maximum/Winner strategy, which consists on taking as answer the output of the rule with maximum activation (Saleh et al., 2018a). However, using this option the information provided by the rest of the rules is lost.

Aggregation operators are mathematical functions that merge a set of numerical arguments into a single one that summarizes them. They are widely used in many knowledge fields, such as sensor data fusion and decision making (Torra and Narukawa, 2007). Fuzzy integrals are one of the most general kinds of aggregation operators found on the literature. In fact, they are usually parameterised families of operators, so they have great flexibility in the aggregation of the inputs. For example, the Choquet integral generalizes both the weighted mean and the OWA operator (Beliakov et al., 2007),(Lucca et al., 2016) and the Sugeno integral generalizes the weighted maximum, weighted minimum and median operators (Grabisch and Labreuche, 2010),(Torra and Narukawa, 2007).

Fuzzy integrals rely on a fuzzy measure (or capacity), which is a set function that indicates the importance of the information sources (i.e., of each group of input sources). Defining a proper fuzzy measure for each problem is a crucial point in order to make a suitable aggregation of the inputs and obtain the correct corresponding outputs.

In this chapter we propose to use other aggregation methods in order to merge the contribution of the different rules that are activated by a certain patient's data. In particular, we study the use of fuzzy integrals and a new way of constructing the fuzzy measure is proposed, based on the confidence score of each of the contributing rules. In the second part of the chapter we study several ways of constructing a fuzzy measure that take advantage of the knowledge stored in the rules of the FRF. This work has been done in collaboration with Professor Torra (currently at Umea University, although he was previously at Skovde University) and Professor Bustince (Univ. Pública de Navarra). A 3-months research stay with Prof. Torra at Skovde and a shorter stay with Prof. Bustince in Pamplona were made during the

development of the thesis to work on the aspects described in this chapter.

One of the key elements in this chapter is the exploitation of the hierarchical structure of the FDTs composing the FRF. Hierarchical structures for aggregation have been studied in many works. In (Sugeno et al., 1995; Murofushi et al., 1997) a hierarchical aggregation based on Choquet integrals was proposed. The sufficient and necessary conditions to make an ordinary Choquet integral decomposable were obtained. Hierarchically $\perp$-decomposable fuzzy measures, a generalisation of $\perp$-decomposable measures, were formalized in (Torra, 1999). More recently, Fujimoto (Fujimoto, 2013) proved that the hierarchical bipolar Sugeno integral can be represented by a hierarchical bipolar Choquet integral.

The rest of the chapter is organized as follows. Section 5.2 presents the main concepts used in this chapter. In section 5.3 we introduce the novel fuzzy measures. In Section 5.4 we discuss the results obtained with several public data sets as well as the results for DR classification. It will be shown that the proposed method outperforms the best results given by the other approaches we have studied (Saleh et al., 2016b; Romero-Aroca et al., 2019; Saleh et al., 2017, 2018b).Finally, section 5.5 concludes the chapter.

5.2 Preliminaries

In this section we define the basic concepts that are used in this chapter.

Definition 5.2.1. A function $agg : [0,1]^\eta \to [0,1]$ is an aggregation function if and only if it fulfills the following properties:

- $agg(x, ..., x) = x$ (Unanimity or idempotency)
- If $(x_1, ..., x_\eta) \leq (y_1, ..., y_\eta)$ then $agg(x_1, ..., x_\eta) \leq agg(y_1, ..., y_\eta)$ (Increasing monotonicity)

Note that some authors use unanimity only in 0 and 1 (e.g. (Barrenechea et al., 2013) and (Lucca et al., 2017)).

Definition 5.2.2. A function $\top : [0,1]^2 \to [0,1]$ is a t-norm function if and only if it fulfills the following properties:

- $\top(x,y) = \top(y,x)$ (Commutativity)
- $\top(x,y) \leq \top(u,v)$ if $(x \leq u)$ and $(y \leq v)$ (Increasing monotonicity)
- $\top(x,\top(y,z)) = \top(\top(x,y),z)$ (Associativity)
- $\top(x,1) = x$ (Neutral element)

Examples of $\top$-norms include minimum and product.

Definition 5.2.3. A function $\perp : [0,1]^2 \to [0,1]$ is a t-conorm function if and only if it fulfills the following properties:

- $\perp(x,y) = \perp(y,x)$ (Commutativity)
- $\perp(x,y) \leq \perp(u,v)$ if $(x \leq u)$ and $(y \leq v)$ (Increasing monotonicity)
- $\perp(x,\perp(y,z)) = \perp(\perp(x,y),z)$ (Associativity)
- $\perp(x,0) = x$ (Neutral element)

Examples of t-conorms include maximum and bounded sum.

Definition 5.2.4. A function $\zeta : [0,1]^2 \to [0,1]$ is a copula function if and only if it fulfills the following properties, for all $x, x\prime, y, y\prime \in [0,1]$ with $x \leq x\prime$ and $y \leq y\prime$:

- $\zeta(x,y) + \zeta(x\prime,y\prime) \geq \zeta(x,y\prime) + \zeta(x\prime,y)$
- $\zeta(x,0) = \zeta(0,x) = 0$
- $\zeta(x,1) = \zeta(1,x) = x$

Definition 5.2.5. A fuzzy measure (also known as non-additive measure) m on a set X with cardinality $\eta = |X|$ is a set function $m : 2^X \to [0,1]$ fulfilling the following properties:

- $m(\emptyset) = 0, m(X) = 1$, (Boundary condition)
- $A \subseteq B$ implies $m(A) \leqslant m(B)$, for all $A, B \subset X$ (Monotonicity)

Fuzzy measures are a way to represent background knowledge about the importance of the sources when we aggregate their values. In that way, they are

used to weight the arguments in aggregation operators like the Choquet and Sugeno integrals. Fuzzy measures can be defined manually or they can be obtained from some domain data. There are quite a few families of fuzzy measures. We review some of them below.

Definition 5.2.6. Let m_{SL} be a fuzzy measure; then it is a Sugeno λ-measure if for some fixed $\lambda >$ -1 it holds that $m_{SL}(A \cup B) = m_{SL}(A) + m_{SL}(B) + \lambda m_{SL}(A) m_{SL}(B)$, for all $A \cap B = \emptyset$.

Definition 5.2.7. Let m_{DP} be a fuzzy measure. We say that m_{DP} is a distorted probability if there exists a probability distribution P on $(X, \wp(X))$ and an increasing function $f : [0,1] \to [0,1]$ such that $m_{DP} = f \circ P$.

In (Weber, 1984), the concept of $\perp$-Decomposable Fuzzy Measure was introduced. The definition is as follows:

Definition 5.2.8. A fuzzy measure m on a set X is a $\perp$-Decomposable Fuzzy Measure if there exists a t-conorm $\perp$ such that, for all $A, B \subseteq X$ with $A \cap B = \varnothing$, it holds that $m(A \cup B) = m(A) \cup (B)$.

Additive measures (when $\perp(x, y) = x + y$), possibility measures (when $\perp(x, y) = \max(x, y)$), and Sugeno λ-measures (when $\perp(x, y) = x + y + \lambda xy$, i.e. the Sugeno-Weber t-conorm) can be seen as examples of $\perp$-decomposable measures. Then, $\perp$-Decomposable measures are distorted probabilities.

Lucca et al.(Lucca et al., 2017) proposed to use a family of distorted probabilities. They were defined in terms of the cardinality of the set of values to be aggregated. This fuzzy measure $m_{PM} : 2^X \to [0,1]$ is defined as follows:

$$m_{PM}(A) = \left(\frac{|A|}{\eta}\right)^q \quad with \quad q > 0 \tag{5.1}$$

We call this measure power mean. For classification problems, the value of q can be optimized for each of the classes considered (Barrenechea et al., 2013).

In this chapter we use the Choquet and Sugeno integrals in order to aggregate the input data with respect to a fuzzy measure. The discrete Choquet integral is defined as follows:

Definition 5.2.9. Let X be a reference set with cardinality η and let m be a fuzzy measure on X; then, the *Choquet integral* of a function $f : X \to \mathbb{R}^+$ with respect to the fuzzy measure m is defined by

$$Choquet(f) = \sum_{i=1}^{\eta} [f(x_{s(i)}) - f(x_{s(i-1)})] \cdot m(A_{s(i)}), \tag{5.2}$$

where $f(x_{s(i)})$ indicates that the indices have been permuted so that $0 \leq f(x_{s(1)}) \leq \cdots \leq f(x_{s(\eta)}) \leq 1$, and where $f(x_{s(0)}) = 0$ and $A_{s(i)} = \{x_{s(i)}, \ldots, x_{s(\eta)}\}$.

In (Lucca et al., 2017) the Choquet-like Copula-based fuzzy integral (CC-integral) is defined. It uses a copula $\bullet$ instead of the product $\cdot$, as usual in the Choquet integral. When $\bullet = prod$ the CC-integral is the Choquet integral. The properties of this extended fuzzy integral were studied in (Lucca et al., 2018) and (Mesiar and Stupňanová, 2018).

Definition 5.2.10. Let X be a reference set with cardinality η and let m be a fuzzy measure on X; then, the *CC-integral* of a function $f : X \to \mathbb{R}^+$ with respect to the fuzzy measure m is defined by

$$CC\text{-}integral(f) = \sum_{i=1}^{\eta} [f(x_{s(i)}) \bullet m(A_{s(i)}) - f(x_{s(i-1)}) \bullet m(A_{s(i)})], \tag{5.3}$$

where $f(x_{s(i)})$ indicates that the indices have been permuted so that $0 \leq f(x_{s(1)}) \leq \cdots \leq f(x_{s(\eta)}) \leq 1$, where $f(x_{s(0)}) = 0$ and $A_{s(i)} = \{x_{s(i)}, \ldots, x_{s(\eta)}\}$, and where $\bullet$ denotes a copula.

Definition 5.2.11. Let m be a fuzzy measure on X with cardinality η; then, the

Sugeno integral of a function $f : X \to [0, 1]$ with respect to m is defined by

$$Sugeno(f) = \max_{i=1,\eta} \min(f(x_{s(i)}), m(A_{s(i)})), \tag{5.4}$$

where $f(x_{s(i)})$ indicates that the indices have been permuted so that $0 \leq f(x_{s(1)}) \leq ... \leq f(x_{s(\eta)}) \leq 1$ and $A_{s(i)} = \{x_{s(i)}, ..., x_{s(\eta)}\}$.

The main steps of Yuan and Shaw's induction method to build a FDT were presented in section 3.3.1.2. A bag of FDTs is used to build a FRF, as explained in section 4.3.1.

Definition 5.2.12. Given a set of rules to classify an object, $RC_{i,k}$ is the *Rule Confidence* of the ith rule about class $C_k \in C$, being C the set of all possibles classes that the classifier is considering. In the case of the induction method we are using, the Rule Confidence for each class is defined as:

$$RC_{i,k} = Truth(C_k|E_i) \tag{5.5}$$

5.3 Methodology

In this section, first we propose two novel fuzzy integrals to merge the conclusions of the rules when a FDT or a FRF is used to classify a new instance. These integrals are based on measures that take into account the confidence of each rule. After that, we present a more complex hierarchically-decomposable fuzzy measure that also takes into account the classification ambiguity at each node of the FDTs.

5.3.1 Fuzzy measures based on rule confidence

Fuzzy measures are used to give background knowledge in relation to the elements to be aggregated (Lucca et al., 2016). In our context, we aggregate data from a set of rules and we have a degree of support for each rule (*rule confidence*). These degrees

define a possibility distribution of the data. These values give important information, so taking them into account while we are making the aggregation process is valuable. In this section we propose the use of two fuzzy measures that will be built from these *rule confidence* values. The first measure is a distorted probability, whereas the second one is a Sugeno λ-measure.

5.3.1.1 Fuzzy measure based on distorted probability

The proposed distorted probability is defined using the following equation:

$$m_{DP}(A) = \left(\frac{\sum_{RC_{j\in A}} RC_j}{\sum_{RC_{i\in RC}} RC_i}\right)^q, with \quad q > 0 \tag{5.6}$$

where the value q needs to be optimised. Different methods can be used to optimize q (e.g. evolutionary algorithms (Lucca et al., 2017), (Barrenechea et al., 2013)). We use here gradient descent and wide search. Note that this fuzzy measure is a distorted probability because $m = f \odot P$, with

$$P_j = \frac{RC_j}{\sum_{RC_{i\in RC}} RC_i}, \quad and \quad f(x) = x^q. \tag{5.7}$$

5.3.1.2 Fuzzy measure based on Sugeno λ-measures

Sugeno λ-measures are defined as follows:

Definition 5.3.1. Let v be a function $v : X \to [0, 1]$ and $\lambda > -1$ such that

$$\begin{aligned} &(1/\lambda)(\textstyle\prod_{x_i\in X}[1 + \lambda v(x_i)] - 1) = 1 && if\ \lambda \neq 0 \\ &\textstyle\sum_{x_i\in X} v(x_i) = 1 && if\ \lambda = 0. \end{aligned}$$

Then, the fuzzy measure $m_{SL}(A)$ for any subset $A \in X$ defined as

$$m_{SL}(A) = \begin{cases} v(x_i) & if\ A = \{x_i\} \\ (1/\lambda)(\prod_{x_i\in A}[1 + \lambda v(x_i)] - 1) & if\ |A| \neq 1 \quad and \quad \lambda \neq 0 \\ \sum_{x_i\in A} v(x_i) & if\ |A| \neq 1 \quad and \quad \lambda = 0 \end{cases} \tag{5.8}$$

is a Sugeno λ-measure.

When only the function v is known, the value of λ can be obtained from the expression above (i.e. $(1/\lambda)(\prod_{x_i \in X}[1+\lambda v(x_i)]-1)=1$) and then it can be used to calculate m_{SL} for any subset of X.

In our proposal the weights $v(x_i)$ are the *rule confidence* values. Therefore, first the *rule confidence* values are used to build the Sugeno λ-measure finding an appropriate λ and then this fuzzy measure is used in the aggregation process.

5.3.2 A hierarchical measure generalizing a Sugeno λ-measure

In this section we present the second contribution of this chapter. We propose to use some of the values obtained during the induction of the FDT (like the classification ambiguity of a node or the rule confidence) to construct a hierarchically-decomposable fuzzy measure (HDFM) that generalizes a Sugeno λ-measure. A HDFM defines a hierarchy (based on the decision tree) and associates a value to each node in the hierarchy (calculated during the construction of the tree). Each node has associated a t-conorm. In our case, this is simply a parameter λ, as we restrict the model to use Sugeno-Weber t-conorms. Note that other t-conorms could be used for this purpose but using a parametric family (as the Sugeno-Weber one) is advantageous. As an alternative, we could have used Yager's family of t-conorms.

The HDFM is calculated during the construction of the decision tree, and it represents important knowledge associated to the rules. Our hypothesis is that taking advantage of the structure of the tree to define the fuzzy measure is relevant for a proper aggregation of the outputs. In other words, including this knowledge in a fuzzy measure may improve the classification accuracy.

In the remaining part of this section we introduce the definition of Hierarchically $\perp$-Decomposable Fuzzy Measure, and we present the method for obtaining the fuzzy measure for each of the nodes of the hierarchy.

5.3.2.1 Hierarchically ⊥-Decomposable Fuzzy Measures

In ⊥-decomposable fuzzy measures, the measure for a set A is defined as the combination through the t-conorm $\perp$ of the measures for the singletons $\{a_i\} \in A$ (Torra and Narukawa, 2007). This combination is homogeneous for all the elements in A as all elements are combined using the same t-conorm. Hierarchically ⊥-Decomposable Fuzzy Measures (HDFM) weaken this constraint by allowing different t-conorms in the combination process. To allow different t-conorms, elements in X must be represented in a hierarchical structure H (i.e. a tree), and then a t-conorm $\perp$ is attached to each node $A \in H$.

The fuzzy measure for a set X is computed recursively using the hierarchy. It uses values $v : X \to [0, 1]$ on the leaves, and a projection of every node A, denoted by $EXT(A)$, that corresponds to the set of leaves associated to A. $EXT(A)$ is computed recursively considering the descendants h_i of A:

$$EXT(A) = \begin{cases} h & if\ |h| = 1 \\ \cup_{h_i} EXT(h_i) & if\ |h| > 1 \end{cases} \tag{5.9}$$

The definition of HDFM follows. See (Torra, 1999; Torra and Narukawa, 2007) for details about HDFM and labeled hierarchies.

Definition 5.3.2. Let $L =< H, \perp, v >$ be a labeled hierarchy; then, the corresponding **Hierarchically ⊥-Decomposable Fuzzy Measure (HDFM)** of a set B is defined as $m(B) = m_{root}(B)$, where m_A for a node A with descendants $\{a_1, ..., a_n\}$ is defined recursively as

$$m_A(B) = \begin{cases} 0 & if\ |B| = 0 \\ v(B) & if\ |B| = 1 \\ \perp_A(m_{a_1}(B_1), ..., m_{a_n}(B_n)) & if\ |B| > 1 \end{cases} \tag{5.10}$$

Here, $B_i = B \cap EXT(a_i)$ for all $a_i \in A$.

We can associate any t-conorm to each node A. In this work we have used the

Sugeno-Weber t-conorm (i.e., $\perp(x, y) = x + y + \lambda xy$). Using this parametric family of t-conorms, we only need to keep a value λ for each node A.

5.3.2.2 Procedure for building a Hierarchically $\perp$-Decomposable Fuzzy Measure based on Classification Ambiguity and Rule Confidence

This section describes our approach to construct a hierarchically-decomposable fuzzy measure. First, in order to exploit the information obtained during the construction of the FDT with the method of Yuan and Shaw, we associate to each node A at level l for class $C_k \in C$ the value $v_{k,l}(A)$ based on the Rule Confidence degree rules denoted as RC_{Ak}.

Moreover, the induction algorithm calculates a value of Classification Ambiguity G_A (Equation 3.6) for each node $A \in H$ of the decision tree (H is the hierarchy). In the beginning of the tree induction (in the root), the classification ambiguity is equal to 1 (the highest value). At each step (when a node is to be split), the attribute that is chosen is the one that reduces the value of the classification ambiguity the most. Therefore, the method to build the decision tree ensures that the classification ambiguity decreases for every successive node in the branch of a tree. Because of that, the values of classification ambiguity in the FDT fulfill the requirements for constructing a HDFM (i.e., if $A_1 \subseteq A_2$ then $m(A_1) \leq m(A_2)$).

The main steps for computing the Hierarchically $\perp$-Decomposable Fuzzy Measure for a given set are described in Algorithm 1. This recursive algorithm finds the λ_A value for each node A taking into account the classification ambiguity of this node, G_A. Consequently, the fuzzy measure $m_A(A) = G_A$ satisfies the property of increasing monotonicity of fuzzy measures. Figure 5.1 illustrates the procedure. We call this measure m_{HDFM}.

Algorithm 1 Algorithm for defining the Hierarchically $\perp$-Decomposable Fuzzy Measures using Classification Ambiguity and Rule Confidence. The algorithm uses a Sugeno-Weber t-conorm with parameter λ_A for each node A.

Inputs:

- A class $k \in C$
- A FDT with a rule confidence value $RC_{A,k}$ for each branch (A is the last node and k is the class).

Notation:

h: number of levels in the tree, leaves are in level = 1
H_l: nodes at depth l in the tree
a_i: direct descendants of node A
G_A: classification ambiguity for node A

for each level $l = 2..h$ **do**
  for each *node* $A \in H_l$ **do**
    if (node A is a leaf) **then**
      $v_{k,l}(A) = RC_{A,k}$.
    else
      1. Calculate λ_A value of Sugeno λ-measure for node A using the classification ambiguity G_A from:

      $(1/\lambda_A)(\prod_{a_i \in A}[1 + \lambda_A v_{k,l-1}(a_i)] - 1) = G_A$

      2. Set $v_{k,l}(A) = G_A$
    end if
  end for
end for

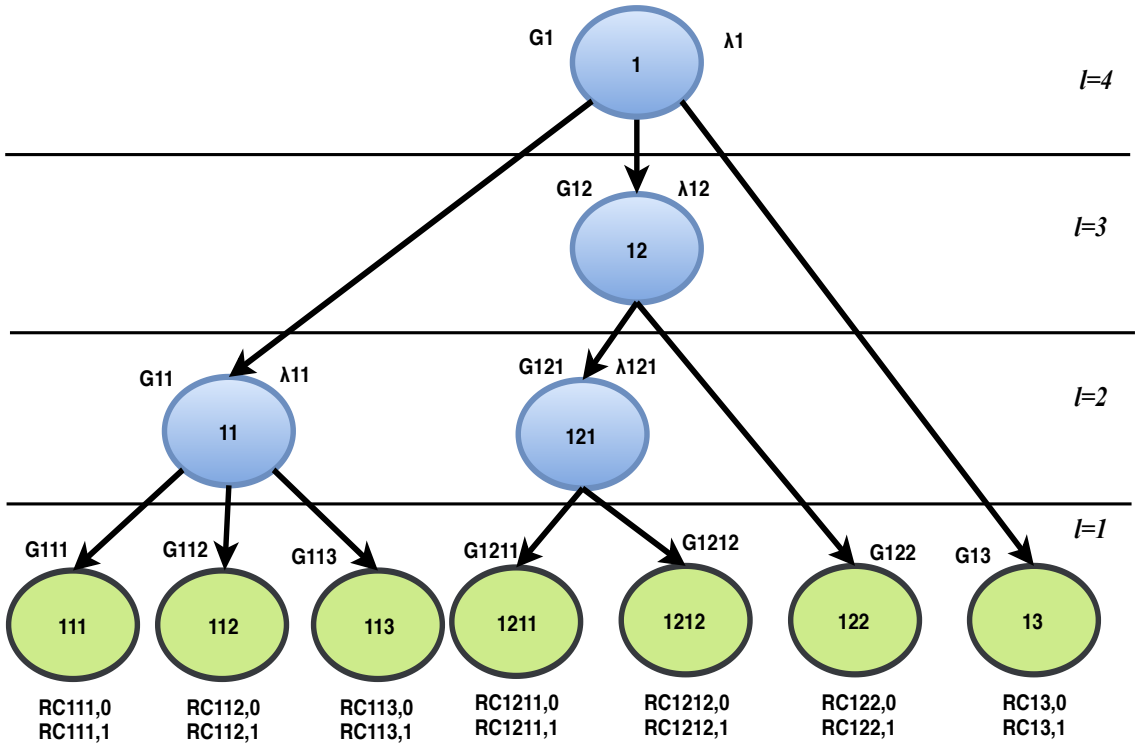


Figure 5.1: Hierarchically $\perp$-Decomposable Fuzzy Measure based on rule confidence for two classes C=0,1 and a hierarchy of 4 levels

5.3.3 Classification using the fuzzy rules

When a new observation u arrives, all rules are applied on the values of this observation, in order to obtain the membership degree to each of the possible conclusion classes $(RMCC_k(u))$. The steps for obtaining these values are the following:

1. Use a t-norm function to calculate the satisfaction degree of each rule $\mu_{R_i}(u)$.
2. Use the product between the satisfaction degree of each rule $\mu_{R_i}(u)$ and the degree of support of the rule (RC_{ik}) to obtain the membership degree to each class k $(RMCC_k(u))$.
3. Calculate a fuzzy measure (distorted probability $m_{DP}(A)$, Sugeno λ-measure $m_{SL}(A)$ or the fuzzy measure m_{HDFM} obtained in Algorithm 1) using the degree of support of the rule (rule confidence).
4. Aggregate the final value of each class using a fuzzy integral (CC-integral

(equation 5.3) or Sugeno integral (equation 5.4)). In the aggregation process, $RMCCs$ from the same class are weighted using fuzzy measures.

5. Compare the aggregation values. The final decision is the class label which has the maximum aggregation value.

5.4 Experimental results

In this section we present the performance of the proposed approaches on several classification datasets commonly used in the literature. Next, we present the application of these measures in the FRF used to assess the risk of developing DR, described in the previous chapter.

5.4.1 Classification Problems Datasets

Fourteen classification datasets have been taken to test and validate the performance of the proposed approach. The datasets correspond to all the binary classification problems available in the KEEL dataset repository (Alcalá-Fdez et al., 2011). Table 5.1 shows the main characteristics of these datasets. The records which have missing values have been removed from the datasets.

To evaluate the performance of the proposed aggregation algorithm and compare the results with the other approaches, we use the accuracy measure, defined in chapter 2.

All FRF models constructed in this work have been trained using 10-fold cross validation. To avoid the effect of the randomness of the FRFs, we repeat the test 10 times and then we calculate the median accuracy. The median has been used because it is robust to the outliers. The parameters for constructing the FRF models, which were analyzed in the previous chapter, are the following: 100 trees, $\alpha = 0.30$, $\beta = 0.70$, $\gamma = 3$.

After constructing the model for each training set, we classify the instances in

No.	Id	Dataset	Instances	Attributes
1	App	Appendicitis	106	7
2	Ban	Banana	5300	2
3	Bnd	Bands	365	19
4	Bup	Bupa	345	6
5	Hab	Haberman	306	3
6	Mag	Magic	1902	10
7	Pho	Phoneme	5404	5
8	Pim	Pima	768	8
9	Rin	Ring	7400	20
10	Sah	Saheart	462	9
11	Spe	Spectfheart	267	44
12	Tit	Titanic	2201	3
13	Two	Twonorm	740	20
14	Wis	Wisconsin	699	11

Table 5.1: Features of the datasets.

Short name	Fuzzy Integral	Fuzzy Measure	Definition
ICMM	CC-integral	distorted probability: m_{DP}	Eq. 5.6
ICMMS	CC-integral	Sugeno λ-measure: m_{SL} with $v = RC_{i,k}$	Eq. 5.8
ISM	Sugeno integral	distorted probability: m_{DP}	Eq. 5.6
ISMS	Sugeno integral	Sugeno λ-measure: m_{SL} with $v = RC_{i,k}$	Eq. 5.8
CHDFM	CC-integral	HDFM based on Sugeno-Weber t-conorms m_{HDFM}	Alg. 1
SHDFM	Sugeno integral	HDFM based on Sugeno-Weber t-conorms m_{HDFM}	Alg. 1
CCPM	CC-integral	power mean m_{PM}	Eq. 5.1

Table 5.2: Aggregation methods and fuzzy measures used in this work.

the testing set using different aggregation methods to merge the conclusions of the fuzzy rules. Table 5.2 shows the description of the 7 aggregation methods that we compare. The copula used in the CC-integral is always the minimum (Lucca et al., 2017; Saleh et al., 2018b).

In Table 5.2, CCPM uses m_{PM} as defined in Eq. 5.1. This is the simplest fuzzy measure since it does not take into account any knowledge about the rules. ICMM and ISM use a fuzzy measure for each output class C_k, which is constructed as a distorted probability from the rule confidence values $RC_{i,k}$ as defined in Eq. 5.5 for $q > 0$. Different methods can be used to optimize the value of q as e.g. evolutionary algorithms (Lucca et al., 2017),(Barrenechea et al., 2013). We use in this work gradient descent and wide search. Note that this fuzzy measure is a

distorted probability because $m = f \circ P$ with $f(x) = x^q$ and

$$P_j = \frac{RC_j}{\sum_{RC_{i \in R}} RC_i}. \tag{5.11}$$

ICMMS and ISMS also use the rule confidence to construct the fuzzy measure for each class using the Sugeno λ-measure, taking $v = RC_{i,k}$. Finally, CHDFM and SHDFM correspond to using CC-integral and Sugeno integrals using the hierarchical Sugeno λ-measure proposed in this dissertation, computed using the definitions and the algorithms given in previous sections.

The accuracy of the classification is given in Table 5.3. There are not large differences in accuracy on the different methods. CCPM is a simpler method based on the power mean, is the best only in two cases. We will now focus on comparing pairs of values with a difference of more than 2 units in the accuracy. Comparing the use of the Sugeno λ-measure m_{SL} instead of the distorted probability m_{DP}, we can see that ICMMS is better than ICMM in 5 cases and ICMM only beats ICMMS in 1 case (Hab). The same comparison between ISM and ISMS shows that ISMS wins 3 times and ISM only 1 (Bup). From these observations, we can see that the use of the rule confidence in the Sugeno λ-measure m_{SL} outperforms the distorted probability m_{DP}.

Regarding the results of using hierarchically decomposable measures, CHDFM obtains equal or larger accuracy than ICMMS in 9 cases and SHDFM beats ISMS in 7 cases, but with smaller differences. The best result for each dataset is highlighted in bold.

The execution time of the results has also been analyzed (Table 5.4). Time includes the whole process from reading the dataset files to printing the results. The experiments have been run on PC with AMD A8-5500B (4 CPUs) 3.2GHz processor, 25 GB RAM memory, and 500 GB Hard disk storage.

Methods using Sugeno λ-measures m_{SL} show faster execution time on all datasets. Time is approximately half than the one employed by the methods based on distorted

No.	Id	ICMM	ICMMS	ISM	ISMS	CHDFM	SHDFM	CCPM
1	App	84.91	83.02	83.96	**86.79**	83.02	80.19	85.84
2	Ban	85.89	86.81	86.85	86.53	86.04	**87.00**	**87.00**
3	Bnd	72.33	71.51	69.86	70.96	71.51	**72.60**	71.06
4	Bup	60.87	70.03	63.48	61.45	**70.43**	61.74	67.25
5	Hab	73.53	69.46	**74.84**	73.53	74.51	72.55	74.48
6	Mag	79.81	79.86	**80.39**	79.81	80.34	**80.39**	80.34
7	Pho	83.44	82.99	83.09	83.09	83.27	82.99	**83.55**
8	Pim	76.17	76.17	76.69	76.30	**77.60**	76.56	76.95
9	Rin	87.97	90.05	87.92	88.04	**90.70**	88.11	90.41
10	Sah	**71.43**	71.00	70.78	70.56	71.00	71.21	70.87
11	Spe	79.03	78.28	**79.78**	80.52	78.65	78.28	79.74
12	Tit	78.96	79.05	79.05	**80.19**	78.83	78.60	78.87
13	Two	85.68	**89.46**	**89.46**	85.81	85.95	85.27	85.68
14	Wis	92.99	**97.22**	94.42	95.11	93.13	95.85	96.71

Table 5.3: Accuracy of the models in the different datasets.

No.	Id	ICMM	ICMMS	ISM	ISMS	CHDFM	SHDFM	CCPM
1	App	2:26	1:21	2:29	1:22	1:23	1:22	2:27
2	Ban	3:45	2:15	3:45	2:17	2:19	2:19	3:43
3	Bnd	2:24	1:45	3:12	1:47	1:56	1:53	3:14
4	Bup	3:23	1:60	3:22	1:61	1:59	1:59	3:21
5	Hab	3:12	1:44	3:08	1:44	1:58	1:57	3:10
6	Mag	5:56	3:24	5:54	3:27	3:38	3:36	5:55
7	Pho	8:19	5:05	8:18	5:05	5:25	5:23	8:19
8	Pim	3:59	2:20	3:57	2:23	2:37	2:33	3:54
9	Rin	3:11	2:13	3:15	2:14	2:24	2:11	3:16
10	Sah	2:00	2:00	2:02	2:01	2:20	2:21	2:03
11	Spe	3:43	2:22	3:43	2:24	2:30	2:30	3:42
12	Tit	2:57	1:31	2:56	1:31	1:35	1:36	2:59
13	Two	4:01	2:32	4:02	2:32	2:39	2:38	4:05
14	Wis	3:28	1:33	3:27	1:34	1:42	1:39	3:31

Table 5.4: The execution time for all methods on the datasets.

probability or power mean, because they do not need to optimize the parameter q. Taking into account that the cost of read/write operations is constant, the temporal improvement is really significant. The proposed methods CHDFM and SHDFM are a bit slower than ICMMS and ISMS because the hierarchical structure of the fuzzy measure needs more time to calculate the Sugeno-Weber parameters λ_A for all nodes A.

HM	Sensitivity	Specificity	Accuracy	TP	FN	FP	TN	Method
79.77	80.42	79.14	79.42	193	47	180	683	ICMM
77.88	76.67	79.14	78.60	184	56	180	683	ICMMS
79.74	78.75	80.76	80.33	189	51	166	697	ISM
78.05	81.67	74.74	76.24	196	44	218	645	ISMS
80.47	**81.25**	**79.72**	**80.05**	**195**	**45**	**175**	**688**	CHDFM
79.77	79.58	79.95	79.88	191	49	173	690	SHDFM
78.31	78.75	77.87	78.06	189	51	191	672	CCPM

Table 5.5: Classification results on DR, with $\alpha = 0.30$, $\beta = 0.70$ and $\gamma = 3$.

5.4.2 Application on the diabetic retinopathy risk assessment problem

The different aggregation methods shown in the previous section have also been tested with data of the diabetic patients collected by Sant Joan de Reus University Hospital. In this case we have a data set slightly larger than the one considered in the previous chapters. The data set has 3346 records labelled with two classes: Class 0 (no DR) and Class 1 (DR). The data set has been divided into a training set with 2243 examples (1605 not suffering from DR and 638 who suffer from DR) and a testing set with 1113 patients (863 not suffering from RD and 240 who suffer from DR).

In this section we study the results achieved by the FRF with the proposed aggregators on the DR testing dataset. We have made a comparison with the different methods given in Table 5.2, as done in Section 5.4.1. Table 5.5 shows the sensitivity, specificity and accuracy of each method. A balancing of these measures is done with the harmonic mean (HM) of specificity and sensitivity.

The basic winner rule (WR) for making decisions is well-known in rule-based models; it simply uses the max t-conorm to aggregate the outputs of the rules. To verify the quality of the aggregation methods proposed in this work in comparison with the state of art, the aggregation method based on Choquet integrals (CC-integral) proposed by Lucca (Lucca et al., 2017) has been reimplemented and used with FRF models. This method uses the power measure in the fuzzy integrals

(see equation 5.3). The rest of the methods correspond to the different versions of Choquet and Sugeno integrals using the Rule Confidence (RC) for the fuzzy measure construction.

By observing the results in table (5.5), ICMM and ISM (using distorted probability measures m_{DP}) achieved achieved a HM score higher than 79%. CCPM (based on power mean m_{PM}) obtained worse results, with HM=78.31%, sensitivity=78.75% and specificity=77.87%. The ISMS method classifies the patients with RD very well. It obtained the highest sensitivity (81.67%) but the specificity score decreases too much. On the contrary, ICMMS decreased the sensitivity values while keeping specificity near 80%.

The new methods based on hierarchical measures perform much better. SHDFM scored over 79% on all the statistical measures. CHDFM obtained the best performance results on the DR testing dataset. HM and accuracy and specificity are all around 80%. The sensitivity of the model is really high (81.25%). This seems to indicate that with the use of HDFM we are able to integrate better the conclusions of the rules in this application domain. It indicates that constructing fuzzy measures with the same structure as the tree can improve the accuracy of the classification made by the set of rules.

5.5 Conclusion

The use of fuzzy measures in aggregation operators shows good performance; thus, *rule confidence* values can play an important role in the aggregation process.

In this work, a Sugeno λ-measure and a distorted probability are used with Choquet and Sugeno integrals. These new aggregation approaches are used in a FRF composed by FDTs. The models with these new aggregation approaches outperform the same models with the max t-conorm aggregation operator.

In comparison with the models that use the same Choquet and Sugeno integrals with a fuzzy measure which utilizes the number of rules, the new approach obtains

better performance results as well. FRFs have shown better performance results than FDTs without any unclassified patients.

Our method applied to aggregate the multiple outputs obtained from the FDTs in a FRF seems to be very relevant to improve the accuracy of previous approaches that consider a flat fuzzy measure. We have proposed a Hierarchically $\perp$-Decomposable Fuzzy Measure (HDFM), which follows the tree structure. The fuzzy measure is constructed using knowledge about the rules (such as the rule confidence RC and the classification ambiguity G of the premises). The results showed that RC and G play a substantial role to improve the aggregation performance. In terms of execution time, the methods using m_{HDFM} are faster than those using distorted probabilities m_{DP}, because we do not need to optimize the parameter q for each class. Therefore, this new way of constructing a fuzzy measure with hierarchical structure is efficient and obtains good performance results.

The application to the DR risk assessment problem has shown a relevant improvement over the other methods. Results obtained in the DR problem show interesting properties of the method. In particular, it permits to reduce the false positives and false negatives in the classification of the patients, which is very important in this problem.

This work is oriented to build a CDSS. The CDSS will be used in the medical centers by family physicians who are not expert ophthalmologists. The goal is to help the physicians to estimate the risk of developing DR of the new patients. The CDSS is already in use in Sant Joan de Reus University Hospital (SJRUH) and the plan is to use it in all Catalonia's medical centres. The next chapter presents the CDSS that is applied in the hospital in more detail.

CHAPTER 6

Retiprogram

6.1 Introduction

The analysis of the image of the retina of a diabetic patient can reveal the presence of lesions in the eye. In Catalonia (as well as in the rest of Spain and many other European countries), annual screening is recommended for diabetic patients. However, this treatment is expensive and it is not feasible to follow this recommendation. Due to the large number of diabetic patients, Catalan health care centers are only able to screen each patient on average every 2.5 years (Saleh et al., 2020). As the incidence of DR is about 10%, most of the tests are unnecessary. Family doctors, who are in charge of continuous monitoring of diabetic people, are not experts in the diagnosis of DR. For this reason, they are not able to decide if it

is not absolutely necessary to perform a screening analysis on a particular patient. The lack of an appropriate filtering procedure causes an excessive increase in the cost of screening. Moreover, patients must be subjected to unnecessary tests, with a loss of time for the patients as well as for the experts that must analyze thousands of eye images every year (Romero-Aroca et al., 2016b).

As described in the previous chapters, we have been working in the last years on the development of an intelligent decision support system that estimates the risk of developing DR using the data available in the EHR of diabetic patients. Depending on the risk level, the system recommends the most appropriate time for the next visit. The goal of developing this CDSS is that family physicians have a tool integrated with their usual software at the hospital that can be executed in order to calculate the degree of DR risk.

6.2 A CDSS based on a Fuzzy Random Forest

Retiprogram is a clinical decision support system composed by a set of fuzzy rules. In particular, the underlying model is the one described in Chapter 4: a FRF, consisting of a set of 100 FDTs.

Machine Learning algorithms have been used to construct the fuzzy rules of each of the trees. In particular, we used a modified version of Yuan and Shaw's induction algorithm to minimize a measure of classification ambiguity based on fuzzy sets, as explained in chapter 3. The inference method applied to classify new cases is the Mamdami procedure. The output of the Mamdami procedure is combined with a hierarchically $\perp$-decomposable fuzzy measure-based aggregation operator based on the Choquet integral (see chapter 5 and (Saleh et al., 2019)). The final classification model achieves an accuracy of 80%, with a sensitivity of 81.3% and a specificity of 79.7%.

6.3 Retiprogram at Hospital Universitari Sant Joan de Reus

The Hospital Universitari Sant Joan de Reus (HUSJR) is the leading health center of Reus. It serves an area of Catalonia with 247,174 inhabitants. The total number of type-2 diabetic patients screened yearly has increased from 5000 to 5500 in the last 10 years, from a total of 15,811 T2DM patients.

After working in close cooperation with the computer engineers at HUSJR, Retiprogram is now included in the toolset of the desktop software of the doctors. In 2019 it was tested by the ophthalmologists of the hospital, and since the beginning of 2020 it may be used by all the doctors of the hospital. A graphical interface has been developed following the style of the rest of the software of the hospital, which helps doctors to be familiar with it and work efficiently with this new system. The frame is as simple as possible, having the input data in the left of the window and the output data at the right, as shown in Figure 6.1.

Most of the 9 input values are automatically retrieved from the EHR that is stored in the hospital database. If their values are missing or must be updated, the doctor introduces manually the new values, and the system performs the calculation, which is made instantly. This tool shows if there is a risk of DR or not and it indicates the degree of confidence in the answer.

In the graphical interface, a red to green bar is displayed on the right side. A red colour means that the system predicts with high certainty that the patient may develop DR, whereas a green bar means that the system predicts with high certainty that the patient does not have a high risk to develop DR. Depending on the answer, the system recommends the date for the next visit (from an immediate visit to the ophthalmologist, if the risk is very high, to a visit in 36 months, if the risk is very low). Table 6.1 presents the relation between the certainty degree of developing DR and the next visit time.

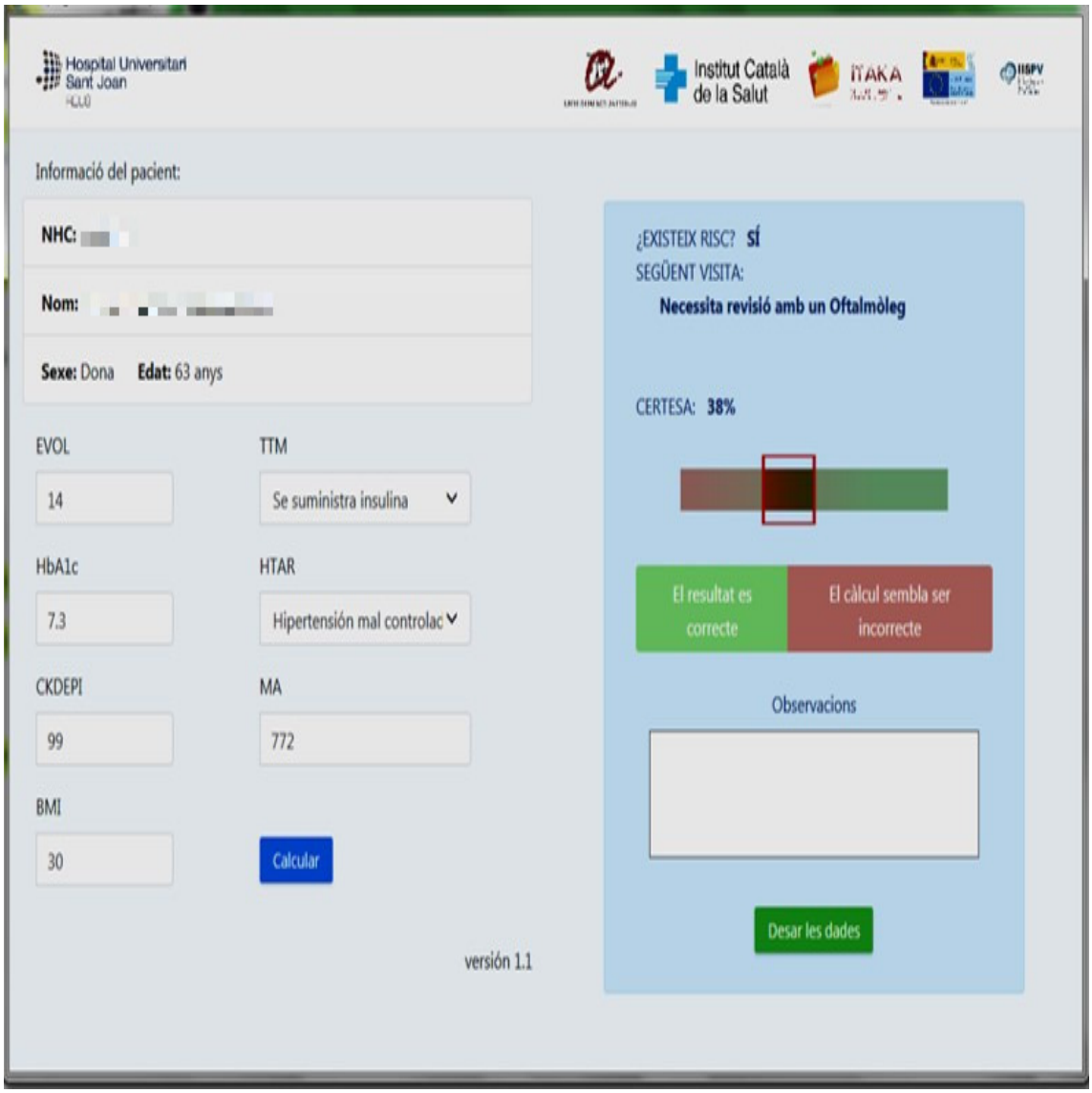


Figure 6.1: Retiprogram at HUSJR.

As the system is not completely perfect in its predictions, the doctor must validate the recommendation. For this reason, the user can introduce comments in order to justify the decision made, especially in the case of not following the recommendation given by the system.

Degree of certainty of developing DR	Time of the next visit
If Certainty < 50	6 months
If $65 >$ Certainty ≥ 50	1 year
If $80 >$ Certainty ≥ 65	2 years
If Certainty ≥ 80	3 years

Table 6.1: Relation between the certainty degree of developing DR and the next visit time.

6.4 Future Work

We are collecting now the feedback from the doctors in the use of the Retiprogram system. After its continuous use in 2020, we will analyze the performance of the system and the comments given by the doctors in the cases where the model gives an unreasonable answer. In another Ph.D. of the ITAKA research group we are already working on the explainability of the output given by the classification system, so that the system can give a justification of its prediction if necessary.

CHAPTER 7

Conclusion and Future Work

In this chapter, we conclude the Ph.D. dissertation work by summarising the main contributions and suggesting some lines of future work.

7.1 Summary of contributions

Diabetic retinopathy risk detection is a hot topic in the Machine Learning research landscape nowadays. This disease can be detected by analysing the retina image or it can be predicted by analysing the personalized risk factors from the EHR data. The key idea of this second option is to avoid using a non-mydriatic fundus camera, which is costly and requires the availability of trained ophthalmologists.

Many researchers have proposed CDSSs that use Computer Vision techniques for

the detection of lesions caused by DR. Classical Computer Vision methods as well as novel Deep Learning techniques are being used for this task. However, there are few researches that use the EHR data for DR risk estimation, as proposed in this work.

This dissertation is a pioneer work in this new line. This work proposes the use of rule-based models to classify the patients on potential developers of DR or not, and to obtain the degree of certainty of the decisions.

In this work, a fuzzy decision tree (FDT)-based model was initially proposed. To avoid the uncertainty of classifying the patient in different classes with a very similar certainty degree, we introduced a parameter to guarantee that the difference of the certainty degree associated to the classes is higher than a certain value.

One of the novel contributions of this dissertation appears in the induction of the FDT. Additional steps have been proposed to guarantee selecting the best attribute to expand the branch of the FDT when there is more than one attribute with the best value. This dissertation has analyzed two different ways of building FDTs, examining different policies that may be used to combine their decisions.

Another contribution of this thesis is the extension of the previous model to an ensemble model, called a Fuzzy Random Forest (FRF). Two different ways of aggregating the results of the FDTs rules were analyzed. To guarantee a minimum degree of certainty when classifying a patient into a class, another threshold parameter was introduced.

One of the fields in Machine Learning that looked promising to improve the proposed models was the use of aggregation functions. In this area we have worked with two well-known world experts, Prof. Torra and Prof. Bustince. With fuzzy rule-based models we have many rules and every rule suggests an assignment of the analyzed object to a certain class. To aggregate the outputs of the FDTs' rules, three aggregation methods have been proposed in this dissertation. Concretely, three fuzzy measures have been proposed and used with the Choquet and the Sugeno integrals to form new aggregation operators. The novel fuzzy measures use the

rule confidence values in their calculations. They are based on Sugeno λ-measure and distorted probabilities. One of the proposed fuzzy measures is a hierarchically $\perp$-Decomposable fuzzy measure. In this fuzzy measure, the hierarchical structure and the values at each node of the FDT are used to build the fuzzy measure.

The work of this dissertation has not been purely theoretical, but it has also had a practical application, with the development of the Retiprogram system, in the context of several FIS funded research projects led by Prof. Moreno and Dr. Valls (URV) and Dr. Romero (S.Joan Hospital). Retiprogram has been in use in the Sant Joan University Hospital in Reus for several months, and we are currently evaluating with the Catalan Health Department the possibility of extending its use to other medical centres in Catalonia. Two forms of the CDSS have been developed: a desktop application and a web service API with an associated GUI. By using the web service it is easy to build mobile and web-based applications to access the CDSS and get the results of any query.

7.2 Future work

The work presented in this Ph.D. thesis opens interesting research lines. The following topics can be pursued to continue the work of this thesis:

- The management of real world medical data is a challenge which we experienced repeatedly through this thesis. We will validate the current approaches using other bigger datasets for DR from all Catalonia in order to examine the performance of the approaches. We need to see if we obtain similar results or if the performance decreases when data from other medical centres is used. This work is already under progress.
- Expand the fusion procedure presented in chapter 3 for the case of FDTs to the case of ensemble models like FRFs, considering mixed forests with trees built either with fuzzy entropy or classification ambiguity.
- The work in this thesis focused on a binary classification problem. In the future

work, the plan is to test the proposed approaches on more complex problems with multiple classes and to study the effect of each of the contributions in this thesis to the different challenges, for example the presence of imbalanced classes.

- Combine the current CDSS based on EHR data with the Computer Vision models that analyse the retina images, to develop the complete framework proposed in the FIS funded research projects. This work is already under way.
- We are collecting during 2020 the feedback from the doctors in the Sant Joan University Hospital in the real use of the Retiprogram system. At the beginning of 2021 the plan is to analyze the performance of the system and the comments given by the doctors in the cases where the model gives an unreasonable answer.
- Propose methods to improve the explainability of the output given by the classification system. There is a new PhD on this topic already being developed in the ITAKA research group.

Appendices

APPENDIX A

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

A.1 Description of the algorithm

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 $q_j \in C$, $\phi : U \rightarrow \mathbb{R}$, 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.

A.1.1 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).

A.1.2 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 and 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 and 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.

A.1.3 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 $C = \{q_1, \dots q_r\}$, we can define the condition part of the rule as a conjunction of elementary conditions on a particular subset of attributes:

$$E = e_{j_1} \wedge e_{j_2} \wedge \dots \wedge e_{j_v}, \tag{A.1}$$

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

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 $q_i \triangleright E$, $j \in \{1, \dots, r\}$. 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:

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

The conditions of definition (A.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.

Appendix A. Dominance-based Rough Set Balanced Rule Ensemble (DRSA-BRE)

APPENDIX B

Prizes and intellectual property

B.1 Prizes

In this appendix, we present the awards obtained with the work in this Ph.D. thesis:

- In the 20th International Conference of the Catalan Association for Artificial Intelligence (CCIA 2017) which took place on 25th - 27th October 2017 in Deltebre (Tarragona), we received the prize for the best article in the conference. The award-winning work is entitled *"Integration of different fuzzy rule-induction methods to improve the classification of patients with diabetic retinopathy"*. The authors of the work are Emran Saleh and the supervisors of his doctoral dissertation, Dr. Aïda Valls, Prof. Antonio Moreno and Dr. Pere Romero (Pere Virgili Health Research Institute). Dr. Romero is the head of

the Ophthalmology Unit of the Sant Joan de Reus University Hospital, and he has working together with the ITAKA group on several FIS funded Spanish research projects on the application of intelligent data analysis and computer vision to improve the early detection of DR. Figure B.1 shows the certificate.

Figure B.1: The best article in CCIA 2017 conference award

- The ITAKA group was awarded the URV 2018 Social Impact of Research prize. This award has been given to the work entitled: *"Design, construction and evaluation of a CDSS for the personalised screening and follow-up of diabetic retinopathy patients"*, which corresponds basically to the research described in this dissertation. As commented previously, this work was developed in close collaboration with the Ophthalmology Unit of Sant Joan University Hospital and the Health Research Institute Pere Virgili. Figure B.2 shows the certificate.

UNIVERSITAT
ROVIRA i VIRGILI
Consell Social

Premi Consell Social
a l'impacte social de la recerca
Convocatòria any 2018

El grup de recerca **ITAKA: Tecnologies Intel·ligents per a la Gestió Avançada del Coneixement,** ha guanyat el premi Consell Social a l'impacte social de la recerca, àmbit ciències experimentals, enginyeria i arquitectura, en la modalitat EXPOST, amb el treball **"Disseny, construcció i avaluació d'un sistema clínic de suport a la presa de decisions per al cribatge i seguiment personalitzat de pacients amb retinopatia diabètica",** presentat per l'investigador **Antonio MORENO RIBAS.**

Joan Pedrerol Gallego
President del Consell Social
setembre 2018

Figure B.2: URV Social Impact of Research 2018 award.

B.2 Intellectual Property

The intellectual property of the Retiprogram system and also of a simplified Web service Risk Calculator (for demo purposes) have been registered at the Benelux Intellectual Property Office with the support of URV and Fundació URV and Institut Investigaci Sanitaria Pere Virgili:

- iDEPOT Number: 069047, Reference: T-2015/014, Name: RETIPROGRAM v1.
- iDEPOT Number: 116715, Reference: IISPV-URV-2/19, Name: RETIRISK-CALC.

The evidences of these registrations are given in the next pages.

B B I E
O B P I

i-DEPOT

i-DEPOT BEWIJS / PREUVE / EVIDENCE

i-DEPOT NUMMER / NUMERO / NUMBER
069047

NAAM / NOM / NAME
FUNDACIÓ URV FUNDACIÓ URV-CENTRE DE TRANSFÈRENCIA I INNOVACIÓ
Avda. Paisos Catalanas 18
43007 Tarragona
Spain

DATUM / DATE
Thu Dec 03 11:18:06 CET 2015

KENMERK / REFERENCE
T-2015/014

Dit elektronische bestand vormt het bewijs dat alle hierin opgenomen gegevens op de aangegeven datum bij het Benelux-Bureau voor de Intellectuele Eigendom (BBIE) werden ingediend en dat deze sindsdien niet zijn veranderd.

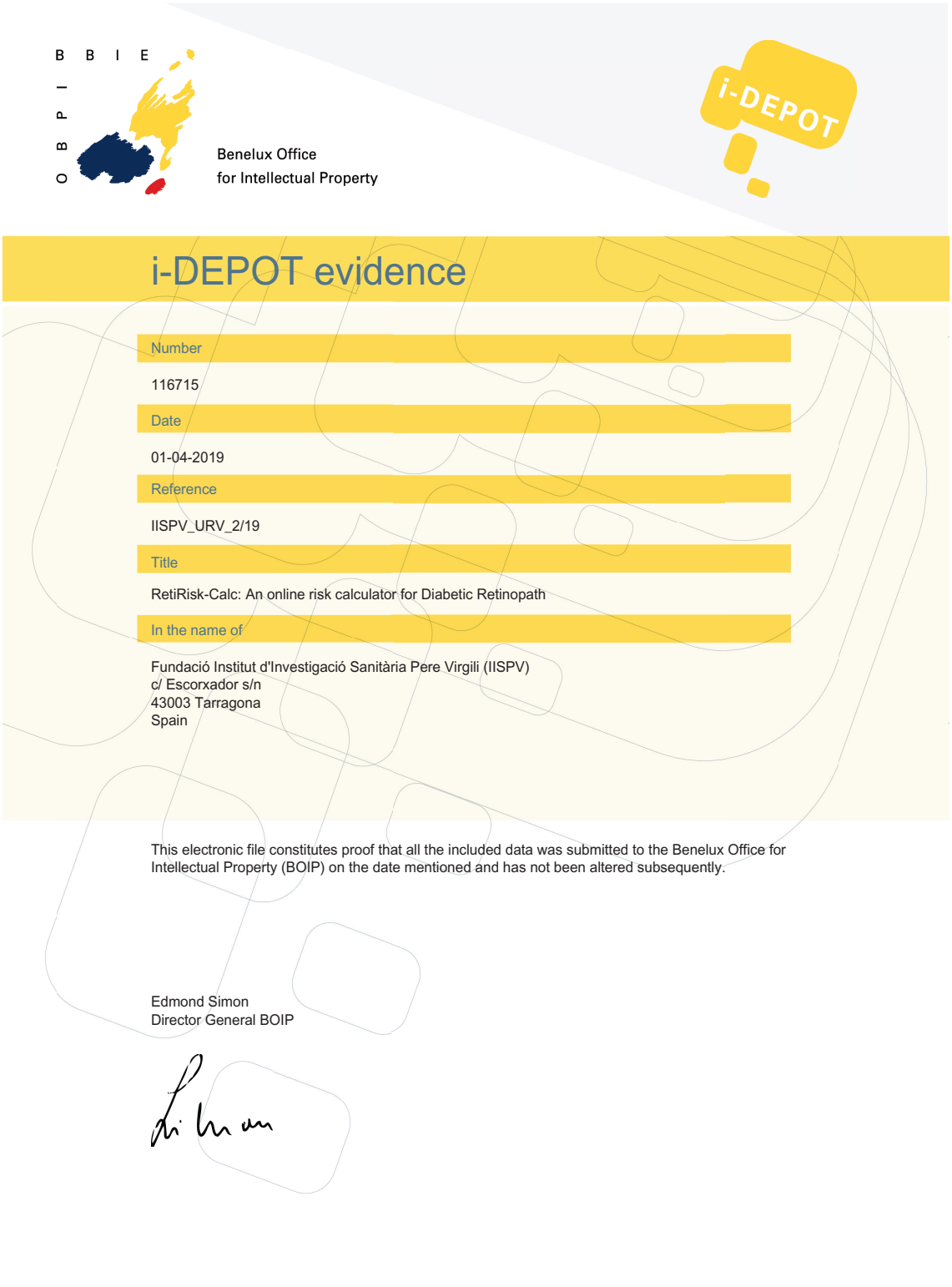
Ce fichier électronique constitue la preuve que toutes les données reprises ont été remises à l'Office Benelux de la Propriété intellectuelle (OBPI) à la date mentionnée et n'ont pas été modifiées depuis.

This electronic file constitutes proof that all the included data was submitted to the Benelux Office for Intellectual Property (BOIP) on the date mentioned and has not been altered subsequently.

Edmond Simon
Directeur-Generaal BBIE / Directeur général OBPI / Director General BOIP

BENELUX-BUREAU VOOR DE INTELLECTUELE EIGENDOM | OFFICE BENELUX DE LA PROPRIETE INTELLECTUELLE | BENELUX OFFICE FOR INTELLECTUAL PROPERTY

Figure B.3: Intellectual Property certificate RETIPROGRAM



B B I E
O B P I

Benelux Office
for Intellectual Property

i-DEPOT

i-DEPOT evidence

Number

116715

Date

01-04-2019

Reference

IISPV_URV_2/19

Title

RetiRisk-Calc: An online risk calculator for Diabetic Retinopath

In the name of

Fundació Institut d'Investigació Sanitària Pere Virgili (IISPV)
c/ Escorxador s/n
43003 Tarragona
Spain

This electronic file constitutes proof that all the included data was submitted to the Benelux Office for Intellectual Property (BOIP) on the date mentioned and has not been altered subsequently.

Edmond Simon
Director General BOIP

1/3

BENELUX OFFICE FOR INTELLECTUAL PROPERTY | WWW.BOIP.INT

Figure B.4: Intellectual Property certificate RETIRISK-CALC

References

Akyol, K., Bayir, S., and Sen, B. (2017). A decision support system for early-stage diabetic retinopathy lesions. *International journal of advanced computer science and applications*, pages 369–379.

Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., and Herrera, F. (2011). Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, 17.

Alickovic, E. and Subasi, A. (2016). Medical decision support system for diagnosis of heart arrhythmia using DWT and random forests classifier. *Journal of medical systems*, 40(4):1–12.

An, S. and Hu, Q. (2012). Fuzzy rough decision trees. In *Rough Sets and Current Trends in Computing*, pages 397–404. Springer.

Balakrishnan, V., Shakouri, M. R., and Hoodeh, H. (2013). Developing a hybrid

predictive system for retinopathy. *Journal of Intelligent & Fuzzy Systems*, 25(1):191–199.

Barrenechea, E., Bustince, H., Fernandez, J., Paternain, D., and Sanz, J. A. (2013). Using the Choquet integral in the fuzzy reasoning method of fuzzy rule-based classification systems. *Axioms*, 2(2):208–223.

Beliakov, G., Pradera, A., and Calvo, T. (2007). *Aggregation functions: A guide for practitioners*, volume 221. Springer.

Beloufa, F. and Chikh, M. (2013). Design of fuzzy classifier for diabetes disease using modified artificial bee colony algorithm. *Computer Methods and Programs in Biomedicine*, 112(1):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., and Słowiński, R. (2012). Inductive discovery of laws using monotonic rules. *Engineering Applications of Artificial Intelligence*, 25(2):284–294.

Błaszczyński, J., Greco, S., Słowiński, R., and Szelkag, M. (2009). Monotonic variable consistency rough set approaches. *International journal of approximate reasoning*, 50(7):979–999.

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

Błaszczyński, J., Słowiński, R., and Szelkag, M. (2011b). Sequential covering rule

induction algorithm for variable consistency rough set approaches. *Information Sciences*, 181(5):987–1002.

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

Bodjanova, Slavka (1993). *Fuzzy Sets and Fuzzy Partitions*, pages 55–60. Springer Berlin Heidelberg, Berlin, Heidelberg.

Bourne, R. R., Jonas, J. B., Flaxman, S. R., Keeffe, J., Leasher, J., Naidoo, K., Parodi, M. B., Pesudovs, K., Price, H., White, R. A., et al. (2014). Prevalence and causes of vision loss in high-income countries and in eastern and central europe: 1990–2010. *British Journal of Ophthalmology*, 98(5):629–638.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *Pattern recognition (ICPR), 2010 20th international conference on*, pages 3121–3124. IEEE.

Cadenas, J. M., Garrido, M. C., and Martínez, R. (2013). Feature subset selection filter–wrapper based on low quality data. *Expert systems with applications*, 40(16):6241–6252.

Cadenas, J. M., Garrido, M. C., Martínez, R., and Bonissone, P. P. (2012). Extending information processing in a fuzzy random forest ensemble. *Soft Computing*, 16(5):845–861.

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

Chalk, D., Pitt, M., Vaidya, B., and Stein, K. (2012). Can the retinal screening interval be safely increased to 2 years for type 2 diabetic patients without retinopathy? *Diabetes care*, 35(8):1663–1668.

Chandrakumar, T. and Kathirvel, R. (2016). Classifying diabetic retinopathy using deep learning architecture. *International Journal of Engineering Research and Technology*, 5(6).

Chang, P.-C., Fan, C.-Y., and Dzan, W.-Y. (2010). A cbr-based fuzzy decision tree approach for database classification. *Expert Systems with Applications*, 37(1):214–225.

Cinelli, M., Coles, S. R., Nadagouda, M. N., Błaszczyński, J., Słowiński, R., Varma, R. S., and Kirwan, K. (2015). A green chemistry-based classification model for the synthesis of silver nanoparticles. *Green Chemistry*, 17(5):2825–2839.

de La Torre, J., Puig, D., and Valls, A. (2018). Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognition Letters*, 105:144–154.

de La Torre, J., Valls, A., and Puig, D. (2020). A deep learning interpretable classifier for diabetic retinopathy disease grading. *Neurocomputing*, 396:465–476.

Dorr, D., Bonner, L. M., Cohen, A. N., Shoai, R. S., Perrin, R., Chaney, E., and Young, A. S. (2007). Informatics systems to promote improved care for chronic illness: a literature review. *Journal of the American Medical Informatics Association*, 14(2):156–163.

Durkin, J. (1993). Expert systems: Catalog of applications. akron, oh: Intelligent computer systems. *Inc. and Macmillan Pub. Co.*

Escorcia-Gutiérrez, J., Torrents-Barrena, J., Romero-Aroca, P., Valls, A., and Puig, D. (2016). Interactive optic disk segmentation ia discrete convexity shape

knowledge using high-order functionals. In *Artificial Intelligence Research and Development*, volume 288, page 39. IOS Press.

Escorcia-Gutiérrez, J., Torrents-Barrena, J., Romero-Aroca, P., Valls, A., and Puig, D. (2018). Constrained convexity shape prior segmentation of retinal cross-related structures. *CARS 2018 Computer Assisted Radiology and Surgery*, 13(Suppl. 1):20–23.

Fan, C.-Y., Chang, P.-C., Lin, J.-J., and Hsieh, J. (2011). A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. *Applied Soft Computing*, 11(1):632–644.

Fujimoto, K. (2013). Obtaining admissible preference orders using hierarchical bipolar sugeno and choquet integrals. *Journal of Advanced Computational Intelligence Vol*, 17(4).

Gadaras, I. and Mikhailov, L. (2009). An interpretable fuzzy rule-based classification methodology for medical diagnosis. *Artificial Intelligence in Medicine*, 47(1):25–41.

Gardiner, E. J. and 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(9):1781–1803.

Giarratano, J. C. and Riley, G. (1998). *Expert systems*. PWS publishing co.

Goldstein, B. A., Navar, A. M., Pencina, M. J., and Ioannidis, J. P. (2016). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, page ocw042.

Grabisch, M. and Labreuche, C. (2010). A decade of application of the choquet and sugeno integrals in multi-criteria decision aid. *Annals of Operations Research*, 175(1):247–286.

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

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

H. Witten, I. and Frank, E. (2013). Data mining :practical machine learning tools and techniques. *The Morgan Kaufmann Series in Data Management Systems*.

IDF (2013). Idf diabetes atlas 6th edition. *International Diabetes Federation*, 46(ISBN:2-930229-85-3).

Jiang, R., Bouridane, A., Crookes, D., Celebi, M. E., and Wei, H.-L. (2016). Privacy-protected facial biometric verification using fuzzy forest learning. *IEEE Transactions on Fuzzy Systems*, 24(4):779–790.

Jin, C., Li, F., and Li, Y. (2014). A generalized fuzzy id3 algorithm using generalized information entropy. *Knowledge-Based Systems*, 64:13–21.

Khan, A. and 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*, pages 334–339. IEEE.

Koley, S., Sadhu, A. K., Mitra, P., Chakraborty, B., and 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.

Kori, A., Chennamsetty, S. S., Alex, V., et al. (2018). Ensemble of convolutional neural networks for automatic grading of diabetic retinopathy and macular edema. *arXiv preprint arXiv:1809.04228*.

Kosko, B. (1986). Fuzzy entropy and conditioning. *Information sciences*, 40(2):165–174.

Kotsiantis, S., Kanellopoulos, D., Pintelas, P., et al. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30(1):25–36.

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

Kulkarni, V. Y. and Sinha, P. K. (2013). Random forest classifiers: a survey and future research directions. *International Journal of Advanced Computing*, 36(1):1144–1153.

La Torre, J. d., Valls, A., and Puig, D. (2016). Diabetic retinopathy detection hrough image analysis using deep convolutional neural networks. In *Artificial Intelligence Research and Development*, volume 288, page 58. IOS Press.

Levashenko, V. G. and Zaitseva, E. N. (2002). Usage of new information estimations for induction of fuzzy decision trees. In *Intelligent Data Engineering and Automated Learning—IDEAL 2002*, pages 493–499. Springer.

Li, F. and Jiang, D. (2011). Fuzzy id3 algorithm based on generating hartley measure. In *Web Information Systems and Mining*, pages 188–195. Springer.

Ligeza, A. (2006). *Logical foundations for rule-based systems*, volume 11. Springer.

Lucca, G., Sanz, J. A., Dimuro, G. P., Bedregal, B., Asiain, M. J., Elkano, M., and Bustince, H. (2017). Cc-integrals: Choquet-like copula-based aggregation functions and its application in fuzzy rule-based classification systems. *Knowledge-Based Systems*, 119:32–43.

Lucca, G., Sanz, J. A., Dimuro, G. P., Bedregal, B., Bustince, H., and Mesiar, R. (2018). Cf-integrals: A new family of pre-aggregation functions with application to fuzzy rule-based classification systems. *Information Sciences*, 435:94 – 110.

Lucca, G., Sanz, J. A., Dimuro, G. P., Bedregal, B., Mesiar, R., Kolesárová, A., and Bustince, H. (2016). Preaggregation functions: construction and an application. *IEEE Transactions on Fuzzy Systems*, 24(2):260–272.

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

Mangasarian, K. (1992). Neural network training via linear programming. *Advances in Optimisation and Parallel Computing*, pages 56–67.

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

Mesiar, R. and Stupňanová, A. (2018). A note on cc-integral. *Fuzzy Sets and Systems*.

Meza-Palacios, R., Aguilar-Lasserre, A. A., Ureña-Bogarín, E. L., Vázquez-Rodríguez, C. F., Posada-Gómez, R., and 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.

Murofushi, T., Sugeno, M., and Fujimoto, K. (1997). Separated hierarchical decomposition of the choquet integral. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 5(05):563–585.

Nakayama, H., Hattori, Y., and 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*, volume 5, pages 924–929. IEEE.

Napierala, K. and Stefanowski, J. (2016). Types of minority class examples and their influence on learning classifiers from imbalanced data. *Journal of Intelligent Information Systems*, 46(3):563–597.

Nikolopoulos, C. (1997). *Expert systems: introduction to first and second generation and hybrid knowledge based systems*. Marcel Dekker, Inc.

Ohno-Machado, L. (2017). Using health information technology for clinical decision support and predictive analytics. *Journal of the American Medical Informatics Association*, 24(1):1–1.

Olafsdottir, E. and Stefansson, E. (2007). Biennial eye screening in patients with diabetes without retinopathy: 10-year experience. *British journal of ophthalmology*, 91(12):1599–1601.

Pałkowski, Ł., Błaszczyński, J., Skrzypczak, A., Błaszczak, J., Kozakowska, K., Wróblewska, J., Kożuszko, S., Gospodarek, E., Krysiński, J., and Słowiński, R. (2014a). Antimicrobial activity and sar study of new gemini imidazolium-based chlorides. *Chemical biology & drug design*, 83(3):278–288.

Pałkowski, Ł., Błaszczyński, J., Skrzypczak, A., Błaszczak, J., Nowaczyk, A., Wróblewska, J., Kożuszko, S., Gospodarek, E., Słowiński, R., and 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., and Wróblewska, J. (2014b). Application of rough set theory to prediction of antimicrobial activity of bis-quaternary imidazolium chlorides. *Fundamenta Informaticae*, 132(3):315–330.

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

Pratt, H., Coenen, F., Broadbent, D. M., Harding, S. P., and Zheng, Y. (2016).

Convolutional neural networks for diabetic retinopathy. *Procedia Computer Science*, 90:200–205.

Quaglio, G., Dario, C., Stafylas, P., Tiik, M., McCormack, S., Zilgalvis, P., d'Angelantonio, M., Karapiperis, T., Saccavini, C., Kaili, E., et al. (2016). E-health in europe: Current situation and challenges ahead. *Health Policy and Technology*, 5(4):314–317.

Quellec, G., Lamard, M., Josselin, P. M., Cazuguel, G., Cochener, B., and Roux, C. (2008). Optimal wavelet transform for the detection of microaneurysms in retina photographs. *IEEE transactions on medical imaging*, 27(9):1230–1241.

Romero-Aroca, P., de la Riva-Fernandez, S., Valls-Mateu, A., Sagarra-Alamo, R., Moreno-Ribas, A., and Soler, N. (2016a). Changes observed in diabetic retinopathy: eight-year follow-up of a spanish population. *British Journal of Ophthalmology*, 100(10):1366–1371.

Romero-Aroca, P., de la Riva-Fernandez, S., Valls-Mateu, A., Sagarra-Alamo, R., Moreno-Ribas, A., Soler, N., and Puig, D. (2016b). Cost of diabetic retinopathy and macular oedema in a population, an eight year follow up. *BMC ophthalmology*, 16(1):136.

Romero Aroca, P., Reyes Torres, J., Sagarra Alamo, R., Basora Gallisa, J., Fernández-Balart, J., Pareja Ríos, A., and Baget-Bernaldiz, M. (2012). Resultados de la implantación de la cámara no midriática sobre la población diabética. *Salud (i) cienc.*, 19(3):214–219.

Romero-Aroca, P., Valls, A., Moreno, A., Sagarra-Alamo, R., Basora-Gallisa, J., Saleh, E., Baget-Bernaldiz, M., and Puig, D. (2019). A clinical decision support system for diabetic retinopathy screening: creating a clinical support application. *Telemedicine and e-Health*, 25(1):31–40.

Saleh, E., Błaszczyński, J., Moreno, A., Valls, A., Romero-Aroca, P., de la Riva-Fernández, S., and Słowiński, R. (2018a). Learning ensemble classifiers for diabetic retinopathy assessment. *Artificial Intelligence in Medicine*, 85:50 – 63.

Saleh, E., Maaroof, N., and Jabreel, M. (2020). The deployment of a decision support system for the diagnosis of diabetic retinopathy into a catalan medical center. In *6th URV Doctoral Workshop in Computer Science and Mathematics*, page 45.

Saleh, E., Moreno, A., Valls, A., Romero-Aroca, P., and de la Riva-Fernandez, S. (2016a). A fuzzy random forest approach for the detection of diabetic retinopathy on electronic health record data. In *Artificial Intelligence Research and Development*, volume 288, page 169. IOS Press.

Saleh, E., Valls, A., Moreno, A., and Romero-Aroca, P. (2017). Integration of different fuzzy rule-induction methods to improve the classification of patients with diabetic retinopathy. In *Recent Advances in Artificial Intelligence Research and Development*, pages 6–15.

Saleh, E., Valls, A., Moreno, A., Romero-Aroca, P., Bustince, H., and Torra, V. (2019). A hierarchically $\perp$-decomposable fuzzy measure-based approach for fuzzy rules aggregation. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 27(Suppl. 1):59–76.

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

Saleh, E., Valls, A., Moreno, A., Romero-Aroca, P., Torra, V., and Bustince, H. (2018b). Learning fuzzy measures for aggregation in fuzzy rule-based models. In Torra, V., Narukawa, Y., Aguiló, I., and González-Hidalgo, M., editors, *Modeling*

Decisions for Artificial Intelligence, pages 114–127, Cham. Springer International Publishing.

Saleh, M. D. and 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(1):186–196.

Sanromà, S., Moreno, A., Valls, A., Romero, P., de la Riva, S., and 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)*, pages 313–318.

Shaw, J. E., Sicree, R. A., and Zimmet, P. Z. (2010). Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes research and clinical practice*, 87(1):4–14.

Sikchi, S. S., Sikchi, S., and Ali, M. (2013). Fuzzy expert systems (fes) for medical diagnosis. *International Journal of Computer Applications*, 63(11).

Singh, H., Naik, A. D., Rao, R., and Petersen, L. A. (2008). Reducing diagnostic errors through effective communication: harnessing the power of information technology. *Journal of General Internal Medicine*, 23(4):489–494.

Skevofilakas, M., Zarkogianni, K., Karamanos, B. G., and 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*, pages 6713–6716. IEEE.

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

Słowiński, R., Greco, S., and Matarazzo, B. (2015). Rough set methodology

for decision aiding. In *Springer Handbook of Computational Intelligence*, pages 349–370. Springer.

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

Su, C.-T., Yang, C.-H., Hsu, K.-H., and 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(6):1075–1092.

Sugeno, M., Fujimoto, K., and Murofushi, T. (1995). A hierarchical decomposition of choquet integral model. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 3(01):1–15.

Szolovits, P. et al. (1995). Uncertainty and decisions in medical informatics. *Methods of Information in Medicine*, 34(1):111–121.

Torra, V. (1999). On hierarchically S-decomposable fuzzy measures. *International journal of intelligent systems*, 14(9):923–934.

Torra, V. and Narukawa, Y. (2007). *Modeling decisions: information fusion and aggregation operators*. Springer Science & Business Media.

Trawiński, K., Alonso, J. M., and Hernández, N. (2013). A multiclassifier approach for topology-based wifi indoor localization. *Soft Computing*, 17(10):1817–1831.

Umano, M., Okamoto, H., Hatono, I., Tamura, H., Kawachi, F., Umedzu, S., and Kinoshita, J. (1994). Fuzzy decision trees by fuzzy id3 algorithm and its application to diagnosis systems. In *Fuzzy Systems, 1994. Proceedings of the Third IEEE Conference on IEEE World Congress on Computational Intelligence*, pages 2113–2118. IEEE.

Wang, X., Yeung, D. S., and 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(2):215–226.

Weber, S. (1984). -decomposable measures and integrals for archimedean t-conorms. *Journal of Mathematical Analysis and Applications*, 101(1):114–138.

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(9):1677–1682.

Xiao, T., Huang, D.-M., Zhou, X., and Zhang, N. (2014). Inducting fuzzy decision tree based on discrete attributes through uncertainty reduction. *Applied Mechanics & Materials*.

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

Yuan, Y. and Shaw, M. J. (1995). Induction of fuzzy decision trees. *Fuzzy Sets and Systems*, 69(2):125–139.

Zhang, L., Wei, H., Zhu, J., de la Cruz, J., Gonzalez, H. J., and 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*, pages 95–98. IEEE.

UNIVERSITAT
ROVIRA i VIRGILI