Artificial Intelligence Research and Development
K. Trejo et al. (Eds.)
© 2025 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA250594

Constructing Fuzzy Linguistic Variables Using Crisp Decision Trees. An Application to Diabetic Retinopathy

Amal ESMAIL [a,1], Jordi PASCUAL-FONTANILLES [a], Eugeni GARCIA [c], Aida VALLS [a,c], Marc BAGET [b,c], and Pedro ROMERO-AROCA [b,c]

[a] *ITAKA, Dept. Enginyeria Informàtica i Matemàtiques*
Universitat Rovira i Virgili, Tarragona, Catalonia, Spain
[b] *Servei d'Oftalmologia, Hospital Universitari Sant Joan de Reus, Catalonia, Spain*
[c] *Institut d'Investigació Sanitària Pere Virgili, Tarragona, Catalonia, Spain*
ORCiD ID: Amal Esmail https://orcid.org/orcid.org/0000-0002-8110-0024, Jordi Pascual-Fontanilles https://orcid.org/orcid.org/0000-0002-7528-5819, Eugeni Garcia https://orcid.org/orcid.org/0009-0001-8305-4797, Aida Valls https://orcid.org/orcid.org/0000-0003-3616-7809, Marc Baget https://orcid.org/, Pedro Romero-Aroca https://orcid.org/0000-0002-7061-8987

Abstract. Diabetic retinopathy is a leading cause of vision loss worldwide. Early detection and understanding of its progression are crucial for preventing serious complications. This work is based on assessing the severity of diabetic retinopathy based on features extracted from eye fundus images. We use LezioSeg, a deep learning-based computer vision model, to segment and locate retinopathy lesions. Specifically, we quantify the number of microaneurysms and hemorrhages in the four retinal quadrants. These features are then used to train Retiprogram, a multiclass classification model based on a Fuzzy Random Forest. The goal is to identify the actual level of Diabetic Retinopathy. A key challenge with Fuzzy Random Forests is the definition of appropriate linguistic variables. This paper proposes a novel method for automatically generating these fuzzy sets from decision values identified at constructing crisp decision trees from a labeled set of examples. These automatically generated fuzzy sets serve as linguistic variables for the Retiprogram. We use crisp Random Forests to evaluate the Retiprogram's performance against both Yuan and Shaw's fuzzification method. The proposed method outperforms the standard both in accuracy and kappa indicators.

Keywords. Fuzzy Logic, Machine Learning, AI Applications, Medical diagnosis

1. Introduction

Diabetic Retinopathy (DR) is a progressive eye disease caused by prolonged hyperglycemia, damaging retinal blood vessels and potentially leading to blindness. It produces lesions such as cotton wool spots, exudates, microaneurysms (MA), and hemor-

[1] Corresponding Author: amalesmailqasem.al-zubairi@urv.cat.

rhages (HM). MA is a well-known indicator of the early stages of DR, while HM is the result of blood vessel leakage, and it appears when DR progresses to more severe stages. Standard scales consider three stages of DR: mild, moderate and severe. The most severe stage, also known as proliferative DR, involves abnormal blood vessel growth, increasing the risk of retinal detachment and vision loss, possibly leading to blindness [1]. Our study addresses the DR severity classification problem. It is a multiclass task that ranges from no DR to the severe stage. Correct classification is crucial to enable timely treatment and prevent vision loss. The novelty of our proposal consists of automatically counting the number of lesions in each of the four quadrants of the ETDRS standard grid [1] and then using these frequencies as features to train a Fuzzy Random Forest model (FRF). The grid dividing the quadrants is centered in the fovea, as shown in Figure 1. The motivation to use a fuzzy model is twofold. First, doctors often make diagnoses qualitatively rather than relying on strict numerical thresholds, so the use of linguistic variables is appropriate. Second, automated lesion detection systems may suffer from segmentation inaccuracies, particularly with small or ambiguous lesions such as MA and HM, so the use of fuzzy logic model helps in the modeling of this uncertainty.

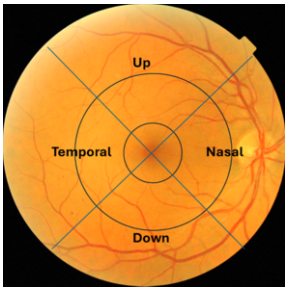


Figure 1. The 4 quadrants of the eye fundus image defined by ETDRS.

Fuzzy Random Forests use fuzzy logic to handle uncertainty in the data. Unlike methods relying on crisp classification, they use flexible boundaries, making them well-suited for complex or unclear data. These models require inputs defined as fuzzy linguistic variables, which consist of a set of labels: $L = \{l_1, l_2, ..., l_r\}$, each one mapped into a numerical reference scale, X, using a fuzzy membership function: $\mu_{l_i}(x) \in [0, 1]$. Fuzzy linguistic variables effectively convert numerical values into descriptive terms like *low*, *moderate*, or *high*, transforming quantitative values into terms commonly used by clinicians, increasing the interpretability of the classification model.

This paper proposes a novel methodology for automatically constructing fuzzy linguistic variables based on the information extracted from a simple crisp decision tree (DT). The proposed algorithm is used to build fuzzy linguistic variables for DR assessment based on lesion counts in the four quadrants of the eye fundus. These variables are then used to train a Fuzzy Random Forest classification model by means of the Retiprogram system [2].

Developing an automatic method to bridge the gap between the quantitative lesion counts provided by an image segmentation model and the qualitative input required by a fuzzy classification model is crucial to improve DR detection.

The remainder of this paper is organized as follows: Section 2 explains the image segmentation method, LezioSeg, and the FRF-based classification model, Retiprogram.

Section 3 presents the proposed methodology for automatically generating fuzzy linguistic variables. Section 4 describes the experimental setup and evaluation metrics. Section 5 discusses the classification results, comparing the proposed method with another. Finally, Section 6 concludes the paper, summarizing our findings and outlining directions for future research.

2. Preliminaries

This work is based on two existing algorithms: LezioSeg and Retiprogram. In this section, we briefly describe them, along with a description of a well-known method for building fuzzy linguistic variables proposed by Yuan and Shaw [3].

2.1. LezioSeg

LezioSeg is an advanced deep learning segmentation model designed for accurately detecting hemorrhages and microaneurysms in eye fundus images [4]. It integrates two multi-scale modules: an Atrous Spatial Pyramid Pooling module in the network's bottleneck for capturing multiscale contextual information, and a Spatial Attention and Temporal (SAT) module in the decoder to merge the features obtained from different resolutions. LezioSeg employs the MobileNet model pretrained on ImageNet as an encoder. The decoder network consists of four blocks, each incorporating Gated Skip Connections to enhance lesion identification by effectively merging encoder-derived features. The final decoder output is improved using the SAT mechanism, combining multiscale data efficiently to obtain a precise segmentation, with lower computational demands compared to traditional architectures like ResNet or DenseNet. The application of affine transformations as data augmentation for generating geometric distortions increased the performance of the segmentation model. Tests on standard IDRiD datased give an F1 score of 66% for HM and 43% for MA, which is comparable to other related works [4].

An algorithm for grouping the pixels of the same lesion, creating contours, and counting lesions is then applied on the mask obtained from LezioSeg to obtain a numerical value for the number of MA and HM.

2.2. Retiprogram

Retiprogram is a binary classification model built upon a FRF [5]. It is based on the work of Yuan and Shaw, who adapted the ID3 algorithm to incorporate fuzzy logic [3], with some extensions presented in [2]. FRF generate a set of fuzzy classification rules induced from a training dataset, allowing the model to classify new instances. Retiprogram employs a user-defined set of fuzzy linguistic variables during the inference of the rules. The model uses three main parameters: the significance level α, which defines the minimum membership degree considered during inference; the truth level threshold β, the minimum truth level for generating fuzzy rules; and γ, a parameter used to determine the number of attributes used in each FDT. The parameters δ_1 and δ_2 were also introduced to identify conflicting cases, which are labeled *Unknown*. In settings such as the medical one, it is preferred to have an unknown prediction in an uncertain case than a wrong prediction. Retiprogram was also adapted for the multiclass classification case [6]. The adaptation was focused on the prediction stage, on the 2-step classification

process, concretely in the detection of conflicting cases, as well as in the aggregation of FDT predictions at the FRF level, which was adapted through weighted voting; finally, a conjunctive aggregation was also introduced to calculate the final class support score.

2.3. Yuan and Shaw method for defining a fuzzy partition

Yuan and Shaw [3] proposed a fuzzification technique that transforms numerical data into fuzzy linguistic terms using triangular membership functions. Given the desired number of terms k, the algorithm converts a numerical attribute A (represented by a sample of values X) into a fuzzy linguistic variable. The output is a partition of X into k fuzzy sets with triangular membership functions centered at m_i. The algorithm starts by evenly initializing k centers across X. Then, these centers are iteratively refined using a self-organizing process based on the Kohonen's algorithm. In each iteration t, a random data point $x^{[t]} \in X$ is selected, and the nearest center $m_c^{[t]}$ is moved closer to it using a decreasing learning rate η:

$$m_c^{[t+1]} = m_c^{[t]} + \eta^{[t]}(x^{[t]} - m_c^{[t]}) \tag{1}$$

The process continues until the total distance $D(X,M)$ between all data points and their nearest centers converges:

$$D(X,M) = \sum_{x \in X} \min_i |x - m_i| \tag{2}$$

The distance metric serves as stopping criteria, indicating that the fuzzy partition has stabilized.

3. Proposed methodology for building fuzzy linguistic variables

The objective of the proposed method is to automatically construct a set of fuzzy linguistic variables optimized for classifying a set of image I into known target classes, C. We focus on classifying eye fundus images based on the counts of MA and HM detected in different eye quadrants, ultimately determining the degree of DR severity according to the ETDRS standard [7]. This results in four classes $C = \{none, mild, moderate, severe\}$ (numbered from 0 to 3). The proposed methodology, shown in Figure 2, has three stages.

In the first stage (Data Collection), LezioSeg is used to segment and count the number of MA and HM for a set of images. The corresponding DR severity level for each image is provided by an expert in ophthalmology.

The second stage (Decision Tree Modelling) involves generating a tabular dataset from the lesion counts obtained in the previous stage. A crisp DT is trained on this dataset, identifying key thresholds for lesion counts that separate the different DR severity classes.

In the last stage, the insights from the trained DT are used to construct the fuzzy partitions. The numerical values at the decision nodes of the DT (i.e. the splitting points used to divide the data) are used to define the boundaries of the fuzzy sets. They are then associated with linguistic labels (e.g., low, medium, high, etc.) used to train the Retiprogram model.

The following subsection formally defines the novel process implemented in stage 3, detailing how DT splitting points are transformed into fuzzy linguistic variables.

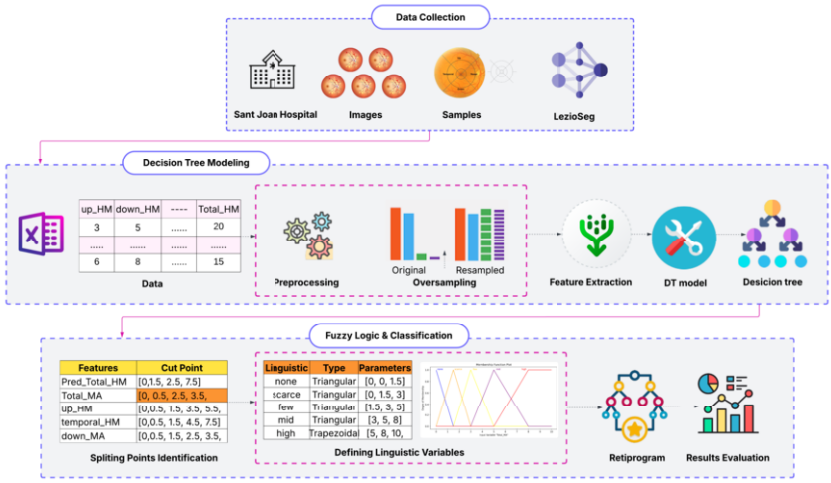


Figure 2. Methodology proposed for building a fuzzy linguistic variable

3.1. Definition of a fuzzy partition with crisp decision trees

Several algorithms exist in the literature to build a crisp Decision Tree from numerical data, such as ID3 and C4.5 [8]. They use a heuristic criterion, commonly entropy-based information gain or Gini impurity, to select the best attribute and split point for splitting the data, aiming to maximize the separation between the target classes. However, the classification using crisp rules has been proven inadequate in certain domains with inherent uncertainty or ambiguity, such as the one we are dealing with in this paper.

We propose to fuzzify the splitting points identified by a crisp DT, transforming them from crisp boundaries into fuzzy ones that define a fuzzy partition. The procedure creates a set of N fuzzy linguistic variables from a set of attributes A, using two steps:

STEP 1: Identification of Splitting Points per attribute A_i:

To identify relevant splitting points, we can train multiple DT classifiers using various combinations of input features. We tested different splitting criteria (entropy, Gini) and tree depths (4 & 5), choosing the combination with the best performance. Each DT is then analyzed to examine four groups of features that will be explained later (Sec:4.1) and extract the split values used in its decision nodes. This process yields a collection of split points $S_i = \{s_{i,1}, ..., s_{i,p_i}\}$ for each feature. If the number of splitting points is larger than the number of desired labels, we select the most frequent ones as representative values. The number of attributes to fuzzify should be large enough to build meaningful decision trees, therefore, $N >= minN$, where $minN$ is application dependent.

STEP 2: Creation of the Fuzzy Partition of each variable:

To construct the fuzzy sets for the new linguistic variable A'_i, we take its corresponding splitting points S_i. Since the procedure is independent for each attribute, we will drop the index i for simplicity. Let us have a set of split points $S = \{s_1, ..., s_p\}$. We aim to obtain $p+1$ terms and fuzzy sets using the following procedure. To define the boundaries of these sets, we extend S with s_0 being the minimum of attribute A, and s_{p+1} being the maximum of A. Let us define the set of central points of each interval defined by the splitting points as $M = \{m_0.., m_p\}$, calculated as $m_i = \frac{s_i + s_{i-1}}{2}$. We will consider

trapezoidal membership functions defined by well-known tuple notation (a,b,c,d). Triangular fuzzy sets are a special case where $c = d$. The membership function values are defined as follows:

- For the first label $i = 0$, representing the smallest term:

$$a_0 = s_0, b_0 = s_0, c_0 = m_0, d_0 = m_1 \tag{3}$$

- For the last label, $i = p$, representing the largest term:

$$a_p = c_{p-1}, b_p = d_{p-1}, c_p = s_{p+1}, d_p = s_{p+1} \tag{4}$$

- For intermediate labels, $i = 1, \ldots, p-1$, we differentiate between two cases based on the distance between the consecutive split points:

  * If the current interval is small $|s_i - s_{i+1}| < \sigma$, we build a triangular fuzzy set:

$$a_i = c_{i-1}, b_i = d_{i-1}, c_i = b_i, d_i = m_{i+1} \tag{5}$$

  * If $|s_i - s_{i+1}| \geq \sigma$, we build a trapezoidal fuzzy set:

$$a_i = c_{i-1}, b_i = d_{i-1}, d_i = m_{i+1}, c_i = d_i - |b_i - a_i|, \tag{6}$$

  Notice that if the next interval to the current label is large $|s_{i+1} - s_{i+2}| < \sigma$, then we must make the fuzzy set symmetric instead of taking the next central point, as follows:

$$d_i = b_i + |b_i - a_i| \tag{7}$$

The threshold σ depends on the range of the attribute A, the desired number of terms and its fuzziness degree.

3.2. An example:

In Figure 3, a portion of a decision tree is shown, where features such as A and B are split at various thresholds. These thresholds divide the range of each attribute into meaningful categories that help distinguish different levels of lesion severity. For example, if the split points for feature A are $S_A = [5, 10, 12, 19]$, these values are used as reference points to define fuzzy sets as explained in the proposed method. The centers of the intervals for this variable are $[2.5, 7.5, 11, 15]$, obtaining the fuzzy sets given in Figure 4.

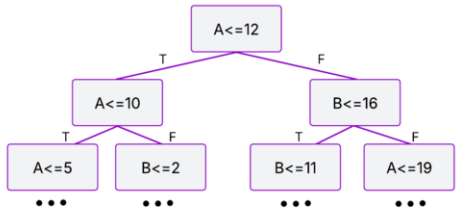
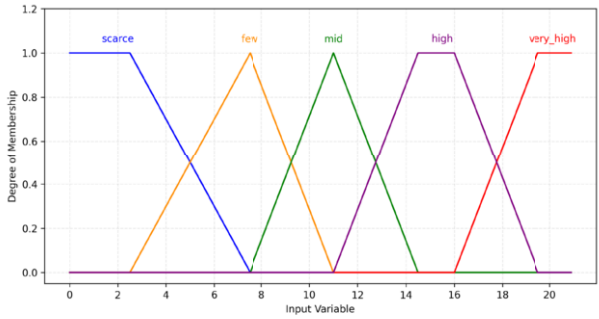


Figure 3. Example of a decision tree using variables A and B



A=[5,10,12,19]		
Name	**Type**	**Parameters**
scarce	Trapezoidal	[0, 0, 2.5, 7.5]
few	Triangular	[2.5, 7.5, 11]
mid	Triangular	[7.5, 11, 14.5]
high	Trapezoidal	[11, 14.5, 16, 19.5]
Very_high	Trapezoidal	[16, 19.5, 20, 20]

Figure 4. Fuzzy Membership Functions obtained with the proposed method for attribute A

4. Experiments and Results

4.1. Dataset Collection and pre-processing

Table 1. Class distribution in training and testing.

Class	Train (70%)	Test (30%)	Total
0 (No DR)	1397	598	1995
1 (Mild)	500	214	714
2 (Moderate)	903	387	1290
3 (Severe)	1016	435	1451
Total	**3816**	**1634**	**5450**

We collected eye fundus images from Sant Joan Hospital in Reus to study and measure retinal lesions for diagnosing and analysing DR. Using LezioSeg , the segmentation was performed, and we quantified lesion in the four major directions around the center of the retina: up, down, temporal, and nasal, to capture their spatial distribution.Table 1 shows the number of images per class in the training and testing datasets. For each image in the training and testing sets, we used LezioSeg to obtain the following 10 variables:

- **Total_MA**: total no. of MA in the eye.
- **Total_HM**: total no. of HM in the eye.
- **temporal_MA, nasal_MA, up_MA and down_MA**: no. of MA in each region.
- **temporal_HM, nasal_HM, up_HM and down_HM**: no. of HM in each region.

We first preprocessed the dataset by removing outliers, specifically abnormally large values considered unrealistic by medical experts. Since the dataset was imbalanced, we used SMOTE (Synthetic Minority Oversampling Technique) [9] to generate additional samples for the underrepresented classes. Oversampling was performed so that all classes matched the size of the majority class (i.e., 1397 samples). This process helped balance the dataset and ensured that the model had enough examples from all stages of DR.

We used four groups of features. The first included all 10 features. The second excluded total MA and HM keeping the eight features of the quadrants. The third and fourth groups were specific to each quadrant, using MA and HM counts from the temporal and

nasal quadrants and the up and nasal quadrants, respectively. For each group, We trained a DT model for each group and then analysed its structure to understand better how the features contributed to the classification.

4.2. Linguistic variables for Diabetic Retinopathy lesion counting

In this section, we compare the results of our proposed method, denoted as PM, with the method defined by Yuan and Shaw (section 2.3), denoted as YS. We focus on two features of our case study: temporal_MA and down_MA. For the (YS) method we generated 6 terms, shown in Table 2 and illustrated in Figure 6. Our method (PM) identified 5 terms for temporal_MA (Table 2 and Figure 5) and 8 terms for down_MA (Table 3 and Figure 7).

Table 2. Fuzzy sets for `temporal_MA_PM` and `temporal_MA_YS`

`temporal_MA_PM`			`temporal_MA_YS`		
Name	**Type**	**Parameters**	**Name**	**Type**	**Parameters**
none	Triangular	[0, 0, 1]	none	Triangular	[0, 0, 0.4]
scarce	Triangular	[0, 1, 2]	scarce	Triangular	[0, 0.4, 3.25]
few	Triangular	[1, 2, 3]	few	Triangular	[0.4, 3.25, 5.81]
mid	Trapezoidal	[2, 3, 7, 8]	mid	Triangular	[3.25, 5.81, 10.36]
high	Trapezoidal	[7, 8, 10, 10]	high	Triangular	[5.81, 10.36, 15.75]
—	—	—	very_high	Trapezoidal	[10.36, 15.75, 30, 30]

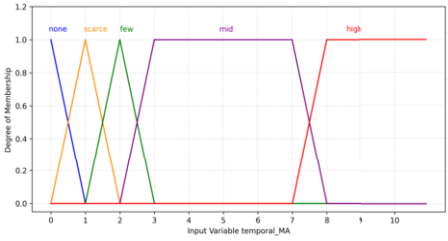


Figure 5. Temporal_MA_PM

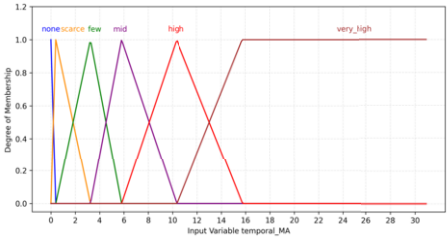


Figure 6. Temporal_MA_YS

Table 3. Fuzzy sets for `down_MA_PM` and `down_MA_YS`

`down_MA_PM`			`down_MA_YS`		
Name	**Type**	**Parameters**	**Name**	**Type**	**Parameters**
none	Triangular	[0, 0, 1]	none	Triangular	[0, 0, 0.76]
scarce	Triangular	[0, 1, 2]	scarce	Triangular	[0, 0.76, 4.07]
few	Triangular	[1, 2, 3]	few	Triangular	[0.76, 4.07, 7.57]
mid	Triangular	[2, 3, 4.5]	mid	Triangular	[4.07, 7.57, 11.75]
almost_high	Triangular	[3, 4.5, 6.5]	high	Triangular	[7.57, 11.75, 18.38]
high	Triangular	[4.5, 6.5, 8.5]	very_high	Trapezoidal	[11.75, 18.38, 30, 30]
very_high	Trapezoidal	[6.5, 8.5, 14, 17]	—	—	—
extremely_high	Trapezoidal	[14, 17, 20, 20]	—	—	—

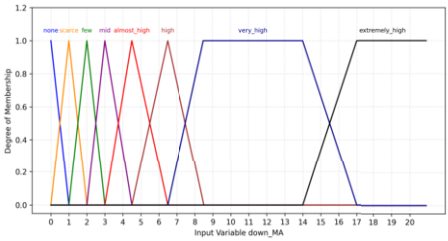


Figure 7. down_MA_PM

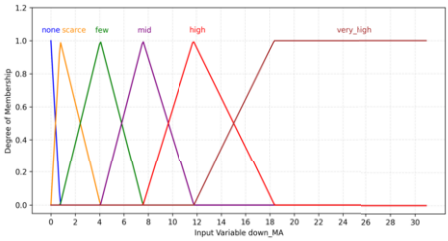


Figure 8. down_MA_YS

5. Comparison with reference method

In this section, we compare the performance of our proposed method (PM) with the method of Yuan and Shaw (YS) in classifying Diabetic Retinopathy. We have used Retiprogram, trained and tested with the data presented in Table 1. The parameters of Retiprogram have been experimentally optimized to: $\alpha = 0$, $\beta = 0.8$, $bagging = 200$, $numberTrees = 100$ and $\gamma = 2$. The parameters used to detect conflicting cases, δ_1 and δ_2, were set to 0 in order to classify all the examples. Table 4 presents the class-wise Precision, Recall, F1-Score and Specificity. Last row corresponds to their weighted average across all classes, accounting for the imbalanced class distribution.

Table 4. Class-wise performance comparison between the (PM) and (YS)

Class	Precision		Recall		F1-Score		Specificity	
	PM	YS	PM	YS	PM	YS	PM	YS
0	0.850	0.912	0.744	0.675	0.793	0.776	0.590	0.564
1	0.386	0.260	0.307	0.293	0.342	0.276	0.852	0.857
2	0.385	0.307	0.594	0.598	0.467	0.406	0.748	0.751
3	0.828	0.825	0.835	0.822	0.831	0.823	0.722	0.721
Average	**0.673**	0.660	**0.675**	0.646	**0.667**	0.635	**0.697**	0.688

Results demonstrate that PM outperforms YS in the average value of all quality indicators. Notably, Recall is particularly relevant in medical diagnosis, as it measures the capacity for retrieving the positive cases. PM has the best recall in classes 0, 1 and 3, while performing the same than YS for class 2, without compromising the specificity scores. From the point of view of the patients, improving recall in class 3 enables timely intervention in the most severe RD stage and helps reduce the risk of vision loss.

Table 5 shows some overall quality measures for the two fuzzy methods using linguistic labels. In all cases, the 10 features representing the number of lesions in different quadrants are used. The proposed method PM achieves the higher classification performance than YS in the three indicators: accuracy (0.673), F1 score (0.667) and especially with the quadratic weighted kappa (0.79), which is a relevant indicator in ordinal classification problems as it penalizes errors when doing classification to distant classes. These results indicate its ability to build more appropriate fuzzy sets for these qualitative medical qualitative variables.

Table 5. Overall classification performance values for the different methods

Method	Accuracy	Kappa	F1 Score
PM in Retiprogram	0.673	0.790	0.667
YS in Retiprogram	0.660	0.768	0.635

6. Conclusions and Future Work

We presented a methodology to construct fuzzy linguistic variables using split points given by a decision tree model. These split points are transformed into descriptive categories, improving data interpretability and providing meaningful inputs for the RF model. The case study of Diabetic Retinopathy has been used to test the method. Using fuzzy linguistic variables help capture the uncertainty in lesion counts, which is crucial for accurate DR severity assessment. As future work, more experimentation is needed to validate this technique and it should be compared with other fuzzification approaches. After, we plan to integrate these fuzzy variables with other clinical risk factors in the Retiprogram software to evaluate the overall impact on DR classification performance.

Acknowledgements

This study is funded by Instituto de Salud Carlos III and the FEDER (PI21/00064), by URV (2023-PFR-URV-114 and 2022PFR-41), as well as by AGAUR consolidated research group (2021PFR-B2-103). First author is funded by Program Martí Franquès under the grant 2023PMF-PIPF-19.

References

[1] Romero-Aroca P, Garcia-Curto E, Pascual-Fontanilles J, Valls A, Moreno A, Baget-Bernaldiz M. Distribution of Microaneurysms and Hemorrhages in Accordance with the Grading of Diabetic Retinopathy in Type Diabetes Patients. Diagnostics. 2024 Jul 17;14(14):1547,doi.org/10.3390/diagnostics14141547.

[2] Saleh E, Błaszczyński J, Moreno A, Valls A, Romero-Aroca P, de la Riva-Fernandez S, Słowiński R. Learning ensemble classifiers for diabetic retinopathy assessment. Artificial intelligence in medicine. 2018 Apr doi:10.1016/j.artmed.2017.09.006

[3] Yuan, Yufei, and Michael J. Shaw. "Induction of Fuzzy Decision Trees." Fuzzy Sets and Systems, vol. 69, no. 2, Jan. 1995, pp. 125–39. ScienceDirect, doi:10.1016/0165-0114(94)00229-Z.

[4] Ali MY, Jabreel M, Valls A, Baget M, Abdel-Nasser M. Lezioseg: Multi-scale attention affine-based cnn for segmenting diabetic retinopathy lesions in images. Electronics. 2023 Dec 8;12(24):4940, doi:10.3390/electronics12244940

[5] Sosnowski, Z. A., and Ł. Gadomer. "Fuzzy Trees and Forests—Review." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2019, e1316. doi:10.1002/widm.1316.

[6] Pascual-Fontanilles J, Lhotska L, Moreno A, Valls A. Adapting a fuzzy random forest for ordinal multi-class classification. InArtificial Intelligence Research and Development 2022 (pp. 181-190). IOS Press,doi:10.3233/FAIA220336

[7] Wilkinson CP, Ferris FL, Klein RE, Lee PP, Agardh CD, Davis M, Dills D, Kampik A, Pararajasegaram R, Verdaguer JT. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology. 2003 Sep;110(9):1677-82.doi:10.1016/S0161-6420(03)00475-5

[8] Costa VG, Pedreira CE. Recent advances in decision trees: An updated survey. Artificial Intelligence Review. 2023 May;56(5):4765-800,doi.org/10.1007/s10462-022-10275-5

[9] Pradipta GA, Wardoyo R, Musdholifah A, Sanjaya IN, Ismail M. SMOTE for handling imbalanced data problem: A review. In2021 sixth international conference on informatics and computing (ICIC) 2021 Nov 3 (pp. 1-8). IEEE,doi.org/10.1109/ICIC54025.2021.9632912