Exploiting the Retinal Vascular Geometry in Identifying the Progression to Diabetic Retinopathy Using Penalized Logistic Regression and Random Forests

Georgios Leontidis, Bashir Al-Diri and Andrew Hunter

Abstract Many studies have been conducted,investigating the effects that diabetes has to the retinal vasculature. Identifying and quantifying the retinal vascular changes remains a very challenging task, due to the heterogeneity of the retina. Monitoring the progression requires follow-up studies of progressed patients, since human retina naturally adapts to many different stimuli, making it hard to associate any changes with a disease. In this novel study, data from twenty five diabetic patients, who progressed to diabetic retinopathy, were used. The progression was evaluated using multiple geometric features, like vessels widths and angles, tortuosity, central retinal artery and vein equivalent, fractal dimension, lacunarity, in addition to the corresponding descriptive statistics of them. A statistical mixed model design was used to evaluate the significance of the changes between two periods: three years before the onset of diabetic retinopathy and the first year of diabetic retinopathy. Moreover, the discriminative power of these features was evaluated using a random forests classifier and also a penalized logistic regression.The area under the ROC curve after running a ten-fold cross validation was 0.7925 and 0.785 respectively.

Keywords Diabetic retinopathy · Diabetes · Penalized · Logistic regression · Random forests · Mixed model

Georgios Leontidis (corresponding author)
University of Lincoln, Brayford pool campus, LN67TS, UK e-mail: gleontidis@lincoln.ac.uk

Bashir Al-Diri
University of Lincoln, Brayford pool campus, LN67TS, UK e-mail: baldiri@lincoln.ac.uk

Andrew Hunter
University of Lincoln, Brayford pool campus, LN67TS, UK e-mail: ahunter@lincoln.ac.uk

1 Introduction

Diabetic retinopathy (DR) is a major disease, affecting the lives of millions of people around the world, leading to blindness, if left untreated or not diagnosed early [3, 17]. It constitutes a complication of diabetes mellitus, although it is not uncommon non-diabetic people to develop background retinopathy. In figure 1, two images can be seen, from the same patient, one during diabetes and one after the first lesions (micro-aneurysm) have appeared in the retina. It is worth pointing out that a normal/non-diabetic image does not seem to have any difference from a diabetic retinal image, since at this stage, the changes occur only to the vascular geometry, which cannot be easily identified.

Retina is a dynamic tissue and a very important, non-invasive window to the blood vessels. Retina processes light through a layer of photoreceptors.The absorbed light is converted into neural signals, in order to be forwarded through the optic nerve head directly to the brain for visual recognition [17]. Each person's retina is unique just like the fingerprints, making it very difficult to compare different retinas, since changes will inevitably and naturally exist. Therefore it is crucial, if someone wants to study the effects that a disease cause to the retinal vasculature, to look at specific segments and regions within the same subjects at different intervals. More details addressing the importance of this approach will be given in the next sections.

The underlying mechanisms that provoke diabetes are more or less known, however it still remains unclear how this sequence of events affects the retina, both structurally and functionally, leading to the development of DR. Diagnosing DR early or identifying diabetic patients with higher risk, can have a big impact on our society and possibly help clinicians deal with the disease earlier and delay the progression, by monitoring the patients more intensively[3].

For the present study, fifty high resolution (3216-by-2316 pixels) fundus images were used, taken from twenty five patients who progressed from diabetes to DR. Our aim is to understand to what extend has the retinal vascular geometry been affected by the progression and proliferation of diabetes, until the moment that the first lesions appear. To accommodate this, two groups were created;one for the period three years before DR and one for the very first year that DR appeared. Therefore we hypothesize that the retina is already adapting to the new underlying conditions, and that especially during the advanced stages of diabetes (few years before DR), these changes can be reliably identified and characterized. The images come from a diabetic screening database in England and all of the ethical guidelines have been followed. It is worth pointing out that in United Kingdom, all the people that are diagnosed with diabetes are entering automatically into the diabetic screening program for annual inspection of their retina. Therefore all the images are labeled and identified by the year they were captured, defining clearly the periods of diabetes, and also the initial appearance of DR.

The chapter is organized in three main sections. In the first section, all the methods, methodologies and tools will be described and analyzed, giving some essential background information of the investigated geometric features and their importance, as well as all the necessary image preprocessing. In the second part, the techniques

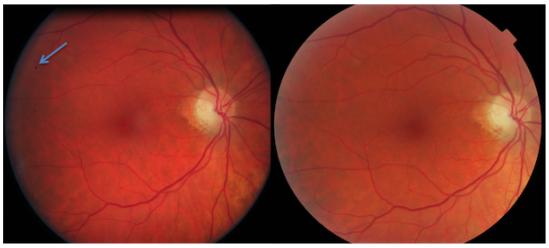

Fig. 1 Two images taken from the same patient. First year of diabetic retinopathy (left) and late stages of diabetes (right). Microaneurysms have already appeared, defining, the beginning of diabetic retinopathy.

for the statistical analysis, feature selection process and the classification approaches will be thoroughly addressed. At the final section the results will be presented, together with the inferences and the implications of the present study, including discussion, limitations, future approaches and conclusions.

2 Related work

Retina includes both very small and very large vessels, which can range from very few μm to more than 100 μm. It can be easily inferred that it is very difficult to compare the retina of different people and include representative and balanced amount of small and large vessels, which will in any case be different among people. During progression of diabetes and also during DR the retinal geometry changes[25].

Most of the studies in the past, investigating either hemodynamic or geometric features, have been focused on the analysis of different groups of people. For instance the oxygen saturation was investigated in different groups of people ranging from normal subjects to proliferative retinopathy, finding significant differences among them [15]. In another study they evaluated the differences between patients with diabetes and DR, using as features only the vessels' widths and angles [12].

Using different subjects, when investigating the human retina, makes it hard to associate any identified changes to diabetes/DR, and not instead to the normal changes that occur to the retina during aging, or between genders, or simply because different retinas, and more importantly different areas of the retina, might also vary [3, 27]. A few follow-up studies have been conducted, studying similar periods of diabetes, without though including in any classification system or evaluating features like central retinal vein/artery equivalent or tortuosity, which is the purpose of this study [18, 4, 20, 19, 21].

3 Methods

As mentioned previously, fifty images in total were analyzed, making sure that all the features can be measured in an equally reliable manner in all of them and thus

ensure that the changes can be attributed to the progression of diabetes. All the methods and tools were carefully chosen, having always as first priority the reliability and accuracy of the measurements, rather than using the fastest or with the fewest human interventions methods. For the image preprocessing, extraction of all the features and for the mixed model design, the software Matlab 2015b was utilized. On the contrary, for the regularized random forests (RRF) and the penalized logistic regression, the open source software "R" was used.

3.1 Features & Tools

A number of features were investigated in this study, which are representative of the whole retinal vasculature. Measuring these geometric features means that many different methods and tools have to be used in all the stages. The main investigated features are the following: a)Vessels' widths, b) Vessels' angles, c)Tortuosity, d) Fractal dimension, e)Lacunarity and f) Central retinal artery and vein equivalents for calculating the arteriovenous ratio as well.

3.1.1 Widths & Angles

Using the tool that was implemented and described in details in a previous study [1], 1200 vessels widths (600 arteries and 600 veins in total for both groups in pixels) and 400 branching angles (in degrees) in the corresponding junctions (200 for arteries and 200 for veins in total for both groups) were measured. Although many state-of-the-art automated tools have been proposed in literature, utilizing many different methods e.g. wavelets and edge location refinement both to segment and measure retinal vessels using image profiles, computed across a spline fit of each detected centerline [5], an infinite active contour model, using an infinite perimeter regularizer and multiple region information [29] or using neighbourhood estimator before filling filter [2], still they cannot be used in large studies for evaluating the progression of the disease. Their consistency and accuracy/precision as well as the measurement errors across datasets with different image quality , do not allow us to find these subtle changes that occur inside the vasculature over time, and which we are trying to identify in the same retinas. Both widths and angles were measured twice by the same observer, yielding an intra-rater reliability of over 90% for the absolute agreement. Therefore both groups of measurements were kept by taking their average.

Empirically, the changes that we are trying to identify as a consequence of the proliferation of diabetes can be as small as 1% of change pre- to post- DR, and in the most extreme cases they can reach up to 7-10 %. Therefore the semi-automated approaches are still preferred, because they let us measure the same junctions over time and be consistent to the accuracy of our measurements. From each junction's vessels' widths, the branching coefficient (BC) is derived and calculated by eq. 1,

$$Branch.Coef. = \frac{W_1^2 + W_2^2}{W_0^2}, \tag{1}$$

Where w1, w2, w0 are the widths of the larger child vessel, smaller child vessel and parent vessel respectively. Furthermore another derivative feature was introduced, as the ratio between the junction angle and the corresponding BC (eq.2).

$$(Angle/BC)_i = \frac{Angle_i}{BC_i}, \tag{2}$$

3.1.2 Tortuosity

In addition to these, tortuosity of the vessels, which is a property of a curve being tortuous i.e twisted, was also included and calculated by the method proposed in[13]. For this purpose the images were segmented, using an algorithm described in [14], and the coordinates of each segment were also extracted (fig. 2), in order to calculate the local tortuosity. The global, image-level tortuosity was then derived by using the mean, median, standard deviation and the third quartile, in a similar way like in a previous study [22].

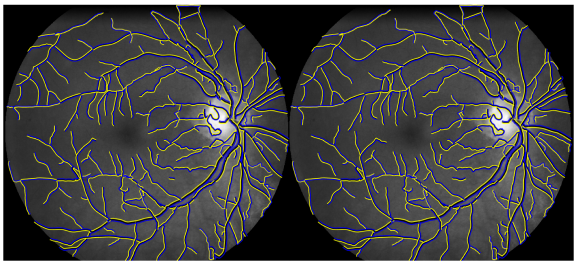

Fig. 2 Segmented images from the same patients before (left) and after diabetic retinopathy (right), used for the evaluation of tortuosity. Vessels edges and centerlines are highlighted.

3.1.3 Fractal Dimension & Lacunarity

Fractal dimension (FD) and lacunarity are another two important features that are included in this study. The former can give us a measure of complexity of a structure, as long as it can be considered a fractal. The latter is a measure of heterogeneity of a fractal structure.

Fractality

Fractals present various degrees of self-similarity in different scales. Human retina has been found to almost be a self-similar structure, thus being possible to be analyzed as such, giving us a measure of complexity, letting us also investigate, whether

it changes during different periods[9]. Its discriminatory power was evaluated within the classification system in conjunction with the other features. Higher values of FD indicate more complex structure.

Lacunarity

Complimentary to the FD, lacunarity was also evaluated, which is a counterpart of FD, describing the gappiness between the structures, or alternatively how the fractals fill the space.

For FD, the well established method of box-counting algorithm (Minkowski - Bouligand dimension) was used [24], based on eq. 3. For this purpose, all the images were segmented [14], obtaining the binary vascular trees, in order to apply the box-counting and gliding box methods. Each image of the same patient was processed, in order to include the same vessels, making sure that any identified differences are due to the proliferation of diabetes and not an error from the algorithm.

$$FractalDim. = \lim_{r \to 0} \frac{LogN(r)}{Log1/r}, \tag{3}$$

in which N(r) refers to the number of boxes of side length r that has to be used to cover a given area in the Euclidean n-space, by using a sequential number of descending size boxes. This occurs in multiple orientations. The final dimension in the 2D space is between 1 and 2 ($1 \leq D \leq 2$)[23].

Lacunarity was estimated using the gliding-box algorithm, for different grid orientations [28]. A unit box of size r is chosen randomly and the number of set points p are counted i.e. the mass. The procedure is repeated with the box centered consecutively for each point within the set, creating a distribution of masses B(p,r). Finally, we get the probability, by converting the distribution into probability distribution $Q(p,r)$, dividing by the total number of boxes (B) of size r (eq.4).

$$Q_{p,r} = \frac{B_{(p,r)}}{B_{(p)}}, \tag{4}$$

Finally, after several transformations, the gliding box equation can be written in terms of the accumulated sum of the mean and the second moments of all boxes (eq.5).

$$L_{GB}(r) = \frac{B_{(r)} \sum_{i=1}^{B_{(r)}} p(i,r)^2}{[\sum_{i=1}^{B_{(r)}} p(i,r)]^2}, \tag{5}$$

where the denominator is the square of the total number of elements in the data set[28].

3.1.4 Arterio-Venous Ratio

Central retinal vein (CRV) and artery (CRA) are the two major vessels of the retina. CRV leaves the optic nerve head 10mm from the eyeball, draining the blood from the capillaries into the superior ophthalmic vein or to the cavernous sinus directly, depending on the individual [7]. On the other side, the CRA branches off the ophthalmic artery, crossing inferior to the optic nerve head within its dural sheath to the eyeball. Since these two vessels cannot be seen in the retinal fundus images, it has been proposed, initially by Parr [26] and then revised by Knudtson [16], a method to estimate the central retinal vein and artery equivalent, CRVE and CRAE respectively, based on the eq.6 and eq.7, derived partly by the branching coefficient that they estimated in normotensive subjects. The region of interest is defined as shown in fig.3, and includes the region where the edges of the vessels course through at 0.5 to 1.0 disc diameters from the optic disc margin. The region between this area

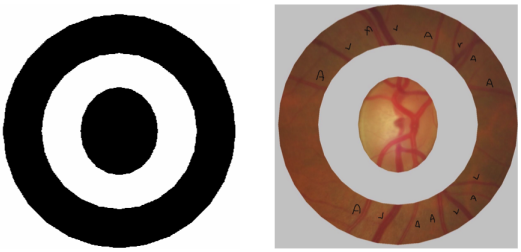

Fig. 3 On the left, the mask as created by our algorithm is shown, after defining the optic disc diameter, and on the right the region of interest, with the veins and arteries labeled, from which the CRVE,CRAE and AVR are calculated.

and the optic disc is excluded, as not having the vessels attained their status inside the retina yet. Within this area, the six largest veins and the six largest arteries are measured, following an iterative procedure of pairing up the largest vessels with the smallest ones, until a final single number is obtained. All the values are entered in eq.6 and eq.7 for arterioles and venules respectively.

The final value for the vein is termed central retinal vein equivalent (CRVE) and the respective final value for the artery is termed central retinal artery equivalent (CRAE). The ratio CRAE/CRVE is known as arterio-venous ratio.

$$Arterioles : \hat{W} = 0.88 * \sqrt{(W_1^2 + W_2^2)} \tag{6}$$

$$Veins : \hat{W} = 0.95 * \sqrt{(W_1^2 + W_2^2)} \tag{7}$$

where ŵ is the estimate of the parent trunk arteriole or venule and w_1,w_2 are the two branches (children).

3.2 Design & Analysis

All of the above features were evaluated separately, using a mixed model design filter, as described in the next subsection [20]. Based on this design, repeated measures analysis of variance (ANOVA) was used, in order to calculate the F-statistic and finally the p-value for each feature. In that way, we try to evaluate whether any observed differences between the two groups, for each feature, are just random observations, or whether they can be attributed to the disease's proliferation. This is also a way of defining the importance of these features and thus make an initial feature selection. It is worth mentioning that, when dealing with features that have a biological meaning, it has to more deeply be investigated, whether they should be included in a classification system, regardless of the result of the statistical analysis. The mixed model based on the repeated measures nature of the analysis, increases the statistical power, requiring fewer subjects to be analyzed [11]. Including matched junctions and the same groups of patients, could lead to the decrease of both the statistical error (difference from the unobserved population mean) and the residuals (difference from the sample mean). In order to make sure that this parametric test is the correct one for the analysis of our data, normality and sphericity tests were run for each feature. For the former, the Shapiro-Wilk test was used, and the null hypothesis that the data are normally distributed was not rejected, regardless of the feature under investigation (p-values ranging from 0.30 to 0.56). Similarly for the sphericity, the Mauchlys test was used, which again failed to reject the null hypothesis that the assumption of sphericity is met (p-values ranged from 0.16-0.39).

Although ANOVA is robust in marginal violations of normality, it still suffers from sphericity, which if present, causes the test to become unstable i.e. leads to an increase of Type I error;that is, the likelihood of detecting a statistically significant result when there is not one.

3.2.1 Mixed Model Filter

As mentioned above, in order to account for the different way that the features are measured, a mixed model factorial/nested design has been developed in MATLAB 2015b version, in which all the local measurements are used in the statistical analysis. As can be seen in fig. 4, in the case of widths and angles, we have multiple measurements within each subject, in a nested formation. That means that all these observations are not independent, and thus that needs to be taken into account. Using this design, each measurement in *P1jM1k*, where 1 is the first case, e.g. pre-DR group, *j* the corresponding patient and *k* the specific measurement, is related only to the corresponding measurement at the same exact junction in *P2jM2k*. This logic is applied in this model,which is then analyzed by ANOVA.

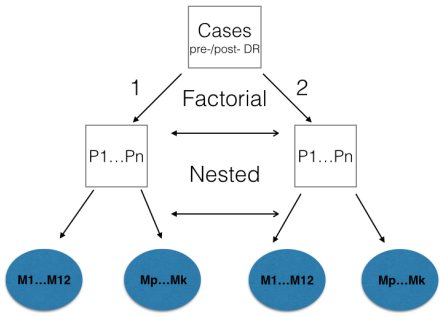

Fig. 4 Mixed model design filter used for the statistical analysis of each feature and for the initial feature selection.

3.3 Classifiers

In order to test the discriminative power of these features, two different approaches were followed. Firstly, a regularized random forests classifier was used, slightly adjusted for the feature selection process, as proposed in [8]. Secondly, a logistic regression model was developed, using both Least Absolute Shrinkage and Selection Operator(Lasso) and ridge regression, as a hybrid penalty for the coefficients of the features (L1- and L2- norms), which is called elastic net regularization described in [10] .

3.3.1 Regularized Random Forests

Random forests is a well-established supervised classifier and very popular in machine learning. It was proposed by Breiman as an improvement to the decision trees' bagging method [6]. It consists of multiple decision trees, each of which is grown on a bootstrap sample, taken from the original training data. The Gini index (Gini(u)) at node u, is defined as

$$Gini(u) = \sum_{c=1}^{c} \hat{p}_c^u (1 - \hat{p}_c^u) \tag{8}$$

where $\hat{p}_c{}^u$,is the proportion of class-c observation at node u. Subsequently, the Gini information gain of Xi for splitting node u,is the difference between the impurity at node u and the weighted average of impurities at each child node of u. This can be seen in eq.9[8].

$$Gain(X_{i,}u) = Gini(X_{i,}u) - w_L Gini(X_{i,}u^L) - w_R Gini(X_{i,}u^R) \tag{9}$$

where u^L and u^R are the left and right children nodes of u respectively.Similarly w_L and w_R are the proportions of instances assigned to the left and right children nodes. The most important part of random forests is the mtry function, in which a random set of features out of P is evaluated. The feature with the highest Gain$(X_{i,}u)$ is used for splitting the node u. The importance score for variable X_i is then calculated,

$$Importance_i = \frac{1}{ntree} \sum_{u \in S_{Xi}} Gain(X_{i,}u) \tag{10}$$

where S_{Xi} refers to the set of nodes split by X_i in random forests with ntree number of tree. In short, the regularized version of random forests (RRF) can select a compact feature subset, by including an additional penalty coefficient, creating a regularized information gain (eq.11) [8]

$$Gain_R(X_{i,}u) = \begin{cases} \lambda \cdot \text{Gain}(\text{X}_{i,}u) & i \notin F \\ \text{Gain}(\text{X}_{i,}u) & i \in F \end{cases} \tag{11}$$

in which F refers to the set of indices of features used for splitting in the previous nodes. The parameter $\lambda \in (0,1]$ is the penalty coefficient. When i$\notin$ F the coefficient penalizes the *ith* feature for splitting node u. Smaller leads to a larger penalty. Regularized random forests uses $Gain_R(X_{i,}u)$ at each node, and adds the index of a new feature to F. For instance a RRF with $\lambda = 1$, has the minimum regularization, however a new feature has to be more informative at a given node than the features that have already been included to the feature subset. The feature subset selected by RRF($\lambda = 1$) is termed the least regularized subset, as it offers minimum regularization. Apart from the feature selection process, the rest of the algorithm is exactly the same as the initially proposed random forests classifier [8].

For the evaluation of the performance of RRF, the Out of Bag error (OOB) was used, which is the internal way of validating the performance of random forests classifier[6]. In addition, ten-fold cross validation was utilized.

3.3.2 Logistic regression with elastic net penalty

In this study, where the response variable is binary, a regularized logistic regression model is used[10]. The difference with the ordinary logistic regression has to do with the penalty parameter applied to the coefficients. In the case of ridge regression, the coefficients of correlated predictors are shrunk towards each other, allowing them to work together. From a Bayesian point of view, the ridge regression works better, if there are many predictors and all have non-zero coefficients.

On the other side the least absolute shrinkage selector operator (Lasso) is to some extend indifferent to very correlated predictors, tending to pick one and discard the rest. The Lasso penalty corresponds to a Laplace prior, which expects many coefficients to be zero or close to zero and a small subset of non-zero coefficients. In the middle of this, elastic net with =1 - ε for small $\varepsilon > 0$, performs similarly to Lasso, removing however any extreme behavior caused by highly correlated predictors. The general formula *Pa* of elastic net, as seen in eq.13, introduces a compromise between ridge and Lasso. As α increases from 0 to 1 for a specific value of parameter λ, the sparsity of the solution in eq.15 (referring to the coefficients equal to zero), increases monotonically from 0 to the sparsity of the Lasso solution. More specifically, assuming that the response variable G = 1,2, then the logistic regression model represents the class-conditional probabilities, through a linear function of the

predictors, which in the logarithmic form is given by eq. 12[10].

$$log\frac{Pr(G=1|x)}{Pr(G=2|x)} = \beta_0 + x^T\beta \tag{12}$$

Where in this case the model is fit by regularized maximum binomial likelihood.

$$P_\alpha(\beta) = \sum_{j=1}^{p}\big[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|\big] \tag{13}$$

Let $p(x_i) = Pr(G=1|x)$ be the probability according to eq.14.

$$Pr(G=1|x) = \frac{1}{1+e^{-(\beta_0+x^T\beta)]}} \tag{14}$$

For an observation i at specific values for the parameters (β_0,β), the penalized log likelihood is maximized (eq.15).

$$\max_{(\beta_0,\beta)\in\mathbf{R}^{(p+1)}}\left[\frac{1}{N}\sum_{i=1}^{N}\big\{I(g_i=1)logp(x_i)+I(g_i=2)log(1-p(x_i))\big\} - \lambda P_\alpha(\beta)\right] \tag{15}$$

Replacing , the log-likelihood part of eq.15 takes the form,

$$l(\beta_0,\beta) = \frac{1}{N}\sum_{i=1}^{N} y_i\cdot(\beta_0 + x_i^T\beta) - log(1+e^{(\beta_0+x_i^T\beta)}) \tag{16}$$

a concave function of the parameters. In this approach, for every value of λ, an outer loop is created for the computation of the quadratic approximation l_Q of eq.16 about the current parameters $(\tilde{\beta}_0,\tilde{\beta})$.

$$l_Q(\beta_0,\beta) = -\frac{1}{2N}\sum_{i=1}^{N} w_i(z_i - \beta_0 - x_i^T\beta)^2 + C(\tilde{\beta}_0,\tilde{\beta})^2 \tag{17}$$

where

$$z_i = \tilde{\beta}_0 + x_i^T\tilde{\beta} + \frac{y_i - \tilde{p}(x_i)}{\tilde{p}(x_i)(1-\tilde{p}(x_i))} \tag{18}$$

$$w_i = \tilde{p}(x_i)(1-\tilde{p}(x_i)), (weights) \tag{19}$$

Finally, the penalized weighted least-squares problem can be solved by eq.20, using the coordinate descent approach[10].

$$\min_{(\beta_0,\beta)\in\mathbf{R}^{(p+1)}}\Big[-l_Q(\beta_0,\beta) + \lambda P_\alpha(\beta)\Big]. \tag{20}$$

A number of sequential nested loops are created :

• Outer loop: Decrement λ.

- Middle loop: New quadratic approximation l_Q for the current parameters $(\tilde{\beta}_0, \tilde{\beta})$.
- Inner loop: Execute the coordinate descent algorithm on the penalized weighted least-squares problem (eq.20).

Further information of the above method is given by Friedman et al. [10].

In the same way as RRF, ten-fold cross-validation was used to evaluate the classifier.

4 Results

This section will present the results of the three different approaches that were previously addressed .

MMF: In which the results of the analysis of every feature are presented, together with some more information about the data.

RRF: In the first part the results of the feature selection process, according to their importance will be shown, followed by the classification results based on the feature subset.

LOG: Similarly to the RRF, in the elastic net logistic regression the first part will be devoted to the selection of α and λ parameters and subsequently the feature subset, and then at the last part, the results of the classification will be shown.

All of the features were scaled (normalized), by centering the data. This was done by subtracting the mean and normalizing it dividing by the standard deviation. Especially with the gradient descent algorithms, like logistic regression, this can be beneficial, as we can achieve better numerical stability and quicker convergence.

The open source software "R" was used both for the RRF and Elastic net logistic regression classifiers, as well as for all the evaluation steps and feature selections.

4.1 Evaluation of features with MMF

In table 1, we can find the results of the analysis using the MMF. As can be seen, some of the features significantly differed across the groups, whereas some others not. In addition to that, no significant results (thus excluded from table 1) were observed in almost any combination of features, when using the mean values, medians or standard deviations (although p-values were between 0.15-0.28), which highlights the superiority of the MMF, in which all the measurements are accounted for as measured.

As can be seen in table 1, arteries' widths and angles, veins' widths, arteries' angles, fractal dimension and tortuosity (standard deviation) are found to differ significantly between the two groups. The rest of them did not appear to do so, however,

Table 1 Mixed Model Analysis of Variance Results

Feature Name	p-value ($\alpha = 0.05$)	F-value (dfn,dfe)[a]	Group Means (SD) (pre-/post- DR)
Arteries Widths	**0.01**	**6.53 (1,299)**	**11.14 (2.20), 10.45 (1.93)**
Arteries Angles	**0.022**	**5.24(1,99)**	**88.45 (8.74), 85.63 (6.93)**
Arteries BC	0.30	1.3 (1,99)	1.24(0.11),1.29(0.12)
Veins Widths	**0.0005**	**16.95(1,299)**	**13.23(2.81),12.17(2.28)**
Veins Angles	0.62	0.24(1,99)	81.72(6.9),81.52(6.62)
Veins BC	0.45	0.85(1,99)	1.12(0.10),1.12(0.11)
Fractal Dim.	**0.024**	**6(1,24)**	**1.628(0.06),1.594(0.06)**
Lacunarity	0.65	0.45(1,24)	0.22(0.04),0.22(0.05)
Tortuosity(SD)	**0.021**	**5.79(1,24)**	**0.074(0.013),0.089(0.02)**
CRVE	0.76	0.10(1,24)	29.13(4.39),28.01(5.53)
CRAE	0.37	0.83(1,24)	20.21(2.87), 19.74(3)
AVR	0.81	0.07(1,24)	0.697(0.10),0.704(0.14)

[a] dfn:degree of freedom numerator, dfe:degree of freedom error term

since all these features reflect functional changes, still remain useful for further investigation and possible inclusion in a classification system.

Interestingly enough, the arteries' widths have been decreased at the first year of DR by almost 6.5% and the angles by 3.5%. Similarly, but only for the widths, veins showed a decrease at the first year of DR by almost 8%.

In fig.5, we can see two examples of how the differences between the post-DR and pre-DR measurements are correlated with the age of the patients, despite the fact that the data are limited for giving us a reliable result. However they can just be used as an indication or a trend of the data.

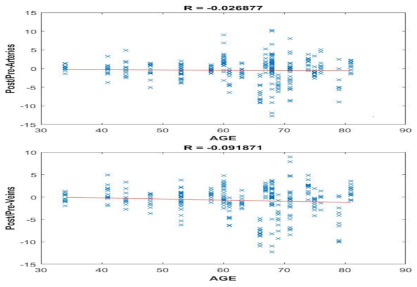

Fig. 5 The plot on the top, shows the differences between the measurements post-DR with the corresponding pre-DR measurements for the arteries. x axis:age, y axis: the individual differences. On the bottom, we find the same plot but for the veins. On top of them is the correlation coefficient parameter R.

4.2 Classification with RRF

All the available features were initially included in the classifier, in order to evaluate their importance. In addition to the features that appear in table 1, for selecting the

feature subset, we included all the original features, including fractal-to-lacunarity ratio and Angle-to-BC ratio, as well as the descriptive statistics of them. In total 20 features were included, with 50 observations in total (25 for each class-balanced design), however fourteen of them were negatively affecting the performance .The classifier had a similar performance when all the initial values for arteries and veins were used, instead of the descriptive statistics, thus the aforementioned balanced structure was chosen.

The final six selected features of our feature subset are a) the mean of arteries' BCs, b) Angle-to-BC ratio of veins, c) Tortuosity, d) Fractal dimension, e) Vein SD and f) Angle-to-BC ratio of arteries. In fig.6 we can see the importance for each of these features. The number of decision trees used for training the classifier was chosen at 5000, although it converged earlier. Choosing more trees than needed, does not affect the performance of the classifier. Larger number of trees produce more stable models and covariate importance estimates, but require more memory and a longer run time. The mtry parameter refers to the number of features available for splitting at each tree node and by default is set as the square root of the total number of features(rounded down).

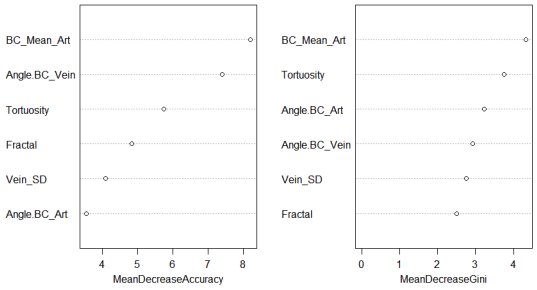

Fig. 6 Mean decrease accuracy shows how much the performance of the classifier will be affected if this feature is removed. A similar measure is the Gini index which is a measure of each feature's importance based on the Gini impurity index, used for the calculation of splits during training.

Finally the performance of the classifier can be seen in fig.7 for the out of bag error and area under the ROC curve. As can be seen, the regularized random forests classifier achieved an OOB error of 22.5% and AUC 0.7925 (Average over all the iterations of the cross-validation). Regarding accuracy, this was at 79.5%.

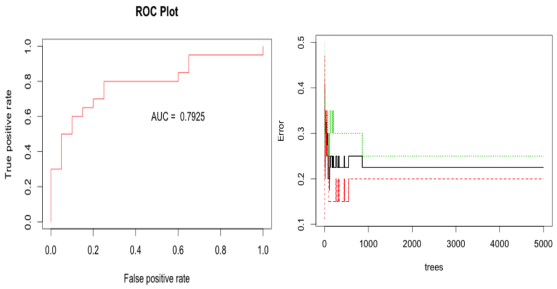

Fig. 7 On the left the ROC curve and the corresponding AUC value can be found. On the right the Out of Bag error for the whole training phase can be seen. The red and green line are the two classes and the black one is the average of them i.e. the final OBB error.

4.3 Penalized logistic regression

As described in the previous section, when running a logistic regression model with elastic net penalty, a few factors have to be taken into account.

- Like in RRF the feature subset has to be selected. This occurs in two steps. The first step includes all the features under investigation. The second step is the final selection between the variables that had the best performance in the first step. In both cases, a ten-fold cross-validation was used in order to calculate the mean square error for the variables for different values of λ, and also of the penalty parameter α, as the compromise between Lasso and ridge regression.
- Secondly, for the selected feature subset and the tuning parameter λ, we run the regression for varying penalties α, ranging from 0 to 1 with 0.1 step.
- After running the cross-validation for all the models, we evaluate which one fits best our data, and therefore define the optimum parameters for λ and α. Having these values set, we validate the performance by reporting the AUC, the accuracy and the ROC curve .

After running the relevant feature selection with the RRF, it was anticipated to obtain a similar feature subset with the logistic regression, since the six selected features were performing quite well. indeed the same six features had the best score. In contrast, the rest fourteen were all together deteriorating the performance of the classifier by about 0.10 of the AUC, having extensive negative impact to the classifier.In fig.8, the cross validation of the different features can be seen, which initially helps us decide which features to discard and then work with the final ones. Secondly it can be inferred how strong should the penalty be, after controlling for

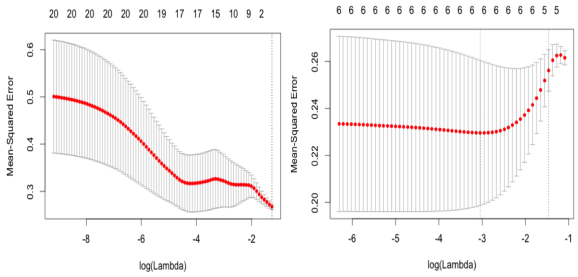

Fig. 8 On the left we can see the feature selection process for all the features, which leads us to the right one, where we can see the final six features based on the their performance according to the mean square error and for different values of λ.The red dotted line is the cross-validation curve, together with the upper and lower standard deviation curves along the λ sequence.

the λ parameter. The best results were obtained for a penalty α=0.2.

Additionally, in fig.9, there is an informative illustration of how the coefficient of each predictor changes along the different λ values. The optimum results were

obtained with λ=0.03 as the tuning parameter that controls the overall strength of the penalty.

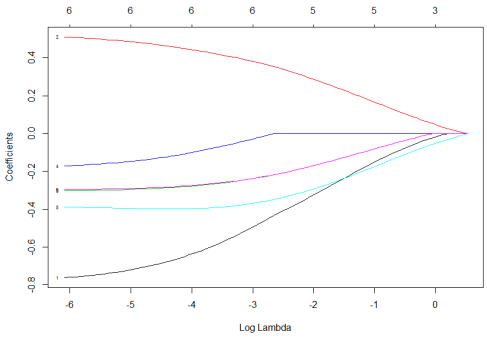

Fig. 9 Plot showing how the coefficients of all the features are adjusted according to the different values of λ that have been applied to each of them. For higher values of λ the predictors are starting moving towards zero. The x axis is the logarithm of λ.

Finally the logistic regression classifier had a similar performance with RRF, as can be seen in fig.10, having an AUC=0.785 and accuracy of 78%.

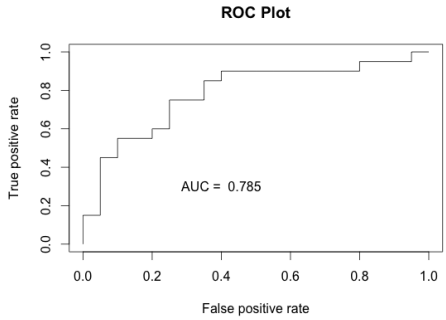

Fig. 10 ROC plot showing the Area under the ROC curve after a ten-fold cross validation. AUC in this case is 0.785. This value is the average over all the iterations of the cross validation.

4.4 Discussion

Taking into account the limited amount of data, as well as the nature of the features, which represent the geometry of the retina and not any other image information, the performance of both classifiers is good enough to let us keep investigating those as well as additional features even further.

Another useful metric of the performance of the classifier is the precision/recall plot (fig.11). Precision is a metric that gives us the positive predictive value of the

classifier, while recall give us the true positive rate. Both these metrics are useful for evaluating a classifier, together with accuracy, AUC and ROC plot.

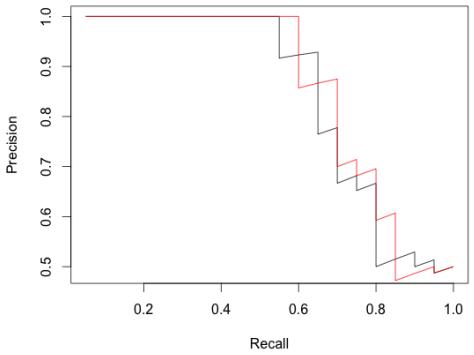

Fig. 11 Precision-Recall plot for both logistic regression (black line) and the RRF classifier (red line). Precision is defined as the $\frac{TruePositive}{TruePositive+FalsePositive}$, whereas recall is the $\frac{TruePositive}{TruePositive+FalseNegative}$.

5 Conclussion & Discussion

Diabetes is a major disease, with millions of people being under medication in order to minimize its consequences. Identifying the changes in the vasculature during the progression of diabetes and measuring them is of paramount importance. Robust and reliable tools are needed for long term studies as well as properly designed experiments, in order to be able to discriminate over the different stages of progression. The alterations are so minor and in such a small scale that sometimes is very hard to measure and identify them. Hence novel tools for extracting information and analyzing data in a larger scale, are crucial for identifying the progression and also create reliable models with valid and robust biomarkers.

In this study, a comprehensive analysis was presented, using many different retinal geometric features and methods. To our best of knowledge, it is the first time that all these features together like CRVE/CRAE, tortuosity, fractal dimension, BC etc. were evaluated and/or utilized inside a classification system, yielding that performance, which is an improvement of approximately 2% from the previous study[20].

As aforementioned, it is a challenging task to extract all these features accurately, evaluate them and more importantly, associate any changes with the progression of diabetes. More data are always needed, in order to identify and investigate all of the possible underlying conditions and variations that occur as the disease progresses. The results of this study give us the boost to extend our investigation in more intervals of diabetes, by including even more data and features. Our immediate next work will include but not limited to building a multiclass system beyond the binary level for different periods of diabetes. Moreover, specific regions inside the retina

will be investigated, focusing also on the bifurcations and the branching patterns of the vasculature.

Acknowledgements This research study was supported by a Marie Sklodowska-Curie grant from the European Commission in the framework of the REVAMMAD ITN (Initial Training Research network), Project number 316990.

References

1. Al-Diri, B., Hunter, A., Steel, D., Habib, M.: Manual measurement of retinal bifurcation features. In Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE. 4760-4764 (2010)
2. Annunziata, R., Garzelli, A., Ballerini, L., Mecocci, A., Trucco, E.: Leveraging multiscale hessian-based enhancement with a novel exudate inpainting technique for retinal vessel segmentation.Biomedical and Health Informatics, IEEE Journal of (2015)
3. Antonetti, D. A., Barber, A. J., Bronson, S. K., Freeman, W. M., Gardner, T. W., Jefferson, L. S., Simpson, I. A.: Diabetic retinopathy seeing beyond glucose-induced microvascular disease. Diabetes. 55(9), 2401-2411 (2006)
4. Avakian, A., Kalina, R. E., Helene Sage, E., Rambhia, A. H., Elliott, K. E., Chuang, E. L., Clark, J.I,Chuang,E.L, Parsons-Wingerter, P.: Fractal analysis of region-based vascular change in the normal and non-proliferative diabetic retina. Current eye research, 24(4), 274-280 (2002)
5. Bankhead, P., Scholfield, C. N., McGeown, J. G., Curtis, T. M.: Fast retinal vessel detection and measurement using wavelets and edge location refinement. PloS one, 7(3), e32435 (2012)
6. Breiman, L.: Random forests. Machine learning. 45(1), 5-32 (2001)
7. Cheung, N., McNab, A. A.: Venous anatomy of the orbit. Investigative ophthalmology visual science. 44(3), 988-995 (2003)
8. Deng, H., Runger, G.: Feature selection via regularized trees. Neural Networks (IJCNN), The 2012 International Joint Conference on IEEE.1-8 (2012)
9. Family, F.,Masters,B.R.,Platt,D.E.:Fractal pattern formation in human retinal vessels.Physica D: Nonlinear Phenomena.38(1),98-103 (1989)
10. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of statistical software. 33(1),1-22 (2010)
11. Guo, Y., Logan, H. L., Glueck, D. H., Muller, K. E.: Selecting a sample size for studies with repeated measures. BMC medical research methodology. 13(1), 100 (2013)
12. Habib, M. S., Al-Diri, B., Hunter, A. Steel, D. H.: The association between retinal vascular geometry changes and diabetic retinopathy and their role in prediction of progression-an exploratory study. BMC ophthalmology, 14(1), 89 (2014)
13. Hart, W. E., Goldbaum, M., Ct, B., Kube, P., Nelson, M. R.: Measurement and classification of retinal vascular tortuosity. International journal of medical informatics. 53(2), 239-252 (1999)
14. Hunter, A., Lowell, J., Ryder, R., Basu, A., Steel, D.: Tram-line filtering for retinal vessel segmentation.Proceedings of the 3rd European Medical and Biological Engineering Conference (2005)
15. Jorgensen, C. M., Hardarson, S. H.,Bek, T.: The oxygen saturation in retinal vessels from diabetic patients depends on the severity and type of vision threatening retinopathy. Acta ophthalmologica. 92(1), 34-39 (2014)
16. Knudtson, M. D., Lee, K. E., Hubbard, L. D., Wong, T. Y., Klein, R.,Klein, B. E.: Revised formulas for summarizing retinal vessel diameters. Current eye research. 27(3), 143-149 (2013)

17. Leontidis, G., Al-Diri, B., Hunter, A.:Diabetic retinopathy: current and future methods for early screening from a retinal hemodynamic and geometric approach.Expert Review of Ophthalmology.9(5),431-442 (2014)
18. Leontidis, G., Al-Diri, B., Hunter, A.:Study of the retinal vascular changes in the transition from diabetic to diabetic retinopathy eye. Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE. 26-30 August (2014)
19. Leontidis, G., Al-Diri, B., Hunter, A.:Retinal vascular geometry: Examination of the changes between the early stages of diabetes and first year of diabetic retinopathy. Science and Information Conference (SAI).709-713. 28-30 July (2015).doi: 10.1109/SAI.2015.7237220
20. Leontidis, G., Al-Diri, B.,Wigdahl, J., Hunter, A.:Evaluation of Geometric Features As Biomarkers of Diabetic Retinopathy for Characterizing the Retinal Vascular Changes During the Progression of Diabetes.Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE. 25-29 August (2015)
21. Leontidis, G., Caliva, F., Al-Diri, B., Hunter, A.:Study of the retinal vascular changes between the early stages of diabetes and first year of diabetic retinopathy.Investigative Ophthalmology Visual Science. 56(7), (2015)
22. Leontidis, G., Wigdahl, J., Al-Diri, B., Ruggeri, A., Hunter, A.:Evaluating tortuosity in retinal fundus images of diabetic patients who progressed to diabetic retinopathy.Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE. 25-29 August (2015)
23. Li,J.,Du,Q.,Sun,C.:An improved box-counting method for image fractal dimension estimation.Pattern Recognition.42(11),2460-2469 (2009)
24. Mandelbrot, B.B.: The fractal geometry of nature. Macmillan. 173, 1983
25. Nguyen, T.T.,Wong,T.Y.: Retinal vascular changes and diabetic retinopathy. Current diabetes reports.9(4),227-283 (2009)
26. Parr, J. C.,Spears, G. F. S.: General caliber of the retinal arteries expressed as the equivalent width of the central retinal artery. American journal of ophthalmology. 77(4), 472-477 (1974)
27. Shimizu, K., Kobayashi, Y.,Muraoka, K.: Midperipheral fundus involvement in diabetic retinopathy. Ophthalmology. 88(7), 601-612 (1981)
28. Tolle, C. R., McJunkin, T. R., Gorsich, D. J.: An efficient implementation of the gliding box lacunarity algorithm. Physica D:Nonlinear Phenomena. 237(3), 306-315 (2008)
29. Zhao, Y., Rada, L., Chen, K., Harding, S.,Zheng, Y.: Automated vessel segmentation using infinite perimeter active contour model with hybrid region information with application to retinal images.Medical Imaging, IEEE Transactions on , 34(9),1797-1807 (2015)