Diabetic Retinopathy Detection using Random Forest

J D Labhade ¹, L K Chouthmol ²

¹E&TC (Signal Processing) Department, SKH Late G N Sapkal COE, Anjaneri, Nasik, jyoti.labhade@gmail.com
²E&TC (Signal Processing) Department, SKH Late G N Sapkal COE, Anjaneri, Nasik, lata_51105@rediffmail.com

Abstract- A great challenge in the biomedical engineering is the non-invasive assessment of the physiological changes occurring inside the human body. Specifically, detecting the abnormalities in the human eye is extremely difficult due to the various complexities associated with the process. Retina is the significant part of the human eye which can reflect the abnormal changes in the human eye. Hence, retinal images captured by digital cameras can be used to identify the nature of the abnormalities affecting the human eye. Retinal image analysis has gained sufficient importance in the research arena due to the necessity for disease identification techniques. Abnormality detection using these techniques is highly complex since these diseases affect the human eye gradually. Conventional disease identification techniques from retinal images are mostly dependent on manual intervention. Since human observation is highly prone to error, the success rate of these techniques is quite low. Since treatment planning varies for different abnormalities, the accuracy of the identification techniques must be significantly high. Lack of accuracy in these techniques may lead to fatal results due to wrong treatment. Hence, there is a significant necessity for automation techniques with high accuracy for retinal disease identification applications. One of the techniques used in this paper is Random Forest with and without sampling. With sampling the accuracy is increased and is 74.93% and without sampling it is 52.05%.

Keywords- Diabetic Retinopathy (DR), microaneurysms, hemorrhage, hard exudates, soft exudates, sensitivity, specificity, accuracy.

I. INTRODUCTION

Figure 1. Normal Eye and Diabetic Retinopathy

Diabetic retinopathy occurs in patients suffering from diabetes, which causes damage to the retina of the eye. This eventually leads to total vision loss. Diabetes is caused due to the body’s inability to store and make use of the sugar level in the blood. Usually there are no early visible
symptoms of the disease and as the disease progresses the presence of micro aneurysms, exudates both hard and soft and new blood vessels can be observed. The main anatomical features in the image, notably the optic disc and the macula. The optic disk is a circular shaped anatomical structure with a bright appearance. The fovea is the very center of the macula, the site of our sharpest vision. The diabetes may cause abnormalities in the retina (diabetic retinopathy), kidneys (diabetic nephropathy), and nervous system (diabetic neuropathy). The diabetes is also a major risk factor in cardiovascular diseases. The diabetic retinopathy typically begins as small changes in the retinal capillaries.

Diabetic retinopathy is of two type’s namely non proliferative and proliferative type. Non proliferative is the early stage of the disease characterized by the presence of micro aneurysms. As the disease progresses the retina is deprived of oxygen and new blood vessels are formed in the retina. These vessels eventually leak and leads to clouding vision. The first detectable abnormalities are microaneurysms (Ma) which are local distensions of the retinal capillary and which cause intraretinal hemorrhage (H) when ruptured [7][8]

Micro aneurysms are small red dots on the retinal surface, which occur due to capillary occlusion leading to lack of oxygen and progression of the disease. They are less than the diameter of the optic vein. Accumulation of proteins and lipids occur in the form of exudates. Vision loss occurs when they occur in the macula.

Exudates appear as yellow or white structures in the retina. There are two types of exudates based on their appearance and occurrence. Hard exudates have well defined boundaries and the soft exudates have unclear boundaries also known as cotton wool spots.

Haemorrhages in the retina occur due to bleeding. Dot haemorrhages lie deep within the retina and reflect leakage of the veins and capillaries. Dot haemorrhages are an indication of diabetic retinopathy.

II. LITREATURE REVIEW

Some of the first automated detection methods for diabetic retinopathy were published by Baudoin et al. [1] to detect microaneurysms from fluorescein angiograms. By using a morphological top-hat transform with linear structuring element at different orientations small round shaped microaneurysms were distinguished from connected elongated structures such as vessels. Although the top-hat transform was very sensitive to microaneurysms, it introduced too many false alarms. Spencer et al. [2] exploited this feature and used the top-hat transform to produce candidate microaneurysms. The true microaneurysms were then pruned by using post-processing based on their earlier work [3] and classification. The candidate microaneurysms segmentation was conducted using a combination of top-hat transform and matched filtering with region growing. To improve the sensitivity of the candidate search a shade correction and dynamic range normalization steps were introduced in the pre-processing. After detection and segmentation of the candidate microaneurysms, the true microaneurysms were pruned from the spurious responses using a rule-based classifier with a number of shape and intensity based features. Kanika Verma et al [4] classification of the different stages of eye disease was done using Random Forests technique based on the area and perimeter of the blood vessels and hemorrhages. Accuracy assessment of the classified output revealed that normal cases were classified with 90% accuracy while moderate and severe NPDR cases were 87.5% accurate.

RF is a powerful machine learning method for classification and regression. The strengths of the RF approach are that: 1) it does not overfit; 2) it is robust to noise; 3) it has an internal mechanism to estimate error rates, called out-of-the-bag (OOB) error; 4) it provides indices of variable importance; 5) it naturally works with mixes of continuous and categorical variables;
and 6) it can be used for data imputation and cluster analysis. These properties have made RF increasingly popular in the last few years, especially in the field of genetics and imaging [4].

III. AUTOMATIC DETECTION OF DR

3.1 Feature Extraction Using MaZda

The totals of 778 images are used. For feature extraction MaZda software is used. MaZda load images in the form of Windows Bitmap, DICOM and unformatted grey-scale image files (raw images) with pixels intensity encoded with 8 or 16 bits. Additionally, there is an option for reading details of image acquisition protocol extracted from the image information header. Image normalization. There are three options: “default” (analysis is made for original image); "+/- 3 sigma" (image mean m value and standard deviation sigma is computed, then analysis is performed for grey scale range between [m-3sigma, m+3sigma]; "1%-99%" (grey scale range between 1% and 99% of cumulated image histogram is taken into consideration during analysis). Defining regions of interest (ROI), then analysis is performed within these regions. Up to 16 regions of any shape can be defined; they can be also edited, loaded and saved as disk files. Additionally, a histogram of defined ROI may be visualized and stored. Image analysis, which is computation of texture feature values within defined ROIs. The feature set (almost 300 parameters) is divided into following groups: histogram-, co-occurrence matrix-, run-length matrix-, gradient, autoregressive model- and Haar wavelet derived features. Displaying image analysis reports, saving and loading reports into disk files. Feature reduction and selection in order to find a small subset of features that allows minimum error classification of analyzed image textures. This is performed by means of two criterions: Fisher coefficient maximization and minimisation of probability of classification error. Selected features can be transferred to B11 program for further processing and/or classification. Image analysis automation by means of text scripts containing MaZda language commands. Scripts allow loading analyzed images and their ROI files, running the analysis and saving report files on disk [6].

Dataset

From the the total 778 images 287 features are extracted using MaZda. After extraction classes are formes as class 0, class 1, class 2, class 3 having 360, 88, 142, 188 respectively. We observe from the above details the data is imbalanced. Imbalanced data sets are a special case for classification problem where the class distribution is not uniform among the classes. Imbalanced data set problem occurs in classification, where the number of instances of one class is much lower than the instances of the other classes. The main challenge in imbalance problem is that the small classes are often more useful, but standard classifiers tend to be weighed down by the huge classes and ignore the tiny ones. In machine learning the imbalanced datasets has become a critical problem and also usually found in many applications such as detection of fraudulent calls, bio-medical, engineering, remote-sensing, computer society and manufacturing industries. After upsampling the dataset the images in all the classes are now 360. The final dataset is 1440 X 287 and Class is 1 X 1440.

3.2 Random forest

The Random Forests algorithm is one of the best among classification algorithms - able to classify large amounts of data with accuracy. Random Forests are an ensemble learning method (also thought of as a form of nearest neighbor predictor) for classification and regression that construct a number of decision trees at training time and outputting the class that is the mode of the classes output by individual trees (Random Forests is a trademark of Leo Breiman and Adele Cutler for an ensemble of decision trees). Random Forests are a combination of tree predictors.
where each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest. The basic principle is that a group of “weak learners” can come together to form a “strong learner”. Random Forests are a wonderful tool for making predictions considering they do not overfit because of the law of large numbers. Introducing the right kind of randomness makes them accurate classifiers and regressors. Single decision trees often have high variance or high bias. Random Forests attempts to mitigate the problems of high variance and high bias by averaging to find a natural balance between the two extremes. Considering that Random Forests have few parameters to tune and can be used simply with default parameter settings, they are a simple tool to use without having a model or to produce a reasonable model fast and efficiently.

Random Forests grows many classification trees. Each tree is grown as follows:

1. If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree.
2. If there are M input variables, a number mM is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning.

IV. RESULTS

Total images used : 778
Total features extracted from MaZDA =287

Before upsampling
Imbalanced dataset shown in the Table 1
Data matrix : 778 X 287
Class : 1 X 778

<table>
<thead>
<tr>
<th>Classes</th>
<th>Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 0</td>
<td>360</td>
</tr>
<tr>
<td>Class 1</td>
<td>88</td>
</tr>
<tr>
<td>Class 2</td>
<td>142</td>
</tr>
<tr>
<td>Class 3</td>
<td>188</td>
</tr>
<tr>
<td>Total</td>
<td>778</td>
</tr>
</tbody>
</table>

After upsampling
Final dataset
Data matrix : 1440 X 287
Class : 1 X 1400

<table>
<thead>
<tr>
<th>Classes</th>
<th>Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 0</td>
<td>360</td>
</tr>
<tr>
<td>Class 1</td>
<td>360</td>
</tr>
<tr>
<td>Class 2</td>
<td>360</td>
</tr>
<tr>
<td>Class 3</td>
<td>360</td>
</tr>
<tr>
<td>Total</td>
<td>1440</td>
</tr>
</tbody>
</table>
Table 3

<table>
<thead>
<tr>
<th></th>
<th>Before up-sampling</th>
<th>After up-sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>52.05</td>
<td>74.93</td>
</tr>
<tr>
<td>Mtry</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Ntree</td>
<td>500</td>
<td>500</td>
</tr>
</tbody>
</table>

CONCLUSION

MaZda package is an efficient and reliable set of tools for analysis of image textures. It provides a complete analysis path for texture images, including feature estimation, statistical analysis of feature vectors, classification and image segmentation. Additional information which includes the list of features produced by MaZda. Use of sampling technique affects the accuracy from the random forest method. However, the scope and direction for further work are to include more instances of retinal images to construct a robust classifier for detecting different stages of diabetic retinopathy (i.e. for training and testing) to achieve higher accuracy. The efficiency of the correct classification can also be improved by extracting more number of features from the images.

REFERENCES
