# Automated “Disease / No Disease” Grading of Age-Related Macular Degeneration by an Image Mining Approach

Yalin Zheng,[1,2] Mohd Hanafi Ahmad Hijazi,[3,4] Frans Coenen[3]

[1] Department of Eye and Vision Science, University of Liverpool, Liverpool, United Kingdom

[2] St. Paul's Eye Unit, Royal Liverpool University Hospital, Liverpool, United Kingdom

[3] Department of Computer Science, University of Liverpool, Liverpool, United Kingdom

[4] School of Engineering and Information Technology, Universiti Malaysia Sabah, Sabah, Malaysia.

**Corresponding Author:** Dr Yalin Zheng, Department of Eye and Vision Science, Institute of Ageing and Chronic Disease, University of Liverpool, 3rd Floor, UCD Building, Daulby Street, Liverpool, L69 3GA. E-mail: yalin.zheng@liv.ac.uk; Telephone: 0151 706 4083; Fax: 0151 706 5802.

Grant information: N/A.

Abstract word account: 247 / 250

Total word count: 3,967 / 3,500

## Abstract

**Purpose:** To describe and evaluate an automated grading system for age-related macular degeneration (AMD) by color fundus photography.

**Methods:** An automated “disease / no disease” grading system for AMD was developed based on image mining techniques. First, image pre-processing was performed to normalize color and non-uniform illumination of the fundus images, to define a region of interest, and to identify and remove pixels belonging to retinal vessels. To represent images for the prediction task, a graph based image representation using quadtrees was then adopted. Next, a graph mining technique was applied to the generated graphs to extract relevant features (in the form of frequent sub-graphs) from images of both AMD and healthy volunteers. Features of the training data were then fed into a classifier generator (Naïve Bayes and Support Vector Machines were used with respect to the evaluation presented later in this paper) for training purposes before employing the trained classifiers to classify new “unseen images”.

**Results:** The algorithm was evaluated on two publically available fundus image datasets (ARIA and STARE) comprising 258 images (160 AMD and 98 normal). Ten-fold cross validation was used. The experiments produced a best specificity of 100% and a best sensitivity of 99.4% with an overall accuracy of 99.6%. Our approach outperformed previous approaches reported in the literature.

**Conclusions:** The proposed technique has demonstrated a proof of concept for an automated AMD grading technique. It has the potential to be further developed as an automated grading tool for future whole scale AMD screening programs.

## Introduction

Age-related macular degeneration (AMD) is the leading cause of irreversible blindness in the developed world.[1] It has a significant impact upon the activities of daily living and the quality of life of patients affected by AMD; it consequently poses a substantial socio-economical burden on society. The prevalence of AMD and its resulting visual impairment and blindness is expected to significantly increase given the world's ageing population.[2] There is mounting evidence that highlights the significance of early diagnosis and treatment to prevent progression to advanced AMD and eventual loss of vision.[2,3]

The diagnosis of AMD is usually based on detecting its characteristic color fundus photographic features, such as drusen and pigment abnormality in the macula, using the Age-related Eye Diseases Study (AREDS) classification system and severity scale.[4,5] With respect to the importance of the detection of features for the diagnosis of AMD, substantial work has been directed at applying image processing and content-based image retrieval techniques to support the diagnosis of AMD, for example the automated segmentation of drusen.[6,7,8,9,10] However, performance of these segmentation-based techniques is still not sufficient for wide-scale clinical application, largely because of the fact that the underlying segmentation techniques are not robust enough for handling feature variations found in fundus images such as quality, color, illumination and so on. In fact, detection of lesions is merely a steppingstone for most medical applications; the objective is to extract useful clinical information for the follow-on decision-making process. The study described here was directed at systems for the automated diagnosis of AMD. Certainly a lesion detection based strategy would be a natural one to pursue, unfortunately this strategy has proved to be challenging and has yet to provide useful results, as noted in previous work on this aspect.[11, 12,13,14]

We advocate an alternative strategy, founded on the concept of image mining, to achieve an automated AMD classification system with a minimal need for segmentation. Image mining does not require a representation that is interpretable by human observers as long as image salient features are captured. Image mining based approach has been successful in categorizing Magnetic Resonance (MR) brain scan images,[15] with a correct selection of image features, the approach was conjectured to performed well in classifying images based on their color information. In this paper we promote the use of spatial context information within images. Our previous work has highlighted the challenge of this strategy, including the representation of images so as to preserve spatial relationships and the selection of appropriate features.[16] Here we propose, describe and evaluate a proof-of-concept image mining technique for disease/no-disease grading of AMD by color fundus photography.

## Methods

### a) Image Dataset

The proposed automated AMD grading system was evaluated using two publically available fundus images datasets, ARIA (http://www.eyecharity.com/aria_online) and STARE (http://www.ces.clemson.edu/~ahoover/stare). The ARIA dataset comprises 161 images (101 AMD and 60 normal) acquired using a Zeiss FF450+ fundus camera at a 50° field with a resolution of 576x768 pixels. The STARE dataset comprises 97 images (59 AMD and 38 normal) taken using a TOPCON TRV-50 fundus camera at a 35° field and with a resolution of 605x700 pixels. These two datasets were merged into a single dataset comprising 258 images (160 AMD and 98 normal). An experienced, accredited grader at the Liverpool Ophthalmic Reading Center has reviewed all the AMD images and split them into three categories: early (14), intermediate (29) and advanced AMD (117) according to the AMD severity scale set out by the AREDS.[4] More specifically, Early AMD (AREDS category 2) is characterized by many small drusen or a few intermediate-sized (63-124 um) drusen or retinal pigmentary abnormalities. Intermediate AMD (AREDS category 3) is characterized by at least one large (>125 um) drusen, numerous medium size drusen, or geographic atrophy that does not extend to the centre of the macula. Advanced AMD (AREDS category 4) can be either non-neovascular or neovascular. Advanced AMD is characterized by drusen and geographic atrophy extending to the centre of the macula.

### b) Image Mining Framework

The proposed framework comprises five stages: *Pre-processing*, *Image decomposition and graph representation*, *Weighted frequent sub-graph mining*, *Feature selection*, and *Classification*.

#### Pre-processing

The objective of the pre-processing stage was to enhance the effectiveness of the classification system by first enhancing the images. The following steps were applied:

i). A "mask image" $I_{background}$ was first defined as proposed in [17] by applying intensity thresholding and morphological operators to the original image $I$ (Fig. 1A): pixels within the circular fundus region of interest were marked as "1" while the rest as "0", as shown in Fig. 1B.

ii). A new image, $I_{color}$ (Fig. 1C), was generated after color normalization of the original image $I$ by using a histogram specification approach.[18]

iii). A common approach proposed by Foracchia et al [19] was then applied to $I_{color}$ to eliminate the illumination variation, as a result $I_{illumination}$ was generated, see Fig. 1D.

iv). A new image, $I_{processed}$ (Fig. 1E), was generated after applying a contrast enhancement technique called Contrast Limited Adaptive Histogram Equalization (CLAHE) [20] to $I_{illumination}$ This was adopted because of its demonstrated superiority over other comparable techniques.[21]

v). Blood vessels in the image $I_{processed}$ were detected by an approach that used wavelet features and a supervised classification technique.[22] The vessel pixels in $I_{vessel}$ (Fig. 1F) and those pixels marked with a "0" (black) in $I_{background}$ (Fig. 1B) were not considered in the subsequent analysis. In this work localization and removal of the optic disc was deliberately omitted as it was observed from our previous experience that this process does not show benefit in terms of classification performance.[23]

## Image Partition / Decomposition

One challenge of image mining is how to represent an image so as to maintain its structural information. Hierarchical trees are often used to represent images due to their ability to focus on the “interesting” parts of the input data, thus permitting an efficient representation of the problem and consequently improving the execution time.[24] Therefore, in this work we used a quadtree representation, the most common hierarchical data structure used in relation to image decomposition. The decomposition commenced by splitting an image into four equal sized quadrants, with the root of the quad-tree representing the entire image. The splitting process continued by further decomposing each quadrant to generate further sub-quadrants, and terminated when a certain level of granularity (or a desired maximum level of decomposition $D_{\max}$ ) was reached or all sub-quadrants were homogeneous. A quadrant is homogeneous if it contains similar pixels values. In this study homogeneity was defined in terms of the similarity between the average intensity value of a quadrant and those of its sub-quadrants. If the difference of average intensities between a quadrant and any of its sub-quadrants divided by its average intensity is less than a predefined threshold, the quadrant is considered homogeneous. A threshold value of 10% was empirically chosen as the default setting in this study. Fig. 2 illustrates the decomposition process of a retinal image.

Throughout the decomposition process the tree data structure was continuously appended to (it is constructed dynamically). Each identified sub-region was represented as a “node” in the tree data structure, whilst the relationship between each sub-region and its parent was represented by the edges. The RGB (Red, Green and Blue) color model was used to extract pixel intensity values, hence three trees were generated initially (one for each channel) and merged on completion.

## Weighted Frequent Sub-graph Mining

On completion of image decomposition the input image set was represented as a collection of trees, see Fig. 3. Each tree was defined as follows: $T = (V, E, L_V, L_E, u)$ where $V$ and $E$ are sets of vertices and edges respectively, $L_V$ and $L_E$ were sets of labels for vertices and edges respectively, while $u$ defined a label mapping function. To extract frequent sub-trees (image features) for classification, a weighted frequent sub-graph (WFSG) mining algorithm was used.[25] Further details of WFSG were presented in Appendix A.

The number of features discovered by the WFSG mining algorithm was determined by two thresholds, $\sigma$ and $\lambda$; where $\sigma$ denotes the minimum node support threshold while $\lambda$ denotes the minimum edge weight threshold. Relatively low $\sigma$ and $\lambda$ values are required in order to extract a sufficient number of features. However, setting threshold values too low may result in large numbers of features, of which many may be redundant and/or ineffective in terms of the desired classification task, as well as adding to the computational cost. Thus, a feature selection process was applied to the discovered features.

## Feature Selection / Reduction

Feature selection is often a desirable process in classification applications as this will serve to improve both the computational efficiency and the classification performance by reducing the data dimensions to only the most appropriate features. For this study, a feature ranking mechanism was employed that used linear Support Vector Machine (SVM) weights to rank features.[26] To generate the weights (to be used for the ranking) the L2-regularized SVM with the L2-loss function (provided in the LIBLINEAR library [27] which can be downloaded in [28]) was employed to rank the set of identified features generated from the previous stage. The resulting list of features was sorted in descending order according to their individual weights (discriminative power). This process allowed us to select the top $K$ features for the subsequent classification task, consequently the size of the feature space ($h$) was reduced by a factor of $h-K$. Again $K$ is a free parameter and its value was tuned for the best classification performance.

## Classifier training and classification

Two different classification techniques were used, Naïve Bayes [29,30] and SVM.[28] Naïve Bayes was selected because: (i) it has been shown to work well and is comparable to other techniques,[29] and (ii) it does not require user defined parameters. SVM was selected because it is recognized as one of the most effective classification methods in machine learning. For the SVM, the LibSVM [28] library was used. A *C*-Support Vector Classification (SVC) formulation of SVM, with a radial basis function (RBF) kernel $k(\mathbf{x}_i - \mathbf{x}_j) = \exp(-\gamma \| \mathbf{x}_i - \mathbf{x}_j \|^2)$ , was employed to generate the SVM classifier. The optimal parameters, such as the soft margin $C$ for C-SVC and the $\gamma$ parameter of the RBF kernel, were determined using the associated grid search strategy.[28]

## Evaluation

The proposed system was evaluated in order to investigate its performance by varying four parameter values: (i) Depth of decomposition ( $D_{\max}$ ), (ii) Minimum node support threshold ( $\sigma$ ), (iii) Minimum edge weight threshold ( $\lambda$ ), and (iv) Number of features selected ( $K$ ). All our experiments were conducted using Ten-fold Cross Validation (TCV). On each TCV iteration one tenth of the data was used as the test set while the remainder was used as the training set. Comparisons were also made with related work reported in the literature. The authors have only been able to identified four instances of previous work on retinal image AMD classification by other research groups: (i) Chaum et al,[11] (ii) Barriga et al,[12] (iii) Brandon and Hoover,[13] and (iv) Agurto et al.[14]

### Metrics

Three commonly used metrics were used to evaluate performance: sensitivity, specificity and accuracy**, and their corresponding 95% confidence intervals (CIs) were also calculated according to the Wilson score method.[31]** Sensitivity (resp. specificity) is a measure of the effectiveness in identifying positive (resp. negative) cases, while accuracy is a metric to indicate the overall classification performance. These metrics are defined as follows:

$$\textit{sensitivity}\ (\mathrm{Se}) = \frac{\text{the number of positive cases (AMD) correctly classified as positive}}{\text{the total number of positive cases}}$$

$$\textit{specificity}\ (\mathrm{Sp}) = \frac{\text{the number of negative cases (normal) correctly classified as negative}}{\text{the total number of negative cases}}$$

$$\textit{accuracy}\ (\mathrm{Acc}) = \frac{\text{the number of cases correctly classified}}{\text{the total number of cases}}$$

## Results

For the experiments on the effect of combinations of different parameter values (e.g. $D_{\max}$ , $\sigma$ , $\lambda$ and $K$ ), our results are shown in Table 1- 3 for $D_{\max}$ values of 5, 6 and 7 respectively. For each $D_{\max}$ , a range of $\sigma$ values from 10 to 90% was used (incremented in steps of 10), while a range of $\lambda$ values from 20% to 80% (incremented in steps of 20) was used. In Table 1-3 only results corresponding to $\sigma$ values from 10 to 50% are shown. Table 1 – 3 show that the SVM classifier produced better results than the Naïve Bayes one with respect to all three $D_{\max}$ values. For $D_{\max} = 5$, the best accuracy using the SVM classifier was 89.3% (sensitivity 92.8%; specificity 83.5%) while for the Naïve Bayes it was 76.1% (sensitivity 80.7%; specificity 68.1%). Note that as $\sigma$ and $\lambda$ were increased the number of features decreased, and consequently the accuracy reduced for both classifiers. The same trends may be observed for $D_{\max} = 6$ and $7$ as well.

**Overall, for the Naïve Bayes classifiers a best accuracy of 79.0% was achieved with $D_{\max} = 6$ , $\sigma = 10\%$ and $\lambda = 40\%$ .** For the SVM classifiers, a best accuracy of 99.6% was observed with settings for $D_{\max} = 6$ and $7$. These all occurred when $\sigma = 10\%$ or $20\%$ while $\lambda$ varied from between 20% to 60%. **The associated sensitivity value is 99.4% (95% CI, 96.6% to 99.9%) and the specificity value is 100% (95% CI, 96.2% to 100%). A sub-analysis has showed that with the SVM approach the sensitivity in detection of early, intermediate and advanced AMD is 100% (95% CI, 78.5% to 100%), 96.6% (95% CI, 82.8% to 99.4%) and 100% (95% CI, 96.8% to 100%) respectively.**

## Discussion

This study is a proof of concept that demonstrates the feasibility of image-mining based classification for automated AMD disease / no disease grading. We have employed two classifier generation techniques, SVM and Naïve Bayes. Our experiments, using two public retinal image databases, produced highly accurate results using SVM classification, with a best accuracy of 99.6% (sensitivity: 99.4%; specificity: 100%). Our SVM approach has also showed promising results in detection of early, intermediate and advanced AMD. The only misclassification of an intermediate AMD image is due to its very poor quality where almost half of the image is in black. This implies that the use of our technique would not miss any patients who need urgent care.

**Our comparative study demonstrated clearly that the proposed framework outperforms the previous work.[11,12,13,14] Our results, and those previously reported in the literature, are presented in Table 4. In our comparison both the SVM and the Naïve Bayes classifiers were tested with $\sigma$ = 10%, $\lambda$ = 20% and $D_{\max} = 7$. The SVM approach yielded better results than the Naïve Bayes classifier and the previous approaches in terms of sensitivity, specificity and accuracy. Brandon and Hoover used the STARE dataset,[13] while the others used data sets that are not publicly available. Table 4 features some missing values because these were not reported in the literature and could not be derived by the authors. The results recorded by Barriga et al.[12] included only sensitivity (75%) and specificity (50%). On the other hand, the method of Chaum et al. was applied in a multi-class setting and hence only accuracy (88%) was reported.[11] In their evaluation, 12 AMD images were classified as "unknown" and excluded from the accuracy calculation. If this number was included as miss-classifications, the accuracy would drop to 75%. Brandon and Hoover only reported the accuracy (90%) and specificity (89%),[13] however, we were able to calculate the sensitivity value (90%). Their evaluation was applied not only to AMD screening (AMD vs. non-AMD) but also to the grade (severity) of the detected AMD. To**

**obtain an overall sensitivity value we summed the total number of AMD images (irrespective of their AMD grades) and counted the number of these images that were correctly classified. The most recent work by Agurto et al. reported detection of AMD with sensitivity (specificity) of 94% (50%) and 90% (50%) for two databases, respectively.[14] The accuracy results can be derived from their reported results of sensitivity and specificity and they were lower than 80%. In contrast to these approaches** our new system can achieve a sensitivity similar to the others but a substantially higher specificity. In clinical practice this improvement would reduce many unnecessary referrals due to false alarms. The evidence from automated disease/no-disease grading of Diabetic Retinopathy (DR) research has shown that the introduction of such systems, even with specificity as low as about 50%, still can lead to cost-effectiveness and reduced overall workload.[32] Our technique can provide comparable sensitivity and much higher specificity; as such our technique represents a considerable advance. **Moreover, our technique has the potential to provide patients with the results at the point of service. It will be able to work without intra- and inter-observer grading variability, tiredness of human graders, and the need for regular training and certification that are required with respect to the human graders employed in manual grading programs.**

All the studies, including ours, only use a relatively small number of images (<500 images) that **may not be well representative of the population to be screened** to address such a challenging problem due to the nature of medical imaging research. **As presented above, the widths of the 95% CI in the detection of early and intermediate AMD are larger than 10% which implies that a larger sample size is needed in order to narrow this down. This limitation suggests that the proposed technique should be further validated by considering large-scale studies before it can be introduced into clinical practice. We envisage that the sample size of such studies has to be carefully considered in order to establish the scalability and generality of the proposed technique and to precisely estimate the level of expected sensitivity. According to**

**Buderer,[33] the sample size is dependent on disease prevalence, expected sensitivity and specificity, and the corresponding width of the CI. For instance, if the prevalence of any AMD is about 10% in the screened population, and the expected sensitivity and specificity is to be ≥ 90% and ≥ 95% respectively. A minimum sample size of about 350 is required to confirm sensitivity larger than 90% when the width of the 95% CI is 5%. If the prevalence is 1%, and all the other requirements are the same as above, then the sample size will becomes about 3500. The latter case may reflect the need for a substantially large sample size for the validation of the program with respect to the detection of subgroups of AMD (e.g. advanced AMD). However, the above results are not necessarily conclusive, the actual required sample size in any future study will have to be determined by the specific application and its performance requirements (i.e., the sample size needed by a validation study for screening people aged over 65 years would be smaller than that for screening people aged over 50 years for the same level of performance). Another important factor with respect to any future validation study is how to establish the reference standard for grading that is crucial for training and validation of the automated grading system. To this end we believe that the proposed strategies developed for automated DR grading can be readily adapted.** In addition some additional components require further development for the current system to become a standalone automated grading systems. For example, image quality is an important factor with respect to the detection of lesions and subsequent diagnosis; an automated image quality assessment mechanism is therefore desirable. It is also expected that further development will make it possible to automatically assess disease severity scales.

In our research we also noted that, due to the nature of image mining techniques, the image representation used for the classification is no longer interpretable by human observers. It would be desirable, with respect to its acceptance and practical use, to allow the model to also be clinically interpretable. This may provide a better way for clinicians to interpret fundus photography and allow clinicians to focus on spatial patterns. This has become a research topic in itself. On the other hand, our argument is that the most important feature of a prediction system like ours is its ability to make correct predictions. No system will be clinically useful if it is transparent to understanding but performs badly. As described above our technique involves graph-mining and feature selection processes in the classifier training phase which may require substantial computing and storage resources when dealing with large datasets, this may be a potential weakness of our technique. However, it is envisaged that with current technical advances in computing this would not be a key issue with respect to scalability and performance.

Over past decades some newly emerging imaging techniques, such as fundus autofluorescence (FAF) and optical coherence tomography (OCT), have become available and showed potential for AMD screening. FAF imaging is a noninvasive imaging technique that allows assessment of the integrity of the retinal pigment epithelium cells.[34] Although it has demonstrated potential for the analysis of distribution patterns of drusen and quantification of geographic atrophy, and as a prognostic tool to predict development of AMD, extensive work is warranted to investigate its use for AMD screening. The advent of OCT has revolutionized diagnosis and treatment of retinal disease.[35] OCT is a noncontact, noninvasive, high-resolution imaging technique that allows cross-sectional images of the retina to be obtained in almost real time and more importantly allows further quantitative analysis of features of the retina.[36,37,38,39] It has been extensively used in the guidelines for follow-up and retreatment of patients with AMD.[40] It appears to be a very promising technique to support AMD screening. However, OCT imaging may not show hemorrhaging, and may miss some abnormalities due to the large gap (or undetected region) between adjacent B-scans. Cost effectiveness may also be an issue, as OCT devices are much more expensive than standard color fundus cameras. It should also be noted that the current AMD severity scale was developed and validated as part of a large scale study (AREDS) using color fundus photograph,[4] effort would be needed to investigate the mapping of this scale between the new emerging techniques and color fundus photography. **Therefore it is believed that both FAF and OCT will help further establish the clinical validation for AMD screening, but may not be feasible for AMD screening alone. A combination of different diagnostic imaging techniques such as OCT, FAF and color fundus may be an optimal solution with respect to future automated screening purposes.** Whatever the case our technique is a generic approach which can be extended to any of the above.

Although the proposed approach has confirmed the technical feasibility of an automated AMD grading system, to the best of our knowledge no such programs exist currently. As suggested by Karnon et al,[41] the major concern for AMD screening is the significant uncertainty about their cost effectiveness, although annual screening from age 60 years onwards appeared to be beneficial at the time of their study. We noted that this conclusion was made without considering the potential benefit of using automated grading systems as at that time automated grading was merely at early stage proof-of concept and no sufficient detail was available for evaluation. Lessons and experience accumulated in DR screening and in particular recent development of automated grading could provide more insight into best practice. As an example, although specificity as high as ours is not achieved in automated disease / no-disease DR grading, models that combines automated and manual grading have demonstrated cost-effectiveness and a reduced overall workload. [32] If there had been an automated AMD grading system with similar performance, the cost-effectiveness of AMD screening would be much improved compared to that observed in 2008.[41] Together with other advances in therapeutic treatment, there would be more weight to support AMD screening. Certainly introduction of a new screening system is a rather complicated process, not only because of the need to satisfy well-established screening criteria,[42] but also with respect to various political, economic and ethical hurdles.[43] An alternative use for an automated AMD screening system, of the form proposed here, is as a “second opinion” generator.[32] We envisage that our approach has great potential for the above activities and lays a foundation for future research and the implementation of an automated screening system**. In addition, the principle and methodology that we propose here may also be adapted for accurate analysis of disease progression, which is important for monitoring disease development and timely treatment**.

In conclusion, this study has demonstrated a powerful image-mining based technique for automated AMD grading whose superior performance warrants further development in order to translate this technique into clinical practice as an automated AMD grading tool.

## Appendix A Weighted Frequent Sub-graph Mining (WFSM)

Let $GD = \{g_1, g_1, \cdots, g_n\}$ be $n$ graphs representing an image dataset of *n* images (one graph for each image). In the context of the WFSM algorithm used [25], each node has a weight defined by the average color intensity value of the region (quadrant) represented by the node. A weight is also assigned to each edge, edge weights are defined as the difference between the average intensity of the child node and that of its parent. The WFSM algorithm extracts frequent sub-tree (image features) for classification purpose. More specifically, a sub-graph, $sg$ , is considered frequent (important) if it satisfies the following two conditions: (i) $N_{wr}(sg) \times \sup(sg) \geq \sigma$, and (ii) $E_{wr}(sg) \geq \lambda$, where $N_{wr}$ denotes the node weighting, $\sup(sg)$ denotes the support (i.e. frequency) of $sg$ , and $\sigma$ denotes the minimum node support threshold; $E_{wr}$ denotes the edge weighting, and $\lambda$ denotes the minimum edge weight threshold. The weightings $N_{wr}$ and $E_{wr}$ are computed as follows:

$$N_{wr}(sg) = \frac{\sum_{k=1}^{|\Delta sg|} w_{node}(g_k)}{\sum_{k=1}^{|G|} w_{node}(g_k)},$$

$$E_{wr}(sg) = \frac{\sum_{k=1}^{|\Delta sg|} w_{edge}(g_k)}{\sum_{k=1}^{|G|} w_{edge}(g_k)},$$

where $|\Delta sg|$ denotes the number of graphs in which $sg$ occurs in the graph dataset $G$, $|G|$ is the number of graphs in the graph dataset $G$, while $w_{node}(g)$ and $w_{edge}(g)$ are the average weights of nodes and edges in $g$ respectively. For full details interested readers should refer to Jiang and Coenen. [25]

The output of the WFSM algorithm is then a set of weighted frequent sub-trees (WFSTs). In order to allow the application of existing classification algorithms to the identified WFSTs, feature vectors were built from them. The identified set of WFSTs was first used to define a feature space. Each image was then represented by a single feature vector comprised of some subset of the WFSTs in the feature space. In this manner the input set can be translated into a two dimensional binary-valued table of size $n \times k$, of which the number of rows, $n$, represents the number of images and $k$ the number of identified WFSTs. An additional class label column will be added for the training data.

## Acknowledgement

Mohd H.A. Hijazi thanks the Ministry of Higher Education Malaysia for their financial support. Dr Y Zheng thanks the Foundation for the Prevention of Blindness for their support.

## References

1. Pascolini D, Mariotti S, Pokharel G, et al. 2002 Global update of available data on visual impairment: A compilation of population-based prevalence studies. *Ophthalmic Epidemiology* 2004;11:67-115.
2. Rein D, Wittenborn J, Zhang X, et al. Forecasting age-related macular degeneration through the year 2050. *Archives of Ophthalmology* 2009;127:533-540.
3. Lamoureux EL, Mitchell P, Rees G, et al. Impact of early and late age-related macular degeneration on vision-specific functioning. *British Journal of Ophthalmology* 2011;2011:666-670.
4. Age-Related Eye Disease Study Research Group. The age-related eye disease study system for classifying age-related macular degeneration from stereoscopic color fundus photographs: AREDS report No. 6. *American Journal of Ophthalmology* 2001;132:668–681.
5. Davis MD, Gangnon RE, Lee LY, et al. The age-related eye disease study severity scale for age-related macular degeneration - AREDS report no. 17. *Archives of Ophthalmology* 2005;123:1484-1498.
6. Sbeh ZB, Cohen LD, Mimoun G, Coscas G. A new approach of geodesic reconstruction for drusen segmentation in eye fundus images. *IEEE Transactions on Medical Imaging* 2001;20:1321-1333.
7. Rapantzikos K, Zervakis M, Balas K. Detection and segmentation of drusen deposits on human retina: Potential in the diagnosis of age-related macular degeneration. *Medical Image Analysis* 2003;7:95-108.
8. Kose C, Sevik U, Gencalioglu O. Automatic segmentation of age-related macular degeneration in retinal fundus images. *Computers in Biology and Medicine* 2008;38:611-619.
9. Kose C, Sevik U, Gencalioglu O. A statistical segmentation method for measuring age-related macular degeneration in retinal fundus images. *Journal of Medical Systems* 2008;34:1-13.
10. Barriga ES, Murray V, Agurto C, et al. Multi-scale AM-FM for lesion phenotyping on age-related macular degeneration. *IEEE International Symposium on Computer-Based Medical Systems*; 2009:1-5.
11. Chaum E, Karnowski TP, Govindasamy VP, Abdelrahman M, Tobin KW. Automated diagnosis of retinopathy by content-based image retrieval. *Retina* 2008;28:1463-1477.
12. Barriga ES, Murray V, Agurto C, et al. Automatic computer-based grading for age-related maculopathy. *Investigative Ophthalmology and Visual Science* 2010;51:E-Abstract 1793.
13. Brandon L, Hoover A. Drusen detection in a retinal image using multi-level analysis. *Proceedings of Medical Image Computing and Computer-Assisted Intervention*: Springer-Verlag; 2003:618-625.
14. Agurto C, Barriga ES, Murray V, et al. Automatic detection of diabetic retinopathy and age-related macular degeneration in digital fundus images. *Investigative Ophthalmology & Visual Science* 2011;52:5862-5871.
15. Elsayed A, Coenen F, Jiang C, Garcia-Finana M, Sluming V. Corpus callosum MR image classification. *Knowledge Based Systems* 2010;23:330-336.
16. Hijazi MHA, Coenen F, Zheng Y. Data mining techniques for the screening of age-related macular degeneration. *Knowledge-Based Systems* 2011;29:83-92.
17. Haar Ft. Automatic localization of the optic disc in digital colour images of the human retina. Utrecht University; 2005.
18. Gonzalez RC, Woods RE. *Digital image processing*: Pearson Prentice Hall; 2008.
19. Foracchia M, Grisan E, Ruggeri A. Luminosity and contrast normalization in retinal images. *Medical Image Analysis* 2005;9:179-190.
20. Zuiderveld K. Contrast limited adaptive histogram equalization. Academic Press Professional, Inc.; 1994:474-485.

21. Hijazi MHA, Coenen F, Zheng Y. Image classification using histograms and time series analysis: A study of age-related macular degeneration screening in retina image data. *Proceedings of 10th Industrial Conference on Data Mining*; 2010:197-209.

22. Soares JVB, Leandro JJG, Jr. RMC, Jelinek HF, Cree MJ. Retinal vessel segmentation using the 2-D gabor wavelet and supervised classification. *IEEE Transactions on Medical Imaging* 2006;25:1214-1222.

23. Hijazi MHA, Coenen F, Zheng Y. Retinal image classification using a histogram based approach. *Proceedings of International Joint Conference on Neural Network 2010 (World Congress on Computational Intelligence 2010)*; 2010:3501-3507.

24. Samet H. The quadtree and related hierarchical data structures. *ACM Computing Surveys* 1984;16:187-260.

25. Jiang C, Coenen F. Graph-based image classification by weighting scheme. *AI2008*; 2008:63-76.

26. Chang Y-W, Lin C-J. Feature ranking using linear SVM. *WCCI2008*; 2008:53-64.

27. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 2008;9:1871-1874.

28. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. http://wwwcsientuedutw/~cjlin/libsvm 2001.

29. Domingos P, Pazzani M. On the optimality of the simple bayesian classifier under zero-one loss. *Journal of Machine Learning* 1997;29:103-130.

30. Witten I, Frank EH. *Data mining: practical machine learning tools and techniques*: Morgan Kaufmann; 2005.

31. Newcombe RG. Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine* 1998;17:857-872.

32. Fleming AD, Philip S, Goatman KA, Prescott GJ, Sharp PF, Olson JA. The evidence for automated grading in diabetic retinopathy screening. *Current Diabetes Reviews* 2011;7:246-252.

33. Buderer NMF. Statistical methodology: I. Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. *Academic Emergency Medicine* 1996;3:895-900.

34. Frank G. Holz, Schmitz-Valckenberg S, Spaide RF, Bird AC (eds). *Atlas of fundus autofluorescence imaging*. 1st ed. Berlin: Springer; 2007.

35. Huang D, Swanson EA, Lin CP, et al. Optical coherence tomography. *Science* 1991;254:1178-1181.

36. Chiu SJ, Li XT, Nicholas P, Toth CA, Izatt JA, Farsiu S. Automatic segmentation of seven retinal layers in SDOCT images congruent with expert manual segmentation. *Optics Express* 2010;18:19413-19428.

37. Jain N, Farsiu S, Khanifar AA, et al. Quantitative Comparison of Drusen Segmented on SD-OCT versus Drusen Delineated on Color Fundus Photographs. *Investigative Ophthalmology & Visual Science* 2010;51:4875-4883.

38. Gregori G, Wang F, Rosenfeld PJ, et al. Spectral Domain Optical Coherence Tomography Imaging of Drusen in Nonexudative Age-Related Macular Degeneration. *Ophthalmology* 2011;118:1373-1379.

39. Chiu SJ, Izatt JA, O'Connell RV, Winter KP, Toth CA, Farsiu S. Validated automatic segmentation of AMD pathology including drusen and geographic atrophy in SD-OCT images. *Investigative ophthalmology & visual science* 2012;53:53-61.

40. Chakravarthy U, Harding SP, Rogers CA, et al. Ranibizumab versus bevacizumab to treat neovascular age-related macular degeneration: one-year findings from the ivan randomized trial. *Ophthalmology* 2012;Epub ahead of print.

41. Karnon J, Czoski-Murray C, Smith K, et al. A preliminary model-based assessment of the cost–utility of a screening programme for early age-related macular degeneration. *Health technology assessment* 2008;12.

42. Wilson J, Jungner G. Principles and practice of screening for disease. *World Health Organization* 1968;22.

43. Abramoff MD, Niemeijer M, Russell SR. Automated detection of diabetic retinopathy: barriers to translation into clinical practice. *Expert Review of Medical Devices* 2010;7:287-296.

## Figure Legends

Figure 1. Illustration of pre-processing steps. A) Original image; B) Image mask; C) Image after color normalization; D) Image after illumination normalization; E) Image after contrast enhancement; F) The identified blood vessels.

Figure 2. Illustration of image decomposition using the quadtree technique.

Figure 3. Illustration of the quadtree data structure.

## Figures

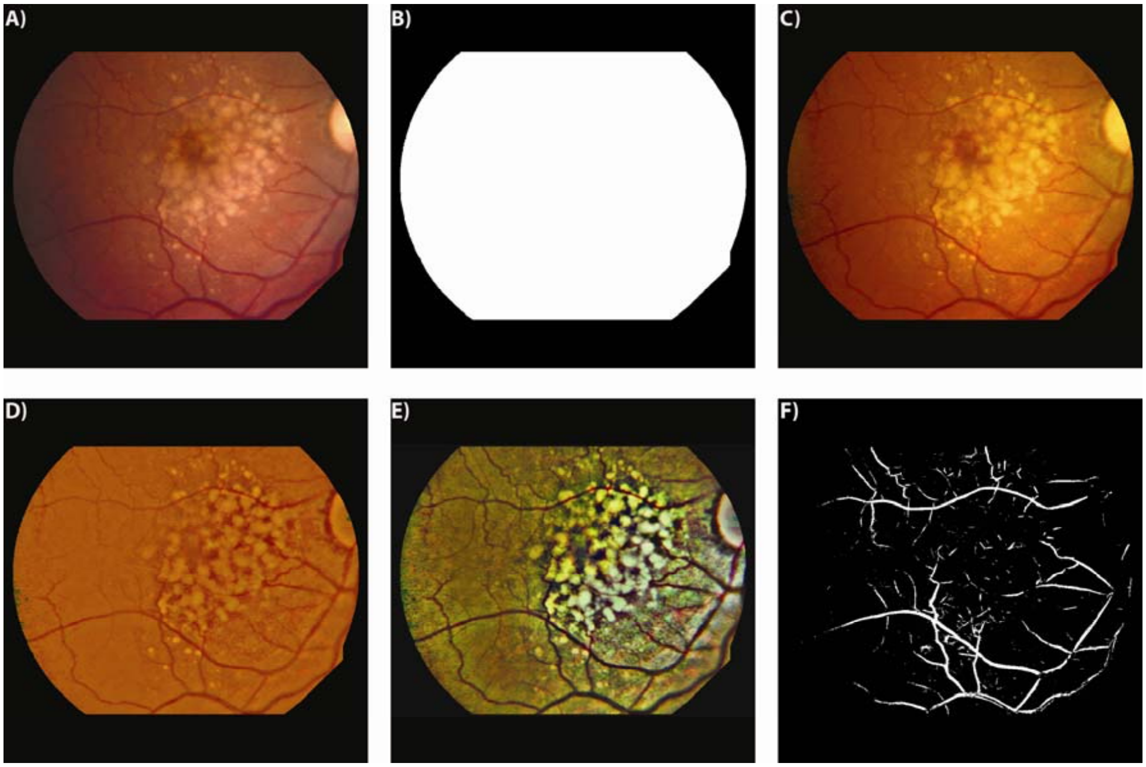

Figure 1. Illustration of pre-processing steps. A) Original image; B) Image mask; C) Image after color normalization; D) Image after illumination normalization; E) Image after contrast enhancement; F) The identified blood vessels.

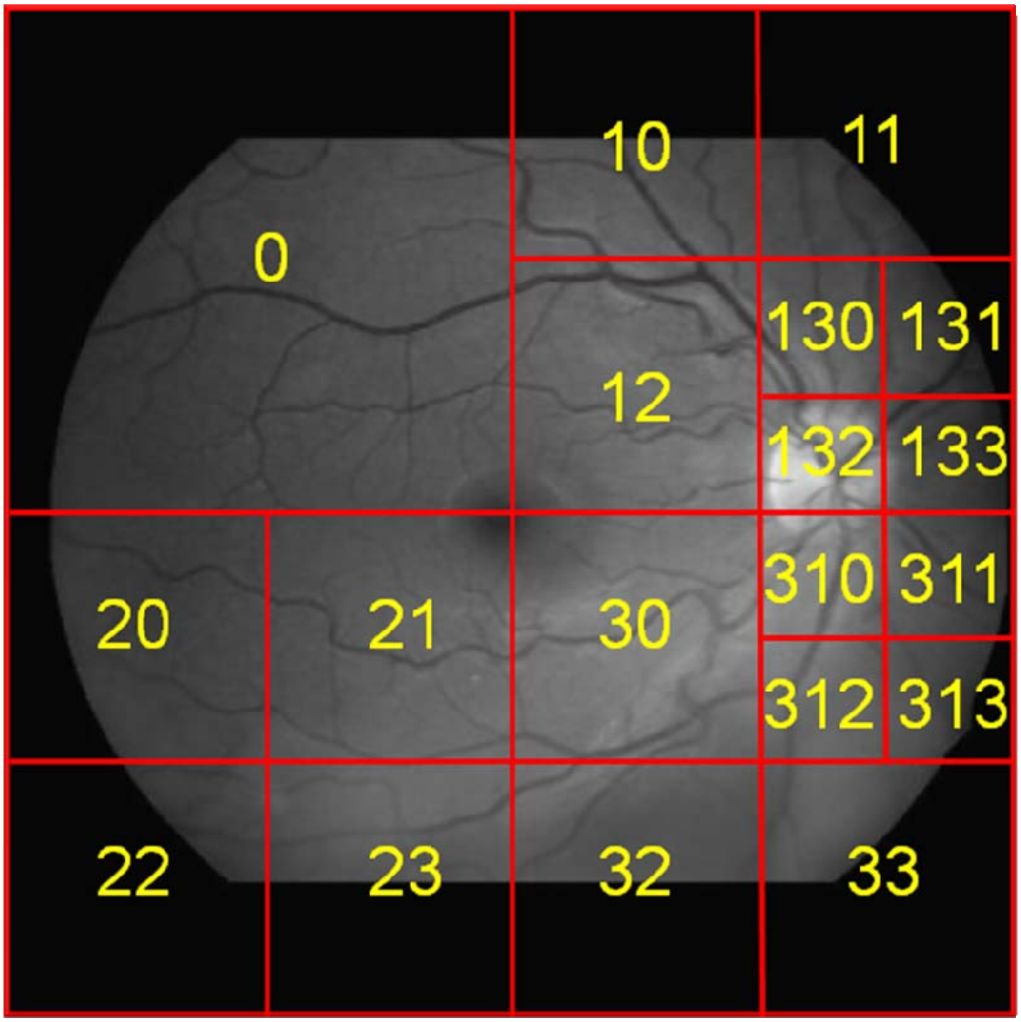

Figure 2. Illustration of image decomposition using the quadtree technique.

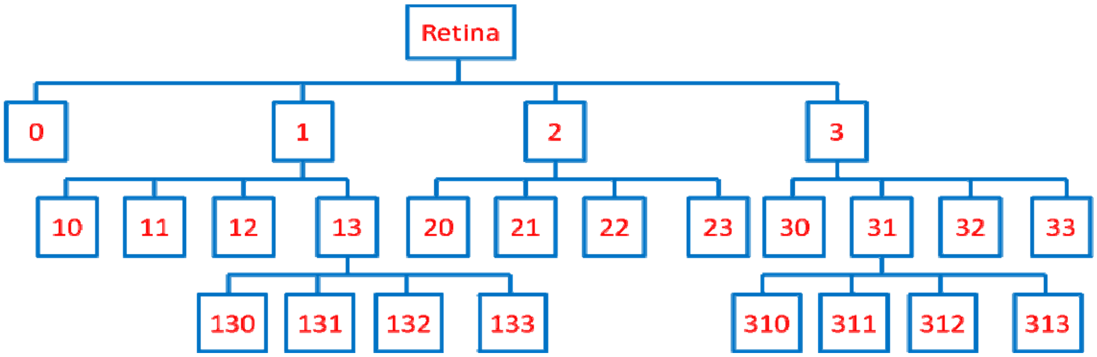

Figure 3. Illustration of the quadtree data structure.

## Table Legend

Table 1. Classification results with $D_{max}$=5.

Table 2. Classification results with $D_{max}$=6.

Table 3. Classification results with $D_{max}$=7.

Table 4. Comparison of results of our proposed approaches with those from previous work.

## Tables

Table 1. Classification results with $D_{max}$=5.

| | | SVM | | | | Naïve Bayes | | | |
|---|---|---|---|---|---|---|---|---|---|
| **minFreq σ (%)** | **minRatio λ (%)** | **Feature size *K*** | **Se (%)** | **Sp (%)** | **Acc (%)** | **Feature size K** | **Se (%)** | **Sp (%)** | **Acc (%)** |
| 10 | 20 | 1000 | 97.0 | 33.7 | 73.4 | 50 | 80.7 | 68.1 | 76.1 |
| | 40 | 200 | 92.8 | 83.5 | 89.3 | 50 | 78.3 | 72.1 | 76.1 |
| | 60 | 200 | 95.8 | 25.6 | 69.7 | 50 | 75.2 | 60.1 | 69.6 |
| | 80 | 50 | 89.9 | 40.7 | 71.6 | 50 | 82.5 | 43.6 | 68.1 |
| 20 | 20 | 200 | 92.8 | 50.8 | 77.2 | 50 | 75.3 | 66.1 | 71.9 |
| | 40 | 200 | 92.8 | 50.8 | 77.2 | 50 | 75.3 | 66.1 | 71.9 |
| | 60 | 200 | 95.8 | 24.6 | 69.3 | 50 | 74.7 | 60.1 | 69.3 |
| | 80 | 50 | 89.9 | 40.7 | 71.6 | 50 | 82.5 | 43.6 | 68.1 |
| 30 | 20 | 200 | 95.8 | 28.6 | 70.8 | 100 | 72.2 | 62.2 | 68.5 |
| | 40 | 200 | 95.8 | 28.6 | 70.8 | 100 | 72.2 | 62.2 | 68.5 |
| | 60 | 200 | 95.8 | 25.6 | 69.7 | 50 | 75.2 | 60.1 | 69.6 |
| | 80 | 50 | 89.9 | 40.7 | 71.6 | 50 | 82.5 | 43.6 | 68.1 |
| 40 | 20 | 50 | 87.8 | 38.8 | 69.5 | 50 | 76.5 | 51.9 | 67.4 |
| | 40 | 50 | 87.8 | 38.8 | 69.5 | 50 | 76.5 | 51.9 | 67.4 |
| | 60 | 50 | 87.8 | 38.8 | 69.5 | 50 | 76.5 | 51.9 | 67.4 |
| | 80 | 50 | 89.9 | 40.7 | 71.6 | 50 | 82.5 | 43.6 | 68.1 |
| 50 | 20 | 100 | 94.0 | 29.7 | 70.0 | 50 | 79.4 | 42.7 | 65.8 |
| | 40 | 100 | 94.0 | 29.7 | 70.0 | 50 | 79.4 | 42.7 | 65.8 |
| | 60 | 100 | 94.0 | 29.7 | 70.0 | 50 | 79.4 | 42.7 | 65.8 |
| | 80 | 50 | 89.9 | 40.7 | 71.6 | 50 | 82.5 | 43.6 | 68.1 |

Table 2. Classification results with $D_{max}$=6.

| | | **SVM** | | | | **Naïve Bayes** | | | |
|---|---|---|---|---|---|---|---|---|---|
| **minFreq σ (%)** | **minRatio λ (%)** | **Feature size *K*** | **Se (%)** | **Sp (%)** | **Acc (%)** | **Feature size *K*** | **Se (%)** | **Sp (%)** | **Acc (%)** |
| 10 | 20 | 1000 | 99.4 | 100.0 | 99.6 | 50 | 80.2 | 73.6 | 77.6 |
| | 40 | 1000 | 98.3 | 96.0 | 97.4 | 50 | 80.1 | 77.2 | 79.0 |
| | 60 | 1000 | 93.4 | 42.7 | 74.6 | 50 | 76.5 | 67.0 | 73.0 |
| | 80 | 200 | 93.3 | 39.9 | 73.5 | 50 | 78.3 | 47.9 | 66.9 |
| 20 | 20 | 1000 | 99.4 | 100.0 | 99.6 | 50 | 79.5 | 72.5 | 76.8 |
| | 40 | 1000 | 99.4 | 100.0 | 99.6 | 50 | 79.5 | 72.5 | 76.8 |
| | 60 | 1000 | 93.4 | 42.7 | 74.6 | 50 | 76.5 | 67.0 | 73.0 |
| | 80 | 200 | 93.3 | 39.9 | 73.5 | 50 | 78.3 | 47.9 | 66.9 |
| 30 | 20 | 1000 | 92.8 | 49.9 | 76.9 | 100 | 74.1 | 66.0 | 71.1 |
| | 40 | 1000 | 92.8 | 49.9 | 76.9 | 100 | 74.1 | 66.0 | 71.1 |
| | 60 | 1000 | 93.4 | 42.7 | 74.6 | 50 | 76.5 | 67.0 | 73.0 |
| | 80 | 200 | 93.3 | 39.9 | 73.5 | 50 | 78.3 | 47.9 | 66.9 |
| 40 | 20 | 400 | 95.3 | 55.2 | 80.3 | 50 | 75.5 | 57.7 | 68.9 |
| | 40 | 400 | 95.3 | 55.2 | 80.3 | 50 | 75.5 | 57.7 | 68.9 |
| | 60 | 400 | 95.3 | 55.2 | 80.3 | 50 | 75.5 | 57.7 | 68.9 |
| | 80 | 200 | 93.3 | 39.9 | 73.5 | 50 | 78.3 | 47.9 | 66.9 |
| 50 | 20 | 200 | 92.3 | 54.9 | 78.4 | 100 | 78.4 | 53.9 | 69.3 |
| | 40 | 200 | 92.3 | 54.9 | 78.4 | 100 | 78.4 | 53.9 | 69.3 |
| | 60 | 200 | 92.3 | 54.9 | 78.4 | 100 | 78.4 | 53.9 | 69.3 |
| | 80 | 200 | 93.3 | 39.9 | 73.5 | 50 | 78.3 | 47.9 | 66.9 |

Table 3. Classification results with $D_{max}$=7.

| **minFreq σ (%)** | **minRatio λ (%)** | **SVM** | | | | **Naïve Bayes** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Feature size *K*** | **Se (%)** | **Sp (%)** | **Acc (%)** | **Feature size *K*** | **Se (%)** | **Sp (%)** | **Acc (%)** |
| 10 | 20 | 4000 | 99.4 | 100.0 | 99.6 | 1000 | 79.5 | 77.5 | 78.7 |
| | 40 | 4000 | 99.4 | 100.0 | 99.6 | 1000 | 78.3 | 77.3 | 77.9 |
| | 60 | 1000 | 99.4 | 100.0 | 99.6 | 100 | 75.8 | 74.5 | 75.4 |
| | 80 | 1000 | 95.8 | 21.5 | 68.1 | 50 | 81.4 | 58.2 | 72.8 |
| 20 | 20 | 1000 | 99.4 | 100.0 | 99.6 | 1000 | 77.7 | 77.5 | 77.6 |
| | 40 | 1000 | 99.4 | 100.0 | 99.6 | 1000 | 77.7 | 77.5 | 77.6 |
| | 60 | 1000 | 99.4 | 100.0 | 99.6 | 100 | 75.8 | 74.5 | 75.4 |
| | 80 | 1000 | 95.8 | 21.5 | 68.1 | 50 | 81.4 | 58.2 | 72.8 |
| 30 | 20 | 4000 | 95.3 | 75.6 | 87.9 | 50 | 78.2 | 73.3 | 76.4 |
| | 40 | 4000 | 95.3 | 75.6 | 87.9 | 50 | 78.2 | 73.3 | 76.4 |
| | 60 | 1000 | 99.4 | 100.0 | 99.6 | 100 | 75.8 | 74.5 | 75.4 |
| | 80 | 1000 | 95.8 | 21.5 | 68.1 | 50 | 81.4 | 58.2 | 72.8 |
| 40 | 20 | 1000 | 97.6 | 97.8 | 97.7 | 50 | 75.9 | 70.2 | 73.8 |
| | 40 | 1000 | 97.6 | 97.8 | 97.7 | 50 | 75.9 | 70.2 | 73.8 |
| | 60 | 1000 | 97.6 | 97.8 | 97.7 | 50 | 75.9 | 70.2 | 73.8 |
| | 80 | 1000 | 95.8 | 21.5 | 68.1 | 50 | 81.4 | 58.2 | 72.8 |
| 50 | 20 | 1000 | 95.3 | 51.0 | 78.7 | 100 | 81.4 | 61.2 | 73.9 |
| | 40 | 1000 | 95.3 | 51.0 | 78.7 | 100 | 81.4 | 61.2 | 73.9 |
| | 60 | 1000 | 95.3 | 51.0 | 78.7 | 100 | 81.4 | 61.2 | 73.9 |
| | 80 | 1000 | 95.8 | 21.5 | 68.1 | 50 | 81.4 | 58.2 | 72.8 |

Table 4. Comparison of results of our proposed approaches with those from previous work.

| **Approach** | **Dataset size** | **Sensitivity** | **Specificity** | **Accuracy** |
|---|---|---|---|---|
| Brandon and Hoover[13] | 97 | 90 | 89 | 90 |
| Chaum et al.[11] | 395 | N/A | N/A | 88 |
| Barriga et al.[12] | 100 | 75 | 50 | N/A |
| Agurto et al.[14] | 392 (Rist Database) | 90 | 60 | 79 |
| | | 94 | 50 | 78 |
| | 395 (UTHSCSA Database) | 90 | 60 | 76 |
| | | 90 | 50 | 76 |
| Proposed Bayes approach | 258 | 79.5 | 77.5 | 78.7 |
| Proposed SVM approach | 258 | 99.4 | 100 | 99.6 |