Automated detection of age-related macular degeneration in color fundus photography

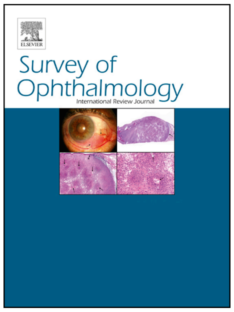
Citation for published version:
Pead, E, Megaw, R, Cameron, J, Fleming, A, Dhillon, B, Trucco, E & MacGillivray, T 2019, 'Automated detection of age-related macular degeneration in color fundus photography: a systematic review', *Survey of Ophthalmology*, vol. 64, no. 4, pp. 498-511. https://doi.org/10.1016/j.survophthal.2019.02.003, https://doi.org/10.1016/j.survophthal.2019.02.003

Digital Object Identifier (DOI):
10.1016/j.survophthal.2019.02.003
10.1016/j.survophthal.2019.02.003

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Survey of Ophthalmology



Automated detection of age-related macular degeneration in color fundus photography: A systematic review

E. Pead, BS, R. Megaw, MD, J. Cameron, PhD FRCOpth, A. Fleming, PhD, B. Dhillon, FRCsED, E. Trucco, PhD, T. MacGillivray, PhD

PII: S0039-6257(18)30207-8

DOI: https://doi.org/10.1016/j.survophthal.2019.02.003

Reference: SOP 6851

To appear in: *Survey of Ophthalmology*

Received Date: 21 August 2018

Revised Date: 31 January 2019

Accepted Date: 4 February 2019

Please cite this article as: Pead E, Megaw R, Cameron J, Fleming A, Dhillon B, Trucco E, MacGillivray T, Automated detection of age-related macular degeneration in color fundus photography: A systematic review, *Survey of Ophthalmology* (2019), doi: https://doi.org/10.1016/j.survophthal.2019.02.003.

Automated detection of age-related macular degeneration in color fundus photography: A systematic review

E. Pead, BS[1], R. Megaw MD[5], J. Cameron, PhD FRCOpth[4], A. Fleming, PhD[3], B. Dhillon, FRCsED [5], E. Trucco, PhD[2]*, T. MacGillivray, PhD[1]*

* These authors contributed equally

VAMPIRE Project, Centre for Clinical Brain Sciences, The University of Edinburgh, Edinburgh, Scotland[1], VAMPIRE Project, Computing (School of Science and Engineering, University of Dundee, UK[2] , Optos plc, Queensferry House, Carnegie Campus, Enterprise Way, Dunfermline[3], MRC Human Genetics Unit, The University of Edinburgh, Edinburgh, Scotland[4], Princess Alexandra Eye Pavilion, Chalmers Street, Edinburgh, Scotland[5],

Emma Pead's PhD studentship is jointly funded by the Scottish Imaging Network, a Platform for Scientific Excellence (SINAPSE) Collaboration, and Optos plc. Dr Alan Fleming is employed by Optos plc. Roly Megaw is funded by the Welcome Trust, The Academy of Medical Sciences and Fight for Sight.

The authors have no conflicts of interest

Corresponding author: E. Pead, The University of Edinburgh, Centre for Clinical Brain Sciences, Chancellor's Building, 49 Little France Crescent, Edinburgh, EH16 4SB

Automated detection of age-related macular degeneration in color fundus photography: A systematic review

Abstract

The rising prevalence of age-related eye diseases, particularly age-related macular degeneration (AMD), places an ever-increasing burden on healthcare providers. As new treatments emerge, it is necessary to develop methods for reliably assessing patients' disease status and stratifying risk of progression. The presence of drusen in the retina represents a key early feature where size, number and morphology are thought to correlate significantly with risk of progression to sight-threatening AMD. Manual labelling of drusen on color fundus photographs by a human is labor intensive and is where automatic computerised detection would appreciably aid patient care. We review and evaluate current artificial intelligence methods and developments for the automated detection of drusen in the context of AMD.

Keywords

Age-related macular degeneration, age-related disorders, artificial intelligence, machine learning, deep learning

1. Introduction

With longer life expectancy, age-related disorders are increasing the burden placed on healthcare providers. In particular, age-related macular degeneration (AMD) is one of the major causes of vision loss in the elderly [28]. AMD currently affects 6 million people in the UK alone [28] and was estimated to have cost the country's economy £155million in 2011 [49]. By 2040, the number of people affected globally by the disease is projected to be 288 million [58]

The earliest phase of AMD is typically observed as presence of (asymptomatic) macular drusen, often incidentally found on examination or fundus imaging. Drusen are small deposits of predominantly lipid, acellular debris that accumulate between the retinal pigment epithelium (RPE) and Bruch's membrane. Whilst the presence of small drusen is not itself diagnostic of AMD, as drusen frequently occur in normal aging, increasing number and size of drusen raise the risk of progression to visually symptomatic AMD. Later signs of AMD, such as pigmentary changes of the RPE that occur prior to the development of geographic atrophy (GA-so-called dry AMD) and exudative abnormalities (so-called wet AMD) enable more established gradings [5] [3] [33] and classification of AMD [2] [28] [32] [34].

Drusen appear as clusters of white or yellow spots in color fundus photographs and broadly exist as two main types, *hard* and *soft*. Hard drusen are round, small, discrete lesions with defined edges whereas soft drusen are less defined and often confluent. Drusen are rarely homogenous in their composition. Because of their yellow color and brightness on color fundus photographs, drusen are distinguishable by the human eye, but computer algorithms to automatically detect them need to be robust to the presence of other similarly bright appearing pathology such as hard exudates. Indistinct borders for drusen appearing in color fundus photographs are challenging for conventional image processing techniques such as edge detection and morphological filtering, and have been discussed in detail in an earlier review [15]. To the best of our knowledge, no reviews cover recent developments, involving the application of artificial intelligence (AI) and deep learning techniques.

AI is a long-standing field of computer science that aims to simulate human intelligence by perceiving its environment and taking appropriate action to achieve a set of goals, which is often one of decision making. Machine learning (ML) is an approach to AI partially inspired by how humans learn [37]. Learning is achieved through examples. If a child is presented with a new object, they will use features such as color, shape and texture so that when they observe the object again they will use what they have learned to identify or categorise it as something they have previously seen. Similarly, many ML classification algorithms use features from training examples to discover or confirm patterns that categorise subsets. When new, unseen data are presented the algorithm can classify which category they belong to (**Figure 1**). These features can be learned by either training from previous examples (i.e. supervised learning) or discovered by the algorithm (i.e. unsupervised learning).

[FIGURE 1]

Figure 1. Illustration of standard supervised machine Learning pipeline. **1)** Image pre-processing is applied to reduce noise and enhance image features. **2)** Features are extracted such as measures of entropy, energy, color and texture of image intensities, and spatial or geometric properties. **3)** Features are grouped into as numerical vectors (forming the image representation) and often undergo a selection process to decide which features best represent the image. **4)** Training phase builds a model that tries to separate the data into the target, distinct classes. **5)** The classifier – the mathematical function – that implements classification and defines the classes. **6)** Testing is performed by classifying unseen data belonging to know classes.

Deep learning (DL) is a subset of ML that is gaining prominence for medical imaging [38][45] and ophthalmology [14] due to increasing reports of high performance for clinical classification and decision making. DL is based on neural networks, a class of algorithms inspired by the human brain. In a neural network, the neurons are organised in layers and implement simple operations on the input data or from the output of previous layers. In a deep neural network, the number of layers is much higher than conventional neural networks (indicatively 10 or more as opposed to 2-3). The connections between the layers are assigned values, called weights, representing connection strengths. Learning the weights is the objective of the training process. Training and testing a deep neural network requires large amounts of labelled data (i.e. known classes).

In this review, we report and evaluate current AI strategies and developments for the automated detection of drusen in the context of AMD (**Figure 2**). Though some recent work has begun to explore the potential for automated drusen detection by optical coherence tomography (OCT), with varied methods and mixed results [10] [27] [50] [56], the focus of this review is on color fundus imaging of the retina.

[FIGURE 2]

Figure 2. Overview of ML methods in discussion and where they are applied at each stage. Deep Convolutional Neural Networks is a DL technique.

2. Methods

2.1 Inclusion and exclusion criteria

We aimed to include all published studies applying AI to automatic drusen detection in color fundus photographs. Inclusion criteria were (1) original study; (2) written in English; (3) validation by performance against at least one manual grader. The following studies were excluded: (1) reviews; (2) nonhuman research; (3) non-English language studies; (4) studies other than color fundus photographs (e.g. OCT); (5) studies that did not feature robust validation, as outlined below.

Validation is the process of showing quantitatively that an algorithm performs correctly, through comparison of its output to a reference standard, for example, manual grading of images by experts

[54]. Any articles that did not include validation were excluded. The performance of an algorithm is typically measured using criteria such as accuracy, sensitivity, specificity and area under ROC (receiver operating characteristic) [24]. Another important aspect is the size of the dataset: the image set an algorithm is tested must be sufficiently large to be representative of the target population, and to be suitable for the number of neural network parameters to be trained. AI methods are not immune to small sample size effects that can contaminate the evaluation of a proposed system. For instance, color fundus photographs can differ in appearance between patients while disease manifestations are also of a varying nature. Considering this, articles validated on less than 50 images were excluded.

2.2 Data Extraction

For all identified studies, an independent reviewer (EP) screened the titles and abstracts. Irrelevant and duplicate articles were removed, and the remaining articles were assessed for agreement with the inclusion and exclusion criteria by full-text review. Data extracted from studies at this stage included title, year of publication, authors, study aim, study type, number of images (training and test), diagnostic criteria, participant selection criteria, method of fundus imaging, algorithm, performance metric(s) results, and conclusions. The most recent papers were hand searched following the same strategy, filtered for the current year (i.e. 2018), and subjected to the same inclusion criteria. A similar strategy was followed for articles cited within the bibliographies of the results.

3. Results

2236 articles were identified in the initial search performed in 2017. Following filtering for AMD, 1318 articles were excluded, such as those featuring diabetic retinopathy (n = 42) and glaucoma (n = 42). From the remaining 918 articles, 834 were excluded as not using color fundus photographs (n = 18), using no imaging (n = 770) or being reviews (n = 34). 73 articles did not meet the selection criteria such as articles not reporting performance (n = 9) or featuring software optimisation (n = 3), hardware reports (n = 2) or fewer than 50 images for validation (n = 12). At the end, 8 papers met all inclusion criteria. One further article was included after searching bibliographies and 5 papers were found by hand search for this current year (2018). The resulting 14 articles were considered in this review. They all applied ML and DL techniques to drusen detection color fundus photographs.

3.1 Study designs and populations

The 14 studies involve 4 publicly available datasets (i.e. ARIA [62], STARE[26], AREDS [2], RetinaGallery [12]), 3 private datasets, 1 sourced from a telemedicine platform and a cohort from an independent study [6]. Some studies contained overlapping report analyses on the same datasets, but use different methods. 4 articles aimed to achieve *disease* or *no disease* classification. Six articles aimed to classify AMD severities according to AREDS [2] or in-house grading criteria (Cologne Image Reading Centre and Laboratory (CIRCLE)). Two articles aimed to classify *Dry AMD* vs. *Normal* images and 1 *Wet AMD* vs. *Dry AMD* or *Normal* (**Table 1**).

3.2 Pre-processing and feature extraction

In automatic detection, pre-processing is a commonly employed step to enhance an image to better facilitate the extraction of features relating to objects of interest. The human eye distinguishes “features” of disease in an image (such as GA and drusen), but AI algorithms need to extract “features” measured from the pixels pertaining to an object (i.e. drusen). In addition, a color fundus photographs typically contains a black border that needs either to be avoided or eliminated because these pixels will not be of any relevance. Retinal landmarks (e.g. the optical nerve boundary, blood vessels and macula) may obstruct features of small objects, so their removal may further improve automatic detection by reducing sources of false targets for drusen detection. A color fundus photographs might also contain artefacts (e.g. from dust particles on the lens) and display areas of uneven illumination that pre-processing can eliminate. The type of pre-processing used in the studies included depended upon the particular features used (**Table 1**).

Pixel values in imaging typically range from 0 (black) to 255 (white) per color channel (e.g. red, green, blue (RGB), or hue, saturation, value (HSV)). In color fundus photographs, drusen appear as small regions of bright pixels. Properties calculated from the image histogram (i.e. a plot of the number of pixels for each intensity value in the range and for each color channel) such as energy, entropy and intensity have all been used as features for classifying whether regions in an image contain drusen or not. Contrast Limited Adaptive Histogram Equalisation (CLAHE) [48] has been used [25] [42] [43] [61] [1] to improve contrast in the image. This well-established technique involves flattening the image histogram of relative color intensities to make the whole image as similar as possible, ultimately enhancing histogram-based features. Two studies utilised a median filter, which is applied after removing the black border to smooth high-frequency noise, but at the cost of reducing contrast [31] [47] . Grivensen and coworkers [20] manually assigned individual pixels a probability that it is part of a drusen and automatically extracted their boundaries using intensity and contrast characteristics to then be used as features for training. Burlina and coworkers [7] obtained training regions of background (no pathology) and testing masks for abnormal areas (candidate drusen) using standard image processing techniques such as median filtering, morphological dilation and thresholding. Garcia-Floriano and coworkers [18] also used mathematical morphology to highlight drusen areas and healthy macular regions. Subsequently, features called Hu moments, a well-recognised tool for object recognition in computer science, were then calculated from each pixel.

Following the pre-processing stage, it is necessary to select which features best perform as descriptors of the object of interest (i.e. drusen) within a classification scheme.

[TABLE 1]

Table 1 Included articles using AI methods for automated detection of AMD.

3.2 Feature selection

Feature selection, reported in 6 articles, is used to select a group from the extracted features or create variables that achieves the best classification performance. This process removes potentially irrelevant or confusing features and avoids model overfitting. In other words, it identifies salient features that can be used to distinguish disease images from healthy ones most effectively. Feature selection returns a numerical feature vector, which is the representation then used to train a classification algorithm (see section 3.3).

Zheng and coworkers [62] used L2-Loss of function, an established FS technique. Their aim was to identify and filter the pixel intensity features that were produced by noise. The resulting list was then ranked and the top features selected to be used for *disease/no disease* classification.

Garcia-Floriano and coworkers [18] used a filter from a feature selection software package [21]. The filter uses correlation-based feature selection that evaluates the predictive capability of features and chooses subsets highly correlated to each class [22].

To assess features that determine whether an image was *Dry* or *No AMD,* Mookiah and coworkers [42] [43] used parametric and non-parametric tests (e.g. t-test and Wilcoxon ranking) to determine the top features achieving the best one-versus-all classification for each class. With each ranked feature incrementally nested into the classification algorithm, they reported in [43] a texture feature (from a Gabor filter) as the highest ranking. In their second paper [42], the best feature was derived using the top energy features (entropy measures and their coefficients and averages) to compute an index for each image. The authors proposed the index value as a method for devising a threshold so that in a virtual clinic the threshold would be used to determine *Dry AMD* from *No AMD*.

In [1], feature selection was achieved combining a shortest-path algorithm, inspired by ant behaviour (ant colony optimisation), with a genetic optimisation algorithm, inspired by mutation and crossover operators in genetics (genetic algorithm). The overall aim was to classify *Dry AMD* and *Wet AMD* from *No AMD*. The highest ranking energy and entropy features were selected according to ANOVA to obtain a p-value. The top 10 features (1 energy, 3 entropy, 6 other non-linear) (**Table 1**) most statistically significant ($p < 0.05$) features were used for classification.

3.3 Classification

Classification uses the features selected to identify the model that best separates the data into the desired classes. A collection of images is typically separated into training and testing sets, where the former is used to develop the model and the latter is used to test it. In the context of AMD, this would test the model’s ability to classify *disease/no disease* or *dry/wet* AMD. To evaluate the accuracy of the classifier, cross-validation is often performed [52]. The algorithm performance is commonly reported in terms of statistics of measures comparing the classifiers decisions against those of one or more human experts (**Table 2-4**). Next, we describe the variety of classifications used in the studies included in this review.

3.3.1 Disease/No disease

Hijazi and coworkers [25], proposed case-based reasoning (CBR) system to develop an automated screening tool to classify 144 color fundus photographs into AMD or normal categories. CBR is a problem-solving technique founded on the observation of how humans use previous examples or information to solve new, but similar, problems. If a CBR system is given a new case, it will use the previous most similar cases in its *case base* to solve the problem. Each image histogram was conceptualized to a set of curves, called a time series, and used to generate a 2 step CBR classification. The first *case* consisted of enhanced green channel images with the blood vessel pixels replaced with null values. The second *case* contained the same but with the further process of removing the optic disc. Histograms and their time series of a collection of unseen graded images were passed to the first case for comparison to the training images. An algorithm called dynamic time warping was used to measure the similarity between the histograms and time series of the testing and training images. If the unseen image was below a certain similarity measure it was then passed to the second case for reassessment. The output is whether the input image is like either the learned time series of an AMD image or a healthy image in the case base. A specificity of 82% was reported for the effectiveness of the classifier in identifying AMD images, 65% specificity for the classifier identifying normal images and 75% accuracy in classifying images as AMD or normal (**Table 2**). This two-pass approach offered a system whereby isolation and segmentation of drusen was not required; however, removal of vessels and the optic disc was needed to improve the accuracy.

Constant false alarm rate (CFAR) detection is an adaptive algorithm that has been used to identify *normal* or *intermediate* AMD in color fundus photographs. CFAR is used in radar systems where true signal and noise signals need to be distinguished to determine origin. This returns a probability that the signal is not a false alarm. Burlina and coworkers [7] adopted such a system on 66 color fundus photographs to separate AMD from healthy images. Training and testing data were constructed from the masks obtained by pre-processing (normal retina tissue mask and edge/artefact mask). The CFAR detector was trained on the RGB and HSV color spaces of each mask, creating the signal which provides a feature for support-vector machine (SVM) classification. SVM classification is a form of ML based on regression where data is projected to a much higher dimensional space to promote linear separability of the target classes. The ability of the classifier to determine whether the image contains interesting (i.e. potentially disease) changes was reported as having a 95% specificity, 95% sensitivity with a positive predictive value (PPV) of 97% and a negative predictive value of 92% (NPV) (**Table 2**).

The same authors in [7] later reported image-mining techniques for *disease/No-disease* classification [61]. In this method, images were represented as quad trees, a form of heirarchical tree data representation, separated by their homogeny that is defined by similar pixel values. In order to extract features of the training image quad trees, a mining algorithm was used to take features from the tree such as the pixel color similarity between parent and child nodes. This returned a set of features that were reduced using an SVM ranking method [16]. To then classify the testing images, machine learning algorithms (Naïve Bayes and SVM) were used. Best detection was reported with SVM. This was then applied to new data to best predict which group the data should lie in. The authors reported 100% specificity, 99.4% sensitivity and 99.6% accuracy. This system required blood vessel removal to improve its accuracy (**Table 2**).

Garcia-Floriano and coworkers [18] used an SVM to classify 70 images into *disease/no-disease* categories. The proposed method was first evaluated on the entire dataset without and without feature selection. They obtained an accuracy of 83.58% for both evaluations. Images where the proposed method failed was due to sub-optimal image quality. Removal of poor quality images and evaluated with feature selection, improved accuracy to 92.16%.

[TABLE 2]

Table 2 Included articles using ML for classification of *disease*/*no-disease.* Performances reported as accuracy (ACC), sensitivity (SEN), specificity (SPEC)

3.3.2 AMD severity

Phan and coworkers [47] attempted to classify AMD severity according to their AREDS categories [5] using visual words, also known as "bag of words". The most salient features in the image were detected and their frequencies counted and binned in to a histogram. This forms a so-called vocabulary that can be used for automated detection of the same words in an unseen image. The authors used SURF (Speeded Up Robust Features) to build the vocabulary from different color spaces (RGB and a color space describing lightness, green-red and blue-yellow called L*a*b) of 279 images, including poor quality images, to build the vocabulary. SVM and Random Forest classifiers were tested with and without feature selection steps. They report the best performance for AMD screening with SVM classifier (AUC 87.7%). For grading the classes of AMD they report {1} vs {2} vs {3} vs {4} accuracy of 62.7%. Accuracy of 75.6% and 72.4% were obtained for {1&2} vs {3} vs {4} and for {1} vs {2&3} vs {4} respectively (Table 3).

Kankanaballi and coworkers [31] also used SURF along with a faster version called Scale-Invariant Feature Transform (SIFT) to extract local features in 2772 AREDS images. These features were taken from the L*a*b color space to generate a vocabulary for a visual words algorithm. They evaluated the performance of the algorithm to correctly classify images into AREDS categories [5] (1) class {1&2} vs {3 & 4}: (2) {1 vs 2} vs {3}: (3) {1} vs {3}: (4): {1} vs {3 & 4} and experimented with 3 dataset designs. A manually selected data set of good quality images (denoted MS). A set of automatically selected [44] good quality images, one where each class of AREDS category was as large as possible (denoted MIPC) and another where AREDS categories was kept equal (denoted EIPC). They reported the highest accuracy for category 1 from MS images of 98.9% accuracy. For images automatically selected, the highest accuracies were 96.1% (category 2 EIPC), 97.1% (category test 3 EIPC) and 97.1% (category 4 MIPC) (**Table 3**).

Grinsven and coworkers [20] segmented drusen so that their location, area and size could be quantified. The overall aim was to distinguish images of low-risk AMD from high-risk AMD. Two observers manually segmented 52 images to provide a reference set for evaluation of automated drusen quantification (set A) and graded 355 images to evaluate automated AMD severity classification (set B). Candidate drusen extraction was achieved by convolving the green channel of the color fundus photographs with Gaussian filters and using their derivatives to train a classifier. The classifier used regression to determine the class of the data point and the pixels filter response, called K-nearest neighbours. The line of regression can be used to assign a probability value that from the filter response of a previously unseen pixel that that it belongs to a lesion. Therefore, neighboring pixels with high probabilities can be grouped into candidate drusen. At this stage, the authors segmented the optic nerve and blood vessels so that any candidate drusen overlapping these anatomical landmarks could be excluded. This produced a probability map of the image where a search-based optimisation method (i.e. dynamic programming) was then used to solve the candidate borders. Subsequently, total drusen area and maximum drusen diameter were quantified and compared to measurements derived from the observers' manual annotations using intraclass correlation coefficients (ICC). Linear discriminant analysis was used to separate candidate drusen from true drusen by extracting over 100 features in different color spaces (Luv, HSI), intensity (RGB contrasts), contextual (Average, SD of pixel probability inside/outside border) and shape (area, perimeter) information. Each image probability map was then binned according to candidate drusen size and used to train a Random Forest classifier. This builds a decision tree whereby the output is whether the image is from a low- or high-risk patient. The authors validated algorithm according to measurement agreeability between algorithm and two graders using ICC. They report ICC's of drusen area and diameter measurements of 0.69 and highest AUC of 0.954 of correct AMD image classification (**Table 3**).

[TABLE 3]

Table 3. Included articles using ML for classification of AMD severity. Equal Number of Images (EIPC), Maximum Number of Images per Class (MIPC), Manually Selected images (MS). Interclass correlation coefficient (ICC) set at 95% Confidence Interval. Kappa scores measure inter rater agreement. Performances reported as area under curve (AUC), sensitivity (SEN), specificity (SPEC) and accuracy (ACC). AMD categories defined using AREDS categories [5] or by in-house grading criteria (Cologne Image Reading Centre and Laboratory (CIRCLE)).

3.3.3 Wet/Dry/No-disease

Using entropy measures as features from wavelet coefficients and from green channel CLACHE enhanced images, detection of Dry AMD using SVM, Naïve Bayes, Probabilistic Neural Networks, k-nearest neighbours and decision trees was proposed by Mookiah and coworkers [42] [43]. This system was trained and tested separately on three datasets (ARIA, STARE and a private dataset). The best performance was reported for a SVM classifier where Gabor, local pixel intensity changes and entropy features ranked best. The highest performances were observed in ARIA and STARE with an accuracy of correctly classifying between *Dry AMD* and *Normal* of 95.7% and 95% respectively [43]. Statistical moments, energy, entropy and Gini index features extracted from discrete wavelet transform (a well-known image denoising technique) also presented the best accuracy for SVM (93.70%) [41]. This system did not require prior segmentation of retinal landmarks and drusen and the use of multiple classifiers provided a degree of discrimination ability of the extracted features (**Table 4**).

SVM was also reported to be the best performing classifier for Pyramid Histogram of Gradients (PHOG) features extracted by particle swarm optimisation (PSO) algorithm, used to detect *Wet* AMD and *Dry* AMD [1]. In a private dataset, 945 images were used for training and testing where the algorithm correctly identified the *Wet* from *Dry* from *Normal* images with 85.12% accuracy. The number of *Wet* AMD images in the data set was imbalanced (21 *Dry* to 1 *Wet*). To compensate for this, synthetic samples was generated by oversampling of the minority class. This produced synthetic features to simulate pathology and balance the dataset. This system did not require any retinal landmark or drusen segmentation steps (**Table 4**).

[TABLE 4]

Table 4. Included articles using ML for classification of wet/dry/*no-disease*. Performances reported as sensitivity (SEN), specificity (SPEC) and accuracy (ACC).

3.4 Deep Learning

DL is a rapidly growing field where conventional ML feature extraction, training and classifiers are replaced with multi-layer neural networks capable of learning latent patterns in the data [37]. Neural network architecture (i.e. the layers) are carefully designed and assembled for the task the network is to perform. Convolution, pooling and fully connected layers are the basic building blocks for the most well known class of neural networks, called convolutional neural networks (CNN). CNN's are considered Deep Convolutional Neural Networks (DCNN) when their architecture typically contains 10 or more convolutional layers. DCNN's require large amounts of often labelled data to train, that may not be available, especially in a healthcare setting. Various methods exist to increase data set size in order to utilise state of the art DL techniques.

Tan and coworkers [55] developed a 14-layer deep convolutional neural network to classify images as disease/no-disease and trained and tested on 1110 images (708 no disease, 402 disease). To increase the size of the data set, data augmentation was used. Images were flipped left, flipped down and flipped left and downwards to increase artificially the size of the dataset. This produced four instances of each image used to train and test the DCNN. They validated the DCNN using 10-fold cross validation reporting an average fold accuracy, sensitivity and specificity of 95.45%, 96.43% and 93.75% respectively.

Pre-trained networks also offer a solution when there is little data, whereby networks already trained to solve a similar task can be re-used (transfer learning). ImageNet is a large general (non-medical) benchmark dataset popularly used to develop DCNN's. Early layers of a DCNN learn lower level features such as edges and colors. The following layers learn higher level features and more image domain specific features to classify the image. Transfer learning is based on the idea that these lower level features may generalize to images different from the training images. For instance, Overfeat is a pre-trained network to detect and localise everyday objects within a non-medical image [51]. Burlina and coworkers [8] assessed the efficacy of the pre-trained DCNN in classification of AMD using OverFeat. With the input of 5600 color fundus photographs from NIH AREDS into the OverFeat network to classify against pairs of AREDS categories [5] {1 & 2} vs {3 &4}; {1 & 2} vs {3} ; {1} and {1} vs {3 & 4} , they reported a preliminary performance of 92% to 95% accuracy. The same experiment was performed in their later work [9] to assess the use of these features to fine tune a SVM classifier and compared the algorithms AREDS grades to a human grader. An input of 5,664 images into the pre-trained Overfeat network was used to obtain a feature vector. These features were then passed to an SVM classifier to classify AMD images as before. They reported a similar performance between class 1 and class 4 and grader with less agreeability between class 2 and class 3, algorithm versus grader.

Ensemble learning is a method where multiple models are combined into one predictive model. Grassmann and coworkers [19] trained six DCNN's from the ImageNet competition independently, [11] [23] [36] [46] [53] [54] to predict AMD severity. Classes were defined as AREDS category (9 classes), late AMD stages (3 classes) and ungradable image (1 class). The results from each DCNN were then used to train a random forest classifier to build a model ensemble. They trained and tested each DCNN and the ensemble on 120,656 color fundus photographs (86,770 training and

21,867 testing). Individual DCNN's achieved accuracies between 57.7% and 61.7%. By combining the DCNN's into an ensemble the overall accuracy was increased to 92.1% for predicting each AMD class. Grassmann and coworkers [18] also used an independent dataset of 5555 [6] to evaluate their algorithm and achieved an accuracy of 34%. Misclassifications were color fundus photographs from healthy individuals incorrectly classified as neovascular AMD. This was due to younger eyes in the KORA dataset (< 40 years old) demonstrating dominant macular reflexes, which was not observed in the training data (> 55 years old). By restricting the analysis to fundus images of eyes 55 years and older they increased the performance to 50% accuracy for predicting AMD severity according to their defined AMD classes. When the algorithm was used to classify early or late AMD, accuracy was improved to 84.2% and correctly classified 94.3% of healthy fundus images.

4. Discussion

Our search highlighted ML as the predominant technique for AMD detection and classification, with most recent papers reporting DL techniques. The primary aim of drusen-related automated image analysis is to support decision-making in the clinic. Rather than detecting individual drusen, image level classification was more common with the aim of computerizing AMD screening and grading systems. Only a single article reported discrete drusen measurement and quantification [20]**.** Manually outlining individual drusen to provide ground truth for algorithm training is very labor intensive and motivates the shortage of ML approaches to individual drusen segmentation. AREDS categories [5]**,** Class 1 and Class 2 AMD are the most difficult to separate because grading relies on drusen counts and measurements that cannot be obtained automatically without the reference data. ML is particularly susceptible to this paradox because they are driven by examples that are assumed to be representative of the population. A newly obtained image may not be similar to any of the examples used to train the model and therefore it may fail to classify it. This effect of data variability was also observed in [19] when the model was evaluated on an independent dataset containing colour fundus photographs with retinopathies not present in the training set and removal improved performance. This raises questions as to how ML would generalise to the clinic.

In terms of translating into the clinic, systems depending on segmentation of retinal landmarks [16] [20] [25] would need reliable and robust detection and segmentation algoithms. Algorithms would also need to be robust to image quality. Comparibly, Kankanballi and colleagues[31] and Phan and colleagues [47] both use a visual words algorithm, but Kankanballi includes poor quality images and achieve lower overall accuracies than Phan who use a larger data set. In Phan [47]**,** the algorithm is tested on datasets with a varying balance of images labelled in the ARED's categories, where highest accuracies are achieved for the more balanced datasets or category contains clear and expected differences between AMD severities (class 1 vs class {3 & 4}). This exemplifies how a classifier can be fine-tuned and stabilised by dataset balance and image quality alone. Additionally Burlina and colleagues [7],use the only algorithm that explicitly states validation on African and Asian eyes, where due to high melanin content, images are darker. This highlights that an algorithm for use in the clinic would also need to be robust to ethnicity.

Interestingly, the single article proposing a *Dry/Wet* classifier yielded good results [1] even with synthetic data. Wet AMD occurs when neovascularisation occurs, with subsequent intra-retinal fluid causing central vison loss. In the clinic, it is now standard practice to use cross-sectional OCT for insight into intra-retinal fluid levels. Presentation of Wet AMD involves a wide spectrum of changes in the retina from normal looking retina to distorted bloody retina. This is a difficult classifier to train and may indicate why there is only a single report of an algorithm using ML to detect Dry from Wet AMD. As DL is becoming state-of-the-art for difficult classification problems, future studies using DL for classifying Wet AMD could yield better results. This would be valuable in the clinic, as Wet AMD requires urgent care.

There is also a clear importance to assess algorithm performance against the expert grader if such systems are to be deployed in a clinical setting. The methods were evaluated on different datasets, which makes levels of performance difficult to compare between algorithms including, for example, variants in pre-processing, feature selection and classification. Methods of pre-processing employed largely depend on the features that need to be enhanced, where the green channel is the most commonly reported input for drusen detection. Texture and color features are predominantly used for AMD grading which is reasonable considering that colour distributions and texture in a diseased image may differ dramatically from that in a normal eye.

ML requires feature design and selection that increase in complexity as the data increases in variability. DL networks exploit underlying patterns that perform well when data complexity and variation increases. Given the variable nature of the human retina, such systems appear more promising for adoption in the clinic. As drusen edges are hard to define, DL may be able to learn subtle patterns within the data to aide in quantifying areas of drusen for detecting disease progression. DL algorithms are producing state-of-the-art results but come at a computational cost. Large amounts of data are required to train the dataset which still requires (some) validation from ground truth. Further development of such algorithms represents a growing and expanding interdisciplinary field for automatic disease detection.

The results of our search identified a number of articles reporting algorithms for detection of DR and glaucoma where drusen can also be present. Fundus imaging has also been utilised to derive biomarkers for systemic conditions, such as hypertension and diabetes [40]. Recently, there are an increased number of reports linking AMD to Alzheimer disease (AD). AD is diagnosed using medical history, psychiatric examination, brain imaging and biomarkers in cerebrospinal fluid (CSF). Definitive classification requires neuropathological changes as seen on post-mortem examination. Characteristic retinal changes have previously been identified in AD, such as a sparser retinal vascular network (inferring altered cerebral vasculature) [41] and thinning of the retinal nerve fibre layer [56] a marker of axonal loss). A key component of AD related deposits in the brain, amyloid β (Aβ), is also found in drusen. Aβ is an aggregate-prone peptide family that aggressively targets neurons [4] and there are an increasing number of reports of amyloid plaques in the retina in AD patients [29] [35] [39] [59]. As the retina is anatomically, embryologically and physiologically linked to the central nervous system, it is perhaps not surprising that these depositions may have implications to neurodegenerative disease of the brain. Indeed, the progression of drusen formation in the peripheral retina has been found to be more prevalent in patients with AD in comparison to age-matched control [13]. These findings were in a small cohort but suggest a promising biomarker for disease-related plaque formation in the brain.

When AMD progresses asymmetrically, patients risk remaining asymptomatic due to maintaining good visual acuity in their healthy eye. The resulting delay in presentation and treatment impacts visual prognosis.

For automated drusen assessment to be applied in the clinic it must go beyond cross-sectional phenotyping and instead relate to real patient visual outcomes. Longitudinal studies will be required to determine if automated image grading, based on drusen detection, can accurately predict disease progression.

Future algorithms involving drusen detection should aim to provide useful quantification to aid screening for AMD. A screening programme should stratify patients according to optimal follow up pathway. In order for automated drusen detection to contribute to the cost-effectiveness of a screening programme for AMD, it must separate individuals with drusen associated with normal aging from patients whose drusen load progresses as well as stratifying patients with mild AMD into those at low risk and at high risk of progression to severe AMD. This would enable the ophthalmologist to select relevant patients for regular follow up, thus improving the efficiency of patient care.

Method of Literature Search

Published studies were identified through systematic searches of EMBASE, PubMed, Web of Knowledge, ScienceDirect, ACM Digital Library and IEEE Xplore. The search terms in the first instance included “*drusen”* and in combination with *“detection”* or *“classification”* or *“identification”* or *“segmentation”* or *“quantification”* or *“measurement”* or *“algorithm”*. Further filtering was conducted on the titles and abstracts based on whether they contain “*age-related macular degeneration”* or *“AMD”*.

References

[1] Acharya U, Hagiwara Y, Koh J, Salatha. Automated screening tool for dry and wet age-related macular degeneration (ARMD) using pyramid of histogram of orientated gradients (PHOG) and nonlinear features. Computational Science. 2017:20:41-51

[2] AREDS Group. A randomized, placebo-controlled, clinical trial of high-dose supplementation and vitamins C and E, beta carotene, and zinc for age-related macular degeneration and vision loss: AREDS report no 8. Arch Ophthalmology. 2001:119:1417-36

[3] AREDS. The age-related eye disease study severity scale for age-related macular degeneration. Arch Opthalmology. 2006:123(11):1484-98

[4] Bennilova I, Karran E, De Strooper B. The toxic Aβ oligomer and Alzheimer's disease: an emperor in need of clothes. Nature Neuroscience. 2012:15(3):349-57

[5] Bird AC, Bressler NM, Bressler SB, Chisholm IH. An international classification and grading system for age-related maculopathy and age-related macular degeneration. Survey of Ophthalmology. 1995:49(5):367-74

[6] Brandi C, Breinlich V, Stark KJ, Enzinger S. Features of Age-Related Macular Degeneration in the General Adults and Their Dependency on Age, Sex, and Smoking: Results from the German KORA Study. PLoS One. 11 (2016) p. e0167181

[7] Burlina P, Freund D, Dupas B, Bressler N. Automatic screening of Age-related macular degeneration and retinal abnormalities. 33rd Annual International Conference of the IEEEMBS. Boston. 2011:3692-6

[8] Burlina P, Freund DE, Joshi N, Wolfson Y. Detection of age-related macular degeneration via deep learning. IEEE 13th International Symposium on Biomedical Imaging (ISBI). 2016:184-88

[9] Burlina P, Pacheco K, Joshi N, Freund D. Comparing humans and deep learning performance for grading AMD: a study in using universal deep features and transfer learning for automated AMD analysis. Computers in Biology and Medicine. 2017:82:80-6

[10] Chen Q, Leng T, Kutzscher L, Ma J. Automated drusen segmentation and quantification in SD-OCT images. Med Image Analysis. 2013:17(8):1058-72

[11] Chen T, Mu L, Li Y. Mxnet: a flexible and efficient machine learning library for heterogeneous distributed systems. ArXiv Prepr. 2015;arXiv:1512

[12] Cohen S. Retina gallery ~full sized retina images. Available: http://retinagallery.com/index.php

[13] Csincsik L, MacGillivray T, Flynn E, Pellegrini E. Peripheral Retinal Imaging Biomarkers for Alzheimer's Disease: A Pilot Study. Opthalmic Research. 2018:59(4):182-92

[14] De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine. 2018. https://doi.org/10.1038/s41591-018-0107-6

[15] Duanggate C, Uyyanovara B. A review of automatic detection and segmentation from retinal images. The 3rd International Symposium on Biomedical Engineering (ISBME). 2008:222-25

[16] Fan RE, Chang KW, Hsieh CJ, Wang XR. LIBLINEAR: A library for the large linear classification. Journal of Machine Learning Research. 2008:9:1871-74

[17] Floriano García A, Sistema A . Integral de análisis para la prevención de ceguera Master of Science Thesis. Mexico City: Centro de Investigación en Computación del IPN; 2011

[18] Garcia-Floriano A, Ferreira-Santiago A, Camacho-Nieto O, Yanez-Marquez C. A machine learning approach to medical image classification: Detecting age-related macular degeneration in fundus images. Computers and Electrical Engineering. 2017. https://doi.org/10.1016/j.compeleceng.2017.11.008

[19] Grassman F, Mengelkamp J, Brandl C, Harsch S. A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. American Academy of Ophthalmology. 2018:1-11

[20] Grivensen M, Lechanteur Y, van de Ven J, Ginneken B. Automatic drusen quantification and risk assessment of age-related macular degeneration on color fundus images. Investigative Ophthalmology & Visual Science. 2013:54:3019-27

[21] Hall M , Frank E , Holmes G , Pfahringer B , Reutmann P , Witten I . The WEKA data mining software: an update. SIGKDD Explor. 2009:11

[22] Hall MA . Correlation-based Feature Selection for Machine Learning PhD Thesis. Hamilton New Zeland: The University of Waikato. 1999

[23] He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. Computing Research Repository (CoRR). 2016: abs/1603.0.

[24] Heneghan C, Flyn J, O'Keefe M, Cahill M. Characterization of changes in blood vessel width tortuosity in retinopathy of prematurity using image analysis. Medical Image Analysis. 2001:6(4):407-29

[25] Hijazi M, Coenen F, Zheng Y. Retinal image classification using histogram based approach. The 2010 International Joint Conference on Neural Networks (IJCNN). Barcelona. 2010:1-7

[26] Hoover A. Structured Analysis of the Retina. Available: http://cecas.clemson.edu/~ahoover/stare/. [Accessed 29 June 2018]

[27] Iwama D, Hangai M, Ooto S, Sakamoto. Automated assessment of drusen using three-dimensional spectral- domain optical coherence tomography. Invest Opthalmol Vis Sci. 2012:53(3):1576-83

[28] Joachim N, Mitchell P, Burlutsky G, Kifley A, Wang JJ. The incidence and progression of age-related macular degeneration over 15 years: the Blue Mountains eye study. 2015:1229(12): 2482-89

[29] Johnson LV, Leitner WP, Rivest AJ, Staples MK. The alzheimers AB-peptide is deposited at sites of the complement activation in pathologic deposits associated with aging and age-related macular degeneration PNAS. 2002:99:11830-11835

[30] Jonas JB, Bourne, RRA, White RA, Flaxman SR. Visual impairment and blindness due to macular diseases globally: a systematic review and meta-analysis. American Journal Of Opthalmology. 2014:159(4):808-15

[31] Kankanaballi S, Burlina P, Wolfson Y, Freund D. Automated classification of severity of age-related macular degeneration from fundus photographs. Investigative Ophthalmology & Visual Science. 2013:54(3):1789-1796

[32] Klaver CC, Ott A, Hofman A. Is age-related macular maculopathy associated with Alzheimer's disease? The Rotterdam study. 1999:120(9):963-68

[33] Klein R, Davis MD, Magli YL. The wisconson age-related maculopathy grading system. Opthalmology. 1991:98(7):1128-34

[34] Klein R, Klein BE, Knudston MD, Meuer SM. Fifteen-year cumulative incidence of age-related macular degeneration: the Beaver Dam eye study. 2007:114(2):253-62

[35] Koronyo-Hamaoui M, Koronyo Y, Liubimov AV. Identification of amyloid plaques in retinas for alzheimer's patients and noninvasive in vivo optical imaging of retinal plaques in a mouse model. Neuroimage. 2011:54:204-17

[36] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. NIPS. 2012:1106-1114

[37] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 521:436-44

[38] Litjens G, Kooi T, Bejnordi BE, Setio AAAS. A survey on deep learning in medical image analysis. Medical Image analysis. 2017:42:60-88

[39] Loffler KU, Edward DP and Tso MO. Immunoreactivity against tau, amyloid precursor protein, and beta-amyloid in the human retina," Invest Opthalmol Vis Sci. 1995:36(1):24-31

[40] MacGillivray TJ, Trucco E, Cameron JC, Dhillon B. Retinal imaging as a source of biomarkers for diagnosis, characterisation and prognosis of chronic illness or long-term conditions. Br J Radiol. 2014:87(1040)

[41] McGrory S, Cameron JR, Pellegrini E, MacGillivray T. The application of retinal fundus camera imaging in dementia. A systematic review. Alzheimer's & Dementia. 2017:6:91-107

[42] Mookiah M, Acharya U, Koh J, Chua CK. Decision support system for age-related macular degeneration using discrete wavelet transform. Med Biol Eng Comput. 2014:52:781-796

[43] Mookiah MRK, Acharya U, Koh J, Chandran V. Automated diagnosis of age-related macular degeneration using greyscale features from digital fundus images. Computers Biology and Medicine. 2014:53:55-64

[44] Niemeijer M, Abramoff MD, Ginneken BV. Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening. Medical Image Analysis. 2006:10(6):888-98

[45] Pellegrini E, Ballerini L, Hernandez M, Chappel F. Machine learning of neuroimaging to diagnose cognitive impairment and dementia: a systematic review and comparative analysis. Neurons and Cognition.2018: arXiv:1804.01961v2 [q-bio.NC]

[46] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011:12:2825e2830.

[47] Phan T, Seoud L, Chakor H, Cheriet F. Automatic screening and grading of age-related macular degeneration from texture analysis of fundus images. Journal of Ophthalmology. 2016:8:1-11

[48] Pizer SM, Amburn EP, Austin JD, Cromarrtie R. Adadptive histogram equalisation and it's variations. Computer Vision, Graphica and Image Processing. 1987:39(3):355-68

[49] RNIB. Key information and statistics. Available: http://www.rnib.org.uk/knowledge-and-research-hub/key-information-and-statistics. [Accessed 3 April 2018]

[50] Schlanitz FG, Baumann B, Spalek T, Schutze C. Performance of automated drusen detection by polarization-sensitive optical coherence tomography. Invest Opthalmol Vis Sci. 2011: 52(7):4571-9

[51] Sermanet P, Eigen D, Zhang X, Mathieu M. Overfeat: integrated recognition, localization and detection using convolutional networks. in International Conference on Learning Representations (ICLR2014), CBLS. April 2014. 2014. http://openreview.net/document/d332e77d-459a-4af8-b3ed-55ba, http://arxiv.org/abs/1312.6229.

[52] Stone M. Cross-validity choice and assessment of statistical predictions. Journal of the Royal Statistics Society. Series B (Methodological). 1974:36(2):111-47

[53] Szegedy C, Vanhoucke V, Ioffe S. Rethinking the Inception Architecture for computer vision. Computing Research Repository (CoRR). 2015:abs/1512.0. Available at:https://arxiv.org/corr/home.

[54] Szegedy C, Wei Liu, Yangqing J. Going deeper with convolutions. IEEE Conf. Comput. Vis. Pattern Recognit IEEE, Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Press. 2015:1063-6919:1e9

[55] Tan JH, Bhandary SV, Sivaprasad S, Hagiwara Y. Age-related macular degeneration detection using deep convolutional neural network. Future Generation Computer Systems. 2018:127-135

[56] Thompson KL, Yeo MJ, Waddell B, Cameron JR. A systematic review and meta-analysis of retinal nerve fiber layer change in dementia, using optical coherence tomography. Alzheimer's Demnt (Amst). 2015:1(2):136-43

[57] Trucco E, Ruggeri A, Karnowski T, Giancardo L. Validating retinal fundus image analysis algorithms: Issues and a Proposal. IOVS. 2013:54(5):3546-59

[58] Wong WL, Xinyi S, Li X, Cheng CM. Global prevelance of age related macuclar degeneration and disease buren projection for 2020 and 2040: a systematic review and meta-analysis. 2014:2(2):p106-16

[59] Yoshida T, Ohno-Matsui K, Ichinose SJ. The potential role of amyloid beta in the pathogenesis of age-related macular degeneration," Clinical Investigation. 2005:115(10):2763-2800

[60] Zhao R, Camino A, Wang J, Hagag AM. Automated detection in dry age-related macular degeneration by multiple depth, enface optical coherence tomography. Biomed Opt Express. 2017:8(11):5049-64

[61] Zheng Y, Hijazi M, Coenen F. Automated disease/no disease grading of age-related macular degeneration by an image mining approach. Investigative Ophthalmology & Visual Science. 2012:53(13) 8310-18

[62] Zheng Y. ARIA. The Foundation for the Prevention of Blindness. Available: https://eyecharity.weebly.com/aria_online.html. [Accessed 29 June 2018]

Reference	Dataset	Fundus Camera (resolution)	Pre-processing	Feature	Output
Hijazi et al 2010 [25]	144 (ARIA)	Not reported	CLAHE Retinal vessels segmented by thresholding and OD segmented using intensity peaks of image (identified by sliding window)	RGB and HSI histogram of each image conceptualised to set of curves (time series)	Disease/No Disease
Burlina et al 2011 [7]	66 (private)	Zeiss FF4 40° FOV (pupils dilated) Images resized to 1000 x 1000	Pyramid decomposition of green channel for regions of high gradient magnitude to create logical masks for training and testing. Areas of high gradient magnitude indicate artefacts and vessels where low gradient magnitude indicate normal retinal tissue	Intensity, colour and gradient features of background (normal retina) and candidate abnormal areas	Disease/No Disease

Zheng et al 2012 [61]	101(ARIA)
	97(STARE)
		TOPCON TRV-50 fundus camera 35 ° field of view (700 x 605)	Mask of whole image to capture circular fundus ROI. Colour normalisation and uneven illumination is applied. CLAHE to enhance contrast. Blood vessels identified using wavelet features.	Image represented as quadtree, separated by their homogeny, defined by similar pixel values. Image mining algorithm returns features	Disease/No Disease

Kankanaballi et al 2013 [31]	2772(NIH AREDS)	Not reported	Green channel smoothed by large median filter. Median filtered image subtracted from original green channel and the result multiplied to increase contrast	SIFT/SURF features of L*a*b colour channel	AMD severity
Grivinsen et al 2013 [20]	407(EUGENDA)	TOPCON TRC 501X 50° field of view Canon CR-DGi (non-mydriatic) 45° field of view	Drusen manually outlined	Each pixel in image assigned probability that it belongs to drusen candidate. Boundary of the candidate extracted using intensity and contrast characteristics	AMD severity
Mookiah et al 2014 [43]	161 (ARIA) 83 (STARE) 540 (KMC)	Carl Zeiss Meditec fundus camera 50 ° field of view (748 x 576) TOPCON TRV-50 fundus camera 35 ° field of view (700 x 605) TOPCON non-mydriatic retinal camera (TRC-NW200) (480 x 364)	CLAHE	Entropy features – Shannon, Kapur, Renyi, Yager Higher Order Spectra (HOS)	Wet/Dry/No Disease
Mookiah et al 2014 [42]	540 (KMC)	TOPCON non-mydriatic retinal camera (TRC-NW200) (480 x 364)	CLAHE	Features for whole image obtained by discrete wavelet transform (DWT) decomposition. Linear features extracted from wavelet coefficients (mean, variance, skewness, kurtosis, Shannon entropy, Renyi entropy, Kapur entropy, relative energy, relative entropy, entropy, Gini index).	Wet/Dry/No Disease
Burlina et al 2016 [8]	5500 (NIH AREDS)	Not reported	Resizing and cropping images to conform to expected OverFeat input network	SURF, SIFT, wavelet features	AMD severity
Phan et al [47]	279 (Telemedicine	Zeiss, DRS, Topcon models	Pre-processing from [31]	Colour Histograms (RGB, L*a*b colour spaces)	AMD severity

2016	Platform)	45° FOV (1400, 2,200,3240 pixels along diameter of image)		Texture - Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), SURF	
Acharya et al 2017 [1]	945 (KMC)	Zeiss FF450 plus mydriatic fundus camera (resized to 480 x 360 from 2588 x 1958	CLAHE	Pyramid of histograms of Orientated Gradients (PHOG) to describe shape and pattern. Features from descriptor: Energy – uniformity of image Entropy features – approximate, fuzzy, Kolmogorov-Sinai, modified multiscale, Permutation, Renyi, Sample, Shannon, Tsallis and wavelet Nonliner features- fractal dimension (D), Hjorth (activity, complexity, mobility parameters), Kolmogorov complexity, largest Lyapunov exponent, Lempel Ziv complexity, relative qualitative analysis (parameters entropy, transitivity, trapping time, recurrence of the 1st type and 2nd type, longest vertical line), Entropy, determinism, laminarity, maximal diagonal line length, averaged diagonal line length, recurrence rate, recurrence time of RQA parameters	Wet/Dry/No Disease
Burlina et al 2017 [9]	5664 (NIH AREDS)	Not reported	Resizing and cropping images to conform to expected OverFeat input network	OverFeat (OF) universal features	AMD severity
Garcia-Floriano et al 2017 [18]	397 (STARE) 70 (RetinaGallery)	Not reported	OD located using [17]. Green channel.	Hu moments were used to describe each object as a measurable quantity calculated from the shape of a set of points	Disease/No Disease
Tan et al 2018 [55]	1110 (KMC)	Zeiss FF450 plus mydriatic fundus camera (2588 x 1958)	Image rescaled to 180 x 180 to conform to network input dimensions	Features learned through Neural Network	Disease/No Disease
Grassman et al 2018 [19]	120,656 (AREDS) 5555 (KORA)	Zeiss FF series fundus camera TOPCON TRC-NW5S 45° fundus camera	Normalisation of colour balance and local illumination by Gaussian filtering. Images resized to 512 x 512 to conform to neural network input dimensions	Features learned through Neural Network	AMD severity

Reference	**Images with *disease* (dataset)**	**Images with *no disease* (dataset)**	**Classifier**	**Reference Standard**	**Performance**
Hijazi et al [25]	86 (ARIA)	56 (ARIA)	Case Based Reasoning (CBR)	Labels from ARIA project	ACC = 75% SEN = 82.00% SPEC = 65.00%
Burlina et al [7]	39 (private)	27 (private)	Constant False Alarm Rate (CFAR)	Graders from JHU Wilmer Eye Institute	SEN = 95% SPEC = 96% PPV (positive predictive value)= 97% NPV (negative predictive value) = 92%
Zheng et al [61]	101 (ARIA) 59 (STARE)	60 (ARIA) 38 (STARE)	Naïve Bayes, SVM	Labels from dataset	SPEC = 100% SENS = 99.4% ACC = 99.6%
Garcia-Floriano et al [18]	34 (STARE) 33 (RetinaGallery)	41 (STARE) 37 (RetinaGallery)	SVM	Labels from STARE and RetinaGallery	ACC = 92.1569% Precision = 0.904 Recall = 0.922 F-measure = 0.921

Reference	Number of images in AMD severity category	Classifier	Reference Standard	AMD category Test	Performance

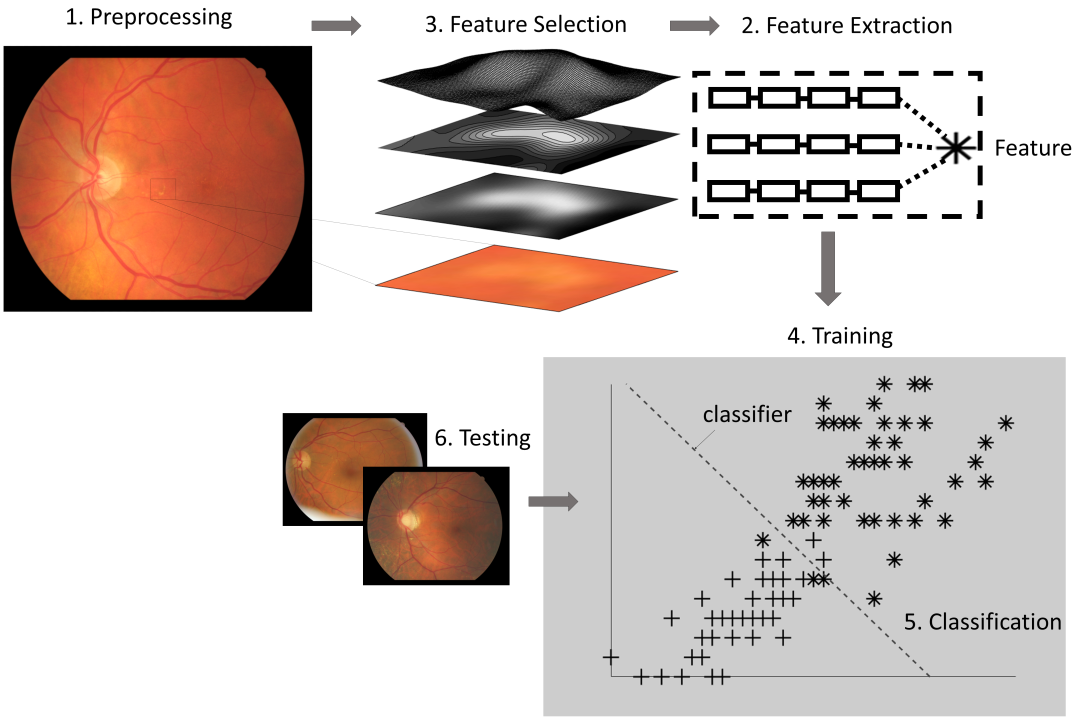
Kankanaballi et al [31]	EIPC:
	• 626 (category 1)
	• 89 (category 2)
	• 715 (category 3)
	• 715 (category 4)
	MIPC:
	• 626 (category 1)
	• 89 (category 2)
	• 1107 (category 3)
	• 950(category 4)
	MS:
	• 180 (category 1)
	• 13 (category 2)
	• 114 (category 3)
	• 78 (category 4)
		Random Forest	Expert Grader	(1) {1 & 2} vs {3 & 4}
				(2) {1 & 2} vs {3}
				(3) {1} vs {3}
				(4) {1} vs {3 &4}
					EIPC: 95.4% (SPEC) 95.5% (SEN) 95.5% (ACC)
					MIPC: 91.6% (SPEC) 97.2% (SEN) 98.9% (ACC)
					MS: 98.4% (SPEC) 99.5% (SEN) 98.9% (ACC)
					EIPC: 96.1% (SPEC) 96.1% (SEN) 96.1% (ACC)
					MIPC: 95.7% (SPEC) 96.0% (SEN) 95.9% (ACC)
					EIPC: 98.6% (SPEC) 95.7% (SEN) 97.1% (ACC)
					MIPC: 96.3% (SPEC) 96.8% (SEN) 96.7% (ACC)
					EIPC: 96.0% (SPEC) 94.7% (SEN) 95.4% (ACC)
					MIPC: 95.4% (SPEC) 97.7% (SEN) 97.1% (ACC)

Grivinsen et al [20]	Set A: • 17 Observer 1 , 20 Observer 2 (No AMD) • 13 Observer 1 , 9 Observer 2 (Early AMD) • 22 Observer 1 , 23 Observer 2 (Intermediate AMD) Set B: • 216 Observer 1 , 218 Observer 2 (No AMD) • 64 Observer 1 , 64 Observer 2 (Early AMD) • 75 Observer 1 , 76 Observer 2 (Intermediate AMD) Average number of drusen: • 130.4 ± 178.1 (Observer 1), 198.5 ± 243.1 (Observer 2) Average size of drusen (µm^2): • 5,873 ± 10,027 (Observer 1), 5115 ± 8257 (Observer 2)	K-nearest Neighbour Linear discriminant classifier Random Forest	2 Observers	Drusen Area: Observer 1 vs Algorithm Observer 2 vs Algorithm Interobserver Drusen Diameter: Observer 1 vs Algorithm Observer 2 vs Algorithm Interobserver Risk Assessment: Observer 1 vs Algorithm Observer 2 vs Algorithm	0.91 (ICC) 0.86 (ICC) 0.87 (ICC) 0.66 (ICC) 0.69 (ICC) 0.79 (ICC) 0.84 (Observer SEN) 0.96 (Observer SPEC) 0.948 (Algorithm AUC) 0.765 (Kappa) 0.85 (Observer SEN) 0.954 (Observer SPEC) 0.954 (Algorithm AUC) 0.760 (Kappa)

Phan et al [47]	Good Quality:
	• 50 (category 1)
	• 43 (category 2)
	• 24 (category 3)
	• 22 (category 4)
	Poor Quality:
	• 29 (category 1)
	• 36 (category 2)
		SVM & Random Forest	2 graders	{1} vs {2} vs {3} vs {4}
				{1&2} vs {3} vs {4}
				{1} vs {2&3} vs {4}
					SVM: 62.7% (ACC)
					Random Forest: 61.7% (ACC)
					SVM: 75.6% (ACC)
					Random Forest: 74.2% (ACC)
					SVM: 72.4% (ACC)
					Random Forest: 69.9% (ACC)

	• 41 (category 3) • 34 (category 4)				

Reference	**Images with *No-disease* (dataset)**	**Images with *AMD*(dataset)**	**Classifier**	**Reference Standard**	**Performance**
Mookiah et al [43]	101 (ARIA) 36 (STARE) 270 (KMC)	60 (ARIA) 47(STARE) 270 (KMC)	Naïve Bayes, K-nearest Neighbours, Decision Tree, Probabilistic neural network, SVM	Ophthalmologist Group	ACC (ARIA) = 95.07% ACC (STARE) = 95.00% ACC (KMC) = 90.19%
Mookiah et al [42]	270 (KMC)	270 (KMC)	Naïve Bayes, K-nearest Neighbours, Probabilistic neural network, SVM	Ophthalmologist Group	ACC = 93.70% SEN = 91.11% SPEC = 96.30%
Acharya [1]	404 (KMC)	517 Dry AMD (KMC) 24 Wet AMD (KMC)	SVM	Ophthalmologist Group	ACC (PSO with SVM) = 85.12% SENS (PSO with SVM) = 87.2% SPEC (PSO with SVM) = 80%


1. Preprocessing
3. Feature Selection
2. Feature Extraction
Feature
4. Training
classifier
6. Testing
5. Classification

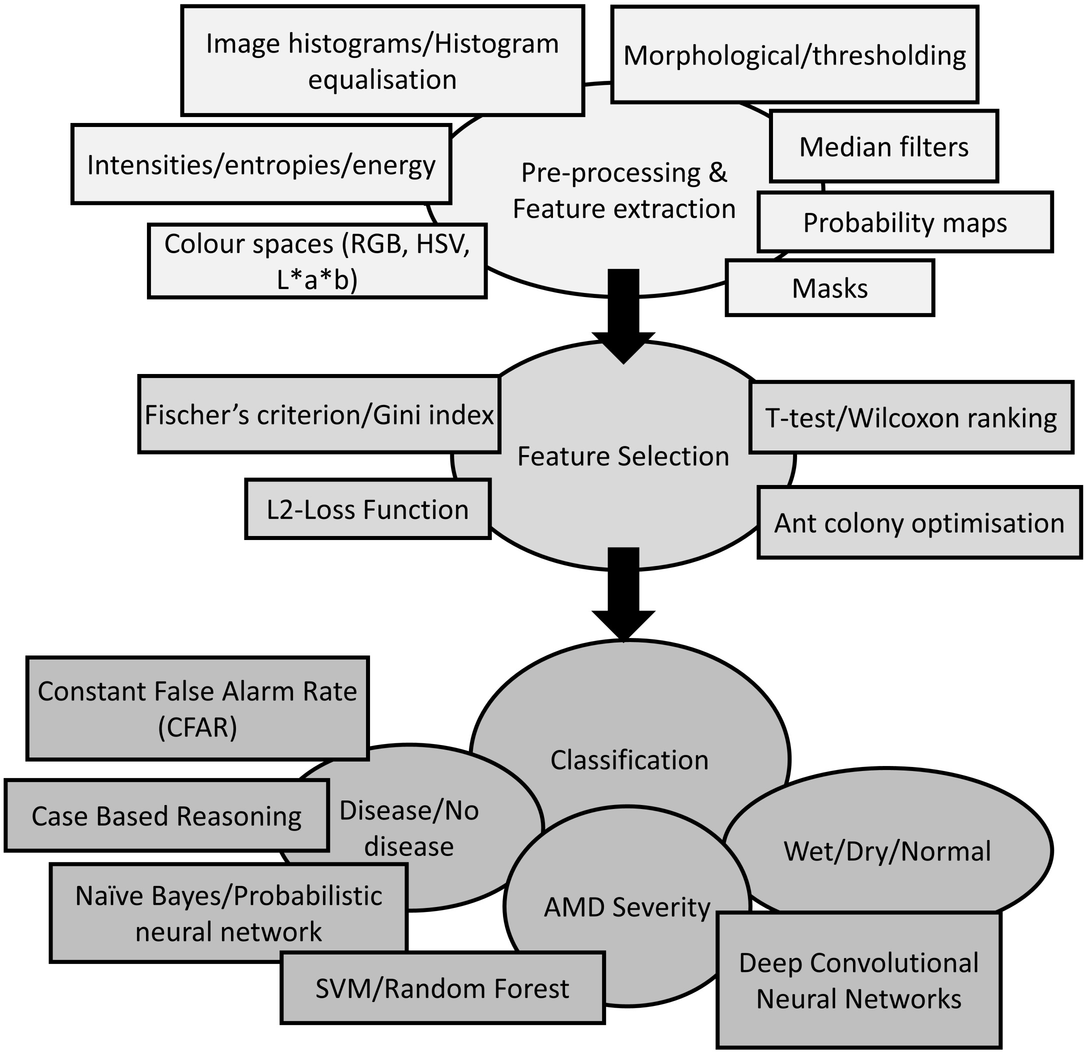


Image histograms/Histogram equalisation
Morphological/thresholding
Intensities/entropies/energy
Median filters
Pre-processing & Feature extraction
Probability maps
Colour spaces (RGB, HSV, L*a*b)
Masks
Fischer's criterion/Gini index
T-test/Wilcoxon ranking
Feature Selection
L2-Loss Function
Ant colony optimisation
Constant False Alarm Rate (CFAR)
Classification
Case Based Reasoning
Disease/No disease
Wet/Dry/Normal
Naïve Bayes/Probabilistic neural network
AMD Severity
SVM/Random Forest
Deep Convolutional Neural Networks