tvst

Article

Deep Learning Applied to Automated Segmentation of Geographic Atrophy in Fundus Autofluorescence Images

Janan Arslan[1,2], Gihan Samarasinghe[3], Arcot Sowmya[3], Kurt K. Benke[4,5], Lauren A. B. Hodgson[1], Robyn H. Guymer[1,2], and Paul N. Baird[2]

[1] Centre for Eye Research Australia, University of Melbourne, Royal Victorian Eye & Ear Hospital, East Melbourne, Victoria, Australia
[2] Department of Surgery, Ophthalmology, University of Melbourne, Parkville, Victoria, Australia
[3] School of Computer Science and Engineering, University of New South Wales, Kensington, New South Wales, Australia
[4] School of Engineering, University of Melbourne, Parkville, Victoria, Australia
[5] Centre for AgriBioscience, AgriBio, Bundoora, Victoria, Australia

Correspondence: Janan Arslan, Centre for Eye Research Australia Ltd, Level 7, 32 Gisborne St, East Melbourne, VIC 3002, Australia.
e-mail: janan.arslan@unimelb.edu.au

Received: August 17, 2020
Accepted: May 23, 2021
Published: July 6, 2021

Keywords: age-related macular degeneration; geographic atrophy; artificial intelligence; deep learning; segmentation; machine learning

Citation: Arslan J, Samarasinghe G, Sowmya A, Benke KK, Hodgson LAB, Guymer RH, Baird PN. Deep learning applied to automated segmentation of geographic atrophy in fundus autofluorescence images. Transl Vis Sci Technol. 2021;10(8):2, https://doi.org/10.1167/tvst.10.8.2

Purpose: This study describes the development of a deep learning algorithm based on the U-Net architecture for automated segmentation of geographic atrophy (GA) lesions in fundus autofluorescence (FAF) images.

Methods: Image preprocessing and normalization by modified adaptive histogram equalization were used for image standardization to improve effectiveness of deep learning. A U-Net–based deep learning algorithm was developed and trained and tested by fivefold cross-validation using FAF images from clinical datasets. The following metrics were used for evaluating the performance for lesion segmentation in GA: dice similarity coefficient (DSC), DSC loss, sensitivity, specificity, mean absolute error (MAE), accuracy, recall, and precision.

Results: In total, 702 FAF images from 51 patients were analyzed. After fivefold cross-validation for lesion segmentation, the average training and validation scores were found for the most important metric, DSC (0.9874 and 0.9779), for accuracy (0.9912 and 0.9815), for sensitivity (0.9955 and 0.9928), and for specificity (0.8686 and 0.7261). Scores for testing were all similar to the validation scores. The algorithm segmented GA lesions six times more quickly than human performance.

Conclusions: The deep learning algorithm can be implemented using clinical data with a very high level of performance for lesion segmentation. Automation of diagnostics for GA assessment has the potential to provide savings with respect to patient visit duration, operational cost and measurement reliability in routine GA assessments.

Translational Relevance: A deep learning algorithm based on the U-Net architecture and image preprocessing appears to be suitable for automated segmentation of GA lesions on clinical data, producing fast and accurate results.

Introduction

Geographic atrophy (GA) is one of two end stages of age-related macular degeneration (AMD)—an age-associated disease of the macula that manifests in those aged 50 years and older. It is responsible for 8.7% of legal blindness globally, affecting approximately 5 million people worldwide, and 9 to 10 million cases are expected by the year 2040.[1,2] The pathogenesis and etiology of AMD and its progression to GA is not completely understood, and drug therapies are currently not available.[3–6] GA is characterized by death of the retinal pigment epithelium (RPE) and photoreceptor cells, as well as loss of the underlying choriocapillaris, which appear as sharply demarcated areas on retinal imaging.[4] Vision loss occurs when atrophic lesions approach the central foveal area, and standard tasks such as reading and recognizing faces become increasingly difficult.[7,8] In the absence of treatment, the condition continues to deteriorate over time, potentially leading to legal blindness (defined as 6/60 vision

in Australia). The rate of irreversible vision loss is also highly variable.[8]

Several imaging modalities are available for the assessment of GA, including fundus autofluorescence (FAF), color fundus photography (CFP), spectral domain optical coherence tomography (SD-OCT), and near-infrared FAF (IR). Currently, the growth of a GA lesion as seen on FAF is accepted as an outcome in clincial intervention trials by the United States Food and Drug Administration[9]; hence, there is much interest and need to accurately define borders and thus lesion size in a timely and efficient manner. Often, FAF is used with semiautomated software to help define the boundaries of a GA lesion.[10] Such software relies heavily on a human user to correctly annotate and identify lesion boundaries within the image.

There are no clinically available complete automation software packages for the extraction of GA information from retinal images. Several groups have, however, developed artificial intelligence (AI)–based automation methods for isolation of GA lesions.[11–30] These studies applied a process known as *semantic segmentation*—the labeling of each pixel within an image and the precise extraction of regions of interest. The range of algorithms tested included region-growing, interactive segmentation using watershed transform, level set approach, geometric active contour model, Fuzzy *c*-means, *k*-nearest neighbor (*k*NN), Chan-Vese model via local similarity factor, convolutional neural networks (CNN), sparse autoencoder deep networks, and an offline/self-learning model. The retinal images used in these GA-AI publications predominantly included FAF imaging. However, SD-OCT, combinations of SD-OCT/FAF, FAF/CFP, and FAF/IR were also used. Among these segmentation algorithms, sensitivity ranged from 0.47 to 0.983, specificity ranged from 0.93 to 0.99, accuracy ranged from 0.42 to 0.995, mean overlap ratio ranged from 0.659 to 0.899, correlation coefficient ranged from 0.82 to 0.998, and the Dice similarity coefficient (DSC) ranged from 0.66 to 0.89.

Although there are many metrics that can be used for the assessment of semantic segmentation algorithms, the DSC is the most suitable because it measures the overlap between machine-generated results and the ground truth (i.e., human annotated images).[31] In the literature, four studies used this metric for evaluation of segmentation algorithms, with results ranging from DSC of 0.66, as reported by Liefers et al.,[32] who described a U-Net–based *encoder-decoder structure,* to a study by Hu et al.,[16] who reported a DSC of 0.89 for the application of the *level set method* on FAF images. In general, an overview of the literature suggests that results from algorithms applied to CFP images produce less promising results when compared with other, more grayscale-based FAF images. For example, sensitivities for CFP-based segmentation algorithms were in the range 0.47 to 0.65; however, for FAF-based algorithms, the range was 0.825 to 0.983. This is due to media opacities and low contrast between atrophic areas and the intact retina in CFPs, which make the detection of GA lesions and their boundaries difficult, even for highly qualified and experienced clinicians and graders.[8,25,33] Furthermore, among the segmentation algorithms already applied in the GA-AI space, a majority of the publications reported relatively small image sample sizes. This was not unusual, given there are many constraints on accessing adequate medical data samples, including challenges on sharing data because of privacy or ethical concerns; the lack of equipment within healthcare systems that makes the sharing of available data challenging; and generally a lack of available cases which could be used to train an AI algorithm, especially in the case of deep learning.[34–36]

In this study, we applied the deep learning U-Net approach to FAF images obtained from GA-affected patients at various stages of progression. The U-Net is a modification of the CNN. It was designed to predict and classify each pixel within an image and thus create a more precise segmentation with fewer training images required.[37] In the past, the U-Net architecture was used for GA segmentation by Wu et al.,[28] who conducted segmentation on SD-OCT and synthesized FAF images, and by Schmidt-Erfurth et al.,[30] who used a residual U-Net model to isolate hyperreflective foci voxels. Liefers et al.[32] described segmentation on CFP using a deep learning model with an encoder-decoder structure with residual blocks, shortcut connections, and contracting-expanding pathways, citing the U-Net developers Ronneberger et al.[37]

Time-series segmentation captures a range of lesion sizes and shapes and provides added variability to reflect real-world clinical settings. For example, it is common for many patients to approach an optometrist or ophthalmologist several months after disease onset, when lesions may have already appeared in a variety of spatial patterns.[38] Therefore an automated segmentation method with the capability of detecting GA lesions at all stages of the disease would be invaluable in a clinical setting.

In this article, the aim was to use the U-Net deep learning approach to isolate lesions in FAF images together with suitable image normalization and preprocessing to address image quality issues. Image segmentation is the first step in the automation of GA assessment because extraction of GA area is required

at patient presentations to chart progression of GA growth over time.

Methods

Study Design and Participants

The study was approved by the Human Research Ethics Committee of the Royal Victorian Eye and Ear Hospital. The study was conducted in accordance with the International Conference on Harmonization Guidelines for Good Clinical Practice and tenets of the Declaration of Helsinki. Ethics approval was provided by the Human Research Ethics Committee (HREC: Project No. 95/283H/15) by the Royal Victorian Eye and Ear Hospital.

Subjects included in this retrospective analysis were AMD participants involved in macular natural history studies from the Centre for Eye Research Australia or from a private ophthalmology practice diagnosed with GA. Cases were referred from a senior medical retinal specialist (R.H.G.) and graded in the Macular Research Unit grading center. Inclusion criteria included being over the age of 50 years, having a diagnosis of AMD (on the basis of the presence of drusen greater than 125 μm) with progression to GA in either one of both eyes. An atrophic lesion was required to be in the macular and not extend beyond the limits of the FAF image at the first visit (i.e., baseline). Participants were required to have foveal centered FAF images and at least three visits recorded over a minimum of two years, with FAF imaging of sufficient quality. Good-quality images were classified as those with minimal or correctable artefacts (e.g., by correction of illumination with pre-processing techniques), and images should encompass the entire macular area and part (i.e., around half) of the optic disc. No minimum lesion sizes were set, because the objective of the study was to be able to automate all lesion sizes. Images contained both unifocal and multifocal lesions. Sampling in the training phase for the algorithm was augmented by time-series segmentation without limitation on lesion size.

Exclusion criteria included participants with neovascular AMD and macular atrophy from causes other than AMD, such as inherited retinal dystrophies, including Stargardt's disease. These patients were excluded based on the determination of a retinal specialist (R.H.G.). Also excluded were patients who had undergone any prior treatment or participated in a treatment trial for AMD. Peripapillary atrophy was not included in the analysis and all participants required atrophy in the FAF image to be included. Poor-quality images were excluded and were classified as images that were not salvageable with preprocessing techniques (e.g., excessive blurriness, shadowing, and contrast issues); images where the optic disc was completely absent; and images where the optic disc was in the center of the image.

FAF images were captured using the Heidelberg Spectralis-OCT (Heidelberg Engineering, Heidelberg, Germany). FAF image files, along with basic demographic data were retrospectively collected in Tagged Image File Format (i.e., TIFF or TIF), and original sizes of images were either 768 × 768 or 1536 × 1536 pixels with 30° × 30° field-of-view. As images were collected retrospectively and from real-time clinical settings, automatic real-time tracking ranged from five to 100.

Outputs from the accompanying software—RegionFinder—were used to compare the machine-generated outputs with the gold standard (i.e., ground truth based on human graders). This included obtaining measurements, such as the total area of growth (mm^2). Additionally, manually drawn annotations (using Wacom Cintiq Pro 13 drawing tablet), separate from RegionFinder, were used to train the AI algorithm. Two graders were involved in the annotation process—a principal grader and a senior grader—for a subset of images to check for consistency. The intraclass correlation coefficient was used as a measure of consistency between the graders.

Image Normalization and Preprocessing

Deep learning algorithms require very large datasets for training on raw images to cover for image acquisition problems that can affect performance. This potential challenge can be addressed by image preprocessing for image standardization, which can greatly reduce the sample size necessary. Problems during FAF image acquisition include illumination (poor uniformity of intensity in the image plane), blurred vision (from involuntary eye movements), physical discomfort (from viewing the blue-light beam), and dark contrasts or "shadowing" (because of vitreous opacities, incorrect adjustments of the camera, or the position of the patient relative to the camera). According to the Heidelberg Engineering's HRA+OCT Spectralis Manual, the built-in real-time eye-tracking system was designed to minimize eye movement artefacts. The manual recommends obtaining and averaging between six to 24 scans to obtain a good-quality FAF image.[39] Poor image quality may create a misinterpretation or might even render the FAF image uninterpretable. Although standardized protocols like the one stipulated by Heidelberg Engineering do exist, noises

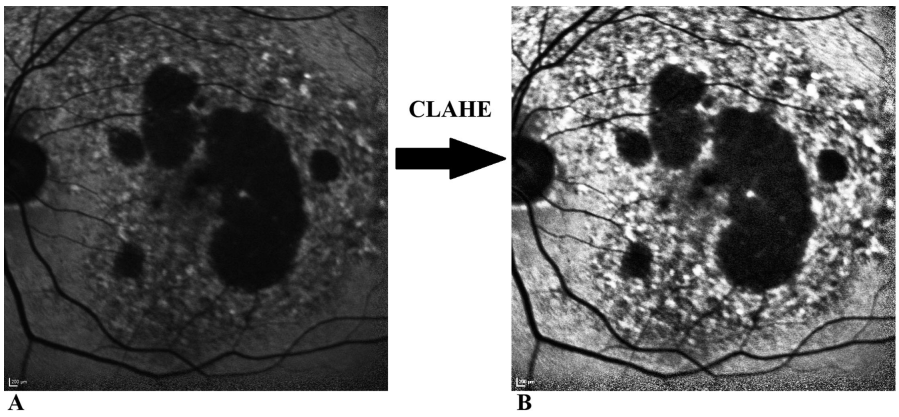

Figure 1. Image preprocessing using the CLAHE technique. (**A**) Original FAF image and (**B**) CLAHE-applied FAF image. The original image was very dark in contrast and illumination. Using the original image would have resulted in contrast-related errors, making it more difficult for the algorithm to distinguish the GA lesions. The CLAHE-applied image shows the dramatic improvement using this technique. CLAHE, Contrast Limited Adaptive Histogram Equalization; FAF, Fundus autofluorescence.

and artefacts may still appear during image acquisition.[40,41]

It is possible to correct for camera characteristics or other issues with image acquisition quality using preprocessing techniques.[42] For this dataset, images with extreme artefacts that could not be salvaged with image preprocessing techniques were removed from the dataset. Extreme artefacts here were defined as the presence of overwhelming artefacts that skewed the extraction of information from the image, such as extreme darkness, blurriness or graininess. For the remaining set of images, residual artefacts were removed using the Contrast Limited Adaptive Histogram Equalization (CLAHE) technique.[43] The CLAHE corrects for different illumination and contrast conditions, as well as improving the edges of objects within the image.[43] Figure 1 illustrates the conversion of an image using CLAHE. Images in their original form (Fig. 1a) can, technically, be trained on a U-Net architecture. However, the resultant model would be flawed, producing contrast-related errors and outputs that do not accurately isolate lesions. By using CLAHE, it is easier to train the algorithm and for the algorithm to produce high quality outputs in its prediction of lesion areas.

Learning Algorithm

A deep learning model was developed using the U-Net architecture together with appropriate image contrast normalization, hyperparameters and training data. The basic U-Net architecture is illustrated in Figure 2 and consists of contracting (left side) and expansive (right side) pathways. The foundation of the architecture is the Fully Convolutional Network. For the contracting pathway, there is a repetitive pattern of two 3×3 convolutions, a rectified linear unit (ReLU) and a 2×2 max pooling operation. At every downsampling step, the number of convolution filters is doubled from the previous step (e.g., step 1 begins with 64 filters, which increases to 128 by step 2). For the expansive pathway, each repeated upsampling step consists of 2×2 convolutions that halve the number of filters, a concatenation with the correspondingly cropped feature map from the contracting path, and finally two 3×3 convolutions followed by a ReLU. A 1×1 convolution is used at the final layer to map each 64-component feature vector to the desired number of classes. The contracting and expansive pathways form a "U" shape, thus aptly giving this architecture its name.[37] In our implementation, the ReLU-aware *He Normal* initialization was used. *He Normal* derives from the research of Glorot and Bengio,[44] who used a scaled uniform distribution for initialization and assumed activations are linear. Proposed by He et al.,[45] this initializer is considered to be more sound for ReLU activation, and involves each layer's weight being initialized in accordance with the size of previous layers.

The adaptive learning rate optimization algorithm, ADAM, was used for stochastic gradient descent and was used with a learning rate of $LR = 3 \times 10^{-5}$. The ADAM optimizer includes bias corrections on both first- and second-order moments.[46] The learning rate is an optimization hyperparameter that adjusts the weight of the algorithm during training. The learning rate was evaluated by assessment of the Dice loss and quality of segmentation outputs. For the dataset, we found larger learning rates, such as $LR = 10^{-4}$,

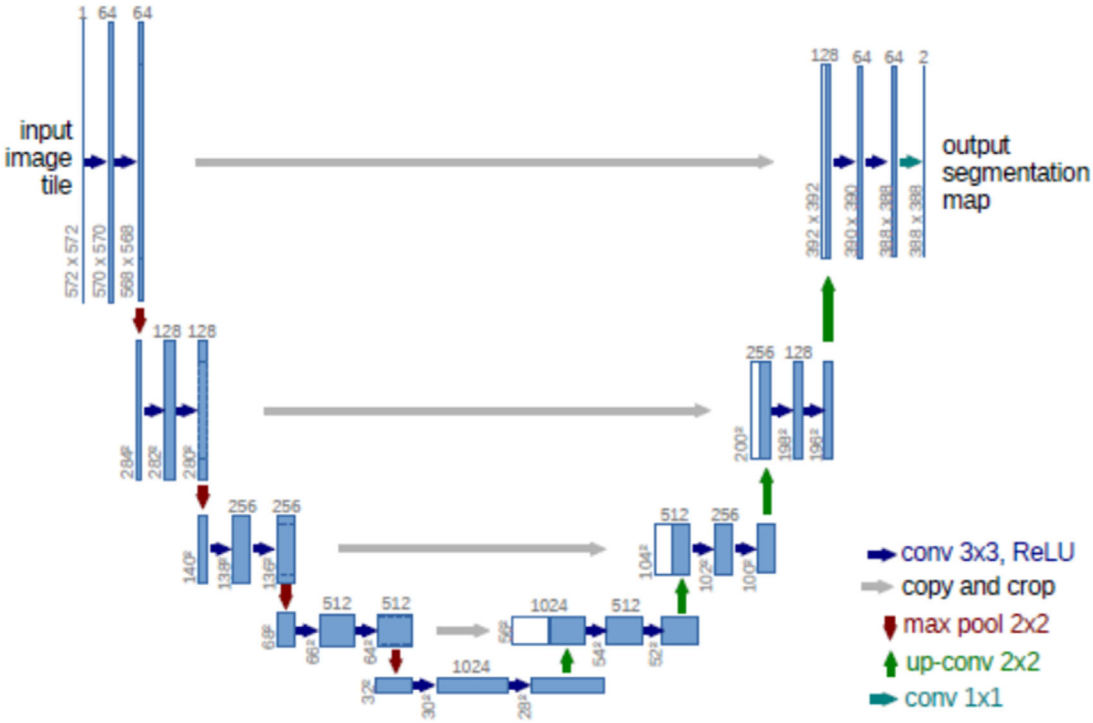

Figure 2. U-Net architecture. The U-Net architecture was created for biomedical image segmentation. The U-Net is aptly named because of the arrangements of the filters in a "U" shape. The contracting pathway (left) consists of two 3 × 3 convolutions, ReLU activation and 2 × 2 max pooling. The expansive pathway (right) consist of 2 × 2 convolution, a concatenation with the correspondingly cropped feature map from the contracting pathway, two 3 × 3 convolutions and ReLU activation (see Ronneberger et al.[37]).

produced suboptimal outputs that did not distinctly characterize GA lesion areas.

The batch size was set to 32, the steps per epoch was set to 175, and the number of epochs selected was 80. Batch sizes typically range from 32 to 512. Note that U-Net designers favor small batch sizes to minimize overhead and maximize graphics processing unit memory.[37] The regularization effect of small batch sizes contributes to the ability to generalize.[47,48] We found 80 epochs to be sufficient in reaching peak model performance and minimizing loss. Learning curves were created using Python's ggplot (http://ggplot.yhathq.com/).

The hardware implementation of the U-Net was carried out on an operating system with an Intel Core i7-7820HQ CPU @ 2.90GHz. All training, testing, and statistics were performed using Keras (https://keras.io/) and Tensorflow (https://www.tensorflow.org/) using NVIDIA Quadro M1200 Graphics Processing Unit.

Training and Validation

The algorithm performance was evaluated by fivefold cross-validation,[49] which is widely used in classification for model assessment and provides estimates of error rates by rotating through subsets of training data and test data (Fig. 3).[50–54]

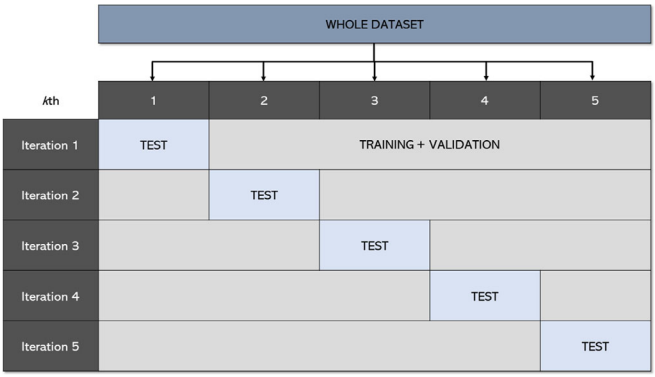

Figure 3. Fivefold cross-validation. This cross-validation was chosen because it has been empirically shown to yield test error rates not excessively influenced by high bias and variances.

Performance Metrics

The following metrics were used for evaluating the performance of the U-Net algorithm: the DSC, DSC loss, sensitivity, specificity, mean absolute error (MAE), accuracy, recall, and precision.[19,55,56]

Bland-Altman plots and coefficient of repeatability (CR) were used to measure the difference between ground truth and segmentation results both visually and numerically, respectively.[57,58] Bland-Altman plots were created using Python's pyCompare (https://pypi.org/project/pyCompare/) and the unit of measure used was pixels. We further compared ground truth with automation segmentation using Spearman's correlation coefficient (ρ) and plotted an appropriate regression line using Python's ggplot.

Although DSC was our primary focus as the metric for segmentation performance, the additional evaluation metrics included were used so that results were comparable with other GA-AI findings. For example, in semantic segmentation, a lower sensitivity would suggest undersegmentation where lesion boundaries would not be captured in detail. Conversely, a lower specificity could indicate oversegmentation where lesions are resolved with too much detail possibly caused by noise or artefacts.

The metrics were computed from pixel-level values for true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). In the context of this study, TP was defined as correctly segmented GA lesion pixels, TN was the correctly identified background pixels, FP was the background pixels mistakenly segmented as GA lesion pixels, and FN was the GA lesion pixels mistakenly identified as background pixels.

The DSC is a spatial overlap index and a validation metric for reproducibility. It measures the agreement between results obtained using ground truth (such as human annotation) and machine-predicted results.[19,55,56]

$$\text{DSC} = \frac{2TP}{2TP + FP + FN} \tag{1}$$

DSC loss is simply denoted as

$$\text{DSC}_{\text{loss}} = 1 - DSC \tag{2}$$

Sensitivity (also known as recall) is defined as the proportion of TP pixels found within the lesions.[56]

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

Specificity is the measure of diagnostic test accuracy and is defined as the proportion of TN pixels found within the background.[56]

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4}$$

The MAE measures closeness of predictions (observed vs. predicted) and is expressed as

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - x_i| \tag{5}$$

where y_i is the prediction and x_i is the true value.[59]

The accuracy of the algorithm is its ability to distinguish different classes (i.e., GA lesion pixel or background pixel).[60]

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Precision represents the proportion of pixels correctly classified as GA lesions.[60]

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

Finally, the coefficient of repeatability (CR) measured the difference between ground truth and automated segmentation outcomes.[57,58]

$$CR = 1.96 \times \sqrt{\frac{\sum (d_2 - d_1)^2}{n}} \tag{8}$$

These metrics were evaluated for every cross-validation fold and compared predicted outcomes to that of human-annotated ground truths. The metrics DSC, sensitivity, specificity, accuracy, and precision all have an outcome range of 0 to 1; the closer to 1 the result, the better the outcome. Conversely, the metrics DSC_{loss} along with MAE should ideally be as close to 0 as possible to indicate that loss and error is minimized.

Qualitative Assessment

While human subjectivity should be accounted for, combining a qualitative assessment along with a quantitative assessment strengthens the evaluation of model prediction. For the purposes of this study, qualitative assessment involves (a) the speed of the algorithm as compared to its human counterpart, and (b) the human visually evaluating machine-generated outcomes and determining whether outputs graphically appear to be accurate.

Results

A total of 702 FAF images from 51 patients with GA secondary to AMD were included in the study, and whose images were manually annotated. The cohort of images and patients was quite large and diverse

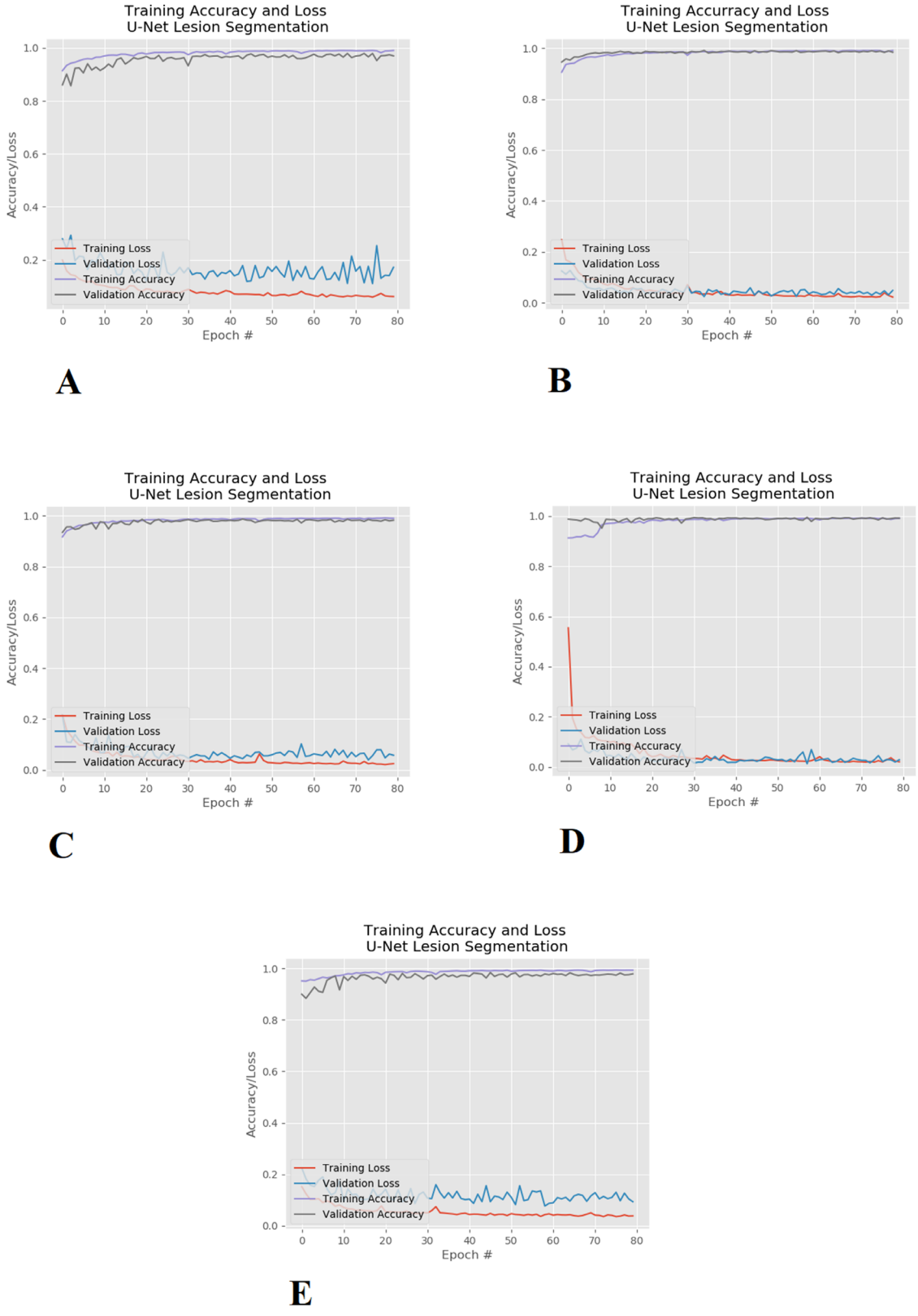

Figure 4. Learning curves with training/validation loss and accuracy across all fivefolds of cross-validation. (**A**) Cross-validation 1. (**B**) Cross-validation 2. (**C**) Cross-validation 3. (**D**) Cross-validation 4. (**E**) Cross-validation 5. The learning curve illustrates a consistent outcome of high accuracy and low loss throughout all fivefolds.

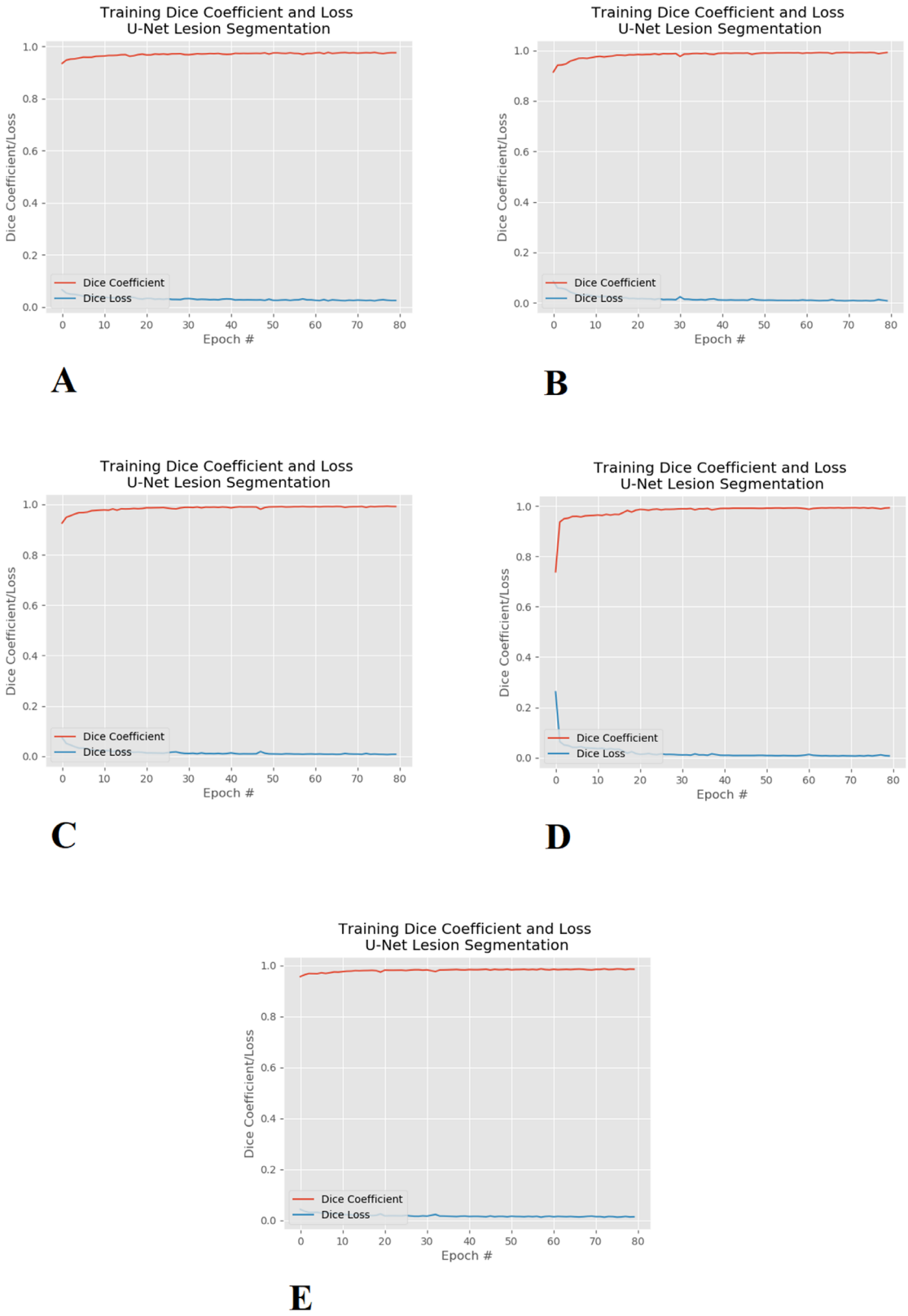

Figure 5. Learning curves for DSC and DSC_{loss} across all fivefolds of cross-validation. (**A**) Cross-validation 1. (**B**) Cross-validation 2. (**C**) Cross-validation 3. (**D**) Cross-validation 4. (**E**) Cross-validation 5. The learning curve illustrates a consistent outcome of high DSC and low loss throughout all 5-folds. DSC, Dice similarity coefficient.

Table 1. Geographic Atrophy U-Net Training Results

	DSC	DSC_{loss}	Sensitivity	Specificity	MAE	Accuracy	Precision
Training Set 1	0.9752	0.0248	0.9949	0.8904	0.0443	0.9904	0.9943
Training Set 2	0.9918	0.0082	0.9948	0.8925	0.0141	0.9903	0.9939
Training Set 3	0.9916	0.0084	0.9951	0.884	0.0147	0.99	0.9934
Training Set 4	0.9931	0.0069	0.9958	0.8913	0.0122	0.9917	0.9949
Training Set 5	0.9855	0.0145	0.9967	0.785	0.0267	0.9935	0.9963
Mean ± SD	0.9874 ± 0.0067	0.0126 ± 0.0067	0.9955 ± 0.0007	0.8686 ± 0.0419	0.0224 ± 0.0121	0.9912 ± 0.0013	0.9946 ± 0.0010

SD, standard deviation.

Table 2. Geographic Atrophy U-Net Validation Results

	DSC	DSC_{loss}	Sensitivity	Specificity	MAE	Accuracy	Precision
Validation Set 1	0.9512	0.0488	0.9901	0.8529	0.0818	0.9702	0.974
Validation Set 2	0.9887	0.0113	0.9887	0.871	0.0207	0.9838	0.9935
Validation Set 3	0.9869	0.0131	0.9938	0.6055	0.0232	0.9824	0.9864
Validation Set 4	0.9947	0.0053	0.998	0.5557	0.0103	0.9921	0.9938
Validation Set 5	0.9678	0.0322	0.9932	0.7452	0.0569	0.979	0.9832
Mean ± SD	0.9779 ± 0.0161	0.0221 ± 0.0161	0.9928 ± 0.0032	0.7261 ± 0.1273	0.0386 ± 0.0267	0.9815 ± 0.0071	0.9862 ± 0.0073

SD, standard deviation.

Table 3. Geographic Atrophy U-Net Test Results

	DSC	DSC_{loss}	Sensitivity	Specificity	MAE	Accuracy	Precision
Test Set 1	0.9835	0.01645	0.9937	0.8504	0.03079	0.9883	0.9938
Test Set 2	0.9916	0.0084	0.9904	0.8613	0.0160	0.9878	0.9970
Test Set 3	0.9783	0.0217	0.9928	0.7207	0.0390	0.9666	0.9706
Test Set 4	0.9820	0.0180	0.9824	0.7104	0.0309	0.9738	0.9880
Test Set 5	0.9548	0.0452	0.9922	0.6060	0.0712	0.9703	0.9693
Average ± SD	0.9780 ± 0.0124	0.0220 ± 0.0124	0.9903 ± 0.0041	0.7498 ± 0.0955	0.0376 ± 0.0184	0.9774 ± 0.0090	0.9837 ± 0.0116

SD, standard deviation.

as compared to others in the GA-AI segmentation space.[11] The cohort consisted of 99 eyes, 49 left eyes (49.5%) and 50 right eyes (50.5%). A total of 359 images were for the left eye and 343 images were for the right eye. The cohort consisted of 38 females (74.5%) and 13 males (25.5%) with an average age of 76.7 ± 8.9 years. Total follow-up time was 61.5 ± 25.3 months. The intraclass correlation coefficient for consistency between the two graders was 0.9855 (95% confidence interval [CI]: 0.9298, 0.9971), showing close agreement between the graders. To suit the requirements of the cross-validation, the images were divided into four parts of 140 images and one part of 142 images by random allocation, with a mix of fast and slow progressors.

Learning curves across all fivefolds with training/validation loss and accuracy are presented in Figure 4 and for DSC and DSC_{loss} in Figure 5. For quantified training outcomes (Table 1), DSC ranged from 0.9752 to 0.9931, DSC_{loss} ranged from 0.0069 to 0.0248, sensitivity ranged from 0.9948 to 0.9967, specificity ranged from 0.785 to 0.8925, MAE ranged from 0.0122 to 0.0443, accuracy ranged from 0.99 to 0.9935, and precision ranged from 0.9934 to 0.9963.

For quantified validation outcomes (Table 2), DSC ranged from 0.9512 to 0.9947, DSC_{loss} ranged from 0.0053 to 0.0488, sensitivity ranged from 0.9887 to 0.998, specificity ranged from 0.5557 to 0.871, MAE ranged from 0.0103-0.0818, accuracy ranged from 0.9702 to 0.9921, and precision ranged from 0.974 to 0.9938.

For quantified test outcomes (Table 3), DSC ranged from 0.9548 to 0.9916, DSC_{loss} ranged from 0.0084 to 0.0452, sensitivity ranged from 0.9824 to 0.9937, specificity ranged from 0.6060 to 0.8613, MAE ranged from 0.0160 to 0.0712, accuracy ranged from 0.9666 to 0.9883, and precision ranged from 0.9693 to 0.9970.

The Bland-Altman plots and CRs (Fig. 6) illustrate graphically that there are minimal differences between the ground truth and segmentation results. The Bland-Altman plots the difference between the ground truth and segmentation output measurements *vs.* the mean of the two measurements. The Bland-Altman showed a bias of (A) −1238.05 pixels (95% CI agreement: −5052.40, 2576.31), (B) −615.99 pixels (95% CI

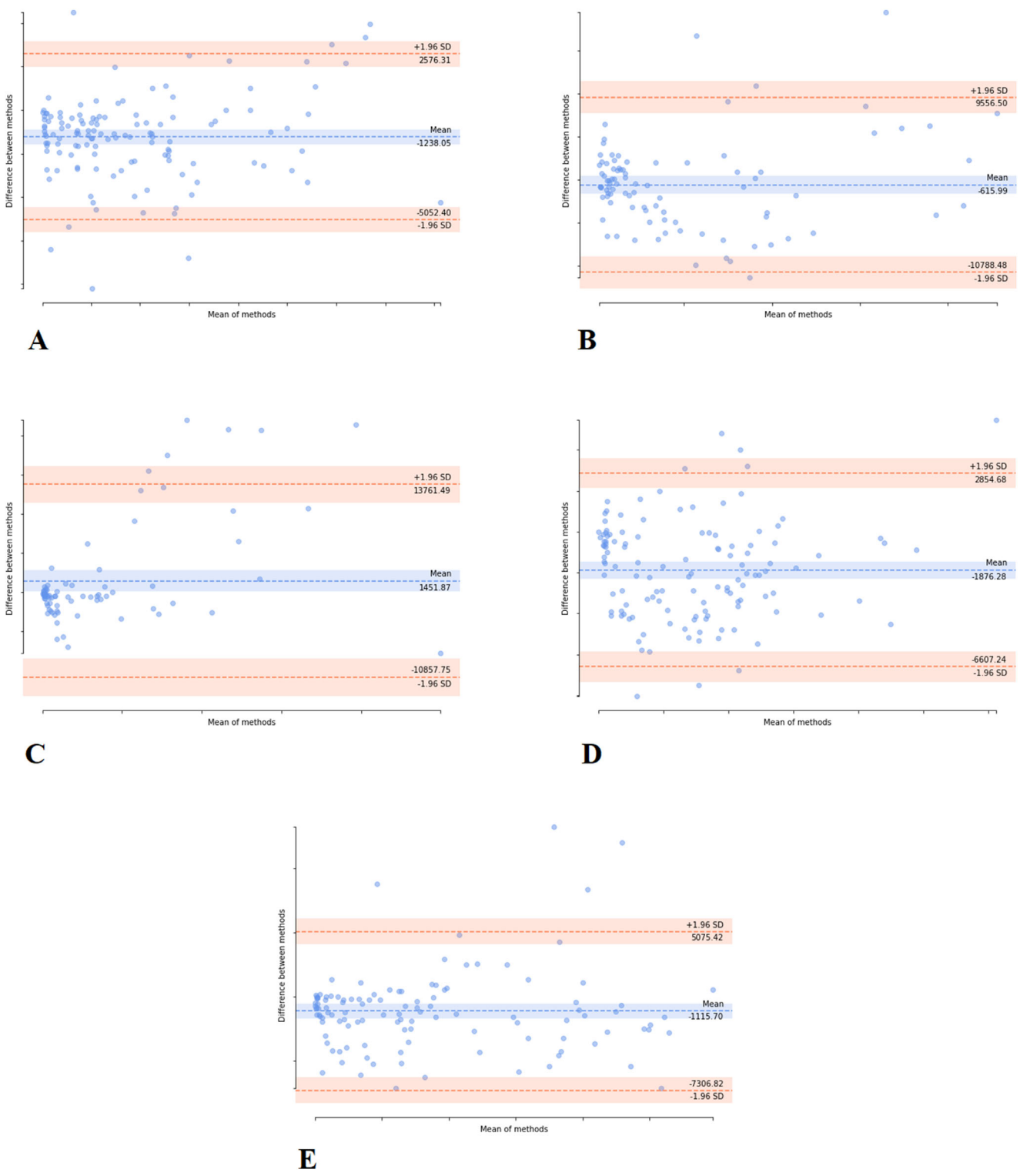

Figure 6. Bland-Altman plots and coefficient of repeatability across all fivefolds of cross-validation (in units of pixels).

agreement: −10788.48, 9556.50), (C) 1451.87 pixels (95% CI agreement: −10857.75, 13761.49), (D) −1876.28 pixels (95% CI agreement: −6607.24, 2854.68), and (E) −1115.70 pixels (95% CI agreement: −7306.82, 5075.42). The coefficients of repeatability were (A) 4520.79 pixels (95% CI: 4030.03, 5148.73), (B) 10243.88 pixels (95% CI: 8941.18, 11994.43), (C) 12634.26 pixels (95% CI: 10961.46, 14914.32), (D.)

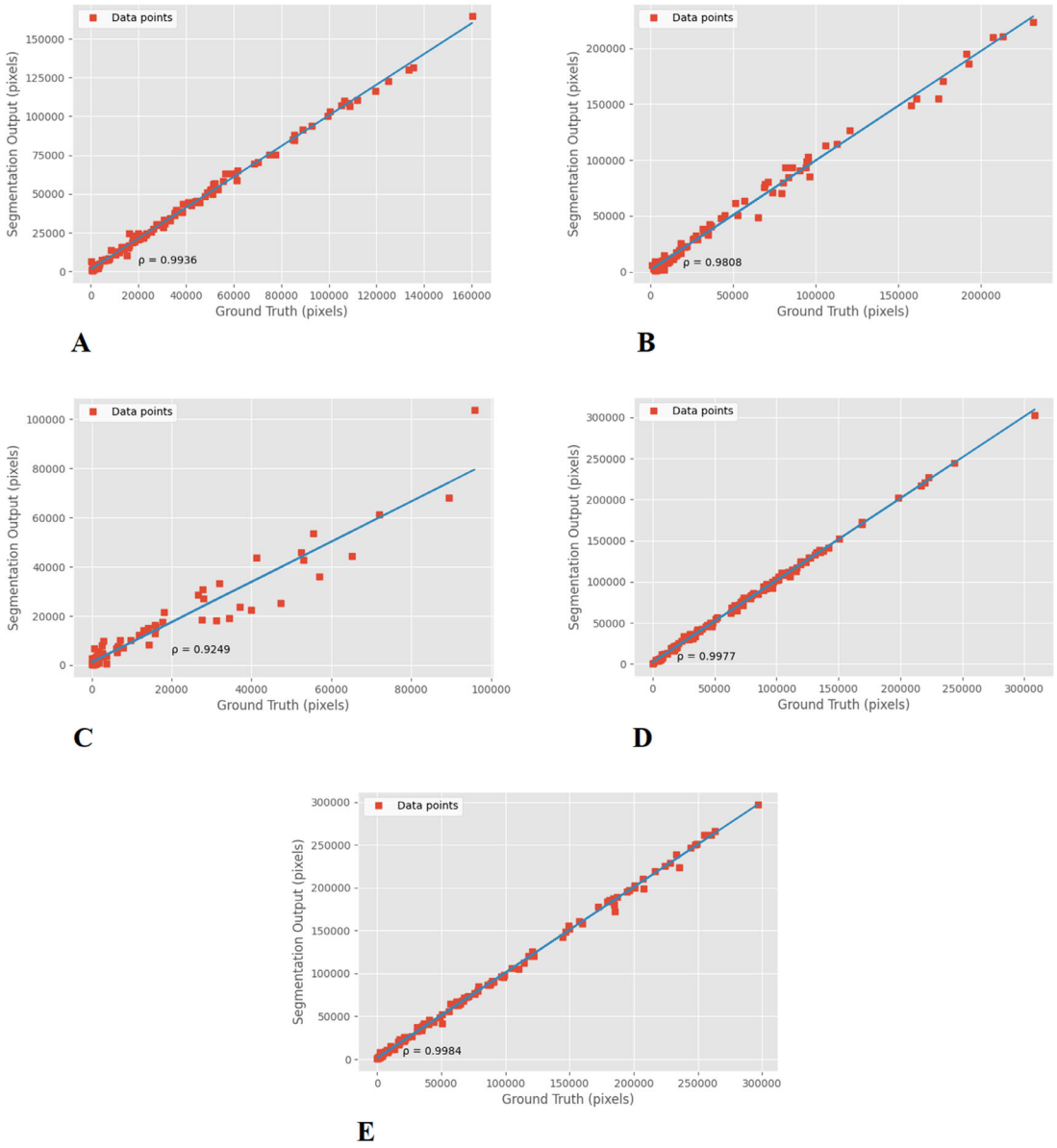

Figure 7. Spearman's correlations and regression lines across all fivefolds of cross-validation (in units of pixels).

5992.17 pixels (95% CI: 5334.85, 6835.69), (E) 6565.97 pixels (95% CI: 5818.85, 7534.94).

The Spearman's correlation coefficient and regression line (Fig. 7) reveals that there is a strong positive correlation between ground truth and segmentation measurements. The Spearman's correlation coefficients were (A) $\rho = 0.9936$ ($P < 0.001$), (B) $\rho = 0.9808$ ($P < 0.001$), (C) $\rho = 0.9249$ ($P < 0.001$), (D) $\rho = 0.9977$ ($P < 0.001$), (E) $\rho = 0.9984$ ($P < 0.001$).

Further to the quantifiable results of the algorithm, we evaluated visually the outputs generated by the U-Net to confirm that lesions were being extracted accurately. Figures 8 and 9 illustrate four sample cases of the U-Net GA lesion output of preprocessed FAF images. The presented cases showed extremely well-outlined lesions. The time in which GA lesions were extracted was compared between humans and the automation method. The average time it took a human, using RegionFinder, to annotate GA lesions was on average 1.04 minutes across all 702 images. The U-Net GA automation takes 6.06 seconds. The qualitative and quantitative assessment coupled illustrate a good performance of the algorithm.

In the current study, the 702 images from 51 persons are comparable with other recent studies using deep learning for GA image segmentation. See, for example, Liefers et al.,[32] who used 409 images for model development and evaluation from two cohorts: 87 images

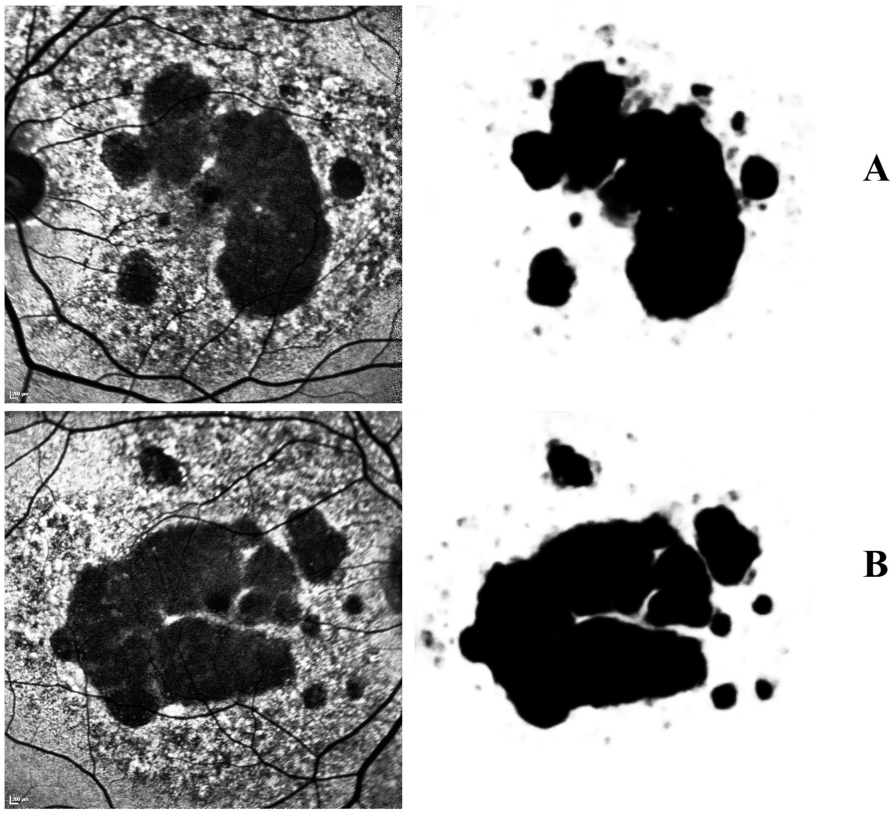

Figure 8. Qualitative assessment of model prediction outcomes. Test cases **A** and **B**. In addition to assessing the performance of U-Net quantitatively, we evaluated the performance by visually assessing the degree of accuracy of U-Net-based lesion segmentation. The test cases presented demonstrate a good segmentation outcome.

Table 4. Geographic Atrophy U-Net Patient-Level Segmentation Results

	DSC	DSC_{loss}	Sensitivity	Specificity	MAE	Accuracy	Precision
Training	0.9931	0.0069	0.9953	0.9634	0.0116	0.9919	0.9951
Validation	0.9892	0.0108	0.9895	0.9318	0.0192	0.9831	0.9916
Test	0.9793	0.0207	0.9792	0.8041	0.0382	0.9633	0.9818

from the BMES study (26 participants, 43 eyes) and 322 images from the RS study (149 participants, 195 eyes). The average time patients were observed in our cohort was quite long (61.5 months), which increased sample variability for model training.

The possibility of overfitting was addressed by (a) appropriate selection of hyperparameters, (b) early stopping in training, (c) batch size selection, and (d) its absence confirmed by results from the fivefold cross-validation—which would have revealed significant disparities between training and test results if the model was overfitted. For example, good results for the training phase, but poor results for the testing phase would be indicative of overfitting. Segmentation and cross-validation were carried out at the patient-level in addition to the image-level, using one image per eye for 99 eyes (Table 4). This is an additional check against overfitting, subject to certain statistical assumptions.[61,62] We found similar outputs and results with patient-level and image-level segmentation. No evidence of overfitting was found.

The current study and experimental results included a number of features that, in combination, distinguish it from other studies: (a) it is a retrospective case study, under clinical conditions, (b) application of the U-Net deep learning architecture to GA segmentation with hyperparameters tuned to a new set of clinical data, (c) use of FAF imagery from the Heidelberg Spectralis instrumentation, and (d) data preprocessing using an optimal normalization process based on CLAHE. This combination of features does not appear to have been reported elsewhere in the research literature.

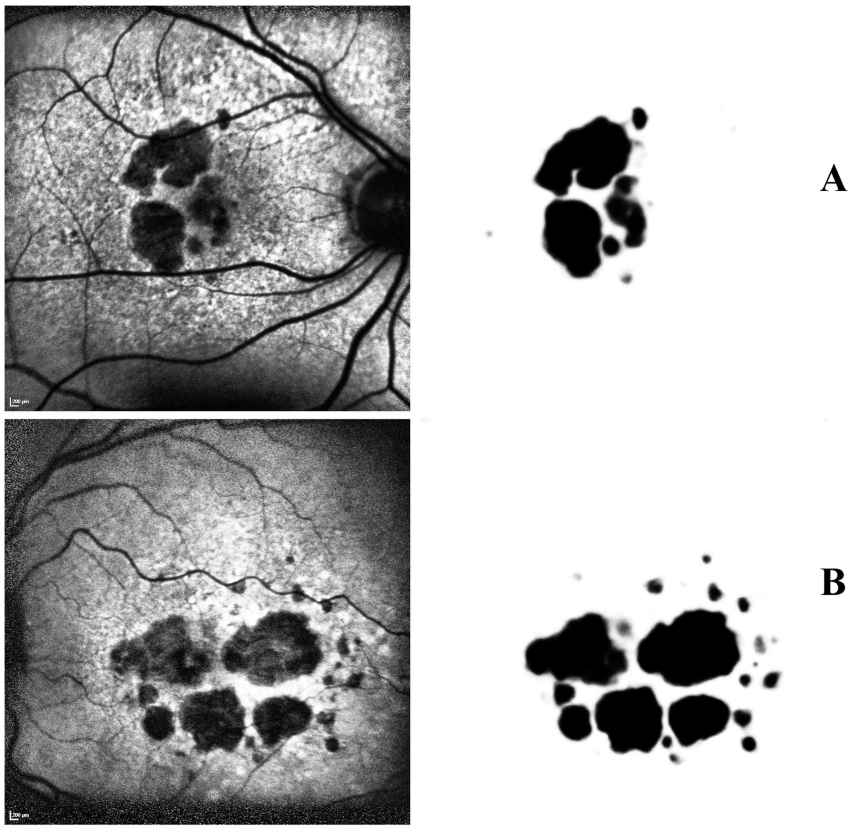

Figure 9. Qualitative assessment of model prediction outcomes. Test cases **A** and **B**. In addition to assessing the performance of U-Net quantitatively, we evaluated the performance by visually assessing the degree of accuracy of U-Net-based lesion segmentation. The test cases presented demonstrate a good segmentation outcome.

Discussion

The automation of GA lesion segmentation is described using a deep learning algorithm based on the U-Net architecture. The algorithm was assessed both qualitatively and quantitatively. Quantitatively, performance was evaluated against the metrics DSC, DSC loss, sensitivity, specificity, MAE, accuracy, recall, and precision. The U-Net performance on FAF images coupled with preprocessing was successful in GA lesion segmentation. Training, validation, and testing scores were very high, particularly for the main metric of interest—the DSC—where the average DSC for training, validation and testing was 0.9874, 0.9779, and 0.9780, respectively. The highest DSC reported in the literature for GA-AI semantic segmentation was 0.89 for the study by Hu et al.,[16] using a level set method on FAF images. The test DSC score of 0.9780 ± 0.0124 produced here compares favorably with other U-Net-based algorithms. Wu et al.[28] reported a DSC score of 0.872 ± 0.66 on SD-OCT and synthesized FAF images, whereas Liefers et al.[32] reported an average DSC score of 0.72 ± 0.26 on CFP. Differences in performance can be ascribed to differences in data quality, different settings for hyperparameters, and the design of image normalization methods.

When comparison is made on the basis of *accuracy* (the more commonly used outcome evaluation), the algorithm compares favorably with scores reported by Hu et al.[20] (i.e., 0.97 accuracy using *k*NN with FAF images) and Ji et al.[28] (i.e., 0.986 and 0.995 accuracies using sparse autoencoders with SD-OCT scans). Our accuracies were 0.9912, 0.9815, and 0.9774 for training, validation, and testing, respectively. Similarly, the sensitivity of the algorithm was very good, with values of 0.9955, 0.9928, and 0.9903 for training, validation and testing, respectively. The highest sensitivity performance reported in the literature was 0.983, by Lee et al.[13] (i.e., watershed transform algorithm with FAF images). In this study, qualitative evaluation by *visual assessment* of machine-generated outputs provided augmentation of scores based on objective metrics. Visually, the U-Net GA automation appears to capture all lesions visually accessible to the human grader. The speed of the automation was far greater

than the speed of human judgement. For the 702 images in the dataset, the speed of automation was ~6.06 seconds to complete each task, whereas the human grader averaged 1.04 minutes.

Metric values (whether DSC, sensitivity, specificity or accuracy) of 0.7–0.8 are typically considered acceptable, whereas 0.8 to 0.9 are considered excellent. Specificities for GA-AI segmentation algorithms ranged from 0.93 to 0.99 in the literature. For this algorithm, the specificities were 0.8686, 0.7261, and 0.7498 for training, validation and testing, respectively. With the results for specificity indicating good performance rather than excellent, there may be a possibility of oversegmentation occurring. Oversegmentation may be due to correctly detecting boundaries of interest within an image (lesion boundaries), but also insignificant boundaries as well. Visually, this may appear as the segmented areas being split up more than necessary. This is in contrast to undersegmentation, where individual segments are merged into singular segments. Future work to address this issue could involve combining the U-Net segmentation algorithm with other AI tools, such as texture discrimination, for improved resolution in spatial analysis.

Some limitations in the study included constraints on the use of the FAF imaging modality, such as image artefacts (e.g., blurriness, shadowing, and poor contrast), discomfort for the patient associated with the blue-light beam, low signal strength, and its potential for toxicity for the retina.[41] Preprocessing techniques were used to standardize FAF images to address some of these issues. Augmentation of machine learning performance in future may be possible by including other imaging modalities, such as gray-scale SD-OCT images during training, which have been used in the past to quantify atrophy of photoreceptors.[63–65] In the current study, the cross-validation approach was used to evaluate performance under controlled experimental conditions. In future, more extensive training and application to external datasets are possible using a model-to-data approach, also known as federated learning, as demonstrated by Mehta et al.[36,66] This process involves exporting the partially trained deep learning model to different institutions for incremental training, while preserving local data privacy.

Conclusion

Estimation of GA area in FAF images is required to evaluate the severity and rate of progression of GA in clinical presentations. To automate this process, a deep learning approach was developed for semantic segmentation and applied to FAF images, with image preprocessing and normalization by CLAHE. The algorithm produced very high accuracy with very high DSC scores, matching or exceeding human performance in all metrics.

The automation results presented in this article are based on application to clinical data and therefore provide support that the algorithm is suitable for application in clinical settings. Automation of diagnostics for GA has the potential to provide savings with respect to patient visit duration, operational cost, and measurement reliability in routine GA assessments.

Acknowledgments

Supported by the National Health & Medical Research Council of Australia (NHMRC) Senior Research Fellowship 1138585 for (P.N.B.); NHMRC Fellowship GNT1103013 (R.H.G.); and RB McComas Research Scholarship in Ophthalmology from the University of Melbourne (J.A.).

Disclosure: **J. Arslan,** None; **G. Samarasinghe,** None; **A. Sowmya,** None; **K.K. Benke,** None; **L.A.B. Hodgson, R.H. Guymer,** Bayer (C), Novartis (C), Roche Genentech (C), Apellis (C); **P.N. Baird,** None

References

1. Resnikoff S, Pascolini D, Etya'ale D, et al. Global data on visual impairment in the year 2002. *Bull World Health Organ*. 2002;82:844–851.
2. Wong WL, Su X, Li X, et al. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *Lancet Global Health*. 2014;2:e106–e116.
3. Boyer DS, Schmidt-Erfurth U, van Lookeren Campagne M, Henry EC, Brittain C. The pathophysiology of geographic atrophy secondary to age-related macular degeneration and the complement pathway as a therapeutic target. *Retina*. 2017;37:819–835.
4. Holz FG, Strauss EC, Schmitz-Valckenberg S, van Lookeren Campagne M. Geographic atrophy: clinical features and potential therapeutic approaches. *Ophthalmology*. 2014;121:1079–1091.
5. Solomon SD, Lindsley K, Vedula SS, Krzystolik MG, Hawkins BS. Anti-vascular endothelial growth factor for neovascular age-related macular degeneration. *Cochrane Database Syst Rev*. 2014;8(8):CD005139.

6. Hobbs RP, Bernstein PS. Nutrient Supplementation for Age-related Macular Degeneration, Cataract, and Dry Eye. *J Ophthalmic Vis Res*. 2014;9:487–493.
7. Bhutto I, Lutty G. Understanding age-related macular degeneration (AMD): relationships between the photoreceptor/retinal pigment epithelium/Bruch's membrane/choriocapillaris complex. *Mol Aspects Med*. 2012;33:295–317.
8. Fleckenstein M, Mitchell P, Freund KB, et al. The progression of geographic atrophy secondary to age-related macular degeneration. *Ophthalmology*. 2018;125:369–390.
9. Csaky KG, Richman EA, Ferris FL, 3rd. Report from the NEI/FDA Ophthalmic Clinical Trial Design and Endpoints Symposium. *Invest Ophthalmol Vis Sci*. 2008;49:479–489.
10. Schmitz-Valckenberg S, Brinkmann CK, Alten F, et al. Semiautomated image processing method for identification and quantification of geographic atrophy in age-related macular degeneration. *Invest Ophthalmol Vis Sci*. 2011;52:7640–7646.
11. Arslan J, Samarasinghe G, Benke KK, et al. Artificial intelligence algorithms for analysis of geographic atrophy: a review and evaluation. *Transl Vis Sci Technol*. 2020;9:57–57.
12. Deckert A, Schmitz-Valckenberg S, Jorzik J, Bindewald A, Holz FG, Mansmann U. Automated analysis of digital fundus autofluorescence images of geographic atrophy in advanced age-related macular degeneration using confocal scanning laser ophthalmoscopy (cSLO). *BMC Ophthalmol*. 2005;5:8.
13. Lee N, Smith RT, Laine AF. Interactive segmentation for geographic atrophy in retinal fundus images. *Conf Rec Asilomar Conf Signals Syst Comput*. 2008;2008:655–658.
14. Devisetti K, Karnowski TP, Giancardo L, Li Y, Chaum E. Geographic atrophy segmentation in infrared and autofluorescent retina images using supervised learning. *Conf Proc IEEE Eng Med Biol Soc*. 2011;2011:3958–3961.
15. Chen Q, de Sisternes L, Leng T, Zheng L, Kutzscher L, Rubin DL. Semi-automatic geographic atrophy segmentation for SD-OCT images. *Biomed Opt Express*. 2013;4:2729–2750.
16. Hu Z, Medioni GG, Hernandez M, Hariri A, Wu X, Sadda SR. Segmentation of the geographic atrophy in spectral-domain optical coherence tomography and fundus autofluorescence images. *Invest Ophthalmol Vis Sci*. 2013;54:8375–8383.
17. Hu Z, Medioni G, Hernandez M, Sadda S. Supervised pixel classification for segmenting geographic atrophy in fundus autofluorescence images. In *Medical Imaging 2014: Computer-Aided Diagnosis*. Bellingham, WA: International Society for Optics and Photonics; 2014;9035:90350G.
18. Ramsey DJ, Sunness JS, Malviya P, Applegate C, Hager GD, Handa JT. Automated image alignment and segmentation to follow progression of geographic atrophy in age-related macular degeneration. *Retina*. 2014;34:1296–1307.
19. Feeny AK, Tadarati M, Freund DE, Bressler NM, Burlina P. Automated segmentation of geographic atrophy of the retinal epithelium via random forests in AREDS color fundus images. *Comput Biol Med*. 2015;65:124–136.
20. Hu Z, Medioni GG, Hernandez M, Sadda SR. Automated segmentation of geographic atrophy in fundus autofluorescence images using supervised pixel classification. *J Med Imaging (Bellingham)*. 2015;2:014501.
21. Niu S, de Sisternes L, Chen Q, Leng T, Rubin DL. Automated geographic atrophy segmentation for SD-OCT images using region-based C-V model via local similarity factor. *Biomed Opt Express*. 2016;7:581–600.
22. Fang L, Cunefare D, Wang C, Guymer RH, Li S, Farsiu S. Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search. *Biomed Opt Express*. 2017;8:2732–2744.
23. Hu Z, Wang Z, Abdelfattah NS, Sadda J, Sadda SR. Automated geographic atrophy segmentation in infrared reflectance images using deep convolutional neural networks. *Invest Ophthalmol Vis Sci*. 2018;59:1714–1714.
24. Hu ZH, Wang ZY, Sadda SR. Automated segmentation of geographic atrophy using deep convolutional neural networks. *Proc Spie*. 2018;10575.
25. Ji Z, Chen Q, Niu S, Leng T, Rubin DL. Beyond retinal layers: a deep voting model for automated geographic atrophy segmentation in SD-OCT images. *Transl Vis Sci Technol*. 2018;7:1.
26. Xu R, Niu S, Gao K, Chen Y. Multi-path 3D Convolution Neural Network for Automated Geographic Atrophy Segmentation in SD-OCT Images. In: *International Conference on Intelligent Computing*; Cham, Switzerland: Springer; 2018:493–503.
27. Yang Q, Dong Y, Tokuda K, et al. Automated geographic atrophy detection in OCT volumes. *Investigative Ophthalmology & Visual Science*. 2018;59:3225–3225.
28. Wu M, Cai X, Chen Q, et al. Geographic atrophy segmentation in SD-OCT images using

synthesized fundus autofluorescence imaging. *Comput Meth Prog Bio*. 2019;182:105101.
29. Xu R, Niu S, Chen Q, Ji Z, Rubin D, Chen Y. Automated geographic atrophy segmentation for SD-OCT images based on two-stage learning model. *Comput Biol Med*. 2019;105:102–111.
30. Schmidt-Erfurth U, Bogunovic H, Grechenig C, et al. Role of deep learning quantified hyperreflective foci for the prediction of geographic atrophy progression. *Am J Ophthalmol*. 2020.
31. Smistad E, Østvik A, Haugen BO, Løvstakken L. 2D left ventricle segmentation using deep learning. *2017 IEEE International Ultrasonics Symposium (IUS)*. 2017:1–4.
32. Liefers B, Colijn JM, González-Gonzalo C, et al. A deep learning model for segmentation of geographic atrophy to study its long-term natural history. *Ophthalmology*. 2020;127:1086–1096.
33. Arya M, Sabrosa AS, Duker JS, Waheed NK. Choriocapillaris changes in dry age-related macular degeneration and geographic atrophy: a review. *Eye Vis (Lond)*. 2018;5:22.
34. Willemink MJ, Koszek WA, Hardell C, et al. Preparing Medical Imaging Data for Machine Learning. *Radiology*. 2020;295:4–15.
35. Benke KK, Arslan J. Deep learning algorithms and the protection of data privacy. *JAMA Ophthalmol*. 2020;138:1024–1025.
36. Mehta N, Lee CS, Mendonca LSM, et al. Model-to-data approach for deep learning in optical coherence tomography intraretinal fluid segmentation. *JAMA Ophthalmol*. 2020;138:1017–1024.
37. Ronneberger O, Fischer P, U-Net Brox T.: *Convolutional Networks for Biomedical Image Segmentation*. Cham: Springer International Publishing; 2015:234–241.
38. Hatami N, Gavet Y, Debayle J. Classification of Time-Series Images Using Deep Convolutional Neural Networks. *The 10th International Conference on Machine Vision (ICMV 2017)*. Vienne, Austria: ICMV Committees; 2017:1710.00886.
39. Heidelberg Engineering Academy. HRA+OCT Spectralis: How to Acquire the Perfect Image. Available at https://www.heidelbergengineering.com.
40. Holz FG, Spaide RF, Bird AC, Schmitz-Valckenberg S. *Atlas of fundus autofluorescence imaging*. New York: Springer; 2007.
41. Yung M, Klufas MA, Sarraf D. Clinical applications of fundus autofluorescence in retinal disease. *Int J Retina Vitreous*. 2016;2:12.
42. Szeliski R. Computer vision algorithms and applications. *Texts in Computer Science*. New York: Springer, 2011:1 online resource (xx, 812 p.).
43. International Conference on Computational Science and Technology. International (4th: 2017: Kuala Lumpur M. *Computational Science and Technology: 4th ICCST 2017, Kuala Lumpur, Malaysia, 29-30 November, 2017*. Singapore: Springer Nature Singapore Pte Ltd.; 2018.
44. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. *International Conference on Artificial Intelligence and Statistics*. 2010:249–256.
45. He K, Zhang X, Ren S, Sun J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classificatio. *arXiv* 2015;1502.01852.
46. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *arXiv* 2017;1412.6980.
47. Zur RM, Jiang Y, Pesce LL, Drukker K. Noise injection for training artificial neural networks: a comparison with weight decay and early stopping. *Med Phys*. 2009;36:4810–4818.
48. Keskar NS, Mudigere D, Nocedal J, Smelyanskiy M, Tang PTKP. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. *arXiv* 2017;1609.04836.
49. Jakubovitz D, Giryes R, Rodrigues MRD. Generalization Error in Deep Learning. In: Boche H, Caire G, Calderbank R, Kutyniok G, Mathar R, P. P (eds). *Compressed Sensing and Its Applications*. Cham: Springer; 2019:153–193.
50. Zheng A. *Evaluating Machine Learning Models*. Sebastopol, CA: O'Reilly Media, Inc.; 2015.
51. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI*. 1995;14:1137.
52. Dabbs B, Junker B. Comparison of cross-validation methods for stochastic block models. *arXiv:160503000* 2016.
53. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: with Applications in R*. New York: Springer; 2013.
54. Kuhn M, Johnson K. *Applied Predictive Modeling*. New York: Springer; 2013.
55. Zou KH, Warfield SK, Bharatha A, et al. Statistical validation of image segmentation quality based on a spatial overlap index. *Acad Radiol*. 2004;11:178–189.
56. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging*. 2015;15:29.
57. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1:307–310.
58. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8:135–160.

59. Dey N, Ashour AS, Shi F, Balas VE. *Soft Computing Based Medical Image Analysis*. London: Academic Press; 2018.
60. Xu Y, Wang Y, Yuan J, Cheng Q, Wang X, Carson PL. Medical breast ultrasound image segmentation by machine learning. *Ultrasonics*. 2019;91:1–9.
61. Little MA, Varoquaux G, Saeb S, et al. Using and understanding cross-validation strategies. Perspectives on Saeb et al. *GigaScience*. 2017;6(5): gix020.
62. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. *Statistics Surveys*. 2010;4:40–79, 40.
63. Reiter GS, Told R, Schranz M, et al. Subretinal Drusenoid Deposits and Photoreceptor Loss Detecting Global and Local Progression of Geographic Atrophy by SD-OCT Imaging. *Invest Ophthalmol Vis Sci*. 2020;61:11.
64. Camino A, Wang Z, Wang J, et al. Deep learning for the segmentation of preserved photoreceptors on en face optical coherence tomography in two inherited retinal diseases. *Biomed Opt Express*. 2018;9:3092–3105.
65. Pfau M, von der Emde L, de Sisternes L, et al. Progression of photoreceptor degeneration in geographic atrophy secondary to age-related macular degeneration. *JAMA Ophthalmol*. 2020;138:1026–1034.
66. Mehta N, Lee CS, Mendonça LSM, et al. Model-to-data approach for deep learning in optical coherence tomography intraretinal fluid segmentation. *JAMA Ophthalmol*. 2020;138:1017–1024.