On the quantitative effects of compression of retinal fundus images on morphometric vascular measurements in VAMPIRE

On the quantitative effects of compression of retinal fundus images on morphometric vascular measurements in VAMPIRE

Muthu Rama Krishnan Mookiah[a,*], Stephen Hogg[a], Tom MacGillivray[b], Emanuele Trucco[a], on behalf of the INSPIRED project

[a] *VAMPIRE project / CVIP, Computing, School of Science and Engineering, University of Dundee, Dundee, UK*

[b] *VAMPIRE project, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK*

Abstract

Background and Objectives: This paper reports a quantitative analysis of the effects of Joint Photographic Experts Group (JPEG) image compression of retinal fundus camera images on automatic vessel segmentation and on morphometric vascular measurements derived from it, including vessel width, tortuosity and fractal dimension.

Methods: Measurements are computed with Vascular Assessment and Measurement Platform for Images of the REtina (VAMPIRE), a specialized software application adopted in many international studies on retinal biomarkers. For reproducibility, we use three public archives of fundus images (Digital Retinal Images for Vessel Extraction (DRIVE), Automated Retinal Image Analyzer (ARIA), high-resolution fundus (HRF)). We generate compressed versions of original images in a range of representative levels.

Results: We compare the resulting vessel segmentations with ground truth maps and morphological measurements of the vascular network with those obtained from the original (uncompressed) images. We assess the segmentation quality with sensitivity, specificity, accuracy, area under the curve and Dice coefficient. We assess the agreement between VAMPIRE measurements from compressed and uncompressed images with correlation, intra-class correlation and Bland-Altman analysis.

Conclusions: Results suggest that VAMPIRE width-related measurements (central retinal artery equivalent (CRAE), central retinal vein equivalent (CRVE), Arteriolar-Venular width Ratio (AVR)), the fractal dimension and arteriolar tortuosity have excellent agreement with those from the original images, remaining substantially stable even for strong loss of quality (20% of the original), suggesting the suitability of VAMPIRE in association studies with compressed images.

Keywords:
Retinal imaging, JPEG compression, Retinal biomarkers, VAMPIRE

1. Introduction and Motivation

This paper reports a quantitative assessment of the effects of Joint Photographic Experts Group (JPEG) compression on morphometric measurements of the retinal vasculature, commonly used in retinal biomarker studies, computed from fundus camera images with the VAMPIRE software tool. VAMPIRE is a validated software application used in many international studies [1, 2, 3, 4, 5].

Our main motivation is the existence of large clinical repositories of retinal scans (tens or hundreds of thousands) in compressed format, typically JPEG. Such repositories represent huge sources of potential data for retrospective studies on retinal biomarkers. Large studies are only possible with automatic or semi-automatic software systems, as manual measurements are too slow and sometime simply impossible (e.g. fractal dimension). However compressed images are normally suitable for visual inspection by experts [6, 7, 8], but not necessarily for automatic software systems. The *quantitative* effects of compression on the values of vascular measurements (candidate biomarkers) computed automatically remain unclear. We address this quantification in this paper and discuss its consequences.

In VAMPIRE and similar systems [9, 2], morphometric measurements of the retinal vasculature are computed from vessel maps obtained by automatic segmentation. We consider therefore the effect of compression on vessel segmentation. VAMPIRE (version 3.2) uses a dense block U-net for segmenting blood vessels. In this study we train the U-net with ground-truth vessel maps traced on retinal images both uncompressed and compressed at various levels. We quantify compression with the *perceptual quality* (PQ), a score between 0 to 100, where 0 compression corresponds to PQ = 100. We then use VAMPIRE to compute 151 measurements from a pre-defined dictionary described elsewhere [10] and sketched in Section 4.4. For each perceptual quality (PQ) level, we record the variations of vessel maps and VAMPIRE measurements. A block diagram of the computation chain is shown in Figure 1.

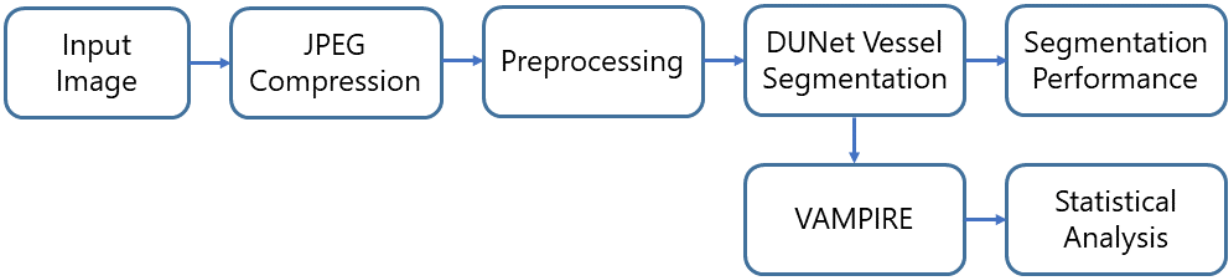

Figure 1: Methodology block diagram.

We assess the segmentation quality with sensitivity, specificity, accuracy, area under the curve and the Dice coefficient. We assess the agreement between VAMPIRE measure-

*Corresponding Author:
Email address: mrkmookiah@dundee.ac.uk (Muthu Rama Krishnan Mookiah)

ments from compressed and uncompressed images with correlation, intra-class correlation and Bland-Altman analysis.

This paper brings two contributions. First, a quantitative analysis of the effects of compressing fundus camera images on the value of measurements (candidate biomarkers) used commonly in studies on retinal biomarkers for systemic conditions. Second, a segmentation network offering good to excellent stability up to aggressive compression and degradation levels for retinal vascular measurements commonly used in biomarker studies. Our results support, for the first time to our best knowledge, the use of automatic algorithms like the Vascular Assessment and Measurement Platform for Images of the REtina (VAMPIRE) with large collections of compressed images, unlocking the potential of large retrospective clinical image sets maintained in JPEG compressed format.

2. Related work

2.1. Retinal imaging and biomarkers

The retina allows easy observation of a rich portion of the human microvasculature using non-invasive and comparatively cheap instruments. Indeed it has been called “a window on the brain" as it develops as a protrusion from the brain itself [11]. This affords the hypothesis that changes in the retinal microvasculature may signal pathological changes in the brain, e.g. stroke and vascular dementia. Associations have also been reported with systemic conditions like cardiovascular disease (CVD), hypertension, dementia and diabetes [12, 9, 13, 14]. The images used in most of the reported biomarker studies are acquired with fundus cameras and OCT scanners, arguably the main imaging modalities in current clinical practice [9, 15, 13] and now offered by optometrist chains, see e.g. [16] for the UK. Other modalities for observing the retina include ultra-wide field of view scanning laser ophthalmoscopy (UWFV SLO) [3], OCT angiography (OCT-A) [17], fluorescein angiography [1, 10, 15] and hyperspectral imaging [18]. For a recent comprehensive overview of imaging in the retina see [19].

2.2. Computing retinal measurements

Algorithms computing features in retinal fundus images can be grouped in two different approaches [20], namely dictionary-based and deep learning.

Dictionary-based approaches aim at quantifying morphological features in a pre-defined dictionary, e.g. vessel width, tortuosity, fractal dimension. This is the approach adopted in VAMPIRE and related systems like SIVA, QUARTZ and IVAN among others [9, 1, 2]. The resulting values are fed to a statistical analysis to evaluate their association with target variables, typically clinical outcomes like major adverse cardiovascular events (MACE) or dementia [9, 21, 22]. *Deep learning approaches*, more recent and increasingly popular, use directly retinal images in deep networks to stratify patient samples according to a target variable (e.g. MACE / no MACE) . No feature extraction is needed. An increasing number of papers address the identification of the image regions most important for the stratification achieved [23, 24].

The two approaches have different characteristics and purposes. Dictionary-based approaches rely on anatomical features immediately understandable to (and indeed suggested by) clinicians, who can relate them to pathophysiological processes. Features not suggested by clinicians (e.g. textural properties) have also been included in these approaches [22]. Hence a pre-defined dictionary confines biomarker discovery to its contents but offers accurate measurements of clearly defined properties. This framework is well suited for statistical studies on the associations between specific measurements and outcomes of interest. Examples of dictionary-based computational systems include VAMPIRE [25, 26, 27], used here, SIVA [28], IVAN [21], ARIA [29] and QUARTZ [30]. Most systems are interactive (semi-automatic), i.e. allow an operator to check intermediate results and correct wrong automatic decisions [2]. The vast majority of studies reported use morphological measurements of the blood vessels network, i.e. width (central retinal artery equivalent (CRAE), central retinal vein equivalent (CRVE) and Arteriolar-Venular width Ratio (AVR)), tortuosity, bifurcation angles and fractal dimension.

Deep learning approaches process retinal images as their only input, typically tens or hundreds of thousands [31, 32, 33, 34]. They do not require a pre-defined dictionary but the image features computed by a deep network are, in general, difficult to interpret clinically and hence to link with meaningful clinical pathways [20]. This is the focus of research of explainable AI [35], a rapidly growing area of research arguably destined to play an increasingly crucial role in AI for healthcare.

We refer the reader to [36] for a comprehensive introduction to recent research in computational retinal image analysis, including clinical motivations and perspectives, AI techniques, and data-related issues.

2.3. Compression and retinal measurements

The accuracy of morphometric retinal measurements is influenced by several factors introducing uncertainty [37]. Such factors include variations of image focus, refractive error [38], image centering [39, 38, 40, 10], image resolution, image compression, acquisition protocols, patient cooperation, racial variations, algorithms and their parameter settings. Understanding the quantitative effects of this uncertainty is clearly very complex but seems crucial to assess the reliability of results of statistical association studies. For instance, the FD [41] has been associated with various clinical outcomes [42], but various authors have reported considerable fluctuations of FD estimates in the presence of varying image noise, changes in vessel segmentation, images resolution and other factors [43, 44, 45].

In this work we use JPEG compression, offered on commercial fundus cameras by the main manufacturers (Zeiss, Topcon, Canon, Nidek). We refer the reader to the website jpeg.org for an overview of JPEG. Compression introduces three primary artefacts: smoothing (blur), ringing around sharp edges, and block effects [46] (Figure 2). The consequences of these artefacts are different for visual inspection and automatic image processing. Several studies have showed that the degradation caused by moderate compression does not make the images unusable for screening and diagnosis when performed by experts, for instance diabetic retinopathy (DR) grading [7, 8]. Baker et al. [8] found no significant effects on the accuracy and reproducibility of DR grading [8] for teleophthalmology even with very strong

JPEG compression (113:1 ratio). Conrath et al. [47] investigated the effect of compression on DR grading by experts with 45 low-resolution TIFF images (800 × 600). Images were compressed with Photoshop using JPEG and JPEG2000 to different image sizes (118KB, 58KB, 41KB and 27KB). They found that image degradation became important only with the highest compression (27KB) and that detection of intraretinal miscrovascular abnormalities (IRMA) and hemorragies and microaneurysms (HMA) was the most degraded.

Modelling the quantitative effects of compression on the accuracy of automatic image processing is still ongoing research. For instance, Lee et al. [48] proposed a semi-automated method for drusen detection and area estimation using JPEG-compressed images. Results showed no significant difference in drusen area measurements between original images and images compressed at 30:1 ratio. The authors did not include comparisons for 80:1 compression, which made quality degradation visible [46]. Akshay [49] studied the effect of five module compression method for retinal vessel segmentation, a format not used by fundus camera, to our best knowledge. They used 160 retinal images and the F-measure for vessel segmentation was lowered 3% from uncompressed to compressed images.

For interest, we mention some reports addressing compression effects in other types of medical images. Daugman et al. [50] studied the accuracy of automatic iris recognition. Their results suggest that only 2% to 3% of the bits in the encoding of retinas (IrisCode) changed even with severe images compression (PQ = 20). López et al. [51] studied variations in morphometric parameters of immunohistological stained cell nuclei with different levels of JPEG compression. Among six morphological parameters, only one (roundness score) was significantly affected by compression. Ozdemir et al. [52] compared different segmentation algorithms applied to JPEG-compressed brain magnetic resonance imaging (MRI) scans. Results comparable with those with uncompressed images were obtained with moderate compression (PQ = 80).

In summary, the image degradation introduced by lossy JPEG compression does not seem to compromise substantially diagnosis and screening performed by clinicians or image readers. However, its quantitative effects on the accuracy of measurements computed automatically and commonly used as candidate retinal biomarkers remains an open question, that we address in this paper.

3. Materials

3.1. Datasets

We use three publicly available datasets, namely Digital Retinal Images for Vessel Extraction (DRIVE), Automated Retinal Image Analyzer (ARIA), and high-resolution fundus (HRF), in the interest of reproducibility.

DRIVE [53] consists of 400 color fundus images from a DR screening program in the Netherlands, acquired with a Canon CR5 non-mydriatic 3-CCD camera with 45° field of view (FOV) with a resolution of 565×584 pixels from diabetic subjects between 25 and 90 years of age. Ground-truth labels from one annotator and additional annotations are provided. We selected 40 images randomly, of which 33 do not show any sign of DR and 7 show early signs of DR.

ARIA [54] was created by St. Paul's Eye Unit in Liverpool, UK. Images were acquired with a Zeiss Meditec fundus camera with 50° FOV and a resolution of 768×576 pixels. ARIA contains 143 color fundus images organized into three classes, namely healthy(n=61), age-related macular degeneration (AMD) (n=23), and diabetic (n=59). Ground truth labels from two experts are provided.

HRF [55] was created by a German-Czech consortium. This dataset consists of 45 colour fundus images acquired with a Canon CR-1 fundus camera with 45° FOV and a resolution of 3504×2336 pixels. The images are organized into three classes, namely healthy(n=14), DR (n=15), and glaucoma (n=15). The ground truth was provided by a single expert.

4. Methods

4.1. JPEG compression

We used the MATLAB JPEG lossy compression software, allowing control of perceptual quality via an integer parameter PQ with values between 0 and 100, where 0 corresponds to the worst image quality (highest compression) and 100 to the highest quality (no compression). A ground-truth vessel map traced manually on the uncompressed, full-quality image is available for all images.

We ran experiments with two protocols: first, training the deep networks with images compressed at different levels (separate compression (SC)), and second, with all images (all levels) together (combined compression (CC)). For the SC protocol, images were compressed at PQ = 10, 20, 35, 55, 80, and 100; for CC, at PQ = 10, 30, 50, 70, and 100.

This resulted in 240 compressed images for DRIVE, 858 for ARIA, and 270 for HRF for the SC protocol, and 200 compressed images for DRIVE, 715 for ARIA, and 225 for HRF for the CC protocol. 100 images were randomly selected for each dataset for CC experiments.

Table 1 relates the PQ levels with the commonly used peak signal-to-noise ratio (PSNR) for all the data sets. Our range spans the compression levels used in many clinical archives. To put these figures in a wider context, typical PSNR values in lossy image and video compression are between 30 and 50dB for 8-bit images (covering our range, see Table 1), and between 60 and 80dB for 16-bit images (not our case).

Table 1: **Compression evaluation results of SC and CC for DRIVE, ARIA, and HRF datasets**

Dataset	**PSNR (dB)**							
	PQ10	*PQ20*	*PQ30*	*PQ35*	*PQ50*	*PQ55*	*PQ70*	*PQ80*
DRIVE	32.04	34.96	35.35	36.47	36.58	38.24	37.75	40.43
ARIA	32.74	36.12	37.72	38.44	39.67	40.11	41.27	42.66
HRF	33.42	37.46	39.58	40.14	41.51	41.89	43.22	44.22

4.2. Pre-processing and training protocol

The DRIVE, ARIA, and HRF training images were pre-processed to eliminate spurious variations [56]. The colour images were converted into grayscale and normalized by subtracting the mean image and dividing each pixel by the image standard deviation. Contrast

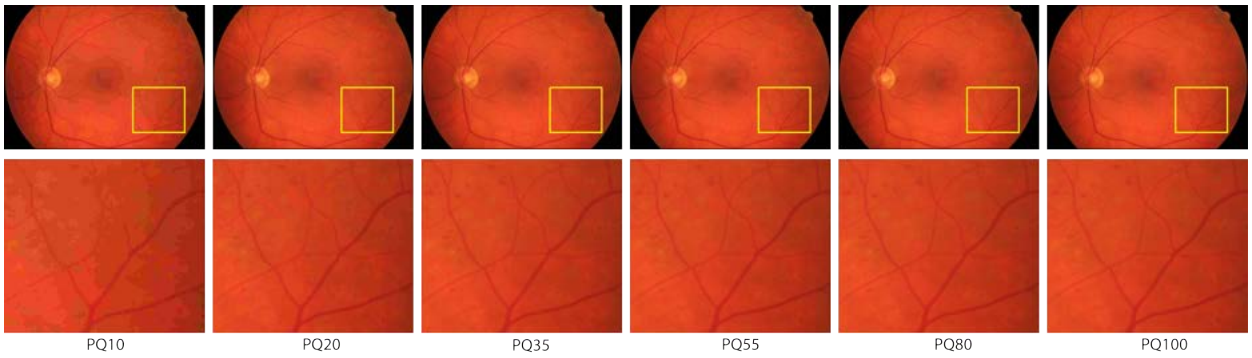

Figure 2: Top row: a fundus image from the HRF dataset and its versions at different compression levels (PQ 10, 20, 35, 55, 80; 100 is uncompressed). Bottom row: enlarged detail (yellow box). Best seen on a computer screen.

was enhanced by contrast-limited adaptive histogram equalization and gamma correction (Figure 3). Each dataset was split into subsets for training, validation and test (DRIVE: 25 training, 5 validation and 10 test; ARIA: 108, 5 and 30; for HRF, 30, 5 and 10). Due to the limited data available, the network is trained with randomly cropped patches [57] (128×128), resulting in 15,000 patches for HRF, 54,000 for ARIA and 12,500 for DRIVE.

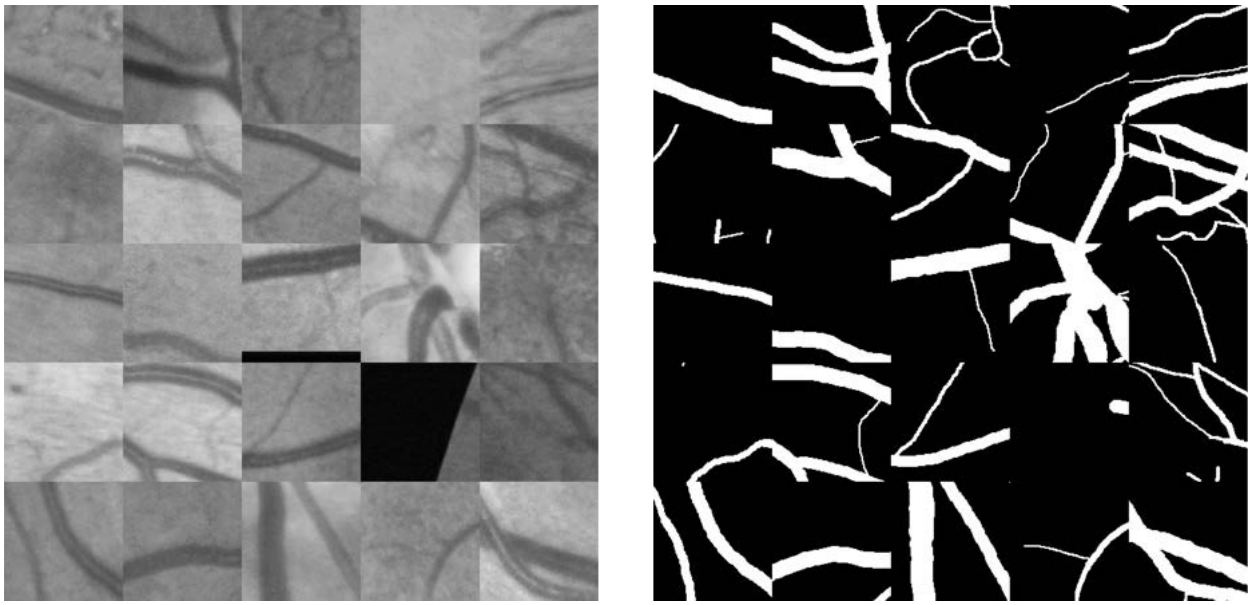

Figure 3: Examples of pre-processed image patches (left) and ground truth maps (right) 128×128 training patches extracted from the HRF dataset.

4.3. Dense U-net for vessel segmentation

We adopted a dense convolutional U-net architecture, the DenseNet model introduced by Ronnenberger et al. [58]. A dense block U-net consists of a typical U-shaped contracting-expanding fully convolutional architecture [59]. Desirable properties of this popular model

were discussed by Huang et al. [60] and include, briefly, addressing the vanishing-gradient problem, strengthening feature propagation, encouraging feature re-using, and limiting the number of parameters. Layers are connected in a feed-forward fashion, with each layer receiving the feature maps from all preceding layers as inputs and feeding its own maps to *all* subsequent layers. Dense blocks learn multiple feature maps. Concatenation connections are added between the contracting and expanding paths to allow the network to use finer resolution features learned earlier for up-sampling. The image undergoes a multi-level decomposition via convolutions and max-pooling followed by an up-sampling process via deconvolution, enabling the network to learn local and global image features at multiple spatial scales [59]. The network computes a vessel probability map subsequently binarized with Otsu's thresholding. An overview of the network architecture is shown in Table 4.

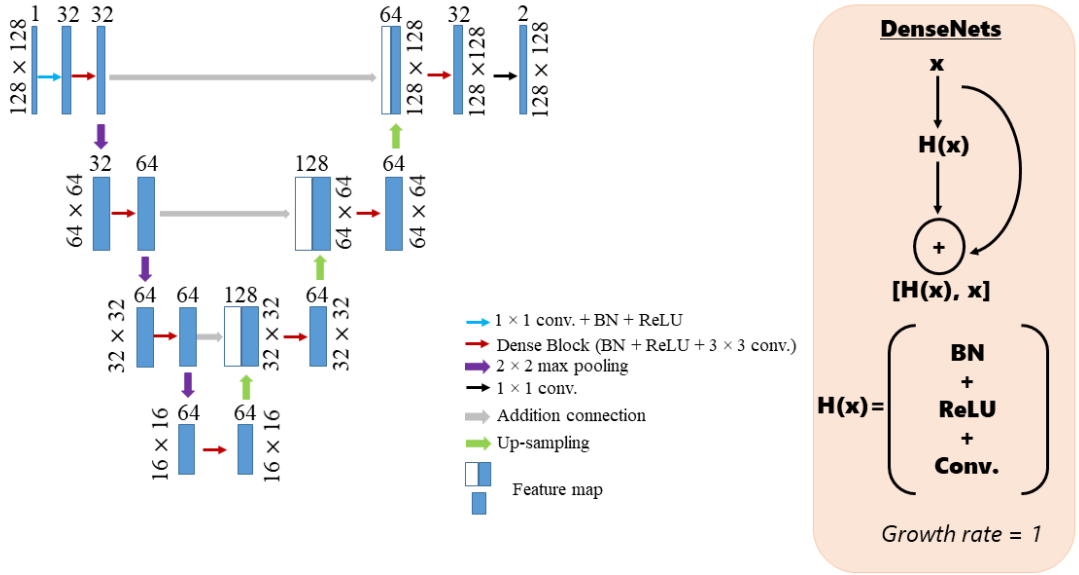

Figure 4: Dense block U-net architecture. Blue boxes represent 3D tensors.

We used Adam optimization with batch updates (25 batches) and momentum term. Weight updates proceeded in batches every 20 epochs; batch normalization layers with momentum term of 0.9 and epsilon of 0.00002 were added in each dense block to standardize the output of subsequent layer. Cross-entropy was used as loss function. The growth rate hyper-parameter was set to 2. The training parameters were optimized based on training performance (training accuracy, training loss, validation accuracy and validation loss). Our model contains 5 millions trainable weights. The dense U-net was implemented in Python 3 using Keras 2.2.4 and run on an Intel Core i7-8700 CPU computer with an NVIDIA TITAN Xp GPU. Part of the implementation is based on Retina-VesselNet [61, 57]. For illustration, training performance with the SC and CC protocols for three quality levels (35%, 55%,

100%) only for HRF datasets are shown in Figure 5; training performance for DRIVE and ARIA are provided as supplementary material (performance-measure-plots, Figure.S1 to Figure.S3). Generalization is good overall (loss curves) and marginally better with the CC protocol.

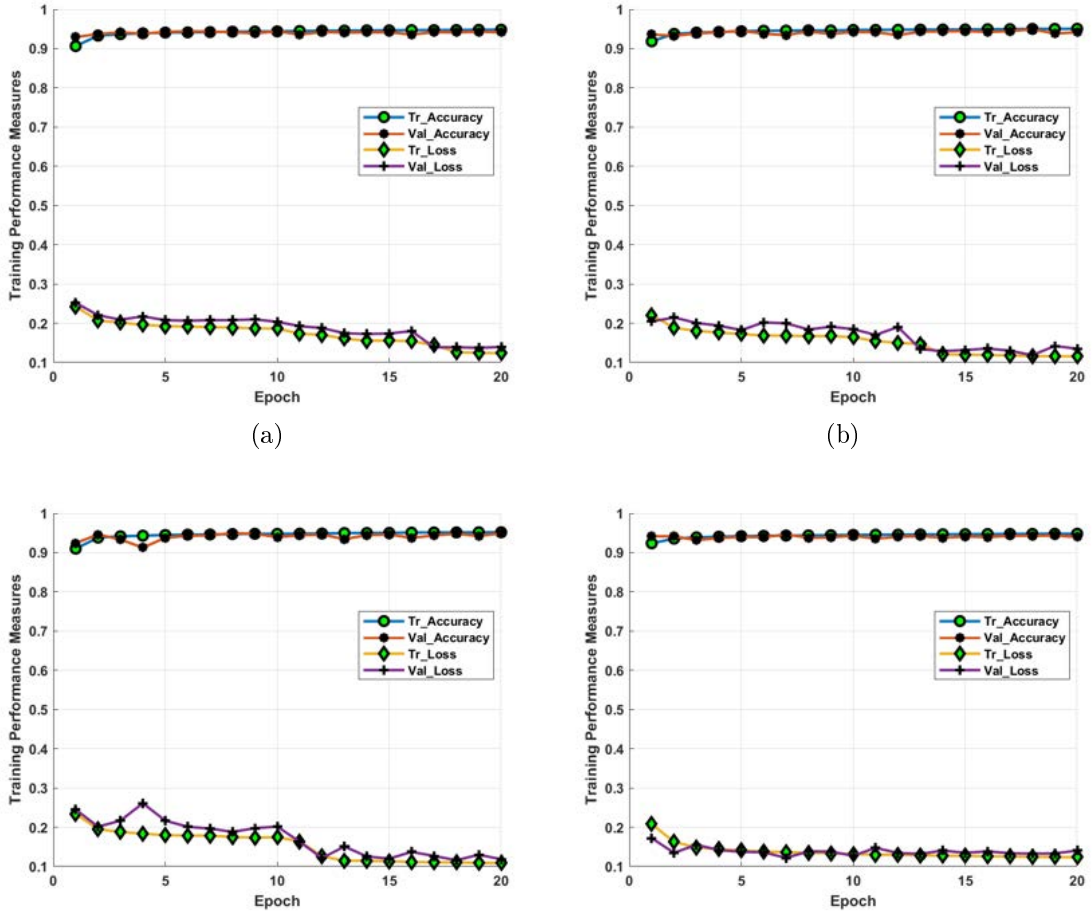

Figure 5: Training/validation accuracy and loss for *three* separate compression levels: (a) PQ35, (b) PQ55, (c) PQ100, and (d) combined compression levels for HRF dataset.

4.4. VAMPIRE measures and statistical analysis

Measurements of the vascular morphometry were performed with the semi-automated VAMPIRE 3.2 software. Detailed definitions have been reported elsewhere [2]; we include a summary here. In all, VAMPIRE computes 149 measurements per image including basic statistics. 39 are width-related: CRAE, CRVE, AVR, basic statistics (mean, median, standard deviation, max, min), width gradients along vessels, average ratio length-diameter at branching points, by arteries and veins; 104 are tortuosity measurements, computed by different algorithms and with the statistics listed above; 6 are fractal dimension coefficients,

computed in Zone C. All measurements are computed by vessel type (arteriole or venule), by region (zone, whole image, quadrant; see supplementary material (VAMPIRE zones and quadrant, Figure.S4) for definitions) and vessel (path, generation). For this study images were measured by author SH, an experienced VAMPIRE operator ($\tilde{1}$0,000 images measured).

VAMPIRE operators take a two-stage training module: (1) learn to operate the software (2-day residential module), (2) measure standard sets with validated results until mistakes in operator choices (e.g. correcting optic disk (OD) position, deleting vessel segments) become uninfluential on the variables most sensitive to operator choices, i.e. OD and fovea locations (primary variables), and width-related measures CRAE, CRVE, and AVR. Operator revalidation is then performed at regular intervals, with durations and batch sizes depending on training progress. Operator alignment is evaluated both qualitatively and quantitatively, e.g. with the intra-class correlation coefficient (ICC) [62] (two-way mixed, consistency agreement, average measure model); batches of 30 images are measured until the ICC exceeds 0.75 [63] for all the above measures.

We used Pearson's correlation, (r), and ICC evaluate the agreement between VAMPIRE measures obtained from images compressed at different levels (PQ10, PQ20, PQ35, PQ55, PQ80) and original (PQ100) images. ICCs has been indicated as more appropriate for assessing *agreement* than Pearson's correlation r, which quantifies linear dependency [64, 65]. Perfect linear dependence does not necessarily reflect good or even moderate agreement, e.g. r may be high even if strong systematic bias is present.

Pearson's r and ICC results are usually interpreted as 0.00–0.49 = poor, 0.50–0.74 = moderate, and 0.75–1.00 = excellent [66]. We also used Bland-Altman plots [67, 68] of the difference of paired measurements (e.g., PQ100 vs PQ10) against their mean. The 95% limit of agreement (LOA) was defined as the mean difference$\pm$1.96$\times$standard deviation (SD).

5. Results

We present experimental results on the effects of compression in two sections, vessel segmentation and VAMPIRE measurements. The former is further divided by data set.

5.1. Vessel segmentation

As stated in Section 4.1, vessel maps were assessed using sensitivity (SE), specificity (SP), accuracy (ACC), DICE similarity coefficient (DSC), and area under the curve (AUC). We assessed resilience to compression with networks trained with the two protocols in Section 4.1. Figure A1, Figure A3, and Figure A5 show examples of segmented DRIVE, ARIA, and HRF images compressed at different PQ levels. Figure A1 shows results for different compression (PQ) levels (Figure A1a), ground truth vessel maps (Figure A1b), and dense U-net segmentation maps (Figure A1c). The figures show significant degradation (blur and ringing artefacts) in the retinal vessels (Figure A1a, PQ10: 32.15dB, 7.9KB, PQ20: 34.82dB, 10.2KB, and PQ35: 36.69dB, 13.7KB). A Portion of Figure A1a is enlarged to show the retinal vessel degradation (Figure A2) for various compression (PQ) levels for the DRIVE dataset. Similar results are observed for ARIA (Figure A3 and Figure A4) and HRF

(Figure A5 and Figure A6). These results are inline with those of Conrath et al. [47]. The results confirm good segmentation (Figure A1c, Figure A3c, and Figure A5c) up to aggressive compression levels (PQ10) compared with ground truth images (Figure A1b, Figure A3b, and Figure A5b).

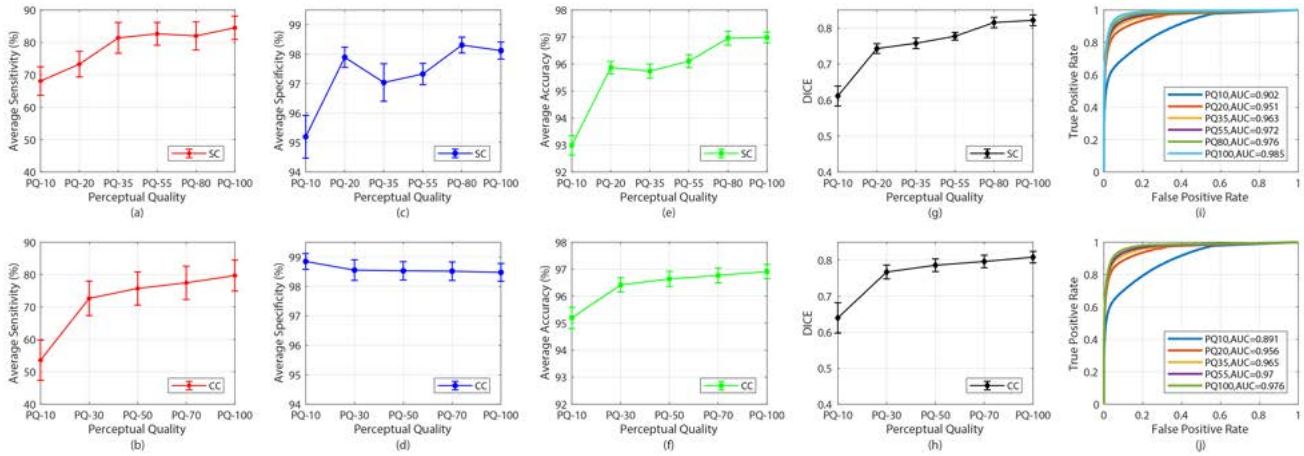

Figure 6: Average segmentation performance for DRIVE dataset. Top row: training images of different PQ levels separately. Bottom row: training with all images. Sections: (a) sensitivity, (b) specificity, (c) accuracy, (d) DICE similarity coefficient, and (e) AUC, all plotted against increasing PQ levels. Uncertainty bars indicate plus/minus standard deviation. The vertical ranges are the same in each column to facilitate comparison of the SC and CC protocols.

Table 2: **Decrease of each evaluation criterion as percentage of the range. SC: training with images at different PQ levels separately. CC: training with all images**

Experiment	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**	**DICE**
SC 10-100	-20	-3	-4	-25
CC 10-100	-5	-1	-1	-8
SC 30-100	-32	-0	-2	-21
CC 35-100	-6	-0.1	-0.2	-4

5.1.1. DRIVE

DRIVE provides ground-truth binary maps of comparable quality by multiple annotators; hence we used only the maps provided by the first annotator. The experimental values of the evaluation criteria for DRIVE images are plotted in Figure 6 and the corresponding numerical values in the Appendix (Tables A2 and A3). The loss of quality as percentage of the value at PQ=100 (uncompressed image) is tabulated in Table 2. When training at different levels separately (top row), all criteria increase with increasing quality (decreasing compression), as expected, apart from a minor decrease at PQ=20 for specificity and accuracy, in both cases within the standard deviation range (uncertainty bars). No considerable

decrease occurs, neither do ROC curves depart from the one obtained with uncompressed images, before strong degradation (PQ $<$ 30, corresponding to aggressive compression).

DICE and AUC are practically unchanged when training with images from compression levels separately and together. Accuracy remains substantially the same but marginally more stable when training with all images. With this training protocol specificity is better and very stable, but sensitivity worsens. This can be explained by the presence of more false positives at PQ35 and PQ55 then at other levels, see Figure 8(e) and (l). Another, important, factor affecting the segmentation is the low resolution (584×565); this is discussed in Section 6.

5.1.2. ARIA

Segmentation results for ARIA images are summarized in Figure 7 (please note the different ranges on the vertical axes). Numerical values are reported in the Appendix (Tables A4 and A5). In general, results confirm those obtained with DRIVE images. Percentage decrease values are not included as entirely comparable with those from DRIVE (Table 2). Similarly, the plots follow the same trends, confirming overall good stability in the same range of low to medium compression rates (PQ $>$ 30). Image resolution in ARIA, like in DRIVE, is low (768×576; see discussion). The same observations made for individual coefficients apply as well, as do those for the comparison of results from the two training protocols.

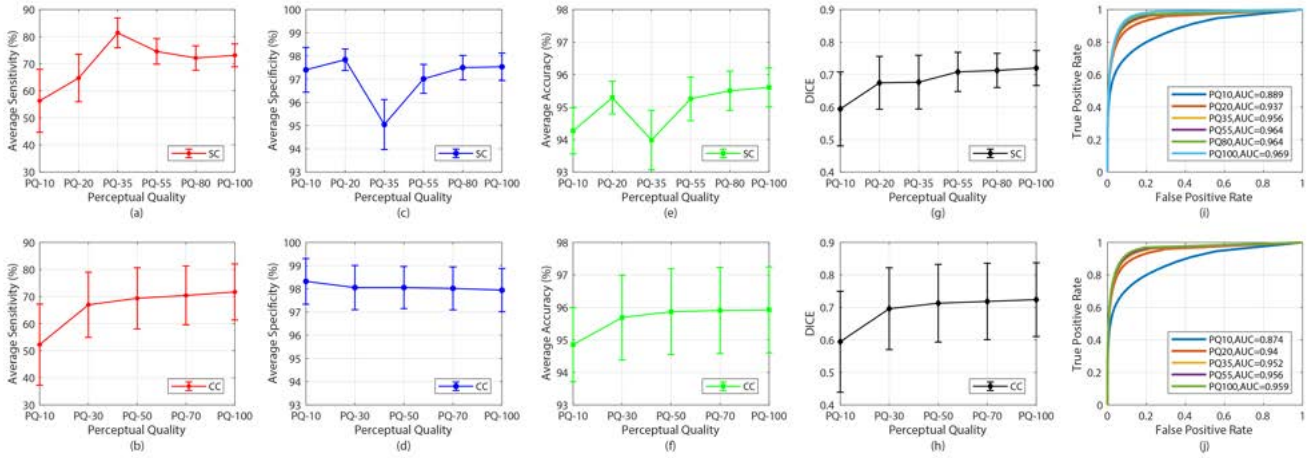

Figure 7: Average segmentation performance for ARIA dataset. Figure organization as per DRIVE figure above.

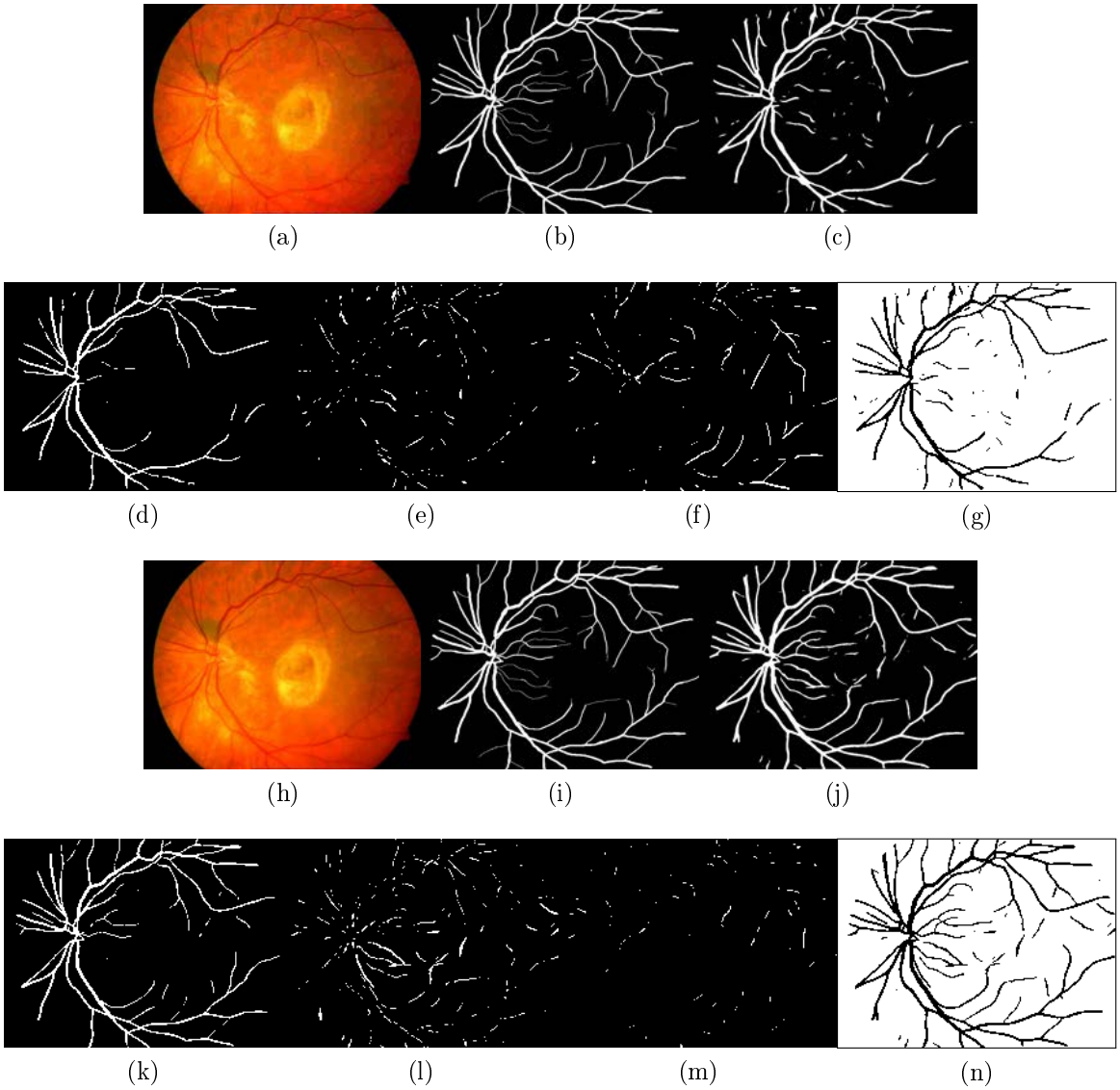

Figure 8: (a) Fundus image at PQ10 (steongest compression). (b) Ground truth vessel map. (c) Vessel map from our Dense U-net. (d) True positive pixels. (e) False positive pixels. (f) False negative pixels. (g) True negative pixels (white). (h) Fundus image with PQ35. (i) Ground truth vessel maps. (j) Vessel map from our Dense U-net on (h). (k) True positive pixels. (l) False positives. (m) False negative pixels. (n) True negative pixels (white).

5.1.3. HRF

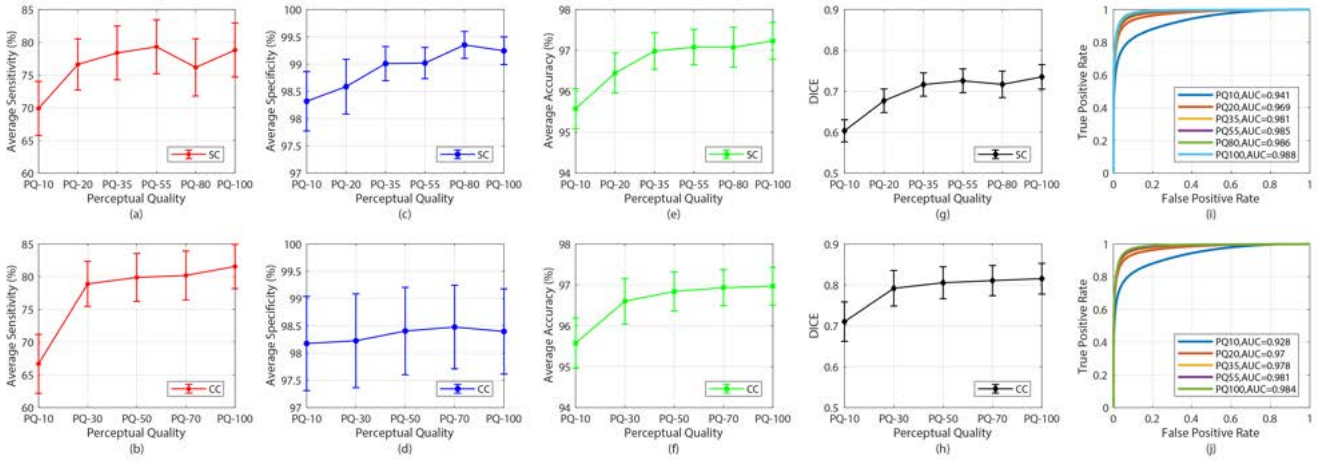

Figure 9: Average segmentation performance for image from the HRF dataset. Figure organization as per DRIVE figure above.

Segmentation results for HRF images are summarized in Figure 9 (please note the different ranges on the vertical axes). Numerical values are reported in the Appendix (Tables A6 and A7).

The resolution of HRF images is much higher (3504×2336) than in ARIA and DRIVE and comparable to the resolution of state-of-the-art fundus cameras. This arguably contributes to overall more regular trends in the graphs of individual criteria, especially noticeable for specificity and accuracy. Results confirm very good stability of segmentation results up to aggressive compression levels. Most peak-to-peak ranges are smaller than for DRIVE and ARIA, with less dramatic decreases for the lowest PQ values (maximum degradation).

5.2. VAMPIRE measurements

The stability of vessel maps against compression is important as vessel maps are the basis for computing morphological vascular measurements. However, from the point of view of biomarker studies, the ultimate question is the stability of retinal measurements against compression. This question is addressed in this section.

As the fundus camera used in recent (and no doubt future) studies yield high-resolution images (above 3000×2000 pixels), we use only the high-resolution HRF images. Given the better results on segmentation of the CC training protocol (Figure 9), we analyzed the stability of VAMPIRE measurements from 6 different versions of the HRF image set, each one compressed at a different level. We consider the twenty VAMPIRE measurements selected by McGrory et al. [14] from the 151 available in VAMPIRE and introduced in Section 4.4. Following the same authors, we log-transform tortuosity measurements to promote gaussianity.

The descriptive statistics of the selected measurements for PQ10 to PQ100 are shown in Table 3. Tables reporting results in full and a description of all retinal parameters are provided in the supplementary material (vampire-retinal-parameters).

Features	PQ10		PQ20		PQ35		PQ55		PQ80		PQ100	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
crae	30.83	2.90	30.74	2.23	29.75	1.97	29.93	2.09	29.62	1.80	30.17	1.97
crve	40.76	2.83	40.79	2.35	40.28	2.69	40.41	2.10	39.28	2.25	40.04	2.72
avr	0.76	0.06	0.75	0.06	0.74	0.05	0.74	0.05	0.76	0.06	0.76	0.06
bstda	1.88	0.60	1.77	0.83	1.62	0.59	1.71	0.83	1.82	0.75	1.61	0.60
bstdv	4.26	1.23	4.40	1.16	4.49	1.17	4.40	1.18	4.31	1.27	4.13	1.10
ba	2.07	1.21	2.61	0.99	3.18	1.39	3.50	1.14	3.25	1.32	2.54	1.23
bv	1.93	0.72	2.39	0.99	2.68	1.09	3.00	1.15	2.39	1.07	1.89	0.96
bca	2.01	0.46	2.07	0.34	2.12	0.31	2.08	0.29	2.12	0.30	1.54	0.45
bcv	1.86	0.19	2.00	0.30	2.09	0.42	2.13	0.35	2.13	0.57	1.29	0.23
afa	0.90	0.08	0.90	0.10	0.92	0.04	0.88	0.10	0.91	0.05	0.71	0.15
afv	0.88	0.07	0.90	0.06	0.90	0.06	0.90	0.05	0.90	0.08	0.69	0.13
ldra	17.43	6.89	16.53	5.12	17.25	5.53	16.57	4.55	18.54	4.16	18.39	9.63
ldrv	17.17	7.33	16.76	6.02	16.24	5.77	17.94	6.43	16.47	6.99	12.31	6.51
doa	1.27	0.12	1.30	0.12	1.32	0.12	1.32	0.13	1.31	0.14	1.32	0.11
d1a	1.26	0.12	1.29	0.12	1.31	0.12	1.31	0.13	1.30	0.14	1.30	0.11
d2a	1.25	0.12	1.28	0.12	1.30	0.12	1.30	0.13	1.29	0.14	1.30	0.11
dov	1.28	0.11	1.30	0.12	1.31	0.10	1.31	0.10	1.31	0.11	1.30	0.11
d1v	1.26	0.11	1.29	0.12	1.30	0.10	1.30	0.10	1.30	0.11	1.29	0.11
d2v	1.26	0.11	1.28	0.12	1.29	0.10	1.29	0.10	1.29	0.11	1.28	0.11
torta	-9.51	0.70	-9.72	0.76	-9.85	0.93	-9.90	0.82	-9.81	0.80	-9.79	0.94
tortv	-10.10	0.41	-10.11	0.61	-10.10	0.48	-10.18	0.56	-10.07	0.49	-9.69	0.58

Table 3: Descriptive statistics for VAMPIRE measurements. crae = central retinal arteriolar equivalent; crve = central retinal vein equivalent; avr = arteriole-to-venule ratio; bstda = standard deviation of arteriolar widths in Zone B; bstdv = standard deviation of venular widths in Zone B; ba = number of first branching points in arterial vessel in Zone C; bv = number of first branching points in venular vessel in Zone C; bca = Average values of branching coefficient for arterial branching points identified in Zone C; bcv = Average values of branching coefficient for venular branching points identified in Zone C; afa = Average values of asymmetry factor for arterial branching points identified in Zone C; afv = Average values of asymmetry factor for venular branching points identified in Zone C; ldra = arteriolar length diameter ratio; ldrv = venular length diameter ratio; d0a–d2a = Fractal dimensions of the arterial vasculature; d0v–d2v = Fractal dimensions of the venular vasculature; torta = arteriolar tortuosity (log transformed); tortv = venular tortuosity (log transformed).

5.2.1. Intra-class correlation analysis

ICC values are shown in Table 4. Width-related measurements (CRAE, CRVE, AVR) and their standard deviations of width estimates showed excellent and significant agreement throughout compression levels (0.70-0.92), with the exception of AVR at the strongest com-

pression (0.59, p = 0.01) and the standard deviation for arteries at the highest compression levels (PQ = 10 and 20). The number of branching points detected were good to high for the first two compression levels (PQ 80 and 55), but mostly low and not significant for stronger compression. ICC values for other branching-related measurements (average branching coefficient, symmetry factor) were generally poor and only occasionally moderate throughout the levels, but none was significant (p > 0.1 for all but two cases). Values of the venular length diameter ratio showed poor to moderate correlation, with only one value significant (p < 0.001). FD values for both arteries and veins indicated excellent agreement at all levels, indeed the best in the table, suggesting that FD estimates of the network complexity are not affected by compression. Arteriolar tortuosity showed moderate to excellent agreement, significant throughout; agreement for venular tortuosity was however lower and significant (p < 0.001) only at the mildest compression level (PQ = 80).

We conclude that, within the boundaries of our data set and protocols, FD measurements are the most reliable with compressed images, followed by width-related measurements. Measurement related to branching and junctions, as well as tortuosity, do not appear to remain sufficiently stable once images are compressed.

Table 4: **ICC analysis of VAMPIRE measures for HRF dataset**. Note: crae = central retinal arteriolar equivalent; crve = central retinal vein equivalent; avr = arteriole-to-venule ratio; bstda = standard deviation of arteriolar widths in Zone B; bstdv = standard deviation of venular widths in Zone B; ba = number of first branching points in arterial vessel in Zone C; bv = number of first branching points in venular vessel in Zone C; bca = Average values of branching coefficient for arterial branching points identified in Zone C; bcv = Average values of branching coefficient for venular branching points identified in Zone C; afa = Average values of asymmetry factor for arterial branching points identified in Zone C; afv = Average values of asymmetry factor for venular branching points identified in Zone C; ldra = arteriolar length diameter ratio; ldrv = venular length diameter ratio; d0a–d2a = Fractal dimensions of the arterial vasculature; d0v–d2v = Fractal dimensions of the venular vasculature; torta = arteriolar tortuosity (log transformed); tortv = venular tortuosity (log transformed).

Features	**PQ100vsPQ10**		**PQ100vsPQ20**		**PQ100vsPQ35**		**PQ100vsPQ55**		**PQ100vsPQ80**	
	ICC	*p-value*	*ICC*	*p-value*	*ICC*	*p-value*	*ICC*	*p-value*	*ICC*	*p-value*
crae	**0.70**	**<0.001**	**0.77**	**<0.001**	**0.90**	**<0.001**	**0.90**	**<0.001**	**0.92**	**<0.001**
crve	**0.70**	**<0.001**	**0.82**	**<0.001**	**0.76**	**<0.001**	**0.90**	**<0.001**	**0.86**	**<0.001**
avr	**0.59**	**0.01**	**0.74**	**<0.001**	**0.72**	**<0.001**	**0.83**	**<0.001**	**0.81**	**<0.001**
bstda	0.33	0.15	**0.58**	**0.01**	**0.81**	**<0.001**	**0.85**	**<0.001**	**0.74**	**<0.001**
bstdv	**0.82**	**<0.001**	**0.85**	**<0.001**	**0.82**	**<0.001**	**0.89**	**<0.001**	**0.87**	**<0.001**
ba	0.22	0.26	0.45	0.06	**0.69**	**<0.001**	**0.70**	**<0.001**	**0.76**	**<0.001**
bv	0.40	0.10	**0.77**	**<0.001**	0.35	0.13	**0.66**	**<0.001**	**0.68**	**<0.001**
bca	-0.33	0.77	-0.04	0.54	-0.14	0.63	0.18	0.31	-0.68	0.91
bcv	-0.04	0.54	-0.62	0.89	0.07	0.43	0.55	0.02	0.38	0.11
afa	0.34	0.15	0.25	0.22	-0.02	0.52	0.48	0.05	-0.45	0.83
afv	0.40	0.09	-0.13	0.62	-0.08	0.58	0.07	0.43	0.30	0.18
ldra	0.09	0.41	0.34	0.14	0.12	0.37	0.43	0.07	-0.24	0.71

Continued on next page

Table 4 – *Continued from previous page*

Features	PQ100vsPQ10		PQ100vsPQ20		PQ100vsPQ35		PQ100vsPQ55		PQ100vsPQ80	
	ICC	*p-value*	*ICC*	*p-value*	*ICC*	*p-value*	*ICC*	*p-value*	*ICC*	*p-value*
ldrv	0.45	0.06	-0.16	0.65	**0.74**	**<0.001**	**0.58**	**0.01**	**0.56**	**0.02**
doa	**0.94**	**<0.001**	**0.96**	**<0.001**	**0.98**	**<0.001**	**0.97**	**<0.001**	**0.93**	**<0.001**
d1a	**0.94**	**<0.001**	**0.97**	**<0.001**	**0.98**	**<0.001**	**0.97**	**<0.001**	**0.93**	**<0.001**
d2a	**0.94**	**<0.001**	**0.97**	**<0.001**	**0.98**	**<0.001**	**0.97**	**<0.001**	**0.93**	**<0.001**
dov	**0.97**	**<0.001**	**0.96**	**<0.001**	**0.96**	**<0.001**	**0.97**	**<0.001**	**0.97**	**<0.001**
d1v	**0.97**	**<0.001**	**0.96**	**<0.001**	**0.96**	**<0.001**	**0.97**	**<0.001**	**0.97**	**<0.001**
d2v	**0.97**	**<0.001**	**0.96**	**<0.001**	**0.96**	**<0.001**	**0.96**	**<0.001**	**0.97**	**<0.001**
torta	**0.70**	**<0.001**	**0.83**	**<0.001**	**0.78**	**<0.001**	**0.91**	**<0.001**	**0.86**	**<0.001**
tortv	**0.54**	**0.03**	0.38	0.11	0.14	0.34	**0.63**	**0.01**	**0.80**	**<0.001**

5.2.2. Correlation Analysis

Values between PQ levels and the reference one (PQ=100, no compression) are shown in Table 5. As with ICC results, width-related values, including the standard deviations of width measurements, showed moderate to high correlation, significant throughout, with the exception of avr and bstda at strong compression levels (low correlation and/or not significant). The number of branching points *ba* showed low to moderate correlation, mostly not significant. Correlation for other branching-related measurements was poor to moderate, and never significant. Length diameter ratios led to poor correlation values (not significant), with occasional good values but not significant. FD measures showed excellent and significant correlation ($r \geq 0.89$, $p < 0.001$) throughout the compression levels, confirming ICC results. Finally, values for arteriolar tortuosity were good to very good and significant; for venular tortuosity, a good and significant correlation was achieved only for the mildest compression (PQ=80).

Table 5: Correlation analysis of VAMPIRE measures for HRF dataset. Variable names as per Table 4.

Features	PQ100vsPQ10		PQ100vsPQ20		PQ100vsPQ35		PQ100vsPQ55		PQ100vsPQ80	
	r	*p-value*	*r*	*p-value*	*r*	*p-value*	*r*	*p-value*	*r*	*p-value*
crae	**0.58**	**<0.001**	**0.64**	**<0.001**	**0.81**	**<0.001**	**0.82**	**<0.001**	**0.85**	**<0.001**
crve	**0.54**	**<0.001**	**0.71**	**<0.001**	**0.61**	**<0.001**	**0.84**	**<0.001**	**0.77**	**<0.001**
avr	**0.42**	**0.03**	**0.59**	**<0.001**	**0.58**	**<0.001**	**0.72**	**<0.001**	**0.68**	**<0.001**
bstda	**0.20**	**0.32**	**0.43**	**0.02**	**0.67**	**<0.001**	**0.78**	**<0.001**	**0.60**	**<0.001**
bstdv	**0.70**	**<0.001**	**0.74**	**<0.001**	**0.69**	**<0.001**	**0.81**	**<0.001**	**0.77**	**<0.001**
ba	0.12	0.54	0.30	0.12	**0.53**	**<0.001**	**0.54**	**<0.001**	**0.62**	**<0.001**
bv	0.26	0.18	**0.63**	**<0.001**	0.21	0.27	**0.50**	**0.01**	**0.52**	**0.01**
bca	-0.14	0.47	-0.02	0.91	-0.07	0.73	0.11	0.59	-0.27	0.16
bcv	-0.02	0.93	-0.24	0.22	0.04	0.85	0.38	0.04	0.25	0.19

Continued on next page

Table 5 – *Continued from previous page*

Features	**PQ100vsPQ10**		**PQ100vsPQ20**		**PQ100vsPQ35**		**PQ100vsPQ55**		**PQ100vsPQ80**	
	r	*p-value*	*r*	*p-value*	*r*	*p-value*	*r*	*p-value*	*r*	*p-value*
afa	0.25	0.20	0.16	0.42	-0.02	0.92	0.35	0.06	-0.33	0.09
afv	0.32	0.10	-0.08	0.69	-0.05	0.80	0.05	0.79	0.20	0.30
ldra	0.05	0.81	0.25	0.21	0.08	0.70	0.36	0.06	-0.14	0.46
ldrv	0.30	0.13	-0.07	0.71	**0.61**	**<0.001**	0.42	0.03	0.39	0.04
doa	**0.89**	**<0.001**	**0.93**	**<0.001**	**0.96**	**<0.001**	**0.95**	**<0.001**	**0.89**	**<0.001**
d1a	**0.90**	**<0.001**	**0.94**	**<0.001**	**0.96**	**<0.001**	**0.95**	**<0.001**	**0.89**	**<0.001**
d2a	**0.89**	**<0.001**	**0.94**	**<0.001**	**0.96**	**<0.001**	**0.95**	**<0.001**	**0.89**	**<0.001**
dov	**0.94**	**<0.001**	**0.93**	**<0.001**	**0.93**	**<0.001**	**0.94**	**<0.001**	**0.94**	**<0.001**
d1v	**0.94**	**<0.001**	**0.92**	**<0.001**	**0.93**	**<0.001**	**0.94**	**<0.001**	**0.94**	**<0.001**
d2v	**0.94**	**<0.001**	**0.92**	**<0.001**	**0.93**	**<0.001**	**0.94**	**<0.001**	**0.94**	**<0.001**
torta	**0.56**	**<0.001**	**0.73**	**<0.001**	**0.64**	**<0.001**	**0.85**	**<0.001**	**0.76**	**<0.001**
tortv	0.39	0.04	0.23	0.23	0.08	0.69	**0.46**	**0.01**	**0.67**	**<0.001**

5.2.3. Bland-Altman analysis

For conciseness we include only a selection of Bland–Altman plots, namely for CRAE, CRVE, AVR, FD (doa, dov) and tortuosity. Figure 10 shows the plots for CRAE, CRVE and AVR. Difference means for CRAE and CRVE are comparable to or smaller than the sub-pixel accuracy achievable with automatic systems estimating vessel width [69] throughout the compression levels. For CRVE, lower compression levels lead apparently to lower mean differences, but these are well within the uncertainty of the width estimation algorithm. In general, the plots confirm good stability for these width-related measurements.

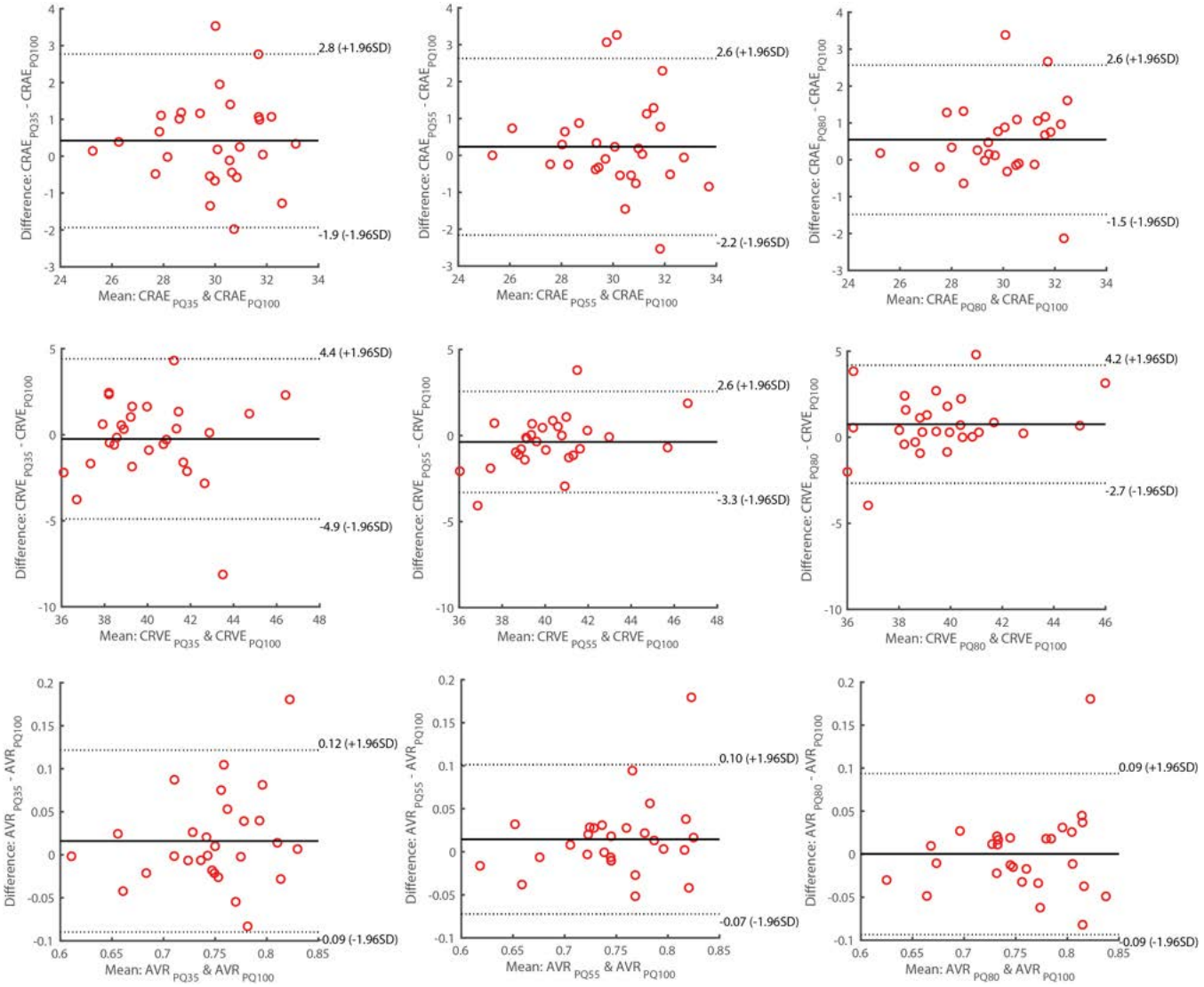

Figure 10: Bland–Altman plots for width-related measurements. Rows are measurements (CRAE, CRVE, AVR), columns pairs of compression levels (PQ100 vs. PQ35, PQ100 vs. PQ55, PQ100 vs. PQ80). The horizontal solid line is the mean value; dotted lines are the limits of the 95% confidence intervals at $\pm$ 1.96 SD. CRAE, CRVE are in pixels; AVR is dimensionless.

FD (doa, dov) plots (Figure 11) confirm the excellent stability of the FD across compression levels suggested by ICC and correlation values, with mean differences either zero or within 0.07 (i.e. at most 16% of the range of the means). This is arguably due to the fact that the FD is a measure of the overall complexity of a structure, hence likely to suffer less from small changes of local properties like vessel width or junction connectivity, that might instead have significant effects on specific measures like CRVE, CRAE, AVR or the count of junctions.

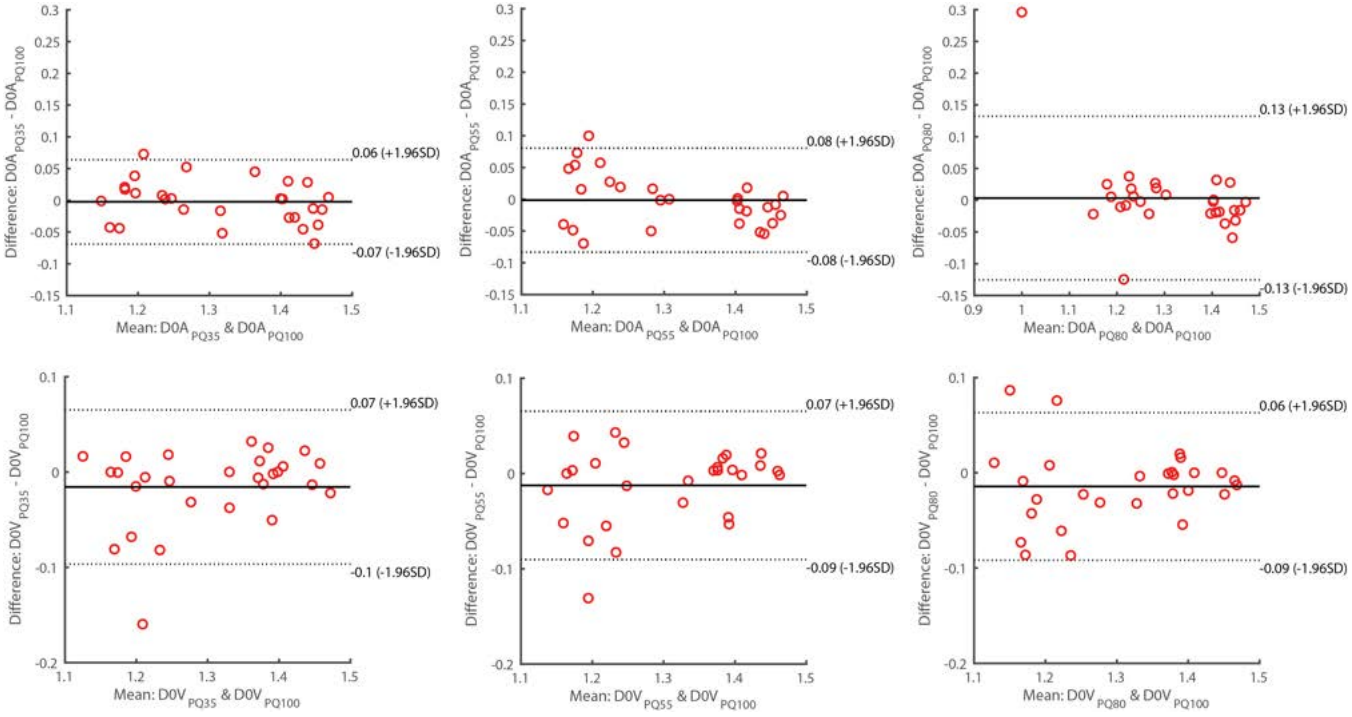

Figure 11: Bland-Altman plots for fractal dimension measurements. Rows are measurements (*doa*, *dov*), columns pairs of compression levels (PQ100 vs. PQ35, PQ100 vs. PQ55, PQ100 vs. PQ80). The horizontal solid line is the mean value; dotted lines are the limits of the 95% confidence intervals at $\pm$ 1.96 SD. FD measurements are dimensionless.

The plots for tortuosity measurements (log(torta), log(tortv) are shown in Figure 12. The confidence interval range, expressed as percentage of the range of the average mean between paired measurements, is between 24% (mildest compression) and 32% (strongest compression) for arteries and between 18% and 28% (strongest compression) for veins.

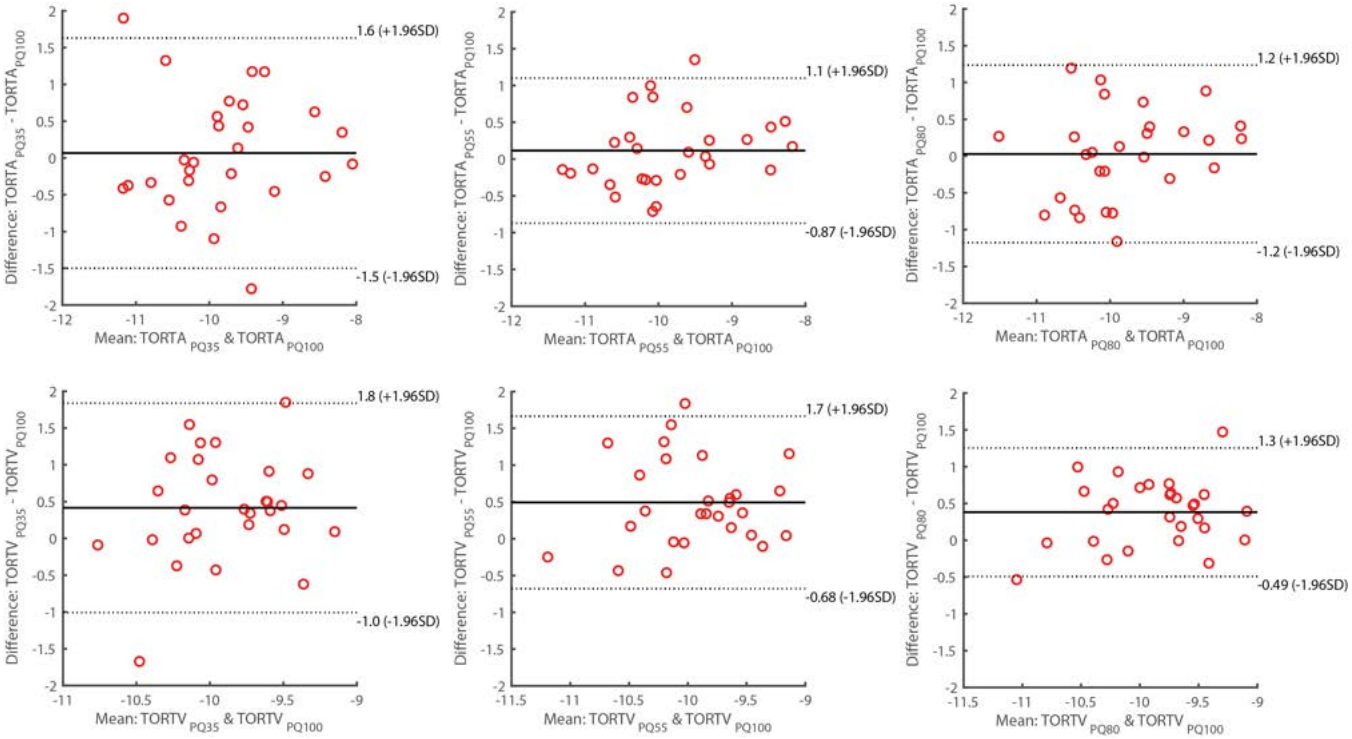

Figure 12: Bland-Altman plots for tortuosity measurements. Rows are measurements (log*torta*, log*tortv*), columns pairs of compression levels (PQ100 vs. PQ35, PQ100 vs. PQ55, PQ100 vs. PQ80). The horizontal solid line represents *mean* value and dotted lines the limits of the 95% confidence intervals at $\pm$ 1.96 SD. Tortuosity measurements are dimensionless.

6. Discussion

Our results show good to excellent stability of several VAMPIRE measurements against compression up to PQ 30-35 (PSNR $\sim$40db). This suggests the suitability of VAMPIRE for biomarkers studies with compressed image sets, given the compression levels normally applied to retinal images.

Figure 5 shows that the segmentation network performs better when trained using all compression levels together (CC protocol) than when trained with separate levels (SC protocol). This was consistent across the three datasets (Figure 6, 7 and 9).

With the SC protocol, the expected increase in false positive pixels for decreasing PQ (increasing compression) is not monotonic with the DRIVE and ARIA datasets. This is most likely caused by the low resolution of these images, amplifying the consequences of ringing artefacts [70] (Figure 8(a) and (h)). Similar observations were reported by Pauli et al. [70], who observed that increased compression ratios led to thicker vessels detected. DSC and AUC results are consistently monotonic from PQ10 to PQ100. Results are very good with HRF images, the resolution of which is comparable to that of contemporary fundus cameras and therefore relevant for applications. DSC and AUC remain high (0.81 and 0.97 respectively) even for strongly compressed images at PQ20 (Table A6). The lowest DSC (0.67) and AUC (0.937) for ARIA dataset (Table A4).

The average performance measures of Dense U-net segmentation is compared with state-of-the-art methods are listed on Table. 6 for DRIVE, ARIA, and HRF. Results shows that Dense U-net performance is comparable with all other published methods. For the DRIVE dataset, sensitivity is higher (84.53%) for PQ level 100 than those of published algorithms and with significant compression (PQ35) the sensitivity is 81.41%. Specificity and accuracy (98.12% and 96.99%) are slightly lower than Park's (98.35% and 97.06%) [71]. For the ARIA dataset, sensitivity, specificity, and accuracy are 81.39%, 97.5%3, and 95.60%, higher than published method [72]. For the HRF dataset, specificity is 99.24%, higher than Zhoa's [73] and accuracy is 97.24%, comparable with Park's (97.61%) [71].

Table 6: **Performance comparison of Dense U-net segmentation with state-of-the-art methods for the DRIVE, ARIA, and HRF dataset.**

Methods	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**
DRIVE			
Azzopardi et al. (2015) [74]	76.55	97.04	94.42
Imani et al. (2015) [75]	75.24	97.25	95.73
Li et al. (2015) [76]	75.69	98.16	95.27
Vega et al. (2015) [77]	74.44	96	94.12
Zhao et al. (2015) [78]	74.20	98.20	95.40
Orlando et al. (2016) [79]	78.97	96.84	–
Liskowski et al. (2016) [80]	74.17	98.04	94.73
Zhou et al. (2017) [81]	80.78	96.74	94.69
Zhao et al. (2018) [73]	80.38	98.15	–
Park et al. (2020) [71]	83.46	**98.36**	**97.06**
Dense U-net (SC-PQ35)	81.41	97.04	95.74
Dense U-net (SC-PQ100)	**84.53**	98.12	96.99
ARIA			
Zhao et al. (2015) [72]	75.10	93	94
Dense U-net (SC-PQ35)	**81.39**	95.05	93.98
Dense U-net (SC-PQ100)	73.09	**97.53**	**95.60**
HRF			
Orlando et al. (2016) [79]	78.74	95.84	–
Zhou et al. (2017) [81]	**80.15**	96.99	95.44
Jiang et al. (2018) [82]	80.10	80.10	96.50
Yan et al. (2018) [83]	78.81	95.92	94.37
Zhao et al. (2018) [73]	80.01	98.23	–
Park et al. (2020) [71]	–	–	**97.61**
Dense U-net (SC-PQ35)	78.38	99.01	96.99
Dense U-net (SC-PQ100)	78.81	**99.24**	97.24

Although not investigated systematically, the VAMPIRE Dense U-net has achieved good results also in the presence of lesions. Figure 13 shows magnified image details with haem-

orrhages, hard exudates, and strong central reflexes. The network is able to classify most vessel pixels correctly even with significant compression, PQ level 35 (Figure 13(g), (h), and (i)). Previous studies have concentrated on the effect of JPEG compression on DR grading [7, 8, 84, 85] and drusen detection [48], with very few considering automated microaneurysm detection [86, 87].

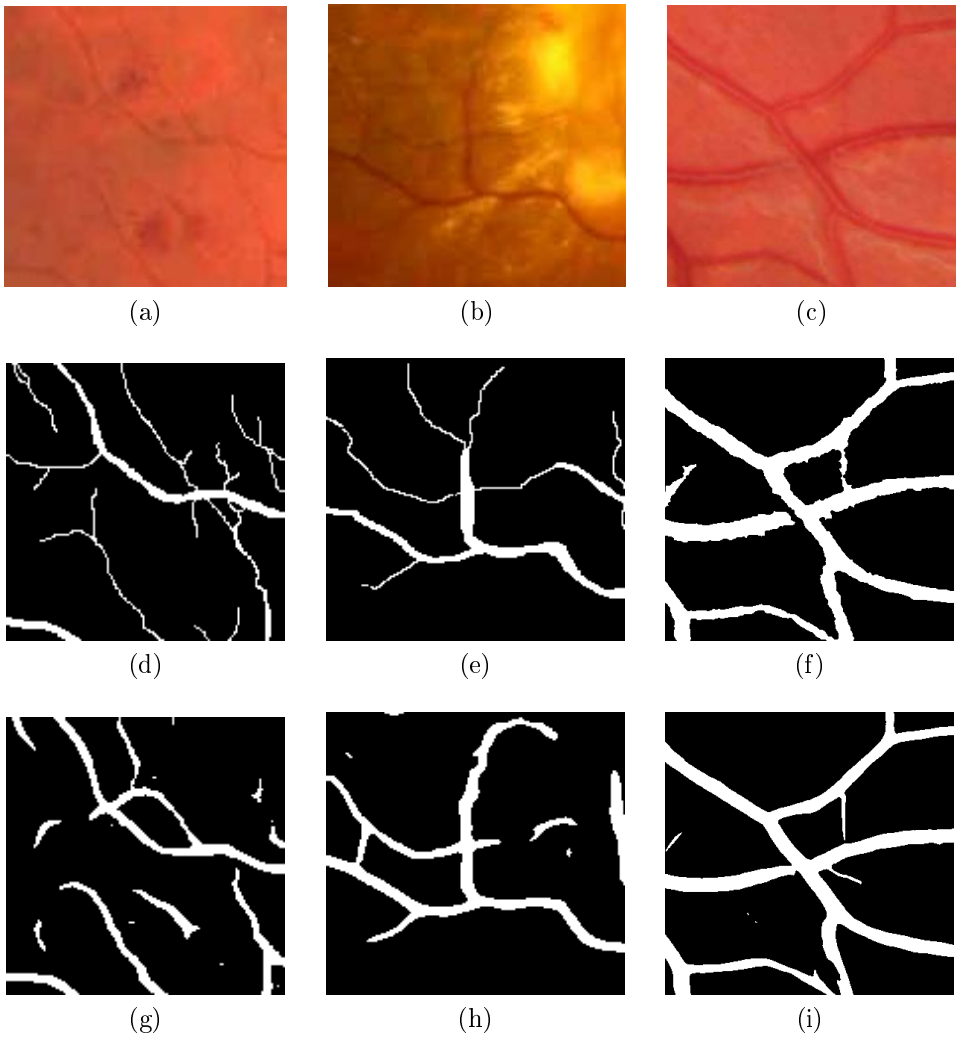

Figure 13: Segmentation details in the presence of lesions. (a) Haemorrhages (DRIVE); (b) hard exudates (ARIA); (c) strong central reflex (HRF); (d), (e), (f) corresponding ground truth; (g), (h), (i) results obtained using Dense U-net with PQ35.

The correlation analysis indicates that several VAMPIRE measurements remain considerably stable against compression. CRAE, CRVE, and AVR showed moderate to excellent ICC agreement with those achieved with uncompressed images (Table 4). Pearson correlation is lower than ICC, which suggests good agreement but not linear dependence.

Arteriolar and venular branching in Zone C had good to moderate agreement with uncom-

pressed images down to PQ35, suggesting that compression upsets the detection of branching points. ICC analysis and Bland-Altman plots showed that fractal dimension (FD) measures had excellent agreement for all compression levels compared with PQ100, highlighting this measure as a prime candidate for association studies with compressed data sets. Arteriolar tortuosity showed better agreement than venular tortuosity in general.

This study has several limitations, that we list in no specific order. First, the need for manually traced vessel maps limited the amount of images usable and was the reason for using low-resolution image sets. The usefulness of results from such images may be considered limited, but becomes significant when considering historical image sets collected when fundus cameras only offered low resolution. Second, we compressed images with the MATLAB JPEG compression function; we did not explore the many parameters of the JPEG algorithm. The conclusions obtained from a systematic analysis of different compression levels seem sufficiently informative for practical use. Third, the range of compression levels covered in our study is representative, to our best knowledge, of those used to store retinal images in clinical centres. Before applying our conclusions to compressed clinical images, one should check that their compression level agrees with our range, e.g. from information included in the image header. Finally, reference VAMPIRE measurements were taken by a single, highly experienced VAMPIRE instructor (more than 10,000 images processed). In our work, operator training has repeatedly shown that the inter-operator variability of certified VAMPIRE operators is modest (ICC > 0.85 approximately).

7. Conclusions

We have studied the variations induced by JPEG compression on vessel segmentation maps and morphometric measurements of the retinal vasculature obtained with the VAMPIRE software tool and three public data sets of fundus retinal images.

To our best knowledge, a detailed analysis of the effects of compression on vessel segmentation and morphometric vascular measures at progressive levels had not yet been reported. A quantitative assessment of such variations provides a statistical estimate of the uncertainty on measurements. This seems important when measurements are used as candidate biomarkers in statistical studies and only compressed images are available, as in the case of some retrospective archives of clinical images, a potentially large source of image data. We have identified VAMPIRE measurements offering good to excellent stability under compression at levels used for retinal fundus camera images.

Looking at the future, a full characterization of uncertainty in retinal measurements would enable us to understand the smallest measurement values needed to detect reliably a given effect, identify measurement ranges that can be used reliably in association studies, and use such ranges in statistical methods taking into account uncertainty in a principled way.

Acknowledgements

This research was funded by the National Institute for Health Research (NIHR) (INSPIRED 16/136/102) using UK aid from the UK Government to support global health

research. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR or the UK Department of Health and Social Care.

We thank our INSPIRED colleagues at the Madras Diabetes Research Foundation and Dr Mohan's Diabetes Speciality Hospitals, and the members of the VAMPIRE and CVIP groups at the universities of Dundee and Edinburgh, for many useful discussions and feedback.

Acronyms

AVR Arteriolar-Venular width Ratio

ACC accuracy

ARIA Automated Retinal Image Analyzer

AMD age-related macular degeneration

AUC area under the curve

CRAE central retinal artery equivalent

CRVE central retinal vein equivalent

CC combined compression

DR diabetic retinopathy

DRIVE Digital Retinal Images for Vessel Extraction

FOV field of view

FD fractal dimension

HRF high-resolution fundus

ICC intra-class correlation coefficient

JPEG Joint Photographic Experts Group

LOA limit of agreement

MRI magnetic resonance imaging

OD optic disk

PQ perceptual quality

PSNR peak signal-to-noise ratio

SD standard deviation

SE sensitivity

SP specificity

SC separate compression

DSC DICE similarity coefficient

UWFV SLO ultra-wide field of view scanning laser ophthalmoscopy

VAMPIRE Vascular Assessment and Measurement Platform for Images of the REtina

Bibliography

[1] T. MacGillivray, E. Trucco, J. Cameron, B. Dhillon, J. Houston, E. Van Beek, Retinal imaging as a source of biomarkers for diagnosis, characterization and prognosis of chronic illness or long-term conditions, The British journal of radiology 87 (1040) (2014) 20130832.

[2] S. McGrory, A. M. Taylor, E. Pellegrini, L. Ballerini, M. Kirin, F. N. Doubal, J. M. Wardlaw, A. S. Doney, B. Dhillon, . J. M. Starr, et al., Towards standardization of quantitative retinal vascular parameters: comparison of siva and vampire measurements in the lothian birth cohort 1936, Translational vision science & technology 7 (2) (2018) 12–12.

[3] L. Csincsik, N. Quinn, T. J. MacGillivray, T. J. Shakespeare, T. J. Crutch, T. Peto, I. Lengyel, The retina as a window to the brain-from eye research to cns disorders, Alzheimer and dementia 15 (7) (2019) P344.

[4] P. Remond, F. Aptel, P. Cunnac, J. Labarere, K. Palombi, J.-L. Pepin, F. Pollet-Villard, S. Hogg, R. Wang, T. MacGillivray, E. Trucco, C. Chiquet, Retinal vessel phenotype in patients with nonarteritic anterior ischemic optic neuropathy, Am Journ Ophthalmology 208 (2019) 178–184.

[5] A. Azanan, S. Mohamad, S. Chandrasekaran, E. S. Rosli, L.-L. Chua, L. Oh, T. F. Chin, T. Y. Y. andR Rajagopal, R. Rajasuriar, T. J. MacGillivray, E. Trucco, N. Ramli, T. A. Kamalden, H. Ariffin, Retinal vessel analysis as a novel screening tool to identify childhood acute lymphoblastic leukemia survivors at risk of cardiovascular disease, Journ Pediatric Hematology/Oncology`doi:10.1097/MPH.0000000000001766`.

[6] R. H. Eikelboom, K. Yogesan, C. J. Barry, I. J. Constable, M.-L. Tay-Kearney, L. Jitskaia, P. H. House, Methods and limits of digital image compression of retinal images for telemedicine, Investigative ophthalmology & visual science 41 (7) (2000) 1916–1924.

[7] R. Newsom, A. Clover, M. Costen, J. Sadler, J. Newton, A. Luff, C. Canning, Effect of digital image compression on screening for diabetic retinopathy, British journal of ophthalmology 85 (7) (2001) 799–802.

[8] C. F. Baker, C. J. Rudnisky, M. T. Tennant, P. Sanghera, B. J. Hinz, A. R. De Leon, M. D. Greve, Jpeg compression of stereoscopic digital images for the diagnosis of diabetic retinopathy via teleophthalmology, Canadian Journal of Ophthalmology/Journal Canadien d'Ophtalmologie 39 (7) (2004) 746–754.

[9] S. McGrory, J. R. Cameron, E. Pellegrini, C. Warren, F. N. Doubal, I. J. Deary, B. Dhillon, J. M. Wardlaw, E. Trucco, T. J. MacGillivray, The application of retinal fundus camera imaging in dementia: a systematic review, Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 6 (2017) 91–107.

[10] M. R. K. Mookiah, S. McGrory, S. Hogg, J. Price, R. Forster, T. J. MacGillivray, E. Trucco, Towards standardization of retinal vascular measurements: on the effect of image centering, in: Computational Pathology and Ophthalmic Medical Image Analysis, Springer, 2018, pp. 294–302.

[11] A. London, I. Benhar, M. Schwartz, The retina as a window to the brain-from eye research to cns disorders, Nature Reviews Neurology 9 (1) (2013) 44.

[12] C. Sabanayagam, T. Y. Wong, Diabetic Retinopathy and Cardiovascular Disease, Karger, 2019.

[13] M. R. K. Mookiah, U. R. Acharya, C. K. Chua, C. M. Lim, E. Ng, A. Laude, Computer-aided diagnosis of diabetic retinopathy: A review, Computers in biology and medicine 43 (12) (2013) 2136–2155.
[14] S. McGrory, L. Ballerini, F. N. Doubal, J. Staals, M. Allerhand, M. d. C. Valdes-Hernandez, X. Wang, T. MacGillivray, A. S. Doney, B. Dhillon, et al., Retinal microvasculature and cerebral small vessel disease in the lothian birth cohort 1936 and mild stroke study, Scientific reports 9 (1) (2019) 1–11.
[15] M. R. K. Mookiah, U. R. Acharya, H. Fujita, J. H. Tan, C. K. Chua, S. V. Bhandary, A. Laude, L. Tong, Application of different imaging modalities for diagnosis of diabetic macular edema: a review, Computers in biology and medicine 66 (2015) 295–315.
[16] A. Jindal, I. Ctori, Bruno, Fidalgo, P. Dabasia, K. Balaskas, J. G. Lawrenson, Impact of optical coherence tomography on diagnostic decision?making by uk community optometrists: a clinical vignette study, Ophthalmic and Physiological Optics 39 (3) 205–205.
[17] M. Diaz, J. Novo, P. Cutrin, F. Gomez-Ulla, M. G. Penedo, O. M, Automatic segmentation of the foveal avascular zone in ophthalmological oct-a images, PLOS-ONE 14 (2) (2019) e0212364.
[18] Y. He, W. Jiao, Y. Shi, J. Lian, B. Zhao, W. Zou, Y. Zhu, Segmenting diabetic retinopathy lesions in multispectral images using low-dimensional spatial-spectral matrix representation, IEEE Journal of Biomedical and Health Informatics 4 (2) (2020) 493–502.
[19] A. R. Harvey, G. Carles, A. Bradu, A. Podoleanu, The physics, instruments and modalities of retinal imaging, 1st Edition, MICCAI Elsevier Series, Academic Press, 2019, Ch. 3, pp. 19–54.
[20] A. Shah, S. Lynch, M. Niemeijer, R. Amelon, W. Clarida, J. Folk, S. Russell, X. Wu, M. D. AbrÃ moff, Susceptibility to misdiagnosis of adversarial images by deep learning based retinal image analysis algorithms, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018, pp. 1454–1457.
[21] T. Y. Wong, F. A. Islam, R. Klein, B. E. Klein, M. F. Cotch, C. Castro, A. R. Sharrett, E. Shahar, Retinal vascular caliber, cardiovascular risk factors, and inflammation: the multi-ethnic study of atherosclerosis (mesa), Investigative ophthalmology & visual science 47 (6) (2006) 2341–2350.
[22] A. E. Fetit, A. Doney, S. Hogg, R. Wang, T. J. MacGillivray, J. M. Wardlaw, F. Doubal, G. J. McKay, S. J. McKenna, E. Trucco, A multimodal approach to cardiovascular risk stratification in patients with type 2 diabetes incorporating retinal, genomic and clinical features, Scientific Reports 9 (1) (2019) 1–10.
[23] R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. McConnell, G. S. Corrado, L. Peng, D. R. Webster, Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning, Nature Biomedical Engineering 2 (3) (2018) 158.
[24] Y. D. Kim, K. J. Noh, S. J. Byun, S. Lee, T. Kim, L. Sunwoo, K. J. Lee, S.-H. Kang, K. H. Park, S. J. Park, Effects of hypertension, diabetes, and smoking on age and sex prediction from retinal fundus images, Scientific reports 10 (1) (2020) 1–14.
[25] A. Perez-Rovira, T. MacGillivray, E. Trucco, K. Chin, K. Zutis, C. Lupascu, D. Tegolo, A. Giachetti, P. J. Wilson, A. Doney, et al., Vampire: vessel assessment and measurement platform for images of the retina, in: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2011, pp. 3391–3394.
[26] E. Trucco, L. Ballerini, D. Relan, A. Giachetti, T. MacGillivray, K. Zutis, C. Lupascu, D. Tegolo, E. Pellegrini, G. Robertson, et al., Novel vampire algorithms for quantitative analysis of the retinal vasculature, in: 2013 ISSNIP Biosignals and Biorobotics Conference: Biosignals and Robotics for Better and Safer Living (BRC), IEEE, 2013, pp. 1–4.
[27] E. Trucco, A. Giachetti, L. Ballerini, D. Relan, A. Cavinato, T. MacGillivray, Morphometric measurements of the retinal vasculature in fundus images with vampire, Biomedical Image Understanding (2015) 91–111.
[28] C. Y. Cheung, W. T. Tay, P. Mitchell, J. J. Wang, W. Hsu, M. L. Lee, Q. P. Lau, A. L. Zhu, R. Klein, S. M. Saw, et al., Quantitative and qualitative retinal microvascular characteristics and blood pressure, Journal of hypertension 29 (7) (2011) 1380–1391.
[29] P. Bankhead, C. N. Scholfield, J. G. McGeown, T. M. Curtis, Fast retinal vessel detection and measurement using wavelets and edge location refinement, PloS one 7 (3) (2012) e32435.

[30] M. M. Fraz, R. Welikala, A. R. Rudnicka, C. G. Owen, D. Strachan, S. A. Barman, Quartz: Quantitative analysis of retinal vessel topology and size–an automated system for quantification of retinal vessels morphology, Expert Systems with Applications 42 (20) (2015) 7221–7234.
[31] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, et al., Clinically applicable deep learning for diagnosis and referral in retinal disease, Nature Medicine 24 (9).
[32] J. De Fauw, J. Ledsam, R.-P. Bernardino., et al., Clinically applicable deep learning for diagnosis and referral in retinal disease, Nature Medicine 24 (2016) 1342–?1350.
[33] P. Burlina, A. Galdran, P. Costa, A. Cohen, A. Campilho, Artificial intelligence and deep learning in retinal image analysis, 1st Edition, MICCAI Elsevier Series, Academic Press, 2019, Ch. 18, pp. 379–404.
[34] L. Zhang, M. Yuan, Z. An, X. Zhao, H. Wu, H. Li, Y. Wang, B. Sun, H. Li, S. Ding, et al., Prediction of hypertension, hyperglycemia and dyslipidemia from retinal fundus photographs via deep learning: A cross-sectional study of chronic diseases in central china, PloS one 15 (5) (2020) e0233166.
[35] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (xai), IEEE Access 6 (2018) 52138–52160.
[36] E. Trucco, T. MacGillivray, Y. e. Xu, Computational Retinal Image Analysis: Tools, Applications and Perspectives, Academic Press, 2019.
[37] E. Trucco, A. Ruggeri, T. Karnowski, L. Giancardo, E. Chaum, J. P. Hubschman, B. Al-Diri, C. Y. Cheung, D. Wong, M. Abramoff, et al., Validating retinal fundus image analysis algorithms: issues and a proposal, Investigative ophthalmology & visual science 54 (5) (2013) 3546–3559.
[38] L. S. Lim, C. Y.-l. Cheung, X. Lin, P. Mitchell, T. Y. Wong, S. Mei-Saw, Influence of refractive error and axial length on retinal vessel geometric characteristics, Investigative ophthalmology & visual science 52 (2) (2011) 669–678.
[39] C. S. Chandler, S. Gangaputra, L. D. Hubbard, N. J. Ferrier, T. W. Pauli, Q. Peng, D. W. Thayer, R. P. Danis, Suboptimal image focus broadens retinal vessel caliber measurement, Investigative ophthalmology & visual science 52 (12) (2011) 8558–8561.
[40] H. M. Pakter, S. C. Fuchs, M. K. Maestri, L. B. Moreira, L. M. Dei Ricardi, V. F. Pamplona, M. M. Oliveira, F. D. Fuchs, Computer-assisted methods to evaluate retinal vascular caliber: what are they measuring?, Investigative ophthalmology & visual science 52 (2) (2011) 810–815.
[41] T. Stosic, B. D. Stosic, Multifractal analysis of human retinal vessels, IEEE transactions on medical imaging 25 (8) (2006) 1101–1107.
[42] S. Lemmens, A. Devulder, K. Van Keer, J. Bierkens, P. De Boever, I. Stalmans, Systematic review on fractal dimension of the retinal vasculature in neurodegeneration and stroke: Assessment of a potential biomarker, Fontiers in Neuroscience 14.
[43] F. Huang, J. Zhang, E. J. Bekkers, B. Dashtbozorg, B. M. ter Haar Romeny, Stability analysis of fractal dimension in retinal vasculature.
[44] S. Krohn, M. Froeling, A. Leemans, D. Ostwald, P. Villoslada, C. Finke, F. J. Esteban, Evaluation of the 3d fractal dimension as a marker of structural brain complexity in multiple-acquisition mri, Human brain mapping.
[45] F. Huang, B. Dashtbozorg, J. Zhang, E. Bekkers, S. Abbasi-Sureshjani, T. T. Berendschot, B. M. ter Haar Romeny, Reliability of using retinal vascular fractal dimension as a biomarker in the diabetic retinopathy detection, Journal of ophthalmology 2016.
[46] M. J. Cree, H. Jelinek, The effect of jpeg compression on automated detection of microaneurysms in retinal images, in: Image Processing: Machine Vision Applications, Vol. 6813, International Society for Optics and Photonics, 2008, p. 68130M.
[47] J. Conrath, A. Erginay, R. Giorgi, A. Lecleire-Collet, E. Vicaut, J. Klein, A. Gaudric, P. Massin, Evaluation of the effect of jpeg and jpeg2000 image compression on the detection of diabetic retinopathy, Eye 21 (4) (2007) 487.
[48] M. S. Lee, D. S. Shin, J. W. Berger, Grading, image analysis, and stereopsis of digitally compressed fundus images., Retina (Philadelphia, Pa.) 20 (3) (2000) 275–281.
[49] S. Akshay, P. Apoorva, Segmentation and classification of fmm compressed retinal images using watershed and canny segmentation and support vector machine, in: 2017 International Conference on

Communication and Signal Processing (ICCSP), IEEE, 2017, pp. 1035–1039.
[50] J. Daugman, C. Downing, Effect of severe image compression on iris recognition performance, IEEE Transactions on information Forensics and Security 3 (1) (2008) 52–61.
[51] C. López, J. J. Martinez, M. Lejeune, P. Escriva, M. T. Salvadó, L. E. Pons, T. Álvaro, J. Baucells, M. García-Rojo, X. Cugat, et al., Roundness variation in jpeg images affects the automated process of nuclear immunohistochemical quantification: correction with a linear regression model, Histochemistry and cell biology 132 (4) (2009) 469–477.
[52] G. Ozdemir, H. Nasifoglu, O. Erogul, Performance comparison of segmentation algorithms for image quality degraded mr images, in: World Congress on Medical Physics and Biomedical Engineering 2018, Springer, 2019, pp. 243–248.
[53] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, B. Van Ginneken, Ridge-based vessel segmentation in color images of the retina, IEEE transactions on medical imaging 23 (4) (2004) 501–509.
[54] M. H. A. Hijazia, F. Coenena, Y. Zhengb, A histogram approach for the screening of age-related macular degeneration.
[55] A. Budai, R. Bock, A. Maier, J. Hornegger, G. Michelson, Robust vessel segmentation in fundus images, International journal of biomedical imaging 2013.
[56] R. Hemelings, B. Elen, I. Stalmans, K. Van Keer, P. De Boever, M. B. Blaschko, Artery-vein segmentation in fundus images using a fully convolutional network, Computerized Medical Imaging and Graphics.
[57] C. Wang, Z. Zhao, Q. Ren, Y. Xu, Y. Yu, Dense u-net based on patch-based learning for retinal vessel segmentation, Entropy 21 (2) (2019) 168.
[58] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W. M. Wells, A. F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, 2015, pp. 234–241.
[59] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.
[60] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.
[61] C. Cong, Retina-vesselnet: a denseblock-unet for retinal blood vessel segmentation (Oct. 2019).
URL `https://github.com/DeepTrial/Retina-VesselNet`
[62] P. K. K. B. M. Bédard, Nancy J. Martin, Assessing reproducibility of data obtained with instruments based on continuous measurements, Experimental Aging Research 26 (4) (2000) 353–365.
[63] D. V. Cicchetti, Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology., Psychological assessment 6 (4) (1994) 284.
[64] J. Lee, D. Koh, C. Ong, Statistical evaluation of agreement between two methods for measuring a quantitative variable, Computers in biology and medicine 19 (1) (1989) 61–70.
[65] P. K. K. B. M. Bédard, Nancy J. Martin, Assessing reproducibility of data obtained with instruments based on continuous measurements, Experimental Aging Research 26 (4) (2000) 353–365.
[66] L. Portney, M. Watkins, Foundations of clinical research: applications to practice, ed 2, upper saddle river, nj, 2000, Pearson, Prentice Hall. Jaeschke R, Singer J, Guyatt GH: Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 10 (4) (1989) 407–415.
[67] J. M. Bland, D. G. Altman, Comparing methods of measurement: why plotting difference against standard method is misleading, The lancet 346 (8982) (1995) 1085–1087.
[68] J. M. Bland, D. Altman, Statistical methods for assessing agreement between two methods of clinical measurement, The lancet 327 (8476) (1986) 307–310.
[69] C. A. Lupaşcu, D. Tegolo, E. Trucco, Accurate estimation of retinal vessel width using bagged decision trees and an extended multiresolution hermite model, Medical image analysis 17 (8) (2013) 1164–1180.
[70] T. W. Pauli, S. Gangaputra, L. D. Hubbard, D. W. Thayer, C. S. Chandler, Q. Peng, A. Narkar, N. J. Ferrier, R. P. Danis, Effect of image compression and resolution on retinal vascular caliber, Investigative ophthalmology & visual science 53 (9) (2012) 5117–5123.

[71] K.-B. Park, S. H. Choi, J. Y. Lee, M-gan: Retinal blood vessel segmentation by balancing losses through stacked deep fully convolutional networks, IEEE Access 8 (2020) 146308–146322.
[72] Y. Zhao, Y. Liu, X. Wu, S. P. Harding, Y. Zheng, Retinal vessel segmentation: An efficient graph cut approach with retinex and local phase, PloS one 10 (4) (2015) e0122332.
[73] H. Zhao, H. Li, S. Maurer-Stroh, L. Cheng, Synthesizing retinal and neuronal images with generative adversarial nets, Medical image analysis 49 (2018) 14–26.
[74] G. Azzopardi, N. Strisciuglio, M. Vento, N. Petkov, Trainable cosfire filters for vessel delineation with application to retinal images, Medical image analysis 19 (1) (2015) 46–57.
[75] E. Imani, M. Javidi, H.-R. Pourreza, Improvement of retinal blood vessel detection using morphological component analysis, Computer methods and programs in biomedicine 118 (3) (2015) 263–279.
[76] Q. Li, B. Feng, L. Xie, P. Liang, H. Zhang, T. Wang, A cross-modality learning approach for vessel segmentation in retinal images, IEEE transactions on medical imaging 35 (1) (2015) 109–118.
[77] R. Vega, G. Sanchez-Ante, L. E. Falcon-Morales, H. Sossa, E. Guevara, Retinal vessel extraction using lattice neural networks with dendritic processing, Computers in biology and medicine 58 (2015) 20–30.
[78] Y. Zhao, L. Rada, K. Chen, S. P. Harding, Y. Zheng, Automated vessel segmentation using infinite perimeter active contour model with hybrid region information with application to retinal images, IEEE transactions on medical imaging 34 (9) (2015) 1797–1807.
[79] J. I. Orlando, E. Prokofyeva, M. B. Blaschko, A discriminatively trained fully connected conditional random field model for blood vessel segmentation in fundus images, IEEE transactions on Biomedical Engineering 64 (1) (2016) 16–27.
[80] P. Liskowski, K. Krawiec, Segmenting retinal blood vessels with deep neural networks, IEEE transactions on medical imaging 35 (11) (2016) 2369–2380.
[81] L. Zhou, Q. Yu, X. Xu, Y. Gu, J. Yang, Improving dense conditional random field for retinal vessel segmentation by discriminative feature learning and thin-vessel enhancement, Computer methods and programs in biomedicine 148 (2017) 13–25.
[82] Z. Jiang, H. Zhang, Y. Wang, S.-B. Ko, Retinal blood vessel segmentation using fully convolutional network with transfer learning, Computerized Medical Imaging and Graphics 68 (2018) 1–15.
[83] Z. Yan, X. Yang, K.-T. Cheng, Joint segment-level and pixel-wise losses for deep learning based retinal vessel segmentation, IEEE Transactions on Biomedical Engineering 65 (9) (2018) 1912–1923.
[84] A. Basu, A. Kamal, W. Illahi, M. Khan, P. Stavrou, R. Ryder, Is digital image compression acceptable within diabetic retinopathy screening?, Diabetic Medicine 20 (9) (2003) 766–771.
[85] C. Stellingwerf, P. Hardus, J. Hooymans, Assessing diabetic retinopathy using two-field digital photography and the influence of jpeg-compression, Documenta ophthalmologica 108 (3) (2004) 203–209.
[86] M. J. Cree, J. A. Olson, K. C. McHardy, P. F. Sharp, J. V. Forrester, A fully automated comparative microaneurysm digital detection system, Eye 11 (5) (1997) 622.
[87] P. Hänsgen, P. Undrill, M. Cree, The application of wavelets to retinal image compression and its effect on automatic microaneurysm analysis, Computer methods and programs in biomedicine 56 (1) (1998) 1–10.

Appendices

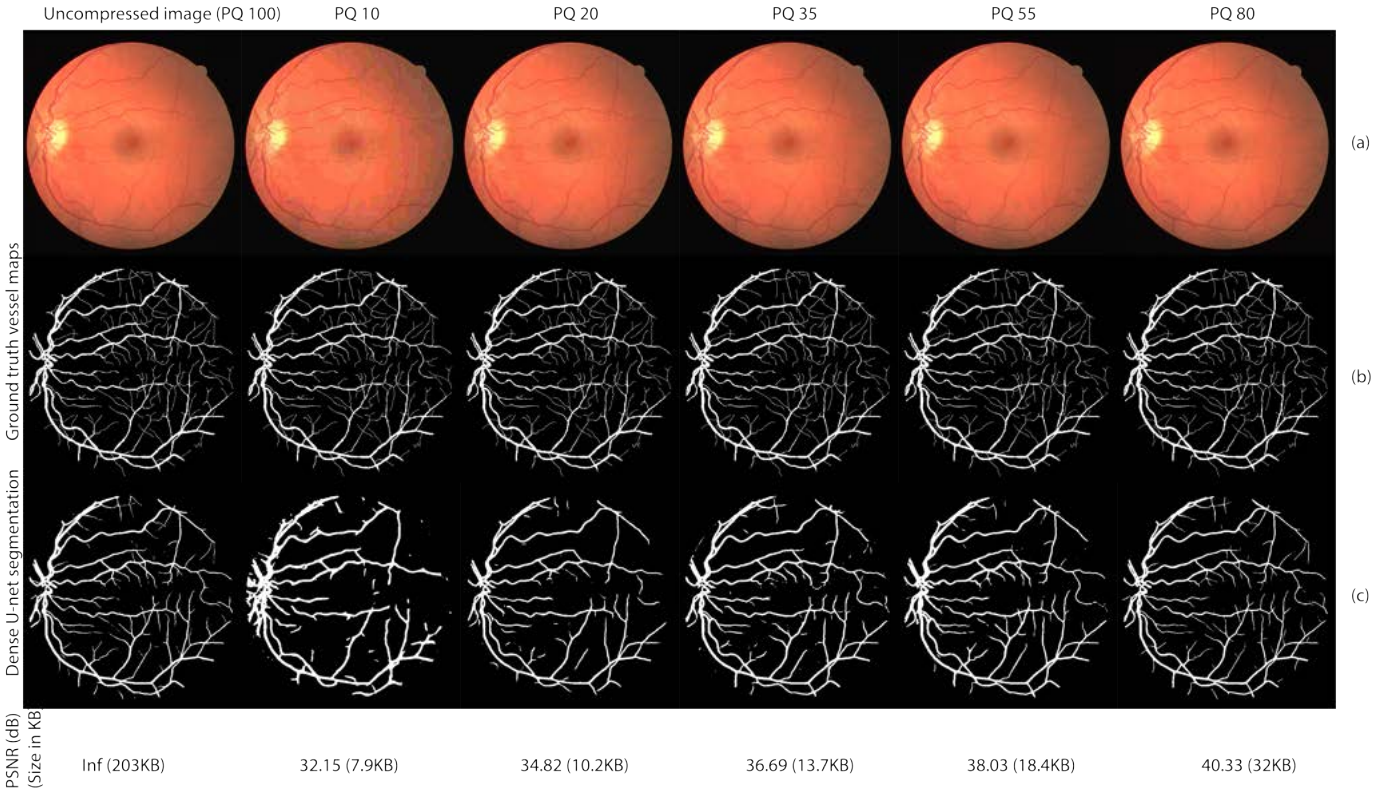

Figure A1: Segmentation results of DRIVE dataset. (a) Images compressed at PQ10, PQ20, PQ35, PQ55, PQ80,and PQ100 is uncompressed. (b) Ground truth vessel maps. (c) Results of Dense U-net segmentation. Best seen on a computer screen.

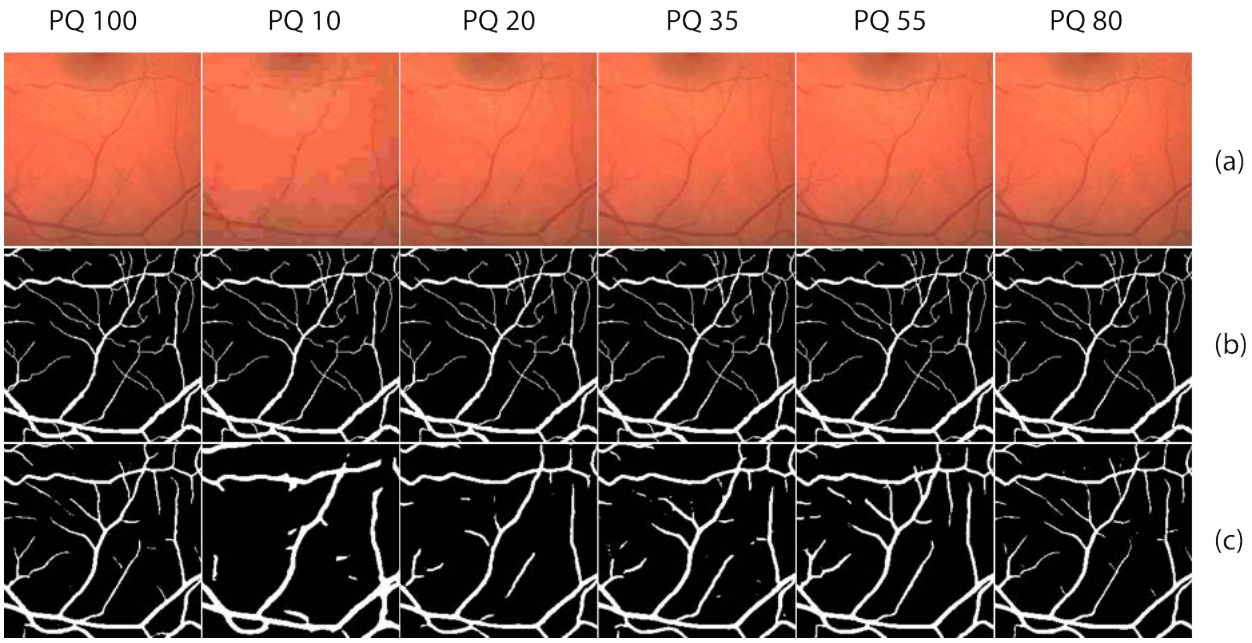

Figure A2: Segmentation results of DRIVE dataset. (a) Images compressed at PQ10, PQ20, PQ35, PQ55, PQ80,and PQ100 is uncompressed. (b) Ground truth vessel maps. (c) Results of Dense U-net segmentation. Best seen on a computer screen.

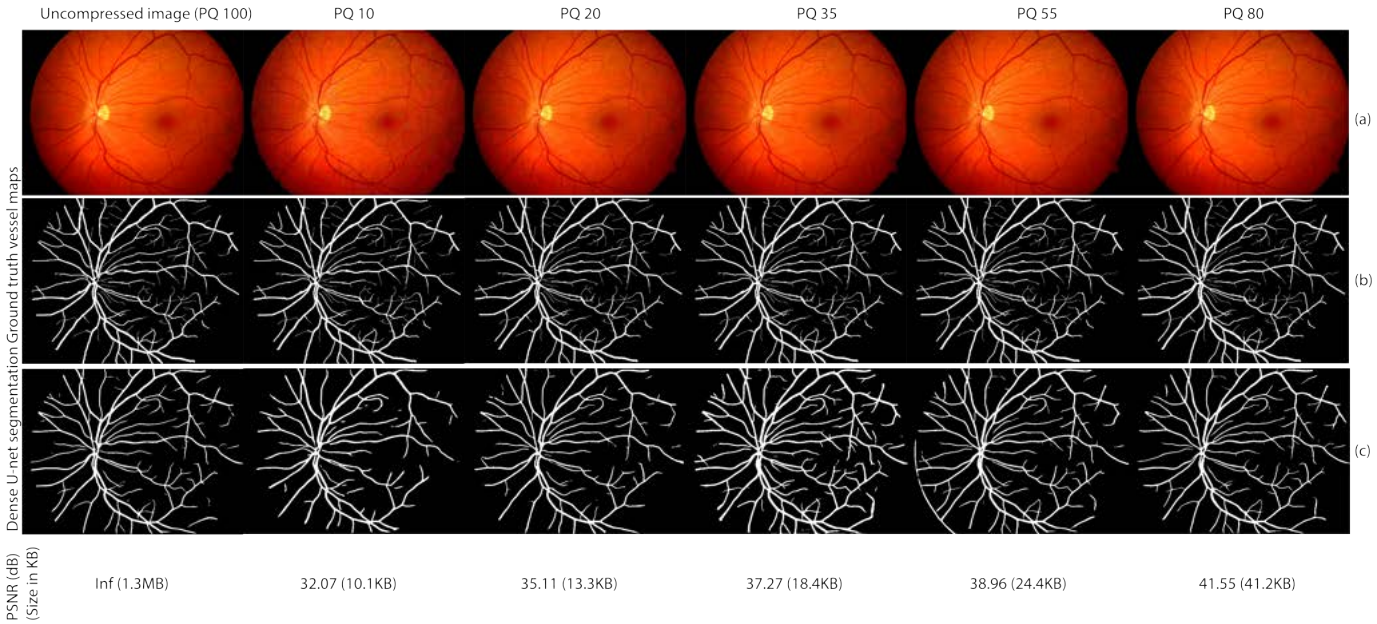

Figure A3: Segmentation results of ARIA dataset. (a) Images compressed at PQ10, PQ20, PQ35, PQ55, PQ80,and PQ100 is uncompressed. (b) Ground truth vessel maps. (c) Results of Dense U-net segmentation. Best seen on a computer screen.

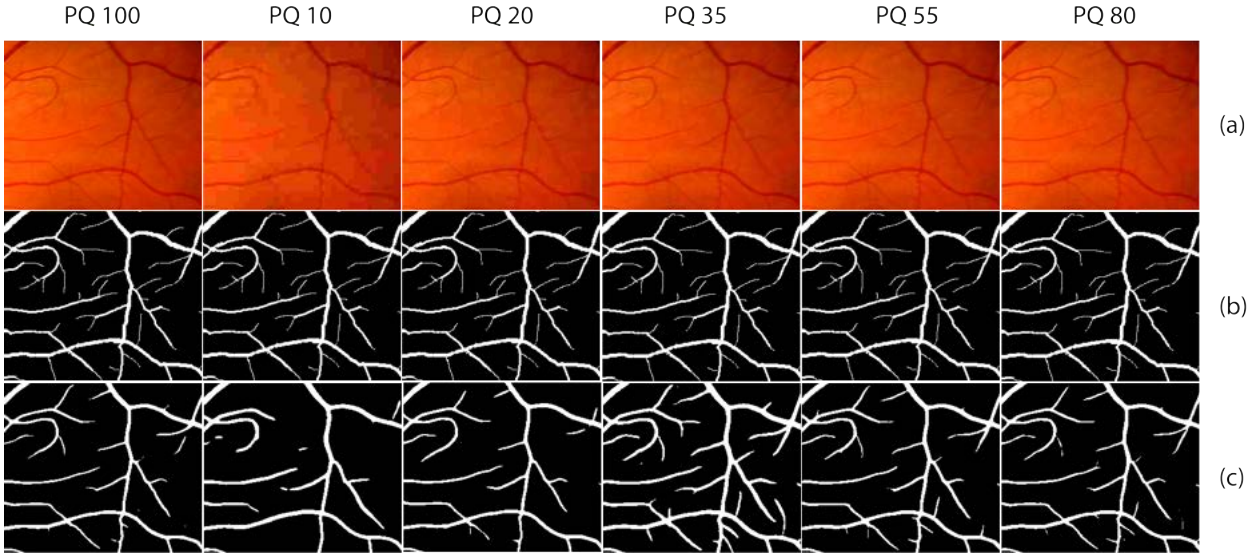

Figure A4: Segmentation results of ARIA dataset. (a) Images compressed at PQ10, PQ20, PQ35, PQ55, PQ80,and PQ100 is uncompressed. (b) Ground truth vessel maps. (c) Results of Dense U-net segmentation. Best seen on a computer screen.

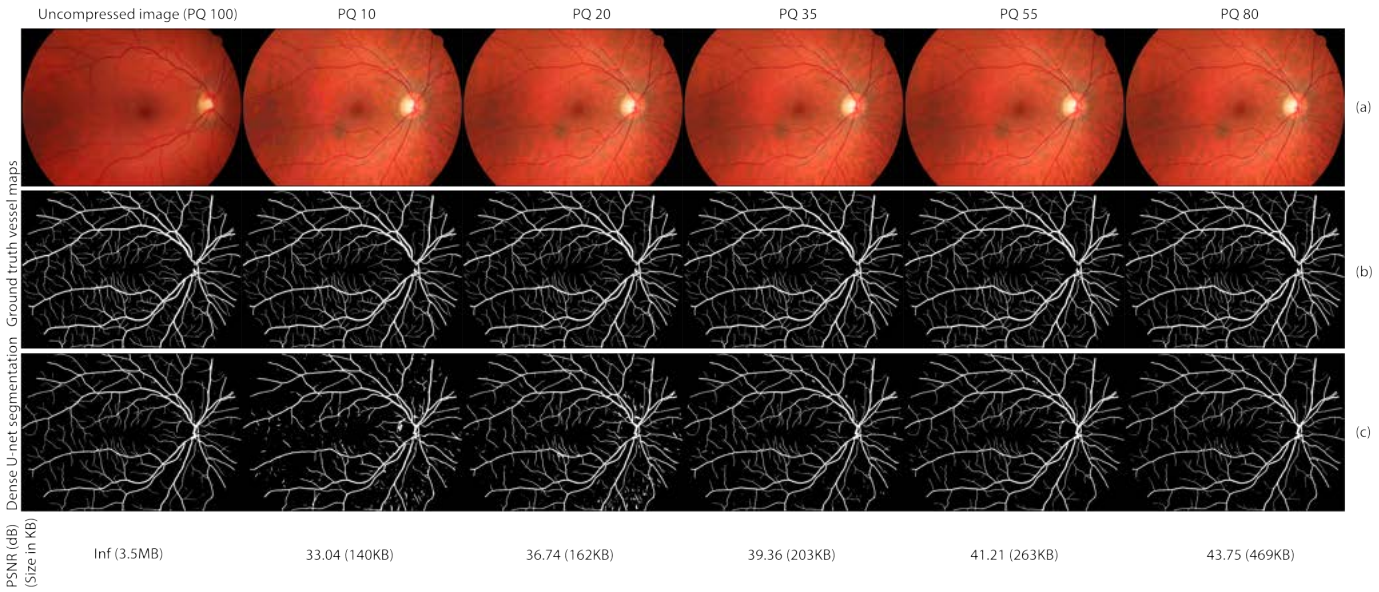

Figure A5: Segmentation results of HRF dataset. (a) Images compressed at PQ10, PQ20, PQ35, PQ55, PQ80,and PQ100 is uncompressed. (b) Ground truth vessel maps. (c) Results of Dense U-net segmentation. Best seen on a computer screen.

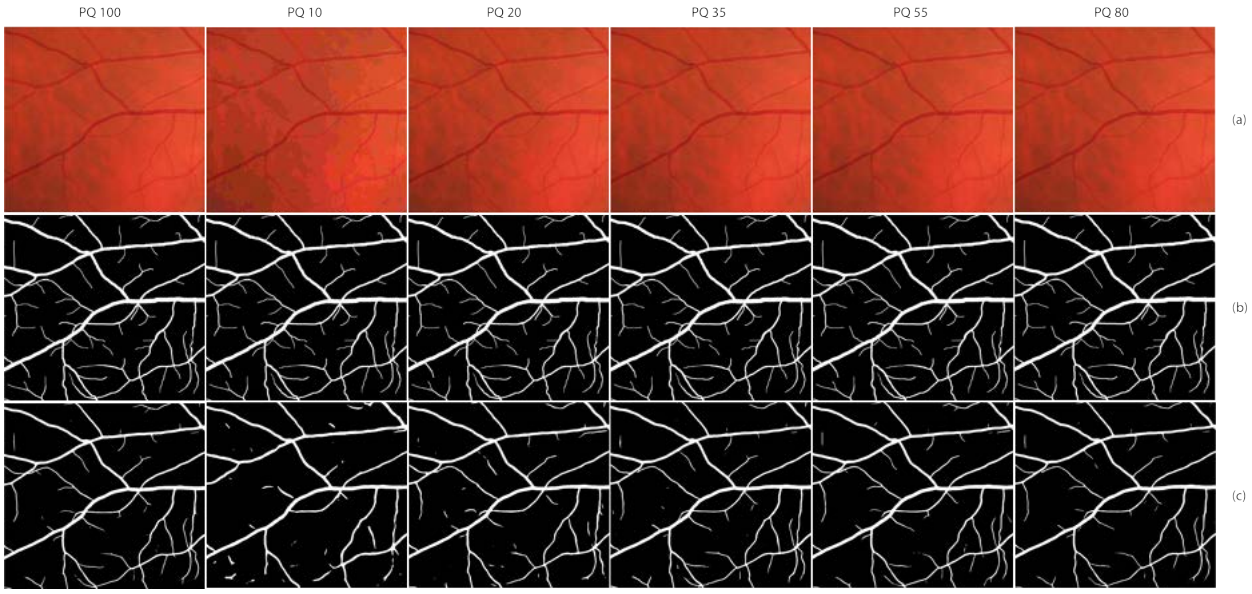

Figure A6: Segmentation results of HRF dataset. (a) Images compressed at PQ10, PQ20, PQ35, PQ55, PQ80,and PQ100 is uncompressed. (b) Ground truth vessel maps. (c) Results of Dense U-net segmentation. Best seen on a computer screen.

Table A1: **A 23-layer architecture of FCN (C - Convolutional, BN - Batch Normalization, MP - Max Pooling, ReLU - Rectified Linear Unit, DB - Dense Block, US - Up Sampling, CN - Concatenate, S - Softmax; * represents the concatenate connection between UP and DB)**

Layer	**Output size**	**Filter size**	**Stride**
Input	128×128×1	–	–
C	128×128×32	1×1	1
BN	128×128×32	–	–
ReLU	128×128×32	–	–
DB***	128×128×32	–	–
MB	64×64×32	–	–
DB**	64×64×64	–	–
MB	32×32×64	–	–
DB*	32×32×64	–	–
MB	16×16×64	–	–
DB	16×16×64	–	–
US*	32×32×64	3×3	2
CN*	32×32×128	–	–
DB	32×32×64	–	–
US**	64×64×64	3×3	2
CN**	64×64×128	–	–
DB	64×64×64	–	–

Continued on next page

Table A1 – *Continued from previous page*

Layer	Output size	Filter size	Stride
US**	128×128×32	3×3	2
CN***	128×128×64	–	–
DB	128×128×32	–	–
C	128×128×2	1×1	1
Reshape	16384×2	–	–
S	16384×2	–	–

Table A2: **Segmentation performance measures of DRIVE dataset for separate compression levels**

PQ Levels	Sensitivity (%)	Specificity (%)	Accuracy (%)	DICE	AUC
PQ10	68.09±4.40	95.20±0.72	92.99±0.35	0.61±0.03	0.902
PQ20	73.30±3.94	97.90±0.35	95.87±0.23	0.74±0.01	0.951
PQ35	81.41±4.75	97.04±0.64	95.74±0.26	0.76±0.02	0.963
PQ55	82.66±3.52	97.33±0.36	96.11±0.24	0.78±0.01	0.972
PQ80	82.02±4.32	98.31±0.27	96.95±0.02	0.82±0.01	0.976
PQ100	84.53±3.61	98.12±0.29	96.99±0.20	0.82±0.02	0.985

Table A3: **Segmentation performance measures of DRIVE dataset for combined compression levels**

PQ Levels	Sensitivity (%)	Specificity (%)	Accuracy (%)	DICE	AUC
PQ10	53.57±6.23	98.85±0.27	95.20±0.40	0.64±0.04	0.891
PQ30	72.66±5.34	98.55±0.35	96.42±0.27	0.77±0.02	0.956
PQ50	75.73±5.12	98.53±0.31	96.64±0.28	0.79±0.02	0.965
PQ70	77.46±5.15	98.52±0.31	96.77±0.27	0.80±0.02	0.970
PQ100	79.70±4.77	98.48±0.30	96.91±0.26	0.81±0.02	0.976

Table A4: **Segmentation performance measures of ARIA dataset for separate compression levels**

PQ Levels	Sensitivity (%)	Specificity (%)	Accuracy (%)	DICE	AUC
PQ10	56.30±11.62	97.40±0.96	94.27±0.71	0.59±0.11	0.889
PQ20	64.70±8.77	97.84±0.46	95.29±0.50	0.67±0.08	0.937
PQ35	81.39±5.49	95.05±1.08	93.98±0.91	0.68±0.08	0.956
PQ55	74.52±4.72	97.01±0.62	95.25±0.67	0.71±0.06	0.964
PQ80	72.11±4.53	97.50±0.52	95.50±0.60	0.71±0.05	0.964

Continued on next page

Table A4 – *Continued from previous page*

PQ Levels	Sensitivity	Specificity	Accuracy	DICE	AUC
PQ100	73.09±4.23	97.53±0.59	95.60±0.60	0.72±0.05	0.969

Table A5: **Segmentation performance measures of ARIA dataset for combined compression levels**

PQ Levels	Sensitivity (%)	Specificity (%)	Accuracy (%)	DICE	AUC
PQ10	52.24±14.97	98.32±0.98	94.85±1.14	0.59±0.16	0.874
PQ30	66.98±12.05	98.05±0.95	95.69±1.31	0.70±0.13	0.94
PQ50	69.37±11.30	98.05±0.91	95.87±1.32	0.71±0.12	0.952
PQ70	70.41±10.83	98.01±0.92	95.91±1.33	0.72±0.12	0.956
PQ100	71.69±10.35	97.94±0.93	95.93±1.32	0.72±0.11	0.959

Table A6: **Segmentation performance measures of HRF dataset for separate compression levels**

PQ Levels	Sensitivity (%)	Specificity (%)	Accuracy (%)	DICE	AUC
PQ10	69.89±4.13	98.32±0.54	95.58±0.49	0.75±0.02	0.941
PQ20	76.62±3.91	98.59±0.50	96.45±0.49	0.81±0.02	0.969
PQ35	78.38±4.11	99.01±0.31	96.99±0.45	0.83±0.02	0.981
PQ55	79.32±4.13	99.02±0.29	97.08±0.43	0.84±0.02	0.985
PQ80	76.16±4.37	99.35±0.25	97.08±0.49	0.83±0.02	0.986
PQ100	78.81±4.14	99.24±0.25	97.24±0.45	0.85±0.02	0.988

Table A7: **Segmentation performance measures of HRF dataset for combined compression levels**

PQ Levels	Sensitivity (%)	Specificity (%)	Accuracy (%)	DICE	AUC
PQ10	66.69±4.50	98.18±0.86	95.57±0.61	0.71±0.05	0.928
PQ30	78.90±3.44	98.23±0.86	96.60±0.56	0.79±0.04	0.970
PQ50	79.89±3.65	98.40±0.80	96.84±0.48	0.81±0.04	0.978
PQ70	80.20±3.73	98.48±0.77	96.93±0.44	0.81±0.04	0.981
PQ100	81.57±3.39	98.40±0.78	96.97±0.46	0.82±0.04	0.984