Journal of Imaging Informatics in Medicine (2024) 37:107–122
https://doi.org/10.1007/s10278-023-00926-6

Intra- and Inter-expert Validation of an Automatic Segmentation Method for Fluid Regions Associated with Central Serous Chorioretinopathy in OCT Images

Mateo Gende[1,2] · Lúa Castelo[1,2] · Joaquim de Moura[1,2] · Jorge Novo[1,2] · Marcos Ortega[1,2]

Received: 15 June 2023 / Revised: 16 October 2023 / Accepted: 16 October 2024 / Published online: 12 January 2024

Abstract

Central Serous Chorioretinopathy (CSC) is a retinal disorder caused by the accumulation of fluid, resulting in vision distortion. The diagnosis of this disease is typically performed through Optical Coherence Tomography (OCT) imaging, which displays any fluid buildup between the retinal layers. Currently, these fluid regions are manually detected by visual inspection a time-consuming and subjective process that can be prone to errors. A series of six deep learning-based automatic segmentation architectural configurations of different levels of complexity were trained and compared in order to determine the best model intended for the automatic segmentation of CSC-related lesions in OCT images. The best performing models were then evaluated in an external validation study. Furthermore, an intra- and inter-expert analysis was conducted in order to compare the manual segmentation performed by expert ophthalmologists with the automatic segmentation provided by the models. Test results of the best performing configuration achieved a mean Dice of 0.868 ± 0.056 in the internal dataset. In the external validation set, these models achieved a level of agreement with human experts of up to 0.960 in terms of Kappa coefficient, contrasting with a value of 0.951 for agreement between human experts. Overall, the models reached a better agreement with either of the human experts than these experts with each other, suggesting that automatic segmentation models for the detection of CSC-related lesions in OCT imaging can be useful tools for assessing this disease, reducing the workload of manual inspection and leading to a more robust and objective diagnosis method.

Keywords Central serous chorioretinopathy · Deep learning · Segmentation · Computer-aided diagnosis · Ophthalmology

Introduction

Central Serous Chorioretinopathy (CSC) is a retinal disease that causes visual impairment characterised by the detachment of the retina due to the accumulation of Subretinal Fluid (SRF) produced by the dysfunction of the retinal pigment epithelium and the hyperpermeability and enlargement of the underlying choroid. Patients with CSC typically experience central vision loss, central scotoma, micropsia or metamorphopsia [1]. This disease was first described in 1866 by Albrecht von Graefe as central recurrent retinitis and involved the detachment of the serous retina, primarily affecting the macular region [2].

Recent research has shed some light on the causes of CSC, pointing towards choroidal vascular hyperpermeability, which can lead to an increase in the hydrostatic pressure beneath the retinal pigment epithelium (RPE), causing it to disintegrate [3]. The balance between oncotic and hydrostatic pressure at the RPE normally results in fluid flowing

✉ Joaquim de Moura
joaquim.demoura@udc.es

Mateo Gende
m.gende@udc.es

Lúa Castelo
lua.castelo@udc.es

Jorge Novo
jnovo@udc.es

Marcos Ortega
mortega@udc.es

[1] Grupo, VARPA, Instituto de Investigación Biomédica de A Coruña (INIBIC), Universidade da Coruña, Xubias de Arriba, 84, 15006 A Coruña, Spain

[2] Centro de investigación, CITIC, Universidade da Coruña, Campus de Elviña s/n, 15071 A Coruña, Spain

from the retina into the choroid. However, in CSC, the increase in hydrostatic pressure within the choroid causes fluid to accumulate beneath the RPE. When the hydrostatic pressure beneath the RPE is high, it pushes the RPE forward, causing a discontinuity in its barrier, leading to the detachment of the RPE and punctate areas of leakage, commonly referred to as "microrips" or "blowouts".

The presence of SRF leakage and accumulation can cause a loss of vision, as seen in the study conducted by Mrejen et al. [4], which researched the long-term causes of vision loss in the CSC. Different medical imaging techniques have been put forward for the diagnosis of this disease including, but not limited, to fluorescein angiography, indocyanine angiography, fundus autofluorescence imaging, Colour Fundus Photography (CFP) and Optical Coherence Tomography (OCT). With the former two being more invasive methods, the latter three are generally favoured due to their non-invasiveness, their lower risk of complication and convenience [3, 5]. These methods allow for a safe and effective monitoring of the disease and an early detection of vision loss.

In particular, OCT is a non-invasive imaging technique that can produce micrometre-resolution cross-sectional and volumetric visualisations of the retinal tissue. Its cross-sectional nature makes it the preferred imaging modality for the diagnosis of the CSC [6] as it enables the visualisation of different layers or sections of the eye [7, 8]. This, in turn, allows the visual inspection of the changes caused by disease progression [9, 10]. Its ability to allow the direct observation of the various layers that make up the macula makes it one of the most widely used techniques for the diagnosis of various ocular pathologies, including age-related macular degeneration [11], diabetic macular oedema [12], cystoid macular oedema [13], as well as CSC [14].

OCT enables an easy visualisation of any alterations related to CSC such as neurosensory detachment, detachment of the pigmentary epithelium, protrusion of the RPE, thickness changes in the posterior retinal surface, granulations on the detached retina, hyperreflective spots, RPE defects, RPE proliferation and subretinal fibrous exudates [10]. One of the most noticeable changes of CSC is the accumulation of SRF, which can appear in OCT imaging as a dark area around the pigmentary epithelium [15]. Early detection of the CSC is crucial for avoiding serious symptoms such as vision loss. However, the diagnosis process of CSC through OCT is slow and time-consuming, as well as subjective and prone to errors or misdiagnosis as it is carried out manually by expert examiners [16]. In this situation, the use of deep learning models for the automatic segmentation in CSC diagnosis could greatly benefit the process and aid the experts in the assessment of this disease.

Deep learning models make use of several consecutive convolutional layers to identify patterns and structures in large datasets [17, 18]. These models are able to automatically learn patterns by means of annotated examples, without the need to formalise the explicit knowledge needed to perform certain tasks. This makes deep learning models especially attractive for fields such as medicine, where the ability of these models to automatically learn how to identify patterns of disease makes them invaluable in the development of new and advanced Computer-aided Diagnosis (CAD) systems.

The use of CAD systems to automate the diagnosis process of CSC can lead to increased efficiency and accuracy, reducing the risk of errors derived from subjective expert assessment. Given the relevance of this topic, several studies have employed fundus imaging for the diagnosis of CSC, such as in the work of Chen et al. [19], where a deep learning model was proposed for the automatic detection of CSC leak points. For this purpose, they employed an attention-gated network architecture, integrating an attention gate with convolutional layers. The results highlighted the performance of deep learning models for the detection of CSC leakage points. However, this study is limited by its reliance on fluorescein angiography imaging, an invasive procedure which requires the use of a contrast die to highlight the blood vessels. Xu et al. [20] developed a deep learning-based architecture for the screening of SRF from CFP images. The network architecture followed a cascade approach in which two separate Convolutional Neural Network (CNN) models are able to determine the presence or absence of the disease, and whether it affects the central foveal region. However, it does not provide a true segmentation map of the presence of fluid. More recently, Yoo et al. [21] used a different approach, training conditional generative models to create the segmentation maps of the area of interest in the lesions with presence of SRF, with results that approach those of the human annotation. In spite of this, the use of generative models for generating segmentation masks has a high risk of generating maps that may look convincing enough to fool the discriminator but have no bearing on the presence of SRF, as the results show. These studies indicate that fundus imaging can be used for the automatic characterisation of the CSC, and highlight the utility of deep learning architectures in extracting the relevant characteristics in this disease, but may require invasive procedures or are otherwise limited in the accuracy of the segmentation outputs they are able to provide.

On the other hand, the advantages of OCT have made this imaging technique increasingly popular for the diagnosis of retinal diseases, particularly for the detection of pathological fluid regions. The high-resolution, cross-sectional images captured by OCT, combined with the use of automatic segmentation techniques, can offer a comprehensive analysis of the affected area, which can be crucial for a precise diagnosis. Previous works have shown the potential of deep learning-based segmentation in OCT images for various retinal diseases, such as serous retinal detachment [22], diabetic macular oedema [23], glaucoma [24], age-related macular degeneration [25] and intra-retinal cysts [26].

Because of this, recent studies have employed OCT to automatically analyse the CSC. Gao et al. [27] presented a study in which they employed an area-constraint fully convolutional network to perform the automatic segmentation of the CSC region in OCT images. The results showed that the model was close to manual segmentation after independent layer segmentation as well as quantitative and qualitative evaluations. However, this methodology was trained and tested only on a small dataset consisting of 10 eyes, 5 of which suffered CSC. Rao et al. [28] conducted a study to automatically segment regions affected by CSC in OCT images using deep learning-based architectures. This methodology relied on a pre-processing stage in order to adapt the images to the architecture, which was trained and validated on a similarly small dataset of only 15 eyes annotated only by a single expert. In the work of de Moura et al. [29], the authors proposed an end-to-end methodology for the automatic identification and segmentation of intra-retinal fluid regions associated with CSC in OCT scans. To achieve this, the authors adapted a fully convolutional architecture inspired by the SegNet architecture [30], while omitting any pre- or post-processing stages. This approach was validated on a larger dataset than the two other approaches. However, the images were only annotated by a single expert, which poses a risk of biasing the results towards that single expert. Pawan et al. introduced a modification to capsule networks based on dilation, residual connections, inception blocks and capsule pooling in order to better adapt the architecture to the segmentation of fluid in images of CSC patients. These changes also reduced the overall complexity of the networks while maintaining competitive performance. Nevertheless, its evaluation is based on the annotations of a single expert, which may pose risks of bias, similarly to the other previous approaches. Indeed, these studies highlight the recent efforts dedicated to the automatic analysis of the CSC in OCT imaging.

Recent works on the automatic segmentation of pathological and CSC-related lesions using OCT have shown promising results, while providing a clear and accurate visualisation of the progression of the disease. Nevertheless, these results are based on the annotations of a single expert which, as previously mentioned, can lead to subjectivity. Moreover, these studies may not fully capture the nuances of the manual diagnosis process. In order to address this limitation, an intra- and inter-expert analysis is necessary, so that the variability and subjectivity typically associated with manual inspection can be properly assessed. This way, a more robust and reliable assessment of the diagnosis can be provided, by taking into account any inconsistencies and potential disagreements among experts.

In this work, we aim to address this crucial challenge by presenting a comprehensive study in the application of deep learning models to the automatic segmentation of SRF regions in OCT images associated with CSC. Complementarily, this study is extended by including an intra- and inter-expert analysis of the best performing models with multiple expert ophthalmologists. These models were trained and validated using a representative dataset of the pathology (specifically designed for this study), along with an external validation dataset used for the intra- and inter-expert analysis. This analysis can provide ground-breaking evaluation of any possible inconsistencies among the automatic segmentation models as well as valuable insight into inter-expert disagreement. The main contributions of this work can be summarised as follows:

- This work presents a comparative analysis of various modular deep learning architectural configurations for fluid segmentation.
- This analysis is complemented by an evaluation of the configurations that produced the best results in an external dataset annotated by two human experts.
- The intra- and inter-expert analysis that was performed revealed that the deep learning models exhibited better alignment with individual human experts, surpassing human inter-expert consistency.
- Intra- and inter-expert analysis can set a new standard for the validation of future studies in the field.

Materials and Methods

In this work, we propose a deep learning-based methodology for the automatic segmentation of fluid regions in OCT images of patients with chronic CSC. The methodology is comprised of two main stages, as displayed in Fig. 1. The first stage involves the training and validation of a series of deep learning models using a representative dataset of OCT images belonging to CSC and healthy control patients. The second stage involves an intra- and inter-expert analysis between the best performing models of the first stage and multiple expert ophthalmologists. Using an external validation dataset, the automatic segmentations produced by the models are compared among themselves and with the manual annotations produced by the human experts. This comparison provides a thorough and comprehensive analysis of the performance of the automatic models, and can offer valuable insight into inter-expert disagreement.

Automatic Segmentation of Fluid Regions

The first stage of the methodology is focused on training, evaluating and comparing the performance of several

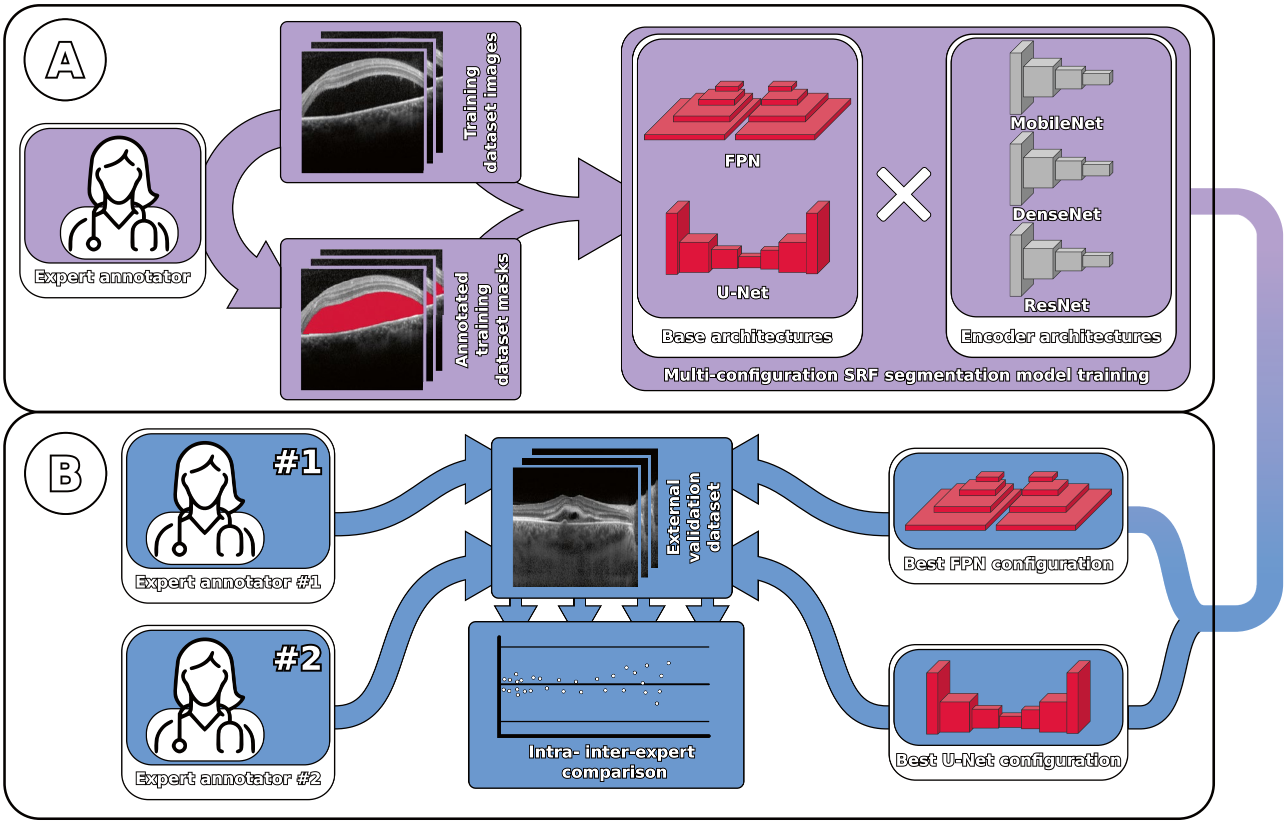

Fig. 1 Summary of the experiments that were performed. **A**: Multi-configuration SRF segmentation model training. Six representative configurations of backbone segmentation network and encoder architectures were trained and validated on an OCT image dataset. **B**: Intra- and inter-expert comparison. The best model configurations for each backbone architecture were selected and compared among themselves and with two expert annotators on an external validation dataset

prominent deep learning architectures for the segmentation of the fluid regions in OCT scans of patients with CSC.

Dataset

A dataset was collected with a total of 557 OCT images from different patients. 303 images correspond to patients of CSC, while 254 display healthy control patients. These images were acquired with a Heidelberg SPECTRALIS® optical imaging platform. These OCT scans are all macula centred, and were extracted from both the left and right eyes using different scanning protocols, including 1- and 7-line scans, the two highest quality scanning protocols most widely used in the assessment of the CSC. Moreover, these images are representative of the inherent variability in terms of severity that can be found in clinical practice, from small isolated cases, to multiple deposits, to larger accumulations of fluid. These scans were manually annotated by a trained expert to accurately segment all the targeted regions of CSC-related pathological fluid. The images range in resolution from 760 × 450 to 1536 × 500. For compatibility with all the models, and to ensure a fair comparison, all images were resized to 512 × 512 pixels during pre-processing. The protocols followed during the development of this project were conducted in accordance with the Declaration of Helsinki, approved by the local Ethics Committee. A representative example of these OCT images, as well as the corresponding annotation indicating the pathological region can be found in Fig. 2.

Methodology

In this study, different CNN architectures were trained and validated to determine the one best suited for the automatic segmentation of SRF regions in OCT images associated with the chronic CSC disease. In order to achieve this goal, two backbone segmentation architectures were trained and validated: Feature Pyramid Network (FPN) [31], and U-Net [32]. These architectures were modularly combined with three different encoder architectures, by substituting and adapting the corresponding encoder part of each architecture. With an aim to explore how models of different complexity adapt to this task, three different encoder

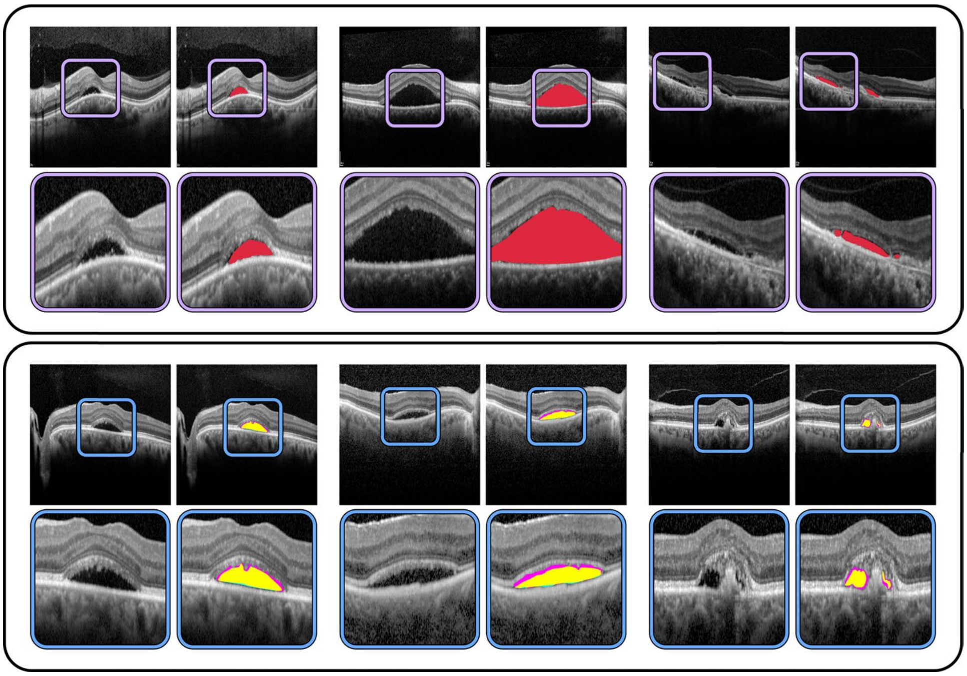

Fig. 2 Representative examples from both datasets, along with a detailed view of the expert annotations. *Top*: First dataset, employed for training and validation of the models, expert annotations shown in red. *Bottom*: Second dataset, employed for the intra- and inter-expert analysis. In cyan, annotation from the first expert. In magenta, annotation from the second expert. In yellow, overlap between the two experts

architectures were selected for this task: MobileNet [33], DenseNet [34] and ResNet [35].

On the one hand, the FPN architecture is a convolutional, pyramid-shaped, top-down neural network with scale-invariant lateral connections originally intended for image classification [31] but later adapted for semantic segmentation [36]. This CNN architecture was specifically developed for focusing on detection at multiple scales, by merging feature maps from lower and deeper layers, and has found application in several related medical image segmentation tasks [37, 38]. This detection at different scales can be of great help to the segmentation of retinal fluid due to the different degrees of affectation that these images can present. Being able to accurately detect small buildups as well as large accumulations can improve the robustness of the models. Figure 3 displays a schematic view of this architecture. On the other hand, the U-Net architecture was specifically developed for medical image segmentation, and has been successfully applied to similar problems (for reference, [39–41]). By using skip connections between different levels of its contracting and

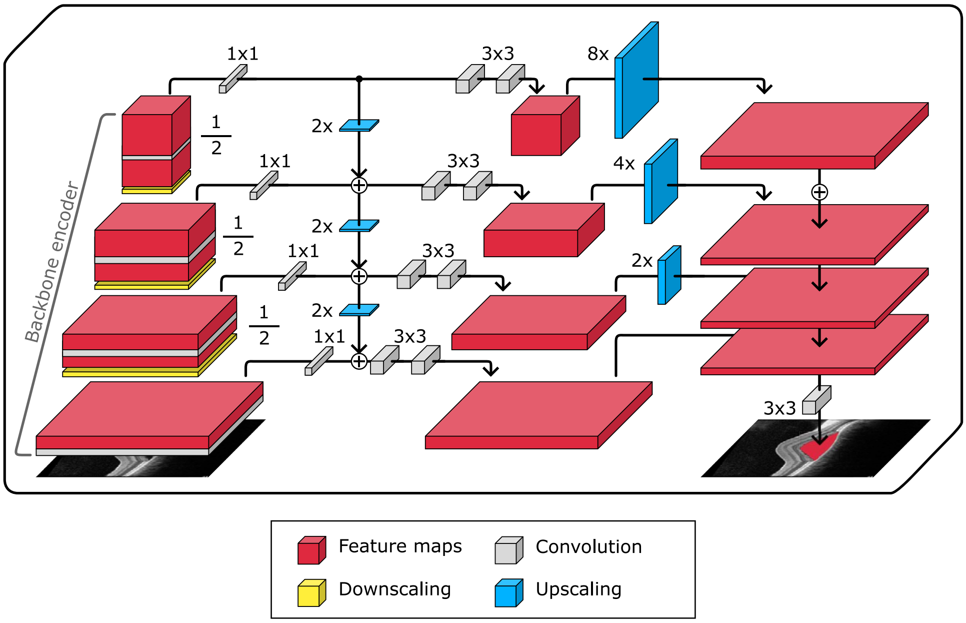

Fig. 3 Base structure of the FPN backbone segmentation architecture. Features are extracted and progressively refined. At the later stages, the extracted features are upscaled and stacked before passing on to the segmentation head which outputs the segmentation mask

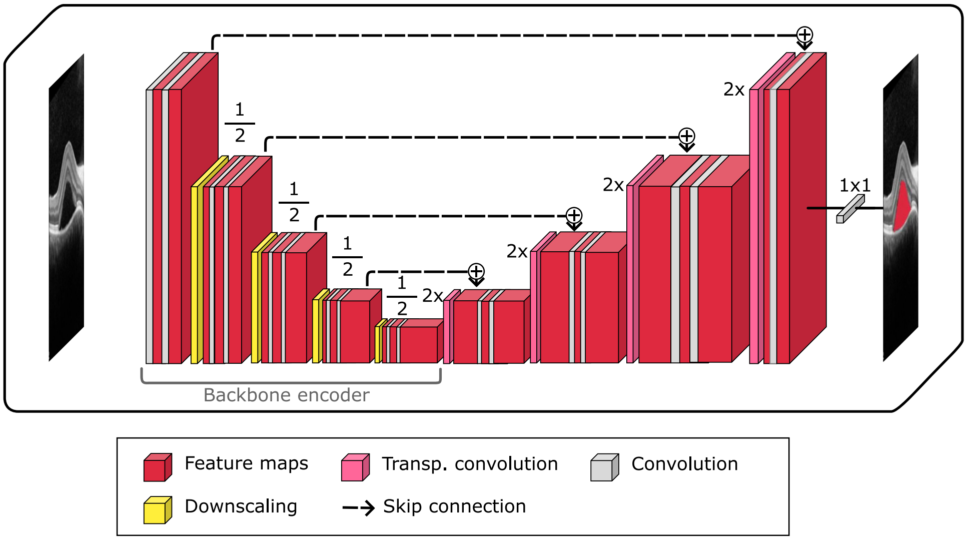

Fig. 4 Base structure of the U-Net backbone segmentation structure. At each scale level, the extracted features are concatenated into the corresponding layers in the later part of the structure, bypassing the deeper levels

expanding path, this architecture enables the transmission of both high- and low-level features to the final layers, enabling a more comprehensive analysis of the information contained in the images, also improving the detection of fluid buildups at different scales. A summarised view of this backbone architecture can be found in Fig. 4.

The three encoder architectures that were selected represent different levels of complexity. The MobileNet-v2 architecture [33] makes use of linearly separable and depth-wise convolutions in order to create a lightweight model. This results in a highly efficient architecture, with the smallest parameter count of those considered in this work. The DenseNet architectures [34] make use of densely connected layers in which the features are transmitted forward and concatenated along the network, using bottlenecks at the end of each dense block for limiting the explosion in the number of features. This allows these models to achieve remarkable depths in terms of layers while avoiding the vanishing gradient problem. The DenseNet-169 architecture was selected for this work as a balance between efficiency and complexity. Finally, the ResNet architectures [35] make use of residual blocks, in which features are transmitted forward via skip connections, adding them to the deeper features instead of using concatenation. These models allow for a greater complexity in terms of trainable parameters. The ResNet-34 architecture was selected for this work, representing the most complex model of those considered. For ease of comparison, a summary of the trainable parameters for each architecture configuration can be found in Table 1.

The experiments that were performed were designed to allow a fair comparison between the various model configurations. To this end, a 5-fold cross-validation strategy was adopted, partitioning the data into 5 subsets. In each partition, 60% of the images were used for training, 20% for validation, and the remaining 20% for testing, ensuring that each image appeared in the test set exactly once for each configuration. Special care was taken to confine images from the same patient to the same set, preventing data leakage and any associated biases to sharing different images from the same patient between sets. This way, the model configurations can be compared fairly among themselves.

Table 1 Number of trainable parameters and multiply and accumulate operations for each configuration of backbone segmentation architecture and modular encoder

	FPN		
	MobileNet-v2	DenseNet-169	ResNet-34
Parameters	4×10^6	15×10^6	23×10^6
Operations	10×10^9	27×10^9	27×10^9
	U-Net		
	MobileNet-v2	DenseNet-169	ResNet-34
Parameters	7×10^6	21×10^6	24×10^6
Operations	14×10^9	38×10^9	31×10^9

In order to better take advantage of the limited amount of available training data, the encoder models were first initialised to a pre-training on ImageNet. Afterwards, each model was trained on its corresponding training set. At this stage, data augmentation was applied in the form of random horizontal flipping. At the end of each training epoch, the models were validated on their corresponding validation set, extracting a loss metric that was used to detect the training stage at which the models could better generalise to images not used during training. For this matter, after a fixed training length, a checkpoint of the models at the stage with the lowest validation loss was selected for testing.

The models were trained using Dice overlap loss [42] due to its performance in similar unbalanced segmentation tasks in medical imaging (for reference, [43–45]) Adam [46] was used for optimisation, with a learning rate of 1×10^{-3}, $\beta_1 = 0.9$ and $beta_2 = 0.999$. These models were trained with a batch size of 16 images for a maximum of 400, which was empirically found to be sufficient for model convergence.

Evaluation

In order to achieve a comprehensive assessment of the performance of the segmentation models, the Accuracy, Recall, Precision, Jaccard index, and Dice coefficient metrics were employed in the evaluation of the models. Collectively, these metrics can provide a thorough validation of how these models perform.

Intra- and Inter-expert Analysis

In order to comprehensively study the subjectivity associated to the manual segmentation of fluid regions in OCT images, as well as to thoroughly validate and assess the robustness of the trained models, an intra- and inter-expert analysis was conducted using a separate dataset consisting of both CSC and control patients. The aim of this analysis is to shed light on the differences that may arise due to expert variability, as well as to provide a benchmark against which the performance of the trained models can be compared.

Dataset

An independent dataset, distinct from the one that was used for the model training and validation, was employed for this analysis. This dataset was comprised of a total of 100 OCT images from different patients, 85 of which displayed signs of CSC and 15 displayed healthy eyes. These images were acquired with the Heidelberg SPECTRALIS® platform, at resolutions ranging from 760 × 450 to 1536 × 500 pixels. As in the previous case, and for compatibility with all the models, these images were resized to a standard size of 512 × 512 pixels during pre-processing. Two different expert annotators were separately asked to manually label the presence and location of CSC-related fluid accumulations for each OCT image. As in the previous case, the dataset was collected after approval from the local ethics committee, following the tenets of the Declaration of Helsinki. Examples of the manual labelling by the experts can be found in Fig. 2.

Methodology

The intra- and inter-expert analysis consists of two parts. The first part is aimed at assessing the robustness and consistency of the trained automatic segmentation models. The second part is aimed at studying the subjectivity and the differences in criteria between the human experts, as well as to validate the performance of the automatic models in this context.

In the first, intra-expert analysis, the different instances of models trained in the first stage are compared among themselves. For each architectural configuration, the five models, each one trained on a cross-validation subset, were separately used to segment this independent dataset. Then, the segmentations produced by these models were compared among themselves, extracting a measurement of how similar the segmentation results are. This, in turn, can allow the comparison of the consistency and robustness of the models when trained using different sets of images. A low variability between the results of different models of a single configuration can indicate a better robustness to training with different samples, and better performance in generalisation to unseen images.

In the second, inter-expert analysis, the segmentations produced by the two experts are compared with the models that produced the highest results for each backbone in the first stage. Three different scenarios were considered in this analysis:

1. *Comparison of manual segmentations produced by the expert annotators*: This comparison was performed to study the impact of differences in criteria when diagnosing this pathology, as well as to set a reference of the magnitude of inter-expert differences.
2. *Comparison of manual segmentations with the automated segmentations produced by the best performing models*: This comparison was conducted in order to validate the segmentation models against an external, unseen dataset, highlighting any potential biases, and comparing the performance of the automated models with inter-expert variability.
3. *Comparison of automated segmentation models*: The automatic segmentation models were also compared among themselves to identify any potential biases resulting in each configuration.

In line with previous work in the segmentation of ophthalmic imaging [47], the total area segmented as SRF was used as a uni-dimensional indicator for each expert's segmentation. This value was calculated for each segmented image and used to create a Bland-Altman plot for each comparison. This plot offers a simple way of assessing bias between mean differences, and of estimating an agreement interval between the experts [48–50].

Evaluation

To provide a comprehensive summary of the intra- and inter-expert comparative analysis, the Limits of Agreement (LoA) of the Bland-Altman plots were calculated at a confidence level of 95% ($\alpha = 1.96$). $LoA = \overline{x} \pm \alpha \times sd$, where $\overline{x}$ and sd denote the average and standard deviation of area segmented as fluid in all images in the set, respectively. These LoA can be used to calculate the amplitude between the upper and the lower LoA as a measurement of the disagreement between the experts, with wider amplitudes showing increased disagreement. Furthermore, the mean difference between the areas segmented by each expert can be used as a measurement of any existing bias, with higher values indicating the first expert tends to over-segment relative to the second one, and smaller values indicating the opposite. Aside from calculating similarity measurements in terms of the total area segmented, the Dice coefficient, Cohen's κ coefficient, and the Mean Square Error (MSE) were also computed as a measurement of the specific similarity between automated and manual segmentation masks.

Software and Hardware Resources

The experiments were performed using the Python language (v.3.8.10). The PyTorch library (v.1.12.1) [51] was used to train and validate the models, while the Segmentation Models Pytorch library (v.0.3.1) [52] was used for model configuration and pre-trained weight acquisition. Statistical calculations were done using the statsmodels (v.0.13.5) and SciPy (v.1.10.1) In terms of hardware, the models were trained and validated using an AMD EPYC 7763 64-Core CPU, with 504GB RAM and two NVIDIA A100 GPUs.

Results and Discussion

Automatic Segmentation of Fluid Regions

The six model configurations were trained and tested as described in Section 2.1.2. Figure 5 displays the training

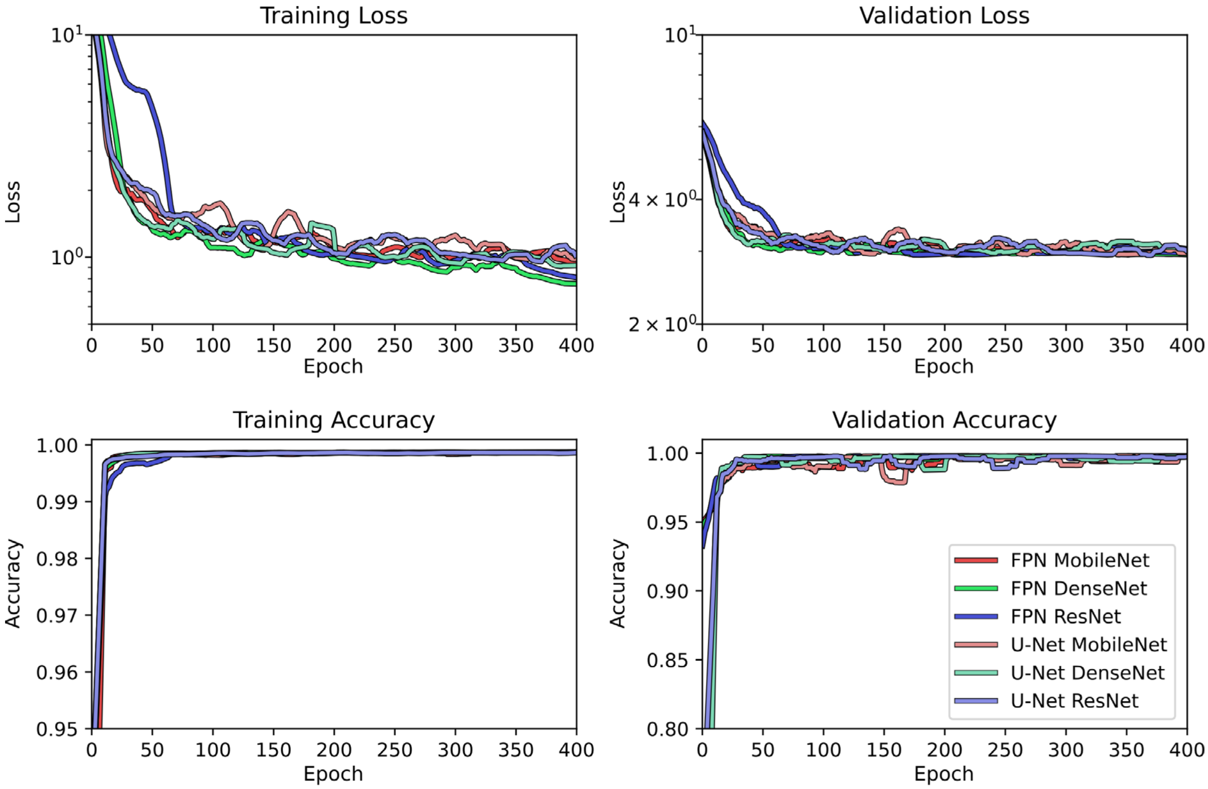

Fig. 5 Training and validation curves describing the average training and validation loss and Accuracy for the models trained across every configuration

and validation loss and Accuracy curves. These curves show that the models converge in validation before the maximum allowed number of epochs. The average epoch at which the models achieved the best results in terms of validation was 253 ± 91. Generally, all the configurations display a similar behaviour during training. The models using the MobileNet encoder architecture display the highest variability during training, and show generally higher loss than the others. The models which incorporated a DenseNet encoder architecture show generally lower training loss than the others, and seem to be able to achieve good results quickly in the early stages of training. The remaining models, which used the ResNet architecture, take longer to adapt, with higher loss in the early stages, but settling in lower values at the later stages, as can be expected from the most complex architecture in terms of trainable parameters.

After selecting the best training stage for each model in terms of the checkpoint with lowest validation loss, the models were evaluated on their corresponding test subset. The results of this test are shown in Table 2. A repeated measures ANOVA test was performed in order to verify whether there are significant differences between the results achieved by the models. Significant differences were found for the Accuracy ($p = 0.002$), Precision ($p = 0.008$) and Recall metrics ($p = 0.032$). Differences could not be considered significant for Jaccard ($p = 0.081$) and Dice ($p = 0.092$) at $\alpha = 0.05$. For the FPN backbone architecture, the best results are achieved by the DenseNet encoder architecture, with a Dice coefficient of up to 0.861 ± 0.069. Conversely, for the U-Net backbone architecture, the best results are obtained by the most complex model using the ResNet encoder architecture, with a Dice coefficient of up to 0.868 ± 0.056. Generally, more complex models seem to achieve better results, with the exception of the combination of FPN and ResNet architectures, which achieve results quite similar to the configuration using the MobileNet encoder. The models belonging to this configuration show higher losses when adapting to the task in the earlier stages.

The lack of a shared publicly available dataset aimed at the segmentation of CSC-related fluid regions precludes a fair comparison between this work and others in the literature. With this in mind, the study by Rao et al. [28] reports values of 0.936 for Precision, 0.890 for Recall, and 0.910 for Dice, using a private dataset and pre-processing stages. The previous work by de Moura et al. [29] achieved values of 0.879 for Jaccard, and a 0.965 for Dice, on a different dataset. The end-to-end configuration with the highest results considered in this work (U-Net with ResNet encoder) achieved average values of 0.832 for Precision, 0.918 for Recall, 0.769 for Jaccard and 0.868 for Dice. While these values are not directly comparable since they are measured against different datasets, they are indicative of these models achieving a performance at least competitive with those of the state of the art [28, 29]. In this scenario, an intra- and inter-expert analysis can provide valuable insight and assess the performance of the models presented in this work. An inter-expert analysis involves the comparison of the results obtained by different models within the same study, rather than by comparing the results with other works. By comparing the results of the architectural configurations within the same study, it is possible to identify the strengths and limitations of each model, as well as to accurately determine which configurations can perform better in specific scenarios. Furthermore, the addition of an intra-expert analysis can shed light on the variability and robustness of the trained models under different training scenarios, ultimately providing a more comprehensive evaluation of the models and highlighting the challenges and opportunities for further research.

Intra- and Inter-expert Analysis

In the intra-expert analysis, the models were evaluated by comparing the segmentation results produced by each model belonging to a configuration among themselves, using the second independent dataset. Thus, for every configuration, the output segmentation maps of every model were

Table 2 Test results for each segmentation architecture configuration

Backbone	Encoder	Par.	Accuracy	Recall	Precision	Jaccard	Dice
FPN	MobileNet	4M	0.998 ± 0.002	0.897 ± 0.167	0.829 ± 0.035	0.752 ± 0.121	0.853 ± 0.086
	DenseNet	15M	0.998 ± 0.002	0.912 ± 0.135	0.825 ± 0.042	0.761 ± 0.101	0.861 ± 0.069
	ResNet	23M	0.998 ± 0.002	0.893 ± 0.185	0.828 ± 0.035	0.747 ± 0.136	0.849 ± 0.098
U-Net	MobileNet	7M	0.997 ± 0.002	0.871 ± 0.179	0.824 ± 0.041	0.729 ± 0.136	0.837 ± 0.095
	DenseNet	21M	0.998 ± 0.002	0.909 ± 0.149	0.827 ± 0.048	0.757 ± 0.136	0.859 ± 0.072
	ResNet	24M	**0.998 ± 0.001**	**0.918 ± 0.123**	**0.832 ± 0.051**	**0.769 ± 0.085**	**0.868 ± 0.056**

Values shown as average ± standard deviation of the models trained in the five cross-validation subsets

Par. denotes the number of parameters

Highest results shown in bold

compared pair-wise with those of every other model. The results were then averaged across all the models belonging to said configuration. This allows the extraction of a single summary value for each metric for every architecture configuration. The results can be found in Fig. 3.

The results show that these models are highly robust, without significant deviations, even with trained with different images and using an external set of images under the same conditions. Between the two base segmentation backbone architectures that were considered, the models that used the U-Net architecture seem to display less variability among themselves than those trained with the FPN architecture, with all U-Net models achieving better results in every metric. This can be indicative that the U-Net architecture is less prone to overfitting to training data than the FPN. Models using the U-Net architecture are generally more complex in terms of trainable parameters (Table 1), which can suggest that more complex architectures may fare better in terms of variability and robustness. This fact is also supported by models using the MobileNet architecture showing greater instability than more complex ones.

In the inter-expert analysis, the human experts and automated models were compared among themselves. The deep learning architecture configurations that produced the highest results were the FPN backbone with DenseNet encoder module, and the U-Net backbone with a ResNet encoder module. Within each configuration, the model with the lowest validation loss was selected to generate the segmentation masks for comparison with the human experts. Figure 6 displays the Bland-Altman plots of the three scenarios that were considered, while Table 4 shows the results from the inter-expert analysis.

These results show that the models correctly adapt to this task, achieving results that fall well within expert variability. In the first comparison scenario, the two human experts were compared. The results show that there is a bias towards the second expert, indicating that they tend to over-segment when compared with the first expert in terms of mean difference of area segmented. This can be seen in the corresponding Bland-Altman plot, where most of the examples tend to describe a descending trend. Comparing the two experts produces a Dice and a κ coefficients of 0.952 and 0.951, as well as a MSE of 0.229. This comparison shows the most significant disagreement of those considered, and is representative of what can be expected from manual inspection in daily clinical practice. Regarding the second scenario, in which the models are compared with the experts, we can see that the models agree more with each of the human experts separately than these experts agree with each other. As established in the first scenario, the first expert may tend to under-segment, while the second one seems to have a tendency to over-segment. This is apparent also in the comparison with the models, where comparisons with Expert 1 yield a positive bias (towards the model segmenting more area) while comparisons with Expert 2 yield a negative bias (towards the second expert). All models achieve higher Dice and κ coefficients, as well as a smaller MSE than the human experts. While the U-Net architecture achieved comparably better results during testing in the first stage, the FPN architecture seems to better align itself with Expert 1. Both architectures align similarly with the second expert. In the third scenario both models were compared among themselves. The corresponding Bland-Altman plot shows that the FPN model tends to over-segment in images with smaller patches of fluid, while U-Net segments more area in more affected images. Nevertheless, with a mean difference of 8, the bias between the models is substantially smaller than in any comparison involving the human experts. Overall, these models seem to provide good balance between the two human experts. Both deep learning-based configurations are able to find a consensus close to either of the human experts, without significant over- or under-segmentation, and finding better agreement among themselves and with either of the experts than either human expert with the other. This highlights the significant subjectivity in the process of manual segmentation, and shows that machine learning models can be used to find a consensus between experts and provide a robust and repeatable assessment of these images for the diagnosis of CSC.

In order to better assess the possible differences between the manual and automatic annotations, Fig. 7 shows some visual examples of the segmentation results. This image

Table 3 Intra-expert analysis results. Metrics are extracted by comparing the segmentations produced by every model against all those produced by the other models within a configuration, then averaging across all models belonging to each configuration. For MSE, lower values are better

		Accuracy	Jaccard	Dice	κ	MSE
FPN	MobileNet	0.999 ± 0.001	0.934 ± 0.029	0.965 ± 0.015	0.965 ± 0.016	0.016 ± 0.005
	DenseNet	0.999 ± 0.001	0.944 ± 0.026	0.971 ± 0.014	0.971 ± 0.014	0.012 ± 0.002
	ResNet	0.999 ± 0.001	0.939 ± 0.033	0.968 ± 0.017	0.967 ± 0.018	0.014 ± 0.001
U-Net	MobileNet	0.999 ± 0.001	0.975 ± 0.030	0.975 ± 0.016	0.972 ± 0.016	0.007 ± 0.001
	DenseNet	0.999 ± 0.000	0.955 ± 0.024	0.977 ± 0.012	0.976 ± 0.013	0.008 ± 0.001
	ResNet	0.999 ± 0.000	0.954 ± 0.019	0.976 ± 0.010	0.977 ± 0.010	0.007 ± 0.002

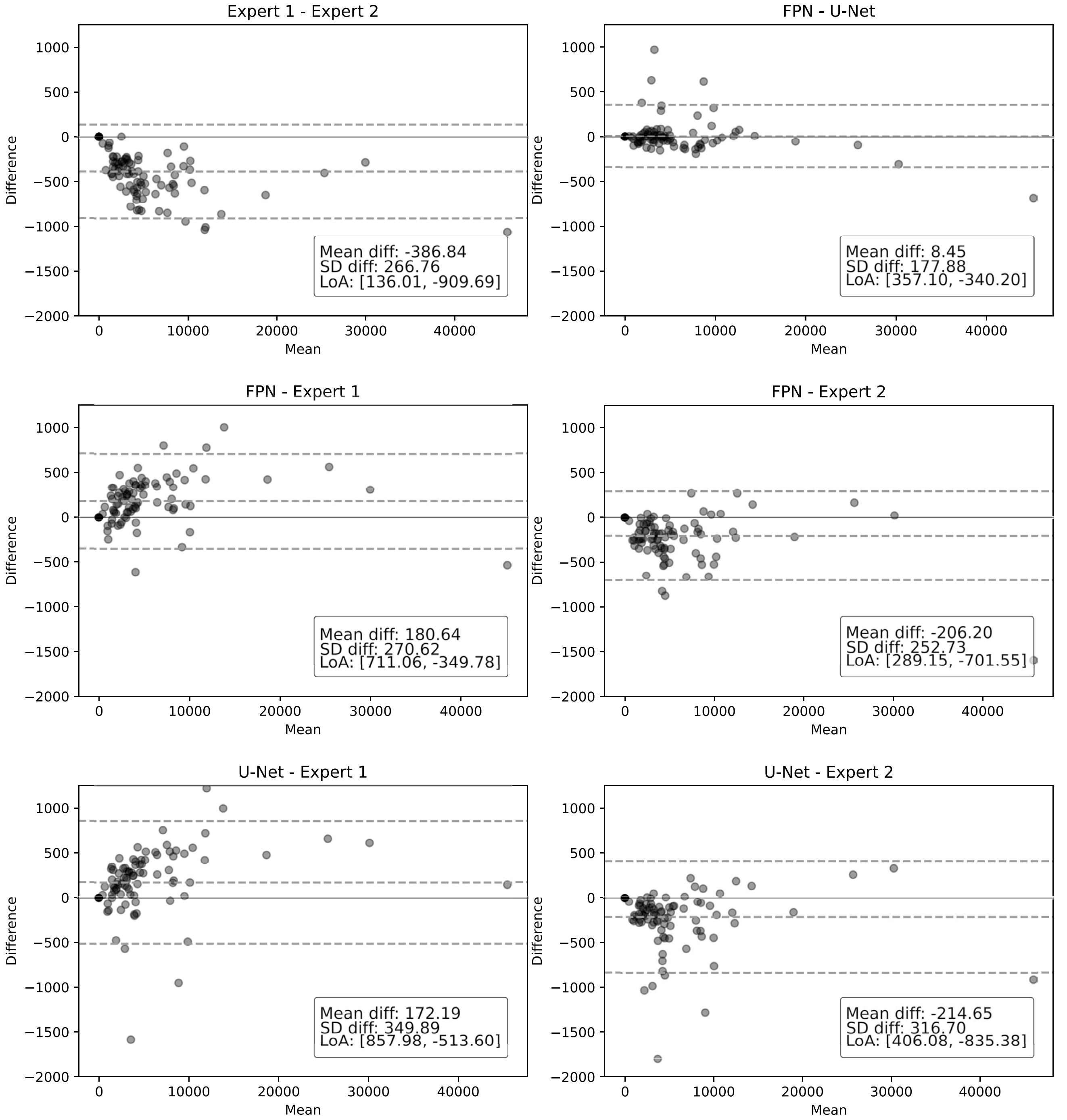

Fig. 6 Bland-Altman plots showing the one-on-one comparison between the human experts and the automated segmentation models. The horizontal axis represents the average area segmented as CSC-related fluid in both images. The vertical axis represents the difference between the area segmented in each image in the first set and its equivalent in the second set

highlights the differences in criteria regarding the boundaries of the segmented areas, as well as the size and location of smaller areas, specially those adjacent to bigger fluid accumulations. For more easily explainable results, Axiom-based Gradient-weighted Class Activation Mapping (XGrad-CAM) [53] can be used to visualise the areas that maximise network activation for the fluid detection. This way, it is possible to see the areas where the model has a higher activation and areas where the activation is lower, producing segmentation results that are easier to interpret (Fig. 8).

While this work presents many strengths, it is essential to address the following limitations. First, the images employed

Table 4 Inter-expert analysis results

			Amp.	MD	Dice	κ	MSE
Expert 1	-	Expert 2	1051	-387	0.952	0.951	0.229
FPN	-	Expert 1	1066	271	0.960	0.960	0.134
FPN	-	Expert 2	996	-206	0.961	0.960	0.140
U-Net	-	Expert 1	1378	172	0.957	0.956	0.132
U-Net	-	Expert 2	1248	-214	0.961	0.960	0.139
FPN	-	U-Net	697	8	0.960	0.959	0.138

Amp. denotes amplitude, as the difference between the higher and lower LoA in the Bland-Altman plot, a higher value indicates higher disagreement

MD is the mean difference between the number of pixels detected by the first and second experts, values further from 0 indicate a bias

in this work are categorised as presenting CSC or healthy tissue, but the actual severity of the CSC images has not been graded. Second, the comparisons in this work have been limited to end-to-end deep learning architectures that have performed well in similar tasks, and other bespoke methods tailored specifically for CSC segmentation may provide better results. Finally, the methods in this work were validated using images from a single platform (Heidelberg SPECTRALIS®).

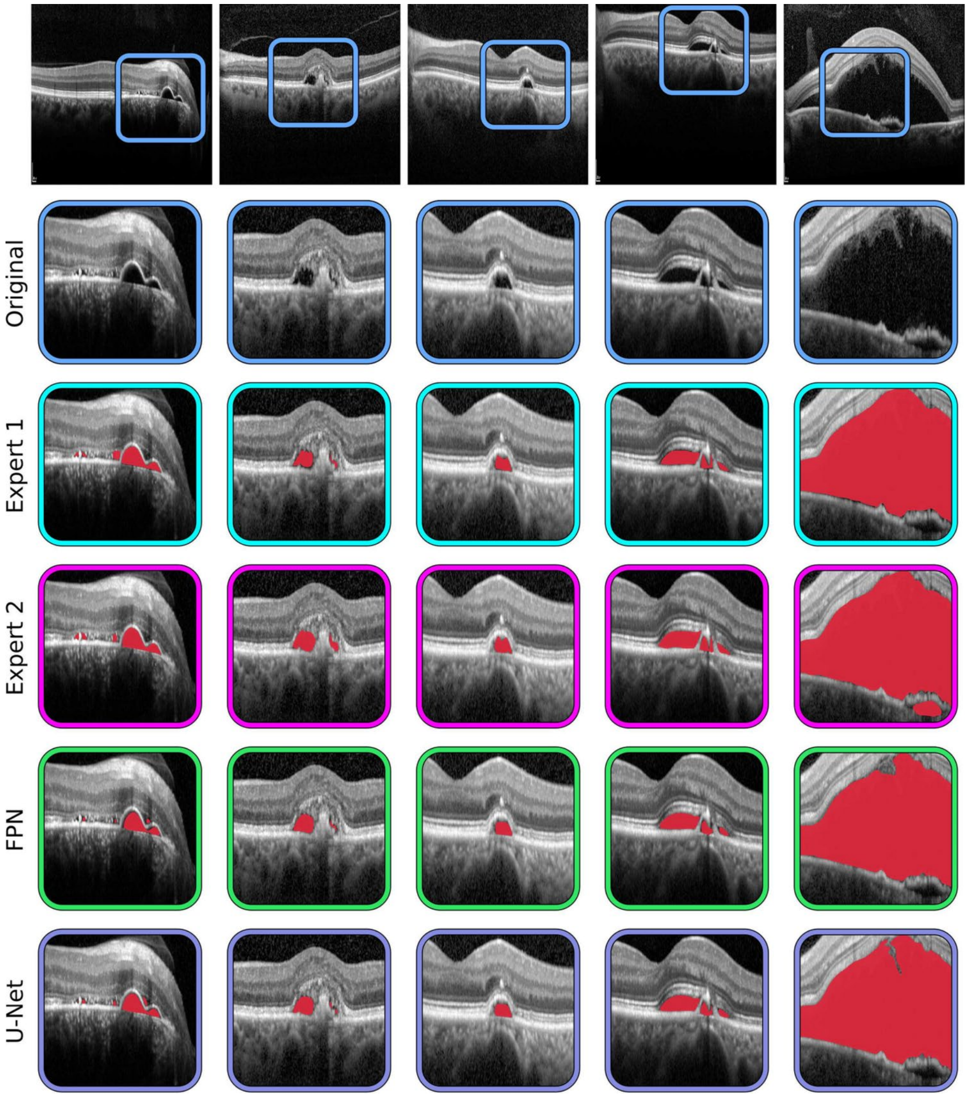

Fig. 7 General and detailed view of the manual annotations and automatic segmentation masks of images from the inter-expert analysis performed using the second dataset. Segmentation masks overlaid with original image for ease of reference

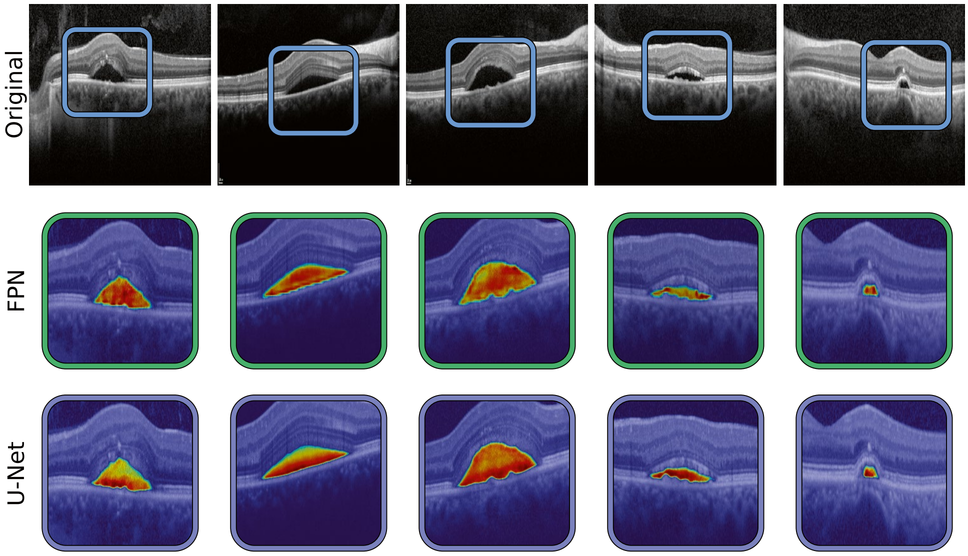

Fig. 8 Heatmap visualisation using XGrad-CAM of the class activation for fluid in the retina generated using the best performing FPN and U-Net configurations

Conclusions

The diagnosis of CSC is typically performed by means of an expert visually inspecting images in search for signs of the disease. This process is subjective, tiring and can lead to errors which, in turn, can translate into a late or even missed diagnosis. In this work, we have designed and validated a methodology for the fully automatic segmentation of CSC-related fluid regions in OCT images. This methodology has been implemented using a series of modular state-of-the-art segmentation architectures, representative of a spectrum of complexity. Along with a thorough comparison of all architectures, studying which are better suited for the segmentation of CSC signs in OCT images, this work is the first in the literature to propose a comprehensive intra- and inter-expert analysis to validate these models. In this study, the models are compared among themselves and with different human experts using an external validation dataset. This study can help measure the variability caused by the natural subjectivity of manual image inspection, and also provide a robust framework with which to validate deep learning models aimed at this task.

The results that were obtained show that these automatic models can perform at least at a level comparable to the experts, finding a balance between them and achieving a higher level of agreement with either of the human experts than those among themselves. The models also show a smaller bias when compared with either of the human experts, and almost none when compared with each other. These findings indicate that while differences in expert criteria may exist, deep learning models can be used to achieve a robust and repeatable segmentation of CSC-related lesions in OCT imaging, finding a consensus among experts and providing an objective and accurate segmentation of the fluid regions. These models can be used to improve the diagnosis process of the CSC, while improving patient care and prognosis thanks to an early and precise assessment of this disease.

Plans for future work involve a more detailed analysis of the different stages at which fluid may accumulate under the RPE in CSC patients. Moreover, future work could focus on including other purpose-specific architectures into the study that may provide other advantages into this task, as well as studying the data efficiency of these architectures. Finally, the inclusion of other imaging platforms to conform to a multi-expert multi-vendor dataset could help provide a more robust validation framework with which to validate future fluid segmentation methodologies.

Author Contribution Mateo Gende: methodology, software, formal analysis, data curation, writing — original draft, writing — review and editing, visualisation. Lúa Castelo: software, formal analysis, data curation, writing — original draft. Joaquim de Moura: conceptualisation, validation, data acquisition, supervision, project administration, writing — original draft, writing — review and editing. Jorge Novo: conceptualisation, validation, data acquisition, supervision, project administration, writing — review and editing. Marcos Ortega: conceptualisation, supervision, project administration, funding acquisition, writing — review and editing.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This research was funded by Instituto de Salud Carlos III, Government of Spain (research project DTS18/00136); Ministerio de Ciencia e Innovación, Government of Spain (research projects PID2019-108435RB-I00, TED2021-131201B-I00, PDC2022-133132-I00); Consellería de Cultura, Educación e Universidade, Xunta de Galicia, through Grupos de Referencia Competitiva (grant number ED431C 2020/24) and predoctoral grant (grant number ED481A 2021/161); and CITIC, Centro de Investigación de Galicia (grant number ED431G 2019/01), and receives financial support from Consellería de Educación, Universidade e Formación Profesional, Xunta de Galicia, through the ERDF (80%) and Secretaría

Xeral de Universidades (20%). The funding sources had no role in the development of this work.

Declarations

Ethics Approval The authors have no relevant financial or non-financial interests to disclose. This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of Universidade da Coruña/Ferrol (2014/437) the 24^{th} of November, 2014.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Competing Interests The authors declare no competing interests.

References

1. R. Kaye, S. Chandra, J. Sheth, C.J.F. Boon, S. Sivaprasad, A. Lotery, Central serous chorioretinopathy: An update on risk factors, pathophysiology and imaging modalities. Progress in Retinal and Eye Research **79**, 100,865 (2020). https://doi.org/10.1016/J.PRETEYERES.2020.100865
2. M. Uyama, H. Matsunaga, T. Matsubara, I. Fukushima, K. Takahashi, T. Nishimura, Indocyanine green angiography and pathophysiology of multifocal posterior pigment epitheliopathy. Retina (Philadelphia, Pa.) **19**, 12–21 (1999). https://doi.org/10.1097/00006982-199901000-00003
3. R. Liegl, M.W. Ulbig, Central serous chorioretinopathy. Ophthalmologica **232**, 65–76 (2014). https://doi.org/10.1159/000360014
4. S. Mrejen, C. Balaratnasingam, T.R. Kaden, A. Bottini, K. Dansingani, K.V. Bhavsar, N.A. Yannuzzi, S. Patel, K.C. Chen, S. Yu, G. Stoffels, R.F. Spaide, K.B. Freund, L.A. Yannuzzi, Long-term visual outcomes and causes of vision loss in chronic central serous chorioretinopathy. Ophthalmology **126**(4), 576–588 (2019). https://doi.org/10.1016/j.ophtha.2018.12.048
5. G. Quin, G. Liew, I.V. Ho, M. Gillies, S. Fraser-Bell, Diagnosis and interventions for central serous chorioretinopathy: review and update. Clin Experiment Ophthalmol **41**(2), 187–200 (2012). https://doi.org/10.1111/j.1442-9071.2012.02847.x
6. A. Daruich, A. Matet, A. Dirani, E. Bousquet, M. Zhao, N. Farman, F. Jaisser, F. Behar-Cohen, Central serous chorioretinopathy: Recent findings and new physiopathology hypothesis. Progress in Retinal and Eye Research **48**, 82–118 (2015). https://doi.org/10.1016/j.preteyeres.2015.05.003
7. S. Aumann, S. Donner, J. Fischer, F. Müller, Optical coherence tomography (oct): Principle and technical realization. High Resolution Imaging in Microscopy and Ophthalmology pp. 59–85 (2019). https://doi.org/10.1007/978-3-030-16638-0_3/FIGURES/19
8. A.F. Fercher, Optical coherence tomography - development, principles, applications. Zeitschrift für Medizinische Physik **20**, 251–276 (2010). https://doi.org/10.1016/J.ZEMEDI.2009.11.002
9. M. Wang, I.C. Munch, P.W. Hasler, C. Prünte, M. Larsen, Central serous chorioretinopathy. Acta Ophthalmologica **86**(2), 126–145 (2008). https://doi.org/10.1111/j.1600-0420.2007.00889.x
10. M.R. Hee, C.A. Puliafito, C. Wong, E. Reichel, J.S. Duker, J.S. Schuman, E.A. Swanson, J.G. Fujimoto, Optical coherence tomography of central serous chorioretinopathy. American Journal of Ophthalmology **120**, 65–74 (1995). https://doi.org/10.1016/S0002-9394(14)73760-2
11. K.K. Bhatia, M.S. Graham, L. Terry, A. Wood, P. Tranos, S. Trikha, N. Jaccard, Disease classification of macular optical coherence tomography scans using deep learning software. Retina **40**(8), 1549–1557 (2020). https://doi.org/10.1097/iae.0000000000002640
12. J. de Moura, G. Samagaio, J. Novo, P. Almuina, M.I. Fernández, M. Ortega, Joint diabetic macular edema segmentation and characterization in OCT images. J Digit Imaging **33**(5), 1335–1351 (2020). https://doi.org/10.1007/s10278-020-00360-y
13. G.R. Wilkins, O.M. Houghton, A.L. Oldenburg, Automated segmentation of intraretinal cystoid fluid in optical coherence tomography. IEEE Transactions on Biomedical Engineering **59**(4), 1109–1114 (2012). https://doi.org/10.1109/tbme.2012.2184759
14. N. Eladawi, M. Elmogy, M. Ghazal, O. Helmy, A. Aboelfetouh, A. Riad, S. Schaal, A. El-Baz, Classification of retinal diseases based on oct images. Frontiers in Bioscience - Landmark **23**, 247–264 (2018). https://doi.org/10.2741/4589/4589.PDF
15. D.S. Maltsev, A.N. Kulikov, J. Chhablani, D.S. Kutik, N.V. Arsenov, [optical coherence tomography in diagnostics and treatment of central serous chorioretinopathy]. Vestnik Oftalmologii **134**, 15–24 (2018). https://doi.org/10.17116/OFTALMA201813406115
16. C. Valverde, M. Garcia, R. Hornero, M. Lopez-Galvez, Automated detection of diabetic retinopathy in retinal images. Indian Journal of Ophthalmology **64**, 26 (2016). https://doi.org/10.4103/0301-4738.178140
17. H. Fujita, AI-based computer-aided diagnosis (AI-CAD): the latest review to read first. Radiol Phys Technol **13**(1), 6–19 (2020). https://doi.org/10.1007/s12194-019-00552-4
18. Y. Lecun, Y. Bengio, G. Hinton, Deep learning. Nature 2015 521:7553 **521**, 436–444 (2015). https://doi.org/10.1038/nature14539
19. M. Chen, K. Jin, K. You, Y. Xu, Y. Wang, C.C. Yip, J. Wu, J. Ye, Automatic detection of leakage point in central serous chorioretinopathy of fundus fluorescein angiography based on time sequence deep learning. Graefe's Archive for Clinical and Experimental Ophthalmology 2021 259:8 **259**, 2401–2411 (2021). https://doi.org/10.1007/S00417-021-05151-X
20. F. Xu, S. Liu, Y. Xiang, Z. Lin, C. Li, L. Zhou, Y. Gong, L. Li, Z. Li, C. Guo, C. Huang, K. Lai, H. Zhao, J. Hong, H. Lin, C. Jin, Deep learning for detecting subretinal fluid and discerning macular status by fundus images in central serous chorioretinopathy. Front. Bioeng. Biotechnol. **9** (2021). https://doi.org/10.3389/fbioe.2021.651340
21. T.K. Yoo, B.Y. Kim, H.K. Jeong, H.K. Kim, D. Yang, I.H. Ryu, Simple code implementation for deep learning–based segmentation to evaluate central serous chorioretinopathy in fundus photography. Trans. Vis. Sci. Tech. & Technology **11**(2), 22 (2022). https://doi.org/10.1167/tvst.11.2.22
22. J. de Moura, J. Novo, S. Penas, M. Ortega, J. Silva, A.M. Mendonça, Automatic characterization of the serous retinal detachment associated with the subretinal fluid presence in optical coherence tomography images. Procedia Computer Science **126**, 244–253 (2018). https://doi.org/10.1016/j.procs.2018.07.258
23. P.L. Vidal, J. de Moura, J. Novo, M.G. Penedo, M. Ortega, Intraretinal fluid identification via enhanced maps using optical

coherence tomography images. Biomed. Opt. Express **9**(10), 4730 (2018). https://doi.org/10.1364/boe.9.004730

24. M. Gende, J. de Moura, J.I. Fernández-Vigo, J.M.M. de-la Casa, J. García-Feijóo, J. Novo, M. Ortega, Robust multi-view approaches for retinal layer segmentation in glaucoma patients via transfer learning. Quantitative Imaging in Medicine and Surgery **0**(0) (2023). https://doi.org/10.21037/qims-22-959
25. C.S. Lee, D.M. Baughman, A.Y. Lee, Deep learning is effective for classifying normal versus age-related macular degeneration oct images. Ophthalmology Retina **1**, 322–327 (2017). https://doi.org/10.1016/J.ORET.2016.12.009
26. G.N. Girish, B. Thakur, S.R. Chowdhury, A.R. Kothari, J. Rajan, Segmentation of intra-retinal cysts from optical coherence tomography images using a fully convolutional neural network model. IEEE Journal of Biomedical and Health Informatics **23**(1), 296–304 (2019). https://doi.org/10.1109/jbhi.2018.2810379
27. K. Gao, W. Kong, S. Niu, D. Li, Y. Chen, Automatic retinal layer segmentation in SD-OCT images with CSC guided by spatial characteristics. Multimed Tools Appl **79**(7-8), 4417–4428 (2019). https://doi.org/10.1007/s11042-019-7395-9
28. T.J.N. Rao, G.N. Girish, A.R. Kothari, J. Rajan, in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE, 2019), pp. 978–988. https://doi.org/10.1109/embc.2019.8857105
29. J. de Moura, J. Novo, M. Ortega, N. Barreira, M.G. Penedo, in *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)* (IEEE, 2021), pp. 1–6. https://doi.org/10.1109/cbms52027.2021.00008
30. V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(12), 2481–2495 (2017). https://doi.org/10.1109/tpami.2016.2644615
31. T.Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 2117–2125
32. O. Ronneberger, P. Fischer, T. Brox, in *Lecture Notes in Computer Science* (Springer International Publishing, 2015), pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
33. M. Sandler, A.G. Howard, M. Zhu, A. Zhmoginov, L. Chen, in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018* (Computer Vision Foundation / IEEE Computer Society, 2018), pp. 4510–4520. https://doi.org/10.1109/CVPR.2018.00474. http://openaccess.thecvf.com/content_cvpr_2018/html/Sandler_MobileNetV2_Inverted_Residuals_CVPR_2018_paper.html
34. G. Huang, Z. Liu, L.V.D. Maaten, K.Q. Weinberger, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2017), pp. 2261–2269. https://doi.org/10.1109/cvpr.2017.243
35. K. He, X. Zhang, S. Ren, J. Sun, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016* (IEEE Computer Society, 2016), pp. 770–778. https://doi.org/10.1109/CVPR.2016.90
36. A. Kirillov, K. He, R. Girshick, P. Dollár. A unified architecture for instance and semantic segmentation (2017). http://presentations.cocodataset.org/COCO17-Stuff-FAIR.pdf
37. S. Ghosh, K. Santosh, in *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)* (IEEE, 2021), pp. 31–36. https://doi.org/10.1109/cbms52027.2021.00013
38. B. Pu, Y. Lu, J. Chen, S. Li, N. Zhu, W. Wei, K. Li, MobileU-Net-FPN: A semantic segmentation model for fetal ultrasound four-chamber segmentation in edge computing environments. IEEE Journal of Biomedical and Health Informatics **26**(11), 5540–5550 (2022). http://doi.org/10.1109/jbhi.2022.3182722
39. R. Yang, Y. Yu, Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. Front. Oncol. **11** (2021). https://doi.org/10.3389/fonc.2021.638182
40. B. Lee, N. Yamanakkanavar, J.Y. Choi, Automatic segmentation of brain MRI using a novel patch-wise u-net deep architecture. PLoS ONE **15**(8), e0236,493 (2020). https://doi.org/10.1371/journal.pone.0236493
41. B. Wu, Y. Fang, X. Lai, Left ventricle automatic segmentation in cardiac MRI using a combined CNN and u-net approach. Computerized Medical Imaging and Graphics **82**, 101,719 (2020). https://doi.org/10.1016/j.compmedimag.2020.101719
42. C.H. Sudre, W. Li, T. Vercauteren, S. Ourselin, M.J. Cardoso, in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Springer International Publishing, 2017), pp. 240–248. https://doi.org/10.1007/978-3-319-67558-9_28
43. A. Yaguchi, K. Aoyagi, A. Tanizawa, Y. Ohno, in *Medical Imaging 2019: Computer-Aided Diagnosis*, ed. by H.K. Hahn, K. Mori (SPIE, 2019), p. 109503G. https://doi.org/10.1117/12.2511438
44. K. Gao, J. Su, Z. Jiang, L.L. Zeng, Z. Feng, H. Shen, P. Rong, X. Xu, J. Qin, Y. Yang, W. Wang, D. Hu, Dual-branch combination network (DCN): Towards accurate diagnosis and lesion segmentation of COVID-19 using CT images. Medical Image Analysis **67**, 101,836 (2021). https://doi.org/10.1016/j.media.2020.101836
45. A. Mehrtash, W.M. Wells, C.M. Tempany, P. Abolmaesumi, T. Kapur, Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Transactions on Medical Imaging **39**(12), 3868–3878 (2020). http://doi.org/10.1109/tmi.2020.3006437
46. D.P. Kingma, J. Ba, in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, ed. by Y. Bengio, Y. LeCun (2015). URL http://arxiv.org/abs/1412.6980
47. R. Mirshahi, P. Anvari, H. Riazi-Esfahani, M. Sardarinia, M. Naseripour, K.G. Falavarjani, Foveal avascular zone segmentation in optical coherence tomography angiography images using a deep learning approach. Sci Rep **11**(1) (2021). https://doi.org/10.1038/s41598-020-80058-x
48. D. Giavarina, Understanding bland altman analysis. Biochem Med **25**(2), 141–151 (2015). https://doi.org/10.11613/bm.2015.015
49. A. Goel, G. Shih, S. Riyahi, S. Jeph, H. Dev, R. Hu, D. Romano, K. Teichman, J.D. Blumenfeld, I. Barash, I. Chicos, H. Rennert, M.R. Prince, Deployed deep learning kidney segmentation for polycystic kidney disease MRI. Radiology: Artificial Intelligence **4**(2) (2022). https://doi.org/10.1148/ryai.210205
50. T.C. Blaney, J.L. Ronsky, E.M. Macri, J.L. Jaremko, G. Kuntze, A. Pakdel, J.L. Whittaker, C.A. Emery, Concurrent validity and reliability of a semi-automated approach to measuring the magnetic resonance imaging morphology of the knee joint in active youth. Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine **236**(7), 1023–1035 (2022). https://doi.org/10.1177/09544119221095337
51. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, in *Advances in Neural Information Processing Systems 32* (Curran Associates, Inc., 2019), pp. 8024–8035. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
52. P. Iakubovskii. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch (2019)
53. R. Fu, Q. Hu, X. Dong, Y. Guo, Y. Gao, B. Li, Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. British Machine Vision Conference (BMVC Oral) (2020). https://doi.org/10.48550/arXiv.2008.02312

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.