Segmentation of retinal structures in optical coherent tomography images

Melinščak, Martina

University of Zagreb

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Martina Melinščak

SEGMENTATION OF RETINAL STRUCTURES IN OPTICAL COHERENT TOMOGRAPHY IMAGES

DOCTORAL THESIS

Zagreb, 2022

University of Zagreb

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Martina Melinščak

SEGMENTATION OF RETINAL STRUCTURES IN OPTICAL COHERENT TOMOGRAPHY IMAGES

DOCTORAL THESIS

Supervisor: Professor Sven Lončarić, PhD

Zagreb, 2022

Sveučilište u Zagrebu

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Martina Melinščak

SEGMENTACIJA RETINALNIH STRUKUTRA U SLIKAMA OPTIČKE KOHERENTNE TOMOGRAFIJE

DOKTORSKI RAD

Mentor: Prof. dr. sc. Sven Lončarić

Zagreb, 2022.

Doktorski rad izrađen je na Sveučilištu u Zagrebu Fakultetu elektrotehnike i računarstva, na Zavodu za elektroničke sustave i obradbu informacija.

Mentor: prof. dr. sc. Sven Lončarić

Doktorski rad ima: 82 stranice

Doktorski rad br.:_____

About the supervisor

Dr. Sven Lončarić was born in Zagreb in 1961. He received Diploma of Engineering and Master of Science degrees in electrical engineering from University of Zagreb Faculty of Electrical Engineering and Computing (FER) in 1985 and 1989, respectively. He received Doctor of Philosophy (Ph.D.) degree in electrical engineering from University of Cincinnati, USA, in 1994. Since 2011, he has been a tenured full professor in electrical engineering and computer science at FER. He was researcher or project leader on a number of research projects in the area of image processing and computer vision. From 2001-2003, he was an assistant professor at New Jersey Institute of Technology, USA. He has been the head of the Image Processing Laboratory at FER. He founded the Center for Computer Vision at University of Zagreb. Prof. Lončarić has been a co-director of the national Center of Research Excellence in Data Science and Cooperative Systems. He authored or co-authored more than 200 scientific publications. Prof. Lončarić is a senior member of IEEE and a member of Croatian Academy of Technical Sciences. He was the Chair of the IEEE Croatia Section. He founded and organized several international scientific conferences. He is an editor, a program committee member and reviewer for a number of scientific conferences and journals. He is a recipient of the National Award for Outstanding Scientist, University of Zagreb „Fran Bošnjaković“ Award, Croatian Academy of Engineering "Rikard Podhorsky" Award, FER Science Award, Fulbright stipend, and an IEEE Croatia Section Award.

O mentoru

Prof. Dr. sc. Sven Lončarić rođen je u Zagrebu 1961. godine. Diplomirao je i magistrirao u polju elektrotehnike na Fakultetu elektrotehnike i računarstva (FER) Sveučilišta u Zagreba, 1985. i 1989. godine. Doktorirao je u polju elektrotehnike na Sveučilištu u Cincinnatiju, SAD, 1994. godine. U zvanje redoviti profesor u trajnom zvanju u polju elektrotehnike i polju računarstva na FER-u izabran je 2011. godine. Bio je suradnik ili voditelj na brojnim istraživačkim i razvojnim projektima u području razvoja metoda za obradu slika i računalnog vida. Od 2001. do 2003. bio je Assistant Professor na Sveučilištu New Jersey Institute of Technology, SAD. Voditelj je istraživačkog laboratorija za obradu slike na FER-u. Osnivač je i voditelj Centra izvrsnosti za računalni vid na Sveučilištu u Zagrebu. Suvoditelj je nacionalnog Znanstvenog centra izvrsnosti za znanost o podatcima i kooperativne sustave. Sa svojim suradnicima publicirao je više od 200 znanstvenih i stručnih radova. Prof. Lončarić član je stručne udruge IEEE i Akademije tehničkih znanosti Hrvatske. Bio je predsjednik Hrvatske sekcije IEEE. Osnivač je i organizator više međunarodnih znanstvenih skupova i ljetnih škola. Bio je urednik, član uredničkih i programskih odbora i recenzent za više međunarodnih znanstvenih skupova i časopisa. Dobitnik je Državne nagrade za znanost, Nagrade „Fran Bošnjaković“ Sveučilišta u Zagrebu, Nagrade "Rikard Podhorky" Akademije tehničkih znanosti Hrvatske, Nagrade za znanost Fakulteta elektrotehnike i računarstva, Fulbrajtove stipendije i Nagrade Hrvatske sekcije IEEE.

Abstract

Optical coherence tomography (OCT) images of the retina provide a structural representation and give an insight into the pathological changes present in age-related macular degeneration (AMD). Due to the three-dimensionality and complexity of the images, manual analysis of pathological features is difficult, time-consuming, and prone to subjectivity. Computer analysis of 3D OCT images is necessary to enable automated quantitative measuring of the features, objectively and repeatedly. As supervised and semi-supervised learning-based automatic segmentation depends on the training data and quality of annotations, a new database of **A**nnotated **R**etinal **O**CT **I**mages (AROI database) has been created. It consists of 1136 images with annotations for pathological changes (fluid accumulation and related findings) and basic structures (layers) in patients with AMD. Inter- and intra-observer errors have been calculated to enable the validation of developed algorithms in relation to human variability. The framework for validation of OCT image segmentation methods consisting of a publicly available database of manually labeled image features and methodology and programming code for standardized measurement of segmentation accuracy is developed thus making the first scientific contribution. In the case of retinal OCT images of patients suffering from AMD, there is a significant deformation of the retinal structure. This causes the methods for retinal layers segmentation without introducing information on pathology significantly impaired. Therefore, an approach for layer segmentation which takes into account information about fluids (joint segmentation of layers and fluids) was chosen. Also, the automatic segmentation with standard U-net architecture and two state-of-the-art architectures for medical image segmentation has been performed to set a baseline for further algorithm development and to get insight into challenges for automatic segmentation. After an analysis of how various architectures and their modifications affect the performance and accuracy of segmentation, an analysis of the effects of various data manipulations (data augmentation, image post-processing) on segmentation performance was performed. Based on all the examined methods, comparison and proposal of two models for segmentation of retinal and intraretinal fluid layers in optical coherence tomography images that give comparable results were made. One model consists of a standard U-net architecture with post-processing using conditional random fields. The second model consists of a neural network architecture that combines the good features of the U-net architecture and the ResNet architecture. The existing state-of-the-art architecture has been adapted to the given problem and improved by adding skip connections. Additionally,

segmentation accuracy was improved by using data augmentation and post-processing using conditional random fields. This is the second scientific contribution. A comparative quantitative analysis of fluid-only segmentation and fluid segmentation with available layer information was performed to gain insight into how much information about layers (fluid position in the retina) contributes to the accuracy of fluid segmentation itself. A study of the efficiency of transfer learning, which is an active and insufficiently researched topic in the field of deep learning with applications in medicine, was conducted, based on which a method for retinal fluid segmentation in optical coherence tomography images was proposed. This is the third scientific contribution. In conclusion, the main challenges of automatic segmentation are presented and the ideas for further research and improvements are given.

Keywords: age-related macular degeneration, optical coherence tomography, annotated retinal OCT images, automatic image segmentation, deep learning

Segmentacija retinalnih struktura u slikama optičke koherentne tomografije

Iako je razvoj tehnologije optičke koherentne tomografije aktivno polje istraživanja od 1991. godine, tek u posljednjem desetljeću počinje značajan razvoj i istraživanje metoda segmentacije retinalnih slike optičke koherentne tomografije. Segmentacija je jedan od najtežih i ujedno nužnih koraka u analizi slika optičke koherentne tomografije. Svakako, važnost automatske segmentacije je prepoznata unutar istraživačke zajednice. Paralelno s napretkom algoritama dubokog učenja u računalnom vidu, u oftalmologiji se povećavaju zahtjevi za automatskom segmentacijom. S jedne strane suočeni smo sa sve većom količinom podataka (jedan volumen optičke koherentne tomografije obično se sastoji od preko 50 B-skenova, a svaki B-sken i po 500000 piksela, ovisno o vrsti uređaja), dok je s druge strane ručna segmentacija dugotrajna i time neprihvatljiva u kliničkoj praksi. Segmentacija retinalne strukture i patoloških biomarkera nužna je za točniju dijagnozu i učinkovitiju terapiju. Svi navedeni čimbenici rezultirali su tijekom posljednjih godina velikim brojem publiciranih radova koji predlažu metode za automatsku segmentaciju. Unatoč svemu, metode automatske segmentacije još uvijek nemaju dovoljnu točnost i mnogo je čimbenika koji ju čine izazovnom.

Trenutno komercijalno dostupni uređaji optičke koherentne tomografije mogu automatski segmentirati slojeve retine u promjenjivom opsegu i prikazivati osnovne parametre kao što su unutarnja ili vanjska debljina retine te elevacija pigmentnog sloja. Iako su radi o korisnim parametrima koji rutinski služe u donošenju kliničkih odluka, sve češće se postavljaju pitanja točnosti automatske segmentacije i ograničene reproducibilnosti i migracije među uređajima. Glavni problem algoritama za segmentaciju suvremenih uređaja optičke koherentne tomografije je nedostatak detaljnih informacija, budući da osnovni parametri debljine i elevacije ne razlikuju specifične temeljne lezije (npr. tekućine retine u različitim odjeljcima).

U uvodnom poglavlju (*„1. Introduction“*) prikazan je kontekst istraživanja i objašnjena nužnost uvođenja automatske segmentacije u oftalmološku kliničku praksu čime bi se omogućila kvantitativna analiza sa svrhom učinkovitije terapije i dijagnostike. U svijetu je preko 300 milijuna pacijenata koji boluju od retinalnih bolesti koje često vode vidnom oštećenju i sljepoći. Istovremeno je oko 230 tisuća oftalmologa koji mogu raditi analizu slika optičke koherentne tomografije. Analiza optičke koherentne tomografije nije nužna samo za

dijagnostiku nego i predikciju/evaluaciju djelovanja terapije, provođenje kliničkih studija itd. što je moguće jedino uz kvantitativna mjerenja za koje je nužna automatska segmentacija. Navedene su dosadašnje prepreke i svi veći izazovi uvođenja automatske segmentacije u kliničku praksu, čime je ujedno obrazložena motivacija te definirani ciljevi istraživanja provedenog u doktorskoj disertaciji.

Senilna makularna degeneracija uzrokuje oštećenje centralnog vida i u kasnoj fazi slijepilo. Senilna makularna degeneracija je vodeći uzrok slijepila kod ljudi starijih od 50 godina u razvijenom svijetu. Prisutan je stalni porast ljudi oboljelih od senilne makularne degeneracije, najvećim dijelom uslijed sve duljeg životnog vijeka, te se predviđa da će do 2040. godine biti 288 milijuna osoba u svijetu oboljelih od senilne makularne degeneracije. Senilna makularna degeneracija je progresivna bolest, dijeli se na ranu, srednju i kasnu fazu. Rana i srednja faza su obično asimptomatske i ne postoji adekvatna terapija. Kasna faza se dijeli na neovaskularnu ili „mokru“ senilnu makularnu degeneraciju i geografsku atrofiju ili „suhu“ senilnu makularnu degeneraciju. U slučaju neovaskularne senilne makularne degeneracije postoji adekvatna anti-VEGF terapija (engl. *anti–vascular endothelial growth factor therapy*). Da bi terapija bila što efikasnija potrebno ju je prilagoditi svakoj osobi – a za to je nužno kvantitativno mjerenje patoloških struktura koje bi bilo znatno unaprijeđeno i olakšano automatskom segmentacijom.

Retinalne slike optička koherentne tomografije (eng. *optical coherence tomography* – OCT) omogućavaju strukturalni prikaz i daju uvid u patološke promjene prisutne kod senilne makularne degeneracije. Za razliku od fundus fotografija koje su dvodimenzionalne, treća dimenzija kod slika optičke koherentne tomografije povezana s dubinom omogućuje mjerenje i vizualizaciju morfoloških značajki (primjerice mjerenje debljine pojedinih slojeva retine, mjerenje volumena retinalnih tekućina i ostalih patoloških struktura). S obzirom na trodimenzionalnost i kompleksnost slika, manualna analiza patoloških značajki je teška, vremenski zahtjevna i podložna subjektivnosti. Stoga je računalna analiza 3D slika optičke koherentne tomografije nužan preduvjet za automatsko kvantitativno mjerenje značajki, na objektivan način u sklopu kliničke prakse.

U drugom poglavlju (*„2. Framework for validation of retinal image segmentation methods“*) detaljno je prikazana prikupljena baza ručno označenih slika. Baza ručno označenih slika je preduvjet za razvijanje metode pomoću dubokog učenja, budući da performanse nadziranih i polu-nadziranih metoda ovise o količini i kvaliteti ručno označenih

slika. U prošlom desetljeću došlo je do brzog razvoja i računalnog vida i oftalmologije i glavni nedostatak na presjeku oba područja je manjak javno dostupnih baza s ručno označenim slikama optičke koherentne tomografije. Stoga su u suradnji s Kliničko bolničkim centrom Sestre milosrdnice prikupljene i anotirane slike optičke koherentne tomografije 24 pacijenta koji boluju od kasne neovaskularne senilne makularne degeneracije. Nova baza anotiranih retinalnih OCT slika (AROI baza) sastoji se od 1136 slika (B-skenova) pacijenata s pripadajućim anotacijama patoloških promjena (akumulacije tekućina i slične patološke promjene) i osnovnih anatomskih struktura (slojeva). Retina je histološki podijeljena u deset slojeva. Kod kasne neovaskularne senilne makularne degeneracije dolazi do značajnih iskrivljenje u strukturi retine uslijed prisutnih patoloških struktura i stručnjacima oftalmolozima je nemoguće pouzdano odrediti svih jedanaest granica odnosno deset slojeva. Zato su odabrane četiri granice koje je bilo moguće odrediti u svim B-skenovima i koje su važne za lokalizaciju patoloških struktura odnosno intraretinalnih tekućina: unutrašnja granična membrana (engl. *internal limiting membrane* – ILM), granica između unutrašnjeg mrežastog sloja (engl. *inner plexiform layer* – IPL) i unutrašnjeg zrnatog sloja (engl. *inner nuclear layer* – INL), unutarnja granica pigmentnog sloja (engl. *retinal pigment epithelium* – RPE) i Bruchova membrana. Od tekućina su označene: odignuće RPE (engl. *pigment epithelial detachment* – PED), subretinalna tekućina (engl. *subretinal fluid* – SRF) i intraretinalna tekućina (engl. *intraretinal fluid* – IRF)

Sve anotacije su napravljene od strane stručnjaka oftalmologa. Također, izračunata je procjena ljudske varijabilnosti kako bi se omogućila usporedba i validacija razvijenih algoritama u odnosu na ljudsku varijabilnost. Time je razvijen radni okvir za provjeru metoda za segmentaciju slika optičke koherentne tomografije koji se sastoji od javno dostupne baze slika s ručno označenim značajkama te metodologije i od programskog koda za standardizirano mjerenje točnosti segmentacije čime je ostvaren prvi znanstveni doprinos.

U uvodnom dijelu trećeg poglavlja (*„3. Methods for retinal layers and joint retinal layers and fluids segmentation in OCT images“*) dan je pregled razvijenih metoda za segmentaciju slojeva i simultanu segmentaciju slojeva i intraretinalnih tekućina. U slučaju retinalnih slika optičke koherentne tomografije pacijenata koji boluju od senilne makularne degeneracije dolazi do značajne deformacije strukture retine čime je onemogućena segmentacija slojeva retine bez uvođenja informacija o patološkim promjenama bez značajnog narušavanja točnosti segmentacije razvijenih modela. Zato je odabran pristup

segmentacije slojeva pri čemu se uzimaju u obzir informacije o tekućinama, odnosno istovremena segmentacija slojeva i tekućina.

Od 2015. godine kada je predstavljena U-net arhitektura, ona i razne modifikacije U-net arhitekture postale su najčešći izbor prilikom semantičke segmentacije u medicini. Zato je provedena automatska segmentacija sa standardnom U-net arhitekturom i dvije napredne i češće korištene arhitekture za segmentaciju medicinskih slika („U-net-like“ i „U-net++“) kako bi se postavila osnova za daljnji razvoj algoritama i dobio uvid u izazove automatske segmentacije. Nakon što je provedena analiza kako razne arhitekture i njihove modifikacije djeluju na točnost i performanse segmentacije, provedena je analiza djelovanja raznih manipulacija nad podacima (augmentacija podataka, obrada slika) na performanse segmentacije. Na osnovu svih ispitanih metoda, napravljena je usporedba i prijedlog dvaju modela za segmentaciju slojeva retine i intraretinalnih tekućina u slikama optičke koherentne tomografije koji daju sumjerljive rezultate. Jedan model se sastoji od standardne U-net arhitekture uz obradu predikcija pomoću uvjetnih slučajnih polja (eng. *conditional random fields*). Drugi model se sastoji od arhitekture neuronske mreže koja kombinira dobre značajke U-net arhitekture i ResNet arhitekture. Postojeća napredna (eng. *state-of-the-art*) arhitektura prilagođena je za dani problem i unaprijeđena dodavanjem veza između enkodera i dekodera (eng. *skip connections*) koje omogućuju istovremeno prikupljanje podatka o lokalizaciji i kontekstu. Dodatno, točnost segmentacije je poboljšana uporabom augmentacije podataka (eng. *data augmentation*) i obradom predikcija pomoću uvjetnih slučajnih polja. Time je ostvaren drugi znanstveni doprinos.

U uvodnom dijelu četvrtog poglavlja (*„4. Methods for retinal fluid segmentation in optical coherence tomography images“*) dan je pregled razvijenih metoda za segmentaciju tekućina u retinalnim slikama optičke koherentne tomografije. Iako je iz literature poznato da simultana segmentacija slojeva i tekućina daje bolje rezultate, napravljena je komparativna kvantitativna analiza segmentacije samo tekućina i segmentacije tekućina uz dostupne informacije o slojevima. Također, provedeno je istraživanje primjene prijenosnog učenja koje je vrlo aktivna i nedovoljno istražena tema na području dubokog učenja s primjenom u medicini.

Posljednjih godina segmentacija medicinskih, a time i oftalmoloških slika provodi se gotovo isključivo metodama dubokog učenja. Jedan od preduvjeta za uspjeh algoritama dubokog učenja je velika količina dostupnih podataka što je medicini rijetko slučaj. U slučaju

učenja pod nadzorom, potrebne su slike koje je stručnjak ručno označio što je vremenski vrlo zahtjevno. Štoviše, označavanje slika nije dio kliničke prakse već se označavanje provodi radi razvoja metoda strojnog učenja. Kako se slike prikupljene iz različitih uređaja optičke koherentne tomografije i različitih generacija uređaja razlikuju, nužno je da algoritmi segmentacije budu što je moguće robusniji. To znači da u slučaju migracije na novi uređaj nije potrebno ponovno označavati veći broj slika, već će se postojeći algoritmi nadograditi s manjim brojem novih slika. Na temelju provedenog istraživanja predložena je metoda za segmentaciju intraretinalnih tekućina u slikama optičke koherentne tomografije uz prijenosno učenje. Time je ostvaren treći znanstveni doprinos.

U petom poglavlju (*„5. Conclusion“*) navedeni su zaključci dobiveni na temelju provedenog doktorskog istraživanja te su dane potencijalne smjernice za nastavak istraživanja. Velike količine podataka i povećanje računalne snage dva su glavna stupa na kojima počiva napredak računalnog vida. Velike baze podataka bile su jedan od glavnih preduvjeta za kontinuirani razvoj novih i poboljšanih modela i na taj način omogućili napredak u računalnom vidu. Nažalost, u oftalmologiji nedostaje jedan od dva glavna stupa, odnosno suočeni smo i o ograničeni s malim skupovima podataka. To jest, barem dok se većina algoritama dubokog učenja oslanja na nadzirane ili polu-nadzirane metode. Jedan od načina zaobilaženje potrebe za velikom količinom podataka bio bi putem dodavanja oftalmološkog znanja budući da modeli automatske segmentacije često griješe u slučaju kada prepoznavanje uzoraka nije dovoljno već je potrebno i dodatno znanje. S obzirom na složenost oftalmološkog znanja, te da se ono teško može predstaviti simboličkom logikom i da ne postoji dobar način da se simbolička logika ugradi u modele dubokog učenja, za sada ne postoji adekvatno rješenje. Drugo rješenje koje se čini vjerojatnijim je generiranje sintetičkih podataka koji se mogu dobiti simulacijom retine i generiranjem slika simulacijom mehanizama bolesti i načina na koji deformiraju retinu, uz istovremeno dodavanje patoloških struktura. Često se veći dio treniranja neuronskih mreža obavlja sa simuliranim podacima ili u simuliranim okruženjima – najveći napredak u umjetnoj inteligenciji kojem smo do sada imali prilike svjedočiti rezultat je simuliranih podataka i treniranja neuronskih mreže u simuliranoj (virtualnoj) stvarnosti. Nažalost, taj dio nije bio izvediv u okviru ovog doktorskog rada, ali je plan za daljnje istraživanje.

Ključne riječi: senilna makularna degeneracija, optička koherentna tomografija, anotirane retinalne OCT slike, automatska segmentacija slika, duboko učenje

Contents

1 INTRODUCTION

Age-related macular degeneration (AMD) is an acquired degeneration of the retina that causes significant central visual impairment in its late-stage and is the leading cause of irreversible blindness in people 50 years of age or older in the developed world [1]. The estimated prevalence of any and late-stage AMD in people aged 45 to 85 years is 8.69% and 0.37%, respectively [2]. The prevalence of late-stage AMD sharply rises to 7.1% in those 75 years or older [3]. Due to the ageing of population, by 2040 an estimated 288 million people will be affected by AMD [2].

AMD is a progressive disease. The stages of AMD are early AMD, intermediate AMD, advanced AMD with geographic atrophy (GA), and advanced AMD with choroidal neovascularization (CNV) as it is shown in Fig. 1.1. The early and intermediate stages are usually asymptomatic, characterized by the accumulation of yellow granular deposits beneath the outermost layer of the retina, the retinal pigment epithelium (RPE) (Fig. 1.2 a). The only known therapy, for now, are antioxidant and mineral supplements that can eventually slow the progress. Advanced or late AMD is defined either by development of atrophy of the RPE and the overlying photoreceptors or by the development of new blood vessels (neovascular membranes) beneath or above the RPE (Fig. 1.2 b). These new vessels tend to leak or rupture, with subsequent exudation or haemorrhage accumulating in different retinal layers [4]. These two forms of advanced AMD, the former usually referred to as geographic atrophy (GA), and the latter as exudative, wet, or neovascular AMD (nAMD), can occur alone or together, either simultaneously or sequentially, and both forms can lead to significant visual impairment [5]. In the case of GA, there is no adequate therapy for now [6]. Intravitreal anti-vascular endothelial growth factor (anti-VEGF) therapy is the mainstay of nAMD treatment [7].

The application of optical technology in medicine and biology has a long history. Since the 18th century, biologists have been using a microscope. Fig. 1.3 shows a drawing from 1891 of the retinal structure by Santiago Ramón y Cajal, Spanish Nobel Prize Awardee and neuroscientist pioneer in the investigation of the microscopy structure of the nervous system including the retina. With the invention of the laser in the early 1960s, physicians were gaining a new powerful tool in surgery. Then follows an invention of optical fibers and an endoscope that allows direct examination of internal organs located deep in the body. Some of

the optical instruments today use the coherent property of light. One of the reasons why optical coherence tomography (OCT) has attracted the attention of scientists and engineers is its potential to become the first diagnostic imaging technology to profile coherent optical features [8].

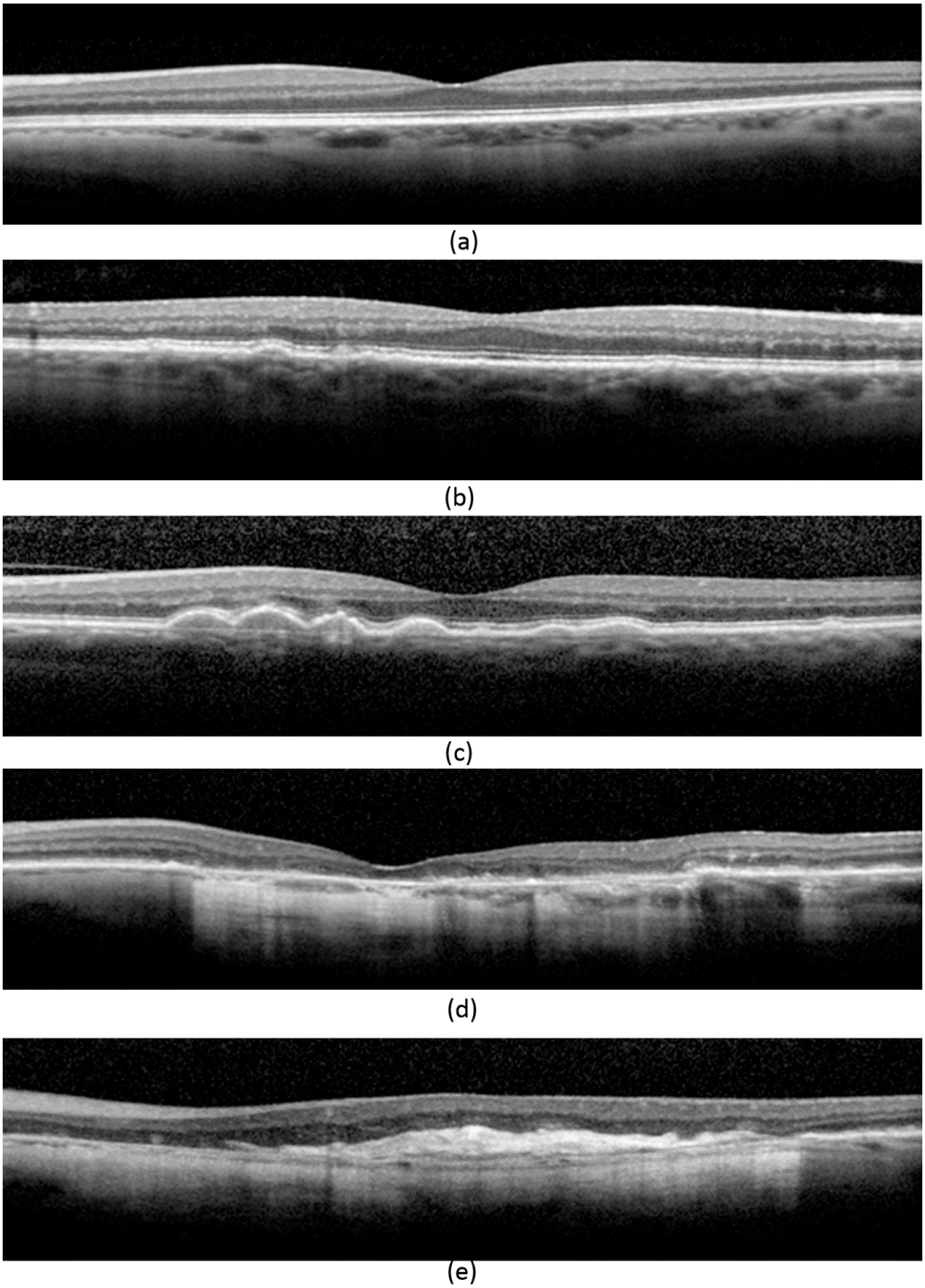

Fig. 1.1 Examples of B-scans showing the stages of AMD: a) no AMD, b) early AMD, c) intermediate AMD, d) advanced AMD with GA, and e) advanced AMD with choroidal neovascularization (CNV). [9]

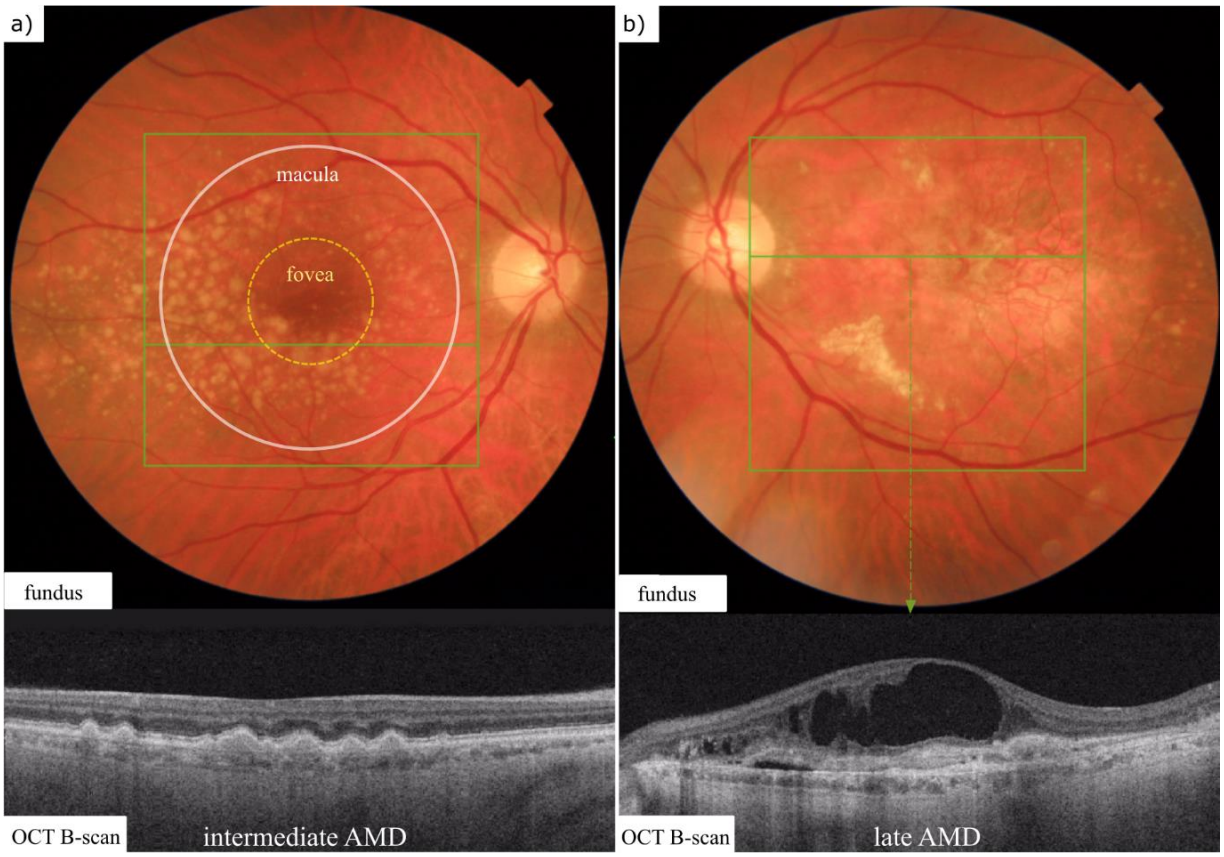

Fig. 1.2 a) Intermediate AMD. Fundus photography showing numerous yellow granular deposits (drusen) in the macular area. OCT showing the location of the drusen beneath the RPE. b) Late (neovascular) AMD. Fundus photography showing new subretinal vessels. [10]

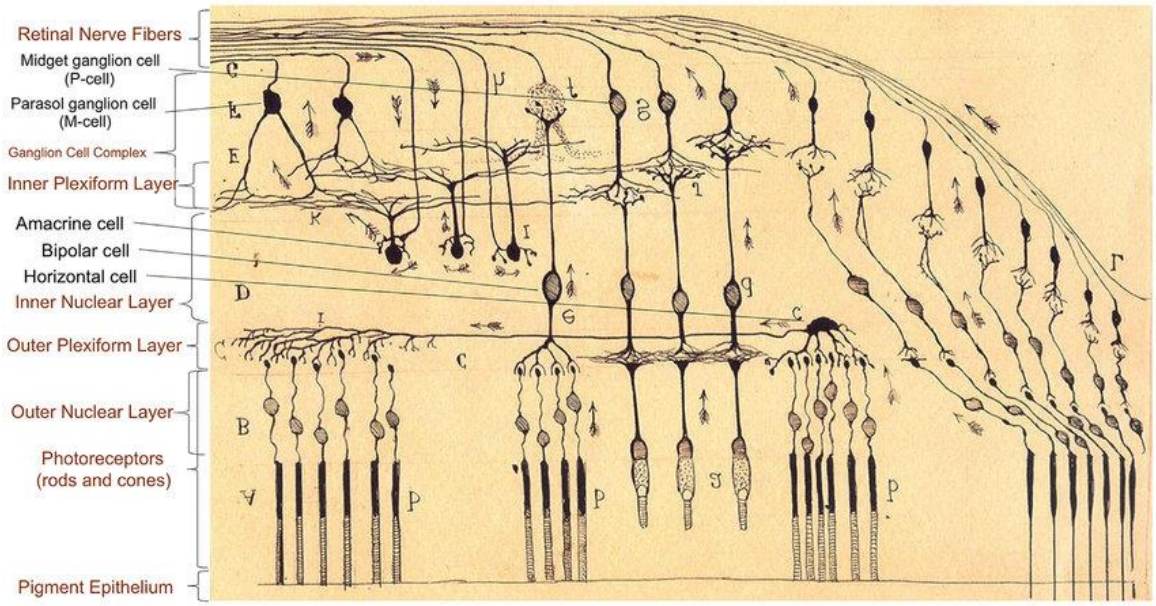

Fig. 1.3 Drawing of the retina by Santiago Ramón y Cajal. [11]

Optical coherence tomography was first presented in 1991 [12] as a method of imaging that allows non-invasive 3D imaging. Collected high-resolution images show the structural composition of the biological tissue and therefore provide the possibility of obtaining different information about the internal structure of objects and showing different aspects of biological tissue such as structural information, blood flow, elasticity parameters, changes in polarization, and molecular composition [12], [13]. As a consequence, OCT imaging found an important role in ophthalmology, where 3D high-resolution imaging enables clinical diagnosis and research of retinal diseases [14]. Morphological features that can be seen and measured from OCT tomograms, such as the thickness of individual retinal layers, shapes, spatial distributions, and optical properties of various lesions and blood vessels, can serve as identifiers in the diagnosis of retinal disease [15].

The principle of optical coherence tomography is to estimate the depth at which a specific feedback signal was generated by measuring its transit time. The feedback signals are mainly caused by the difference in the refractive index when passing from one tissue to another. Feedback from deeper tissues can be differentiated from feedback generated closer to the surface because light takes longer to reach the sensor. As the total retinal thickness is 300 to 500 μm, the differences in passing times are very small and can only be measured by interferometry [12]. Optical coherence tomography uses light from a diode that is in wavelength close to infrared, analogous to obtaining a 2D image using ultrasound. Although the sound has an advantage in terms of depth of penetration into tissue, light with much shorter wavelengths has an advantage in obtaining images of much higher resolution [16]. The use of light leads to an improvement in resolution from one to two orders of magnitude [17]. Typically, a beam splitter serves to split light into two rays: light reflected from retinal tissue that makes up the sample arm and light reflected from the reference mirror and called the reference arm. The energy of the interferogram between these two rays is converted into image intensity via a photosensor (Fig. 1.4). The depth of the scan (commonly referred to as the A-scan using ultrasound terminology) is displayed at different intensities resulting from backscattering at different depths. For 3D imaging, as with scanning laser ophthalmoscopy (SLO), the light beam is moved along the retina mainly using galvanic mirrors that change position in two directions (along the x and y-axis) resulting in an A-scan tomographic image for each x and y position [18].

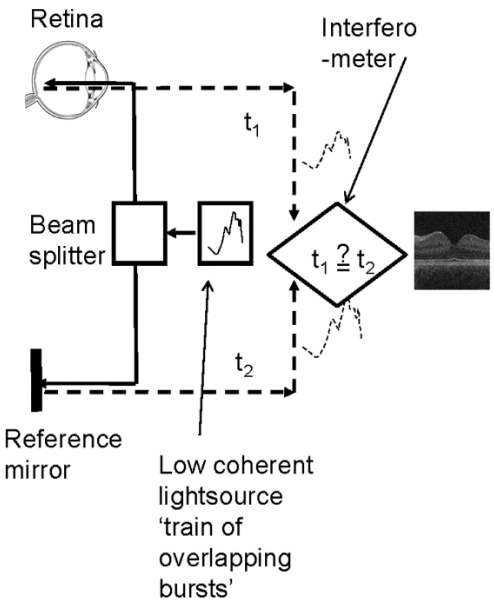

Fig. 1.4 Schematic diagram of the OCT system. [18]

Various approaches have been developed to maximize the number of A-scans obtained per unit time since the reduction of artifacts due to displacement, patient comfort, and achieving high resolution depends on short intervals of A-scans. Three main OCT approaches have been developed:

- Temporal OCT or time-domain OCT (TD-OCT) in which the reference mirror moves mechanically in different positions resulting in different time delays of the reference beam. Because the speed at which the mirror can be mechanically moved is limited, only about a thousand A-scans can be recorded per second. The envelope of the interferogram determines the intensity at all depths [12].
- Time encoded frequency domain OCT or swept-source OCT (SS-OCT) – instead of moving the reference beam, the light is rapidly modulated around the central wavelength. The photosensor is used to measure the correlogram for each wavelength center over time. Fourier transform of the spectral interferogram is performed to determine the depth of the tissue at a particular position [19].
- Spectral-domain OCT (SD-OCT) is similar to time-coded frequency OCT, but instead of modulating light wavelengths, a broadband light source (wider than temporal OCT) is used and the interferogram is spectrally parsed mainly using CMOS or CCD linear sensors. Fourier transform is also applied to the intensity spectral correlogram to

determine the depth of each individual signal [19]. The image quality of spectral OCT is much better than that of temporal OCT [20].

There are mainly two types of scans performed to examine patients' eyes: a scan centered around the macula (yellow spot) associated with central vision and a scan centered around the optic nerve head (ONH) associated with peripheral vision (Fig. 1.5) [20].

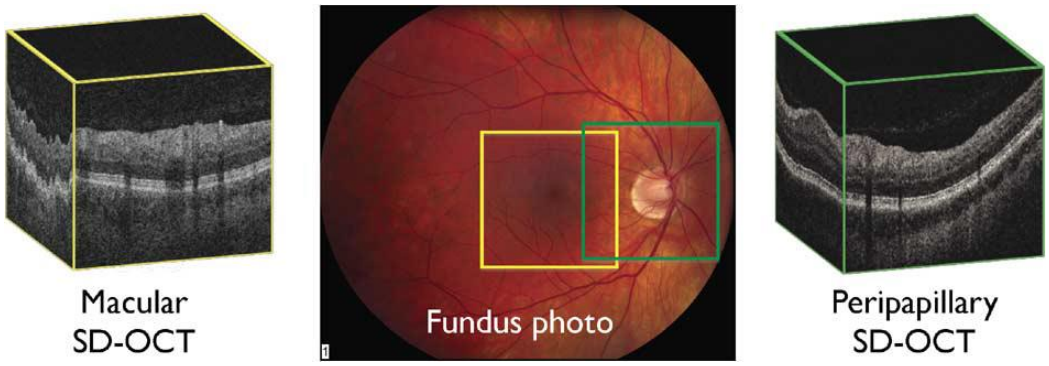

Fig. 1.5 Typical scan areas (shown in the fundus image) of the SD-OCT system. Left: Macular scan centered around the macula (marked in yellow). Right: Peripheral scan located around ONH (marked in green). [12]

Although the development of OCT technology has been an active field of research since 1991, it is only in the last decade that significant development and research of OCT image segmentation methods has begun. Segmentation is one of the most difficult and at the same time necessary steps in the analysis of OCT images. Anyhow, the importance of automatic segmentation has been recognized in the research community. Simultaneously with the progress of deep learning algorithms in computer vision, in ophthalmology, there are increasing requirements for automatic segmentation. We are faced with the growing amount of data (a single OCT volume usually consists of over 50 B-scans, depending on the type of OCT device), manual segmentation is time-consuming (and therefore infeasible in clinical practice), and at the same time segmentation of retinal structure and pathological biomarkers are necessary for more accurate diagnosis and more effective therapy. All the issues resulted in a large number of published papers for automatic segmentation in recent years. It is predominantly an outcome of introducing U-net architecture [21] in 2015 which accomplished good results in a small data regime.

Artificial intelligence and deep learning (DL) algorithms have had many breakthroughs in recent years partially due to large datasets like ImageNet [22]. Given that there are no similar datasets (in terms of diversity, scale, and open access) in medicine, transfer learning is

often used. However, it has not been widely adopted in ophthalmology [23]. It is still not clear enough how much transfer learning can help in the case where the source dataset consists of natural images (as the ImageNet dataset) and the target dataset consists of medical images (entirely different from natural images) [24]–[26], or how suitable the neural network architectures developed on ImageNet will be for medical images [27].

Nowadays, OCT imaging is increasingly present in ophthalmic clinical practice. OCT is invaluable for diagnosing AMD and guiding AMD treatment as it provides high-resolution, pseudohistological cross-sectional images of the retina and choroid. In nAMD, OCT is used to detect and visualize specific lesions such as intraretinal fluid (IRF), subretinal fluid (SRF), subretinal hyperreflective material (SRHM), and retinal pigment epithelial detachment (PED). These changes represent exudates (IRF, SRF, some cases of SRHM, some forms of PED), haemorrhage (some cases of SRHM, some forms of PED), or neovascular membranes and fibrosis (some cases of SRHM and some forms of PED) [28]. Previous studies have shown these lesions can serve as OCT biomarkers for visual function or therapy response [29]–[34]. In comparison to 2D analyses of central B-scans, volumetric analyses of these pathologic lesions in the entire macular area might more precisely predict these outcomes [33], [34]. Automatic segmentation of retinal layers and of the IRF, SRF, SRHM, and PED (from this point on altogether referred to as fluids) is crucial for detecting and characterization of AMD objectively and in a reproducible way. Also, a reliable automatic OCT system for segmentation is crucial for further development of diagnosing retinal disease. Due to the quantitative analysis of pathological changes, the therapy effectiveness could be predicted [35], [36]. The occurrence of intra- and subretinal fluids is an important biomarker that plays a major role in (re-) treatment decisions and is prognostic of visual rehabilitation [37]. Quantitative analysis also allows the prediction of the transition from the middle to the late phase and the prediction of whether in the case of one diseased eye, the disease will affect the other eye [38], [39]. Recent studies confirm the correlation between individual fluids and the success of anti-VEGF therapy, making computer segmentation and quantitative analysis even more important [40]–[42].

Currently, commercially available OCT software algorithms automatically perform retinal layer segmentation to a variable extent and derive basic parameters such as inner or outer retinal thickness, RPE elevation, or central subfield thickness. Although helpful and routinely used to inform clinical decisions, issues of susceptibility to segmentation errors and

limited inter-device reproducibility have been raised [43], [44]. However, the main issue of these algorithms is the lack of detailed information, as these basic thickness and elevation parameters do not distinguish specific underlying lesions (e.g., retinal fluids in different compartments).

For the purpose of developing methods for segmentation of pathological changes present in nAMD as well as basic structures, it was necessary to collect a database of images with manually annotated fluids and layers. In chapter 2 the creation of the publicly available Annotated Retinal OCT Images (AROI) database is described. It is given an overview and analysis of the current state of development of the AROI database, with plans for further improvements and new features. It currently consists of 1136 annotated B-scans (from 24 patients suffering from nAMD) and associated raw high-resolution images. The existing research in the field indicates that the approach in which layers and fluids are jointly segmented aims to take advantage of the interdependence of fluids and layers and thus achieves the best possible segmentation [45]–[50]. Therefore, annotations for pathological changes (fluid accumulation and related findings) and basic structures (layers) have been provided. Also, results for intra- and inter-observer errors are presented to enable the validation of developed algorithms in relation to human variability. In Chapter 3 different methods for retinal layers and joint retinal layers and fluid segmentation in OCT images are explored. It is divided into two major parts: model-centric approach and data-centric approach. In model-centric approach is explored how different architectures affect segmentation accuracy. To set a baseline for deep learning (DL) methods, results for automatic segmentation with standard U-net architecture [21] and two state-of-the-art architectures (U-net-like [51] and U-net++ [52]) in medical image segmentation are presented. The results indicate that there are major challenges for automatic segmentation in the case of severe pathologies. Also, more advanced architectures do not perform better, or at least not much better – major challenges could not be resolved solely with enhanced U-net architectures. To tackle problems like poor image quality, huge class imbalance, a small number of images that are complex and difficult even for experts, in the second part is investigated how data manipulation affects segmentation accuracy. It is shown how data augmentation, and post-processing can boost performance, and which are limits. Based on all the research conducted, a comparison of methods is made and those that give the best results are proposed as models for joint retinal layers and fluids segmentation. In Chapter 4 methods

for retinal fluid segmentation in OCT images are explored. It is investigated how information about fluid localization increase segmentation performance. Also, transfer learning in case of intraretinal fluid segmentation is explored. A method for retinal fluid segmentation in OCT images based on transfer learning is proposed. In the final chapter, interpretability is discussed as there are rising concerns about applying AI models in medicine and whether it is ethical. A brief discussion is given both from a technical and clinical perspective. In conclusion, is given a review of the field, which are major obstacles to introducing AI in clinical practice, and ideas on how they could be overcome.

The main contributions of this thesis are as follows:

- Framework for validation of OCT image segmentation methods consisting of a publicly available database of manually labeled image features and methodology and programming code for standardized measurement of segmentation accuracy
- Method for retinal layers segmentation in optical coherence tomography images
- Method for retinal fluid segmentation in optical coherence tomography images

2 FRAMEWORK FOR VALIDATION OF RETINAL IMAGE SEGMENTATION METHODS

2.1 Overview of existing databases

Despite great progress in computer vision and medical image segmentation, a major shortcoming is the lack of publicly available databases of annotated images. In medical image segmentation supervised or semi-supervised methods are still predominant, and their accuracy depends on the quality and scale of annotated data. Most of the published methods are developed on the datasets of images in which no significant pathological changes are present, so it is questionable to what extent the methods are applicable in the case of severe disease. Also, it is difficult to evaluate and compare different methods as there is a lack of publicly available databases of manually labelled images, and results depend significantly on the number of processed images, quality of images (related to the type of OCT device), present disease, etc.

The lack of publicly available databases hampers the validation of algorithms and leads to reproducibility/replication and therefore reliability crises. A recent study was done by Wintergerst et al. (2020) [53] attempted to replicate Chen et al. algorithm (2013) [54] (well-known algorithm for drusen segmentation in OCT images) according to the report in the original paper (as code is not open source). After replication on a different dataset, inferior results were obtained. Even with improvements of the algorithm, originally reported results could not be achieved. To my knowledge, it is the first such replication study that raises concerns about reliability/reproducibility. Another problem with algorithms trained on small datasets with a lack of diversity is their considerable bias and no possibility of generalization. The reliability of published results is paramount for applying automatic segmentation and AI into clinical practice. Also, the algorithms applied in clinical practice should generalize well, and that could be achieved by training on large and diverse databases (covering all eye diseases as well as healthy persons; with variability in demographic characteristics; different types of OCT devices) – and this can be managed by a public repository with databases collected from various research groups. As labelling is time-consuming and requires expertise, this can be achieved quicker by pooling efforts across research groups.

The common practice of algorithm validation is through community challenges due to standardized benchmarks (same dataset and same validation metrics). In 2017 “The Retinal OCT fluid Detection and Segmentation Benchmark and Challenge” (RETOUCH) [55] was organized. The benchmarking resource database consisted of OCT volumes from 112 patients suffering from age-related macular degeneration (AMD) and retinal vein occlusion (RVO). OCT volumes were acquired from three types of OCT devices: Topcon, Spectralis, and Cirrus. Three types of retinal fluids were annotated: subretinal fluid (SRF), pigment epithelial detachment (PED), and intraretinal fluid (IRF). [56] To my knowledge, it has been the largest publicly available database up to date. Considering the same dataset and the same metric for the evaluation of segmentation and detection, this was a notable improvement in the evaluation of different methods. All applied methods (eight teams were participating) for automatic segmentation and detection of fluids were based on deep learning methods (segmentation was usually performed with the popular U-net architecture [21]) in combination with other machine learning (ML) methods and image analysis methods. Details can be found on the challenge website [55] and in the accompanying paper [56]. Unfortunately, only the training set is still publicly available. As the challenge is finished, the test set is no more available, and therefore no further comparison of algorithms is possible.

Besides the RETOUCH challenge database, the only publicly available dataset, which some authors use for evaluation of their methods, is the DUKE dataset [57] containing 110 B-scans of 10 DME (diabetic macular edema) patients acquired with the Spectralis OCT.

One of the common problems is that most of the publicly available databases have never become popular due to lack of visibility, accessibility, transparency, absence of data descriptions, etc. In a recent review paper, Khan et al. (2020) [58] give an extensive overview of publicly available databases in ophthalmology (up to May 2020). According to their study, there are 94 open-access datasets. AMD is present in 16% of datasets, 19% of all datasets include images from OCT or OCT angiography, and 35% of datasets include manually annotated segmentations. It leads the authors to raise serious concerns about “data poverty” and argue about challenges in data collection.

2.2 AROI database

2.2.1 Data collection

For the purpose of the study, the database of manually **A**nnotated **R**etinal **O**CT **I**mages (AROI database) was collected. In collaboration with Sestre milosrdnice University Hospital Center (Zagreb, Croatia), images were collected and annotated by an ophthalmologist. Selection criteria included patients aged 60 years and older diagnosed with nAMD, irrespective of their anti-vascular endothelial growth factor (anti-VEGF) therapy status, with no significant media opacities precluding adequate retinal imaging, and with no other retinal disorders. The concurrent presence of geographic atrophy (GA) of any extent was not an exclusion criterion, since nAMD and GA occur simultaneously or sequentially in a significant number of patients with advanced AMD. From April 2018 to June 2018, 24 consecutive patients were included in the study. Macular SD-OCT volumes were recorded with the Zeiss Cirrus HD OCT 4000 device. Image quality was checked by an ophthalmologist. The overall signal strength of 6/10 or more and the absence of any focal shadow artifacts or out of register artifacts was a prerequisite for further analysis. Each OCT volume consisted of 128 B-scans, spaced 47.24 μm apart, with a resolution of 1024 x 512 pixels (pixel size 1.96 μm x 11.74 μm) (Fig. 2.1).

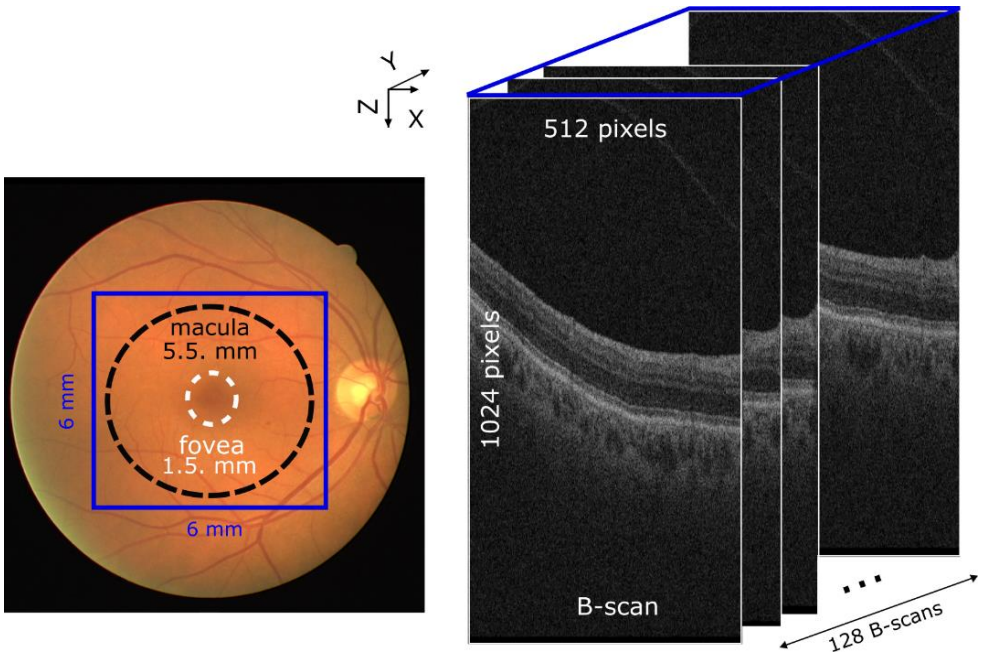

Fig. 2.1 Fundus image and macular OCT volume. [59]

All annotations were done by an ophthalmologist in GNU Image Manipulation Program (GIMP) [60]. The database consists of 1136 annotated B-scans out of a total of 3072 B-scans for 24 patients (37% of B-scans were annotated). Annotations were not provided for each B-scan if adjacent scans were similar. In the fovea – a central part of macula (Fig. 2.1), a larger number of B-scans were annotated as visual acuity mostly depends on the pathological changes in that area, while more eccentric B-scans were annotated at the ophthalmologist's discretion: in case the adjacent B-scans were deemed similar, the annotations were skipped (performed for every 2nd to 10th scan, depending on the extent and complexity of pathological changes). Fig. 2.2 shows the number of annotated B-scans for each individual patient. The average number of annotated B-scans per patient is 47.3 (with a standard deviation of 25.7). The distribution of the number of annotated B-scans along the vertical (Y) axis is as well shown in Fig. 2.2. It is seen that the major number of annotated scans is around the foveal center (ordinal number 62).

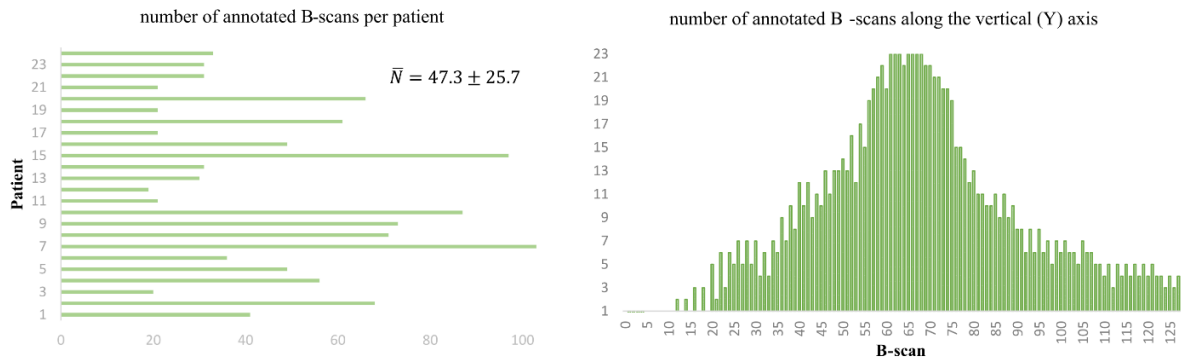

Fig. 2.2 The number of annotated B-scans for each individual patient (left) and the distribution of the number of annotated B-scans along the vertical (Y) axis (right). [59]

Images from OCT devices could be exported in IMG format, and in the open-access database, they are available in more common and accessible PNG format. Also, images are grouped so that for each patient there are all raw B-scans, with annotated ones provided separately. The name of each image is associated with the patient ordinal number (1-24) and B-scan ordinal number (0-127) (e.g., patient12_123.png) as is shown in Fig. 2.3. B-scans which are not annotated are also available as adjacent scans provide contextual information that can be used in the case of developing a 3D segmentation method. At this stage, the images are not divided into training, validation, and test set – division could be optimized depending on the developing method. However, splitting B-scans from the same patient over training, validation, and test set is not recommended as adjacent B-scans are similar, and

therefore in the test set, there will be “already seen” B-scans in the training set (causing overestimated evaluation of the method).

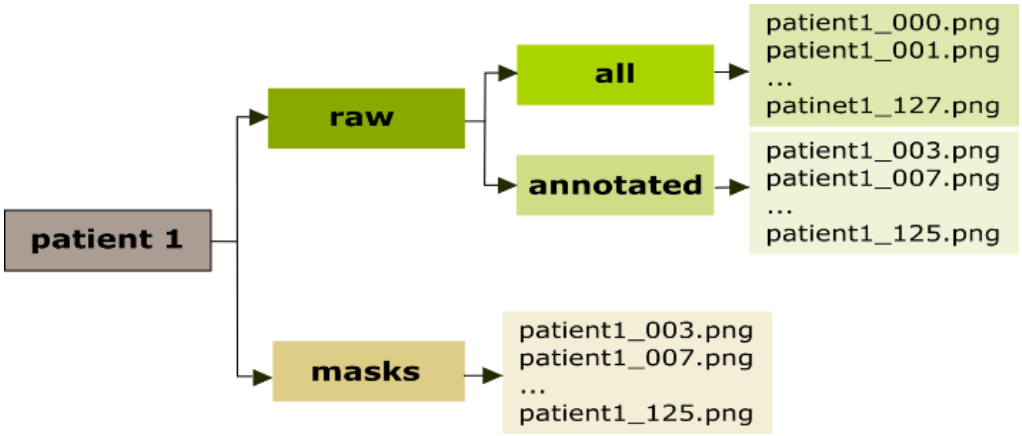

Fig. 2.3 Organization of images in the database. [59]

The database has its limitation: images were collected from only one type of OCT device and from patients suffering from one type of disease. However, developing algorithms on as large and as diverse databases as possible would provide for more robust algorithms that could be implemented in commercial OCT device software. Hopefully, open access will become a common practice for the majority of research groups in the future and online image collections and repositories will contribute to building a single database covering various diseases, various types of devices, and various retinal structures annotated from different experts.

Training ML algorithms on databases that are not diverse enough would lead to considerable bias which is unacceptable in clinical practice. The AROI database is limited in scale and diversity, but the goal was not to build a broad and comprehensive database as that is not a feasible task for a single research group. In agreement with an expert ophthalmologist, we have limited ourselves to a narrow segment of the population, one type of disease, and one type of OCT device, but at the same time, significant efforts to secure the quality of annotations have been made.

The main challenge to the introduction of automatic segmentation and, more generally, AI into clinical practice is not in developing more accurate architectures of neural networks but in overcoming the lack of well-annotated open access data.

The AROI database is available at: https://ipg.fer.hr/ipg/resources/oct_image_database [61] and at Open Science Framework (OSF): https://doi.org/10.17605/OSF.IO/5WYR3 [62]

Data collection adhered to the tenets of the Declaration of Helsinki and the standards of Good Scientific Practice of Sestre milosrdnice University Hospital Center (Zagreb, Croatia). All patients signed informed consent, the images are anonymized and do not contain any additional information about patients. The presented study was approved by the Ethics Committee of the Sestre milosrdnice University Hospital Center (EP-3272/18-11) and the Faculty of Electrical Engineering and Computing (Zagreb, Croatia).

2.2.2 Annotations of retinal layers and fluids

All annotations were done by one expert ophthalmologist. The lesions of interest and annotated in this study were the intraretinal fluid (IRF), subretinal fluid (SRF), subretinal hyperreflective material (SRHM), and retinal pigment epithelial detachment (PED) (Fig. 2.4). Although some of these lesions can reflect processes such as fibrotic scars or fibrovascular membranes, as they most often reflect exudation or haemorrhage, the term “fluids" was liberally used here to refer to all these lesions. Additionally, as SRF and SRHM share the same location characteristics, and often reflect the same exudative process, the distinction between SRF and SRHM was not performed in this study, and they were annotated jointly as SRF.

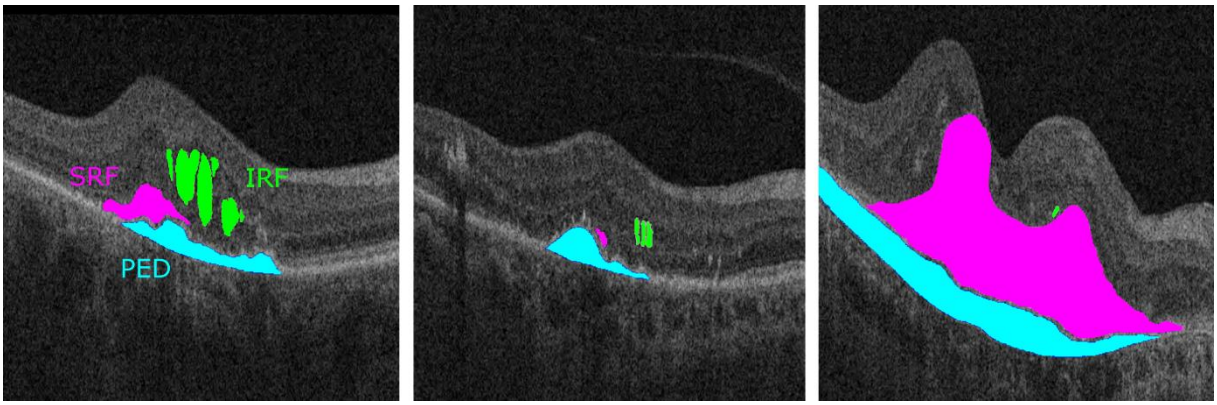

Fig. 2.4 Examples of annotated fluids: pigment epithelial detachment (PED), subretinal fluid (SRF), and intraretinal fluid (IRF). Images are cropped and only the ROI is visible. [59]

To detect IRF, SRF, and PED, the knowledge of their location within or outside specific retinal layers can be used to facilitate their detection and differentiation. The retina is

histologically divided into ten layers (Fig. 2.5): 1. internal limiting membrane (ILM), 2. retinal nerve fiber layer (RNFL), 3. ganglion cell layer (GCL), 4. inner plexiform layer (IPL), 5. inner nuclear layer (INL), 6. outer plexiform layer (OPL), 7. outer nuclear layer (ONL), 8. external or outer limiting membrane (ELM or OLM), 9. photoreceptor layer, and 10. retinal pigment epithelium/Bruch's membrane complex (RPE/BM). The layers visible in the OCT scan have been correlated to these histological layers (Fig. 2.6), with the exception of a few additional zones observed in the photoreceptor layer for which the exact histological counterpart is not yet defined [63].

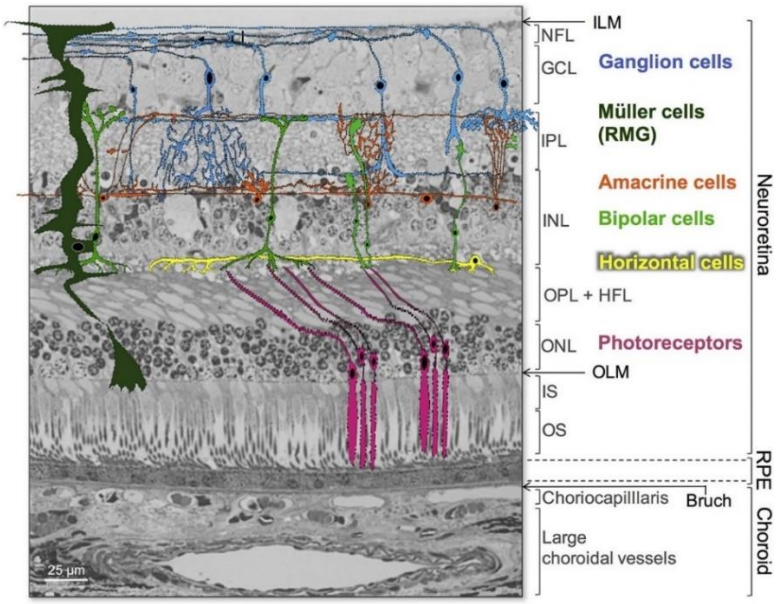

Fig. 2.5 Histological structure of the retina. [64]

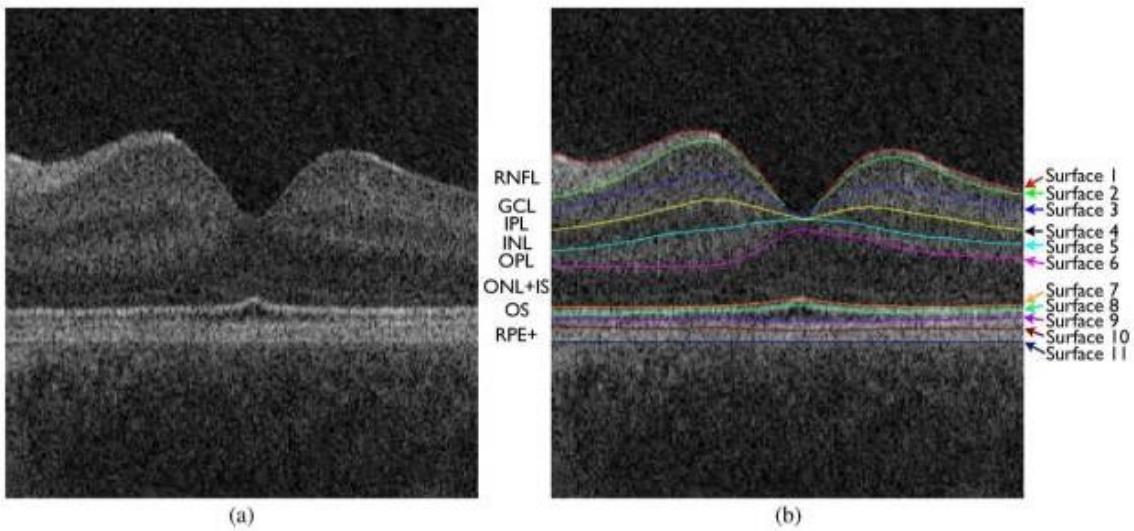

Fig. 2.6 Ten layers (eleven boundaries) are visible in OCT B-scan. [65]

The IRF is a localized extracellular intraretinal fluid accumulation seen as a hyporeflective area located anywhere between ILM and ELM, the SRF is a subretinal fluid accumulation seen as a hyporeflective area between the photoreceptor layer and RPE (the SRHM is seen in the same location but as a homogenously or inhomogeneously hyperreflective area), and the PED is either a hyporeflective accumulation of fluid or hyperreflective accumulation of other material between RPE and BM (the RPE/BM complex is separated in the case of PED) [28]. However, as pathological processes such as those in AMD can lead to extensive changes in retinal structure (Fig. 2.7), it is often impossible for an ophthalmologist reading the OCT scan to determine all 11 boundaries or 10 layers reliably. In agreement with the expert ophthalmologist were selected boundaries that can be readily determined in virtually all images and that are also relevant for localizing the observed fluids: ILM, the inner boundary of RPE, BM, and the boundary between IPL and INL (IPL/INL) (Fig. 2.8). While ILM, the inner boundary of RPE, and BM were chosen as pertinent for IRF, SRF, and PED localization, the IPL/INL boundary was chosen as it could be used to locate the foveal center in case of an eccentric scan (in patients with poor fixation) or loss of normal foveal depression (in patients with extensive foveal edema, elevation, traction, or parafoveal atrophy). As the foveal center consists only of ILM, ONL, photoreceptor layer, and RPE/BM, while other layers taper towards the foveal center, the center can be defined as the point of least distance between the ILM and any of these other layers, including the IPL/INL boundary.

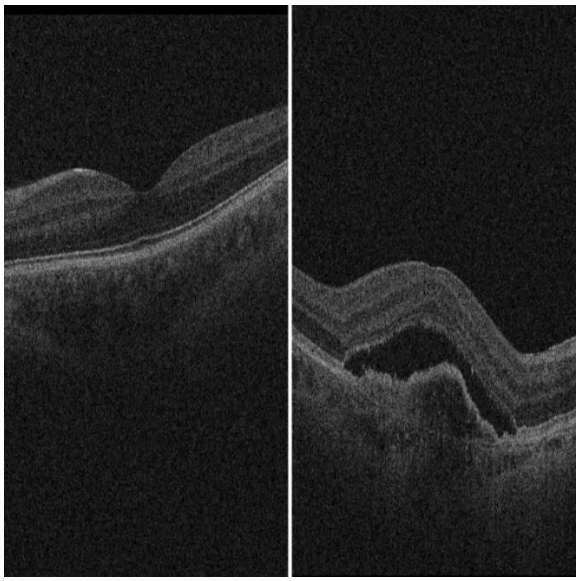

Fig. 2.7 Example of a B-scan (in the fovea, the central part of the macula) for the healthy person (left) and patient suffering from nAMD (right). [59]

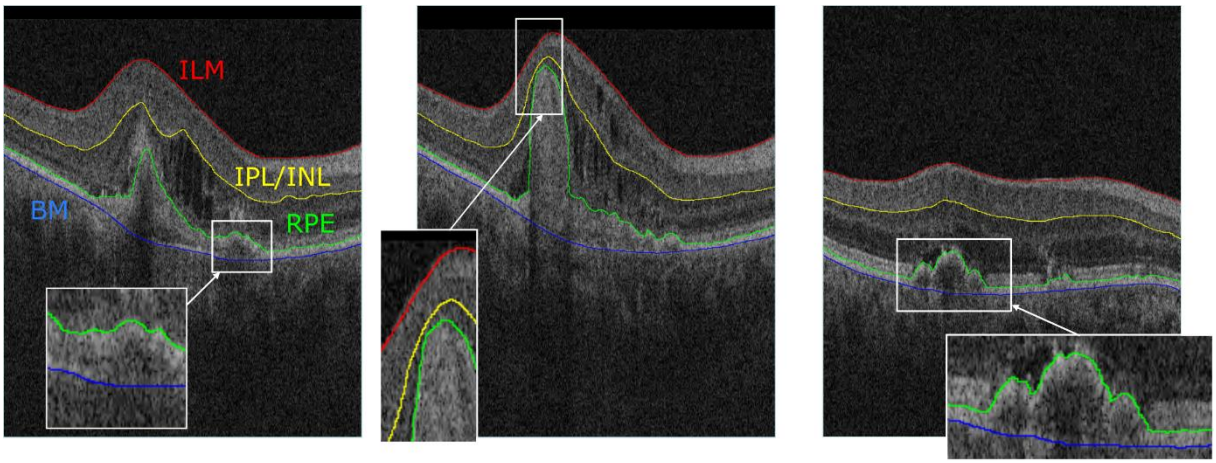

Fig. 2.8 Examples of annotated boundaries between layers: internal limiting membrane – ILM, the boundary between the inner plexiform layer and inner nuclear layer (IPL/INL), retinal pigment epithelium (RPE), and Bruch's membrane (BM). Images are cropped and only the ROI is visible. [59]

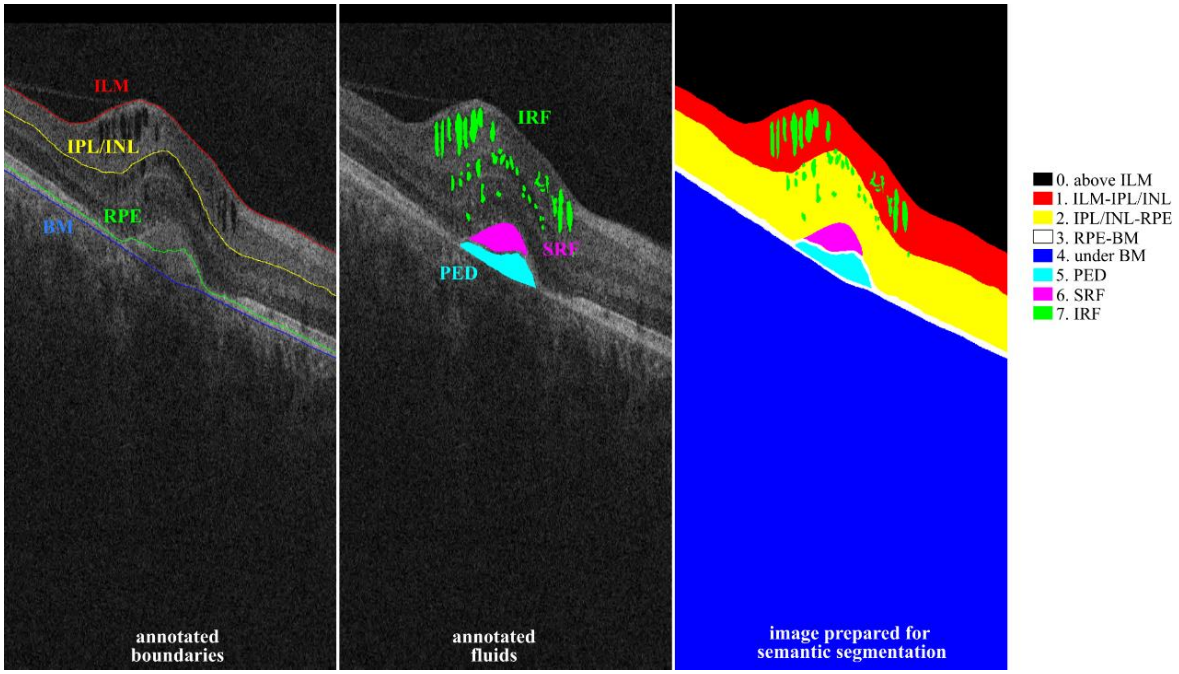

Fig. 2.9 From left to right: example of the image with annotated boundaries, an image with annotated fluids, and an image prepared for semantic segmentation (with eight classes). [10]

Joint segmentation of layers and fluids should lead to better accuracy due to the dependence of layers and fluids. To determine only the annotated boundaries, it is more appropriate to make semantic segmentation first and subsequently to determine the boundaries. For that reason, images are prepared for semantic segmentation (Fig. 2.9) with eight classes: area above the retina (vitreous), three retinal surfaces (the term "surface" is used to label the area between annotated boundaries and to distinguish them from 10 retinal layers

since they do not match; surface usually covers more layers), area below the retina (choroid) and three retinal fluids (PED, SRF, IRF).

A significant class imbalance is one of the reasons that make automatic segmentation challenging and should be considered when developing the algorithm for automatic segmentation. Shares for each class in the total number of pixels are shown in Table 2.1. The background (area above ILM and under BM) occupies as much as 83.26% while the IRF occupies only 0.12% in the total number of pixels. The higher segmentation errors are expected for classes that are less represented. Out of a total of 1136 scans, PED, SRF, and IRF are present in 1014 (89.26%), 648 (57.04%), and 229 (20.16%) B-scans, respectively.

Table 2.1 Percentage of the number of pixels for each class in the total number of pixels.

Above ILM	ILM - IPL/INL	IPL/INL - RPE	RPE - BM	Under BM	PED	SRF	IRF
35.07%	5.10%	7.89%	1.07%	48.19%	1.50%	1.05%	0.12%

2.2.3 Intra- and inter-observer error

Considering the lack of standardization in measuring inter- and intra-observer variability in ophthalmology there are a lot of rising questions about how to properly measure inter- and intra-observer variability (how to choose the size of the dataset, should images be chosen randomly, or to best present the whole dataset, how many observers to choose, etc.) According to Popović et al. (2017) [66], it would be better to choose three annotators. In my opinion, it would be preferable if they are not from the same department and the same hospital (the greater probability of the introducing same bias), but unfortunately, it was not manageable in this study.

For the purpose of calculating inter- and intra-observer error, annotations were additionally made for 75 B-scans (randomly chosen from the existing dataset). To calculate intra-observer error, the same expert who made all annotations (1st expert) made a re-annotation with no reference to previous annotations, and with a time delay (3 months after finishing the first annotations) for the 75 B-scans. To calculate inter-observer error, annotations were made by another expert (2nd expert) for the same 75 B-scans. Fig. 2.10 shows two examples: an example of good and bad matching. Differences of opinion among

experts can be observed. This can partly be explained by the poor quality of images (a large amount of speckle-noise), and by the fact that in the face of extensive disturbance of normal retinal structure, certain OCT findings cannot be reliably discerned or localized without normal anatomic landmarks. Also, with suboptimal image quality, the annotated changes could be confounded with other pathological phenomena (hyperreflective foci, pseudodrusen, outer retinal tubulations, etc.).

To determine the intra-observer error, the differences between "ground truth" (original annotations from 1st ophthalmologist) and re-annotations from the same ophthalmologist were calculated. To determine the inter-observer error, the differences between "ground truth" and annotations from 2nd ophthalmologist were calculated. The Dice score was used to evaluate results as it is a similarity measure often used as a metric in the segmentation of medical images. The Dice score ranges from 0 to 1, where 1 indicates perfect matching. It is calculated according to Eq. (1) where TP is true positive, FP false positive, and FN false negative.

$$DSC = \frac{2TP}{2TP+FP+FN} \qquad (1)$$

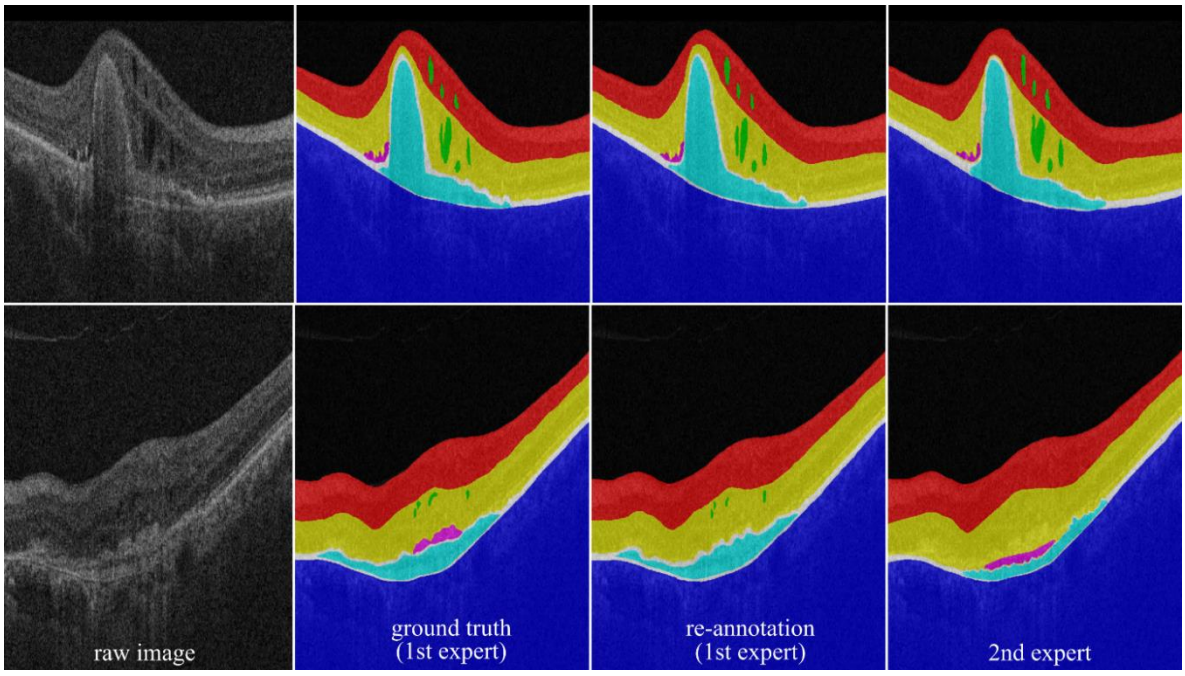

Fig. 2.10 From left to right: raw image, ground truth (annotations from 1st expert), re-annotations from 1st expert with time delay, annotations from 2nd expert. In the first row, there is an example with good matching. In the second row, there is an example with some differences even between annotations from the same expert (in annotations of PED and SRF) and between two experts (different annotations for fluids). Images are cropped and only the ROI is visible.

In Table 2.2 there are reported Dice scores (mean and standard deviation) for each class and for the inter- and intra-observer error. The values are the lowest for IRF and the area between RPE and BM which are represented by only 0.12% and 1.07% pixels, respectively (Table 2.1). One of the factors that contribute to the complexities of automatic segmentation is already mentioned significant class imbalance.

Table 2.2 The Dice score (mean and standard deviation) in the inter-observer and the intra-observer case.

	Above ILM	ILM - IPL/INL	IPL/INL - RPE	RPE - BM	Under BM	PED	SRF	IRF
Inter-observer	0.982 (0.072)	0.950 (0.111)	0.948 (0.112)	0.699 (0.129)	0.989 (0.114)	0.860 (0.301)	0.876 (0.366)	0.735 (0.280)
Intra-observer	0.998 (0.003)	0.973 (0.008)	0.970 (0.117)	0.778 (0.092)	0.998 (0.001)	0.912 (0.242)	0.924 (0.331)	0.844 (0.140)

As the Dice score is not an appropriate metric in case of a high-class imbalance (the Dice score for regions above the ILM and below BM will always be close to one), results for evaluation in case of converting the layer-segmentation task into a boundary detection problem are also provided. Variations in the assessment of retinal boundaries are presented with mean square error (MSE) and mean absolute error (MAE) with belonging standard deviations (Table 2.3). Values are shown in pixels where each pixel corresponds to 1.96 μm along the axial (Z) axis.

Table 2.3 Intra- and inter-observer error in boundaries assessment.

	Inter-observer error		**Intra-observer error**	
	MSE	MAE	MSE	MAE
ILM	5.87 (3.68)	1.86 (0.64)	2.10 (1.43)	1.02 (0.36)
IPL/INL	20.46 (47.71)	2.65 (1.19)	5.93 (4.80)	1.70 (0.59)
RPE	51.74 (148.15)	2.96 (2.30)	8.28 (9.33)	1.80 (0.79)
BM	12.77 (17.79)	2.54 (1.59)	5.17 (6.68)	1.53 (0.78)

The intra- and inter-observer errors were calculated to enable the validation of algorithms for automatic segmentation. However, it is still not clear what level of segmentation accuracy we need in clinical practice as manual segmentation is rarely performed in clinical practice. There is probably no universal rule, and the required accuracy

of segmentation will depend on the purpose – whether it is a diagnosis or prediction of the outcome of anti-VEGF therapy or prediction of another eye disease.

3 METHODS FOR RETINAL LAYERS AND JOINT RETINAL LAYER AND FLUID SEGMENTATION IN OCT IMAGES

3.1 Related prior work

So far, a number of methods for retinal layer segmentation have been developed. Initially, 2D segmentation was usually performed on each B-scan, and 3D segmentation was performed by carrying out 2D segmentation on each B-scan, and thus volume segmentation was obtained without using features given by spatial connections. One of the reasons for performing 3D segmentation in this way was that first generations of OCT devices had worked in the time domain and not the frequency domain – B-scans were spatially too distant to observe them as a volume.

3.1.1 Overview of methods for retinal layers segmentation

Methods for retinal layers segmentation can be divided into methods based on pixel values, methods based on active contours [67], methods based on graph theory [68], methods based on graph theory and dynamic programming [69], and methods based on diffusion maps [70]. Initial works that describe 3D segmentation are based on graph search [68], [71].

It is worth mentioning that Iowa Institute for Biomedical Imaging have developed an OCT explorer [72] – out of the box tool for OCT volume segmentation. To my knowledge, it is the only such software developed so far. Fig. 3.1 shows how the interface looks like. All eleven boundaries can be determined. Fig. 3.2 and Fig. 3.3 show two examples with raw B-scan, 'ground truth' from the AROI database, and segmentation performed with OCT Explorer. Only the boundaries labeled in the AROI database are shown (ILM, IPL/INL, RPE, BM) on segmentation done by OCT explorer. In another example where a larger retinal deformity is present (PED is between RPE and BM, and BM is not visible due to geographic atrophy) OCT Explorer is not able to determine BM even close to accurate (in all similar cases BM monitors RPE with some small distances). Segmentation is based on graph theory [65], [73].

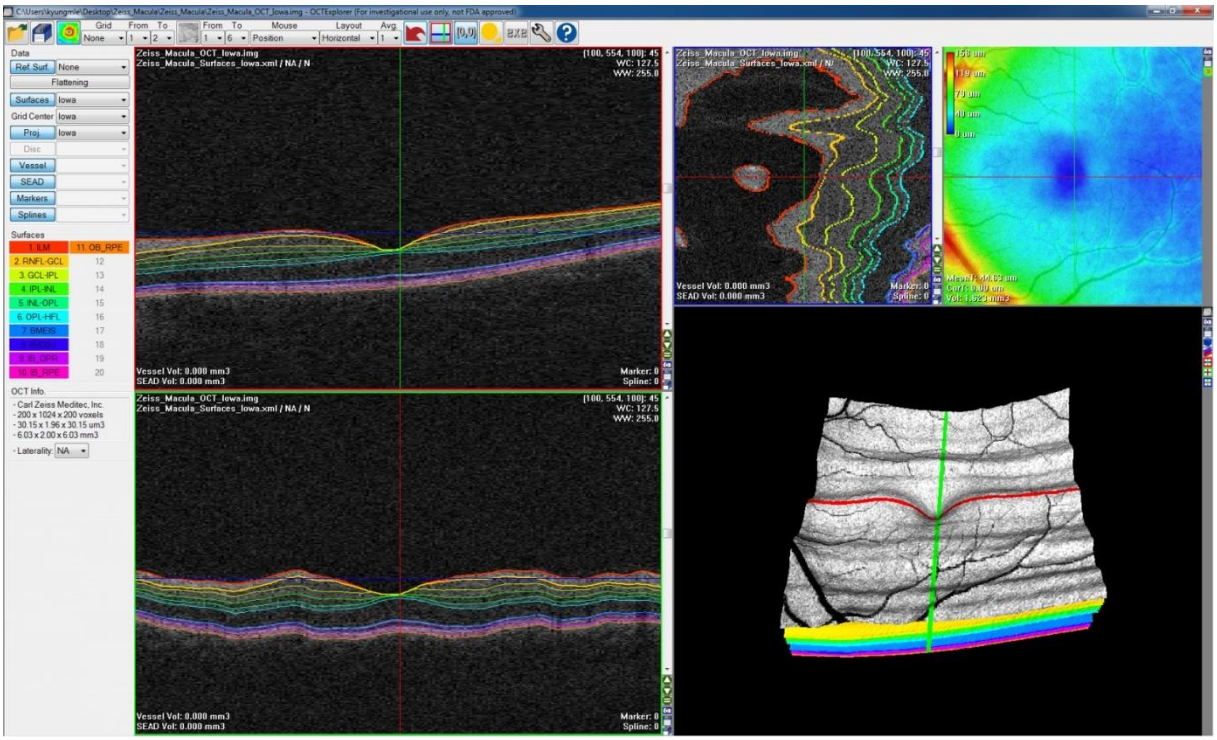

Fig. 3.1 OCT Explorer interface.

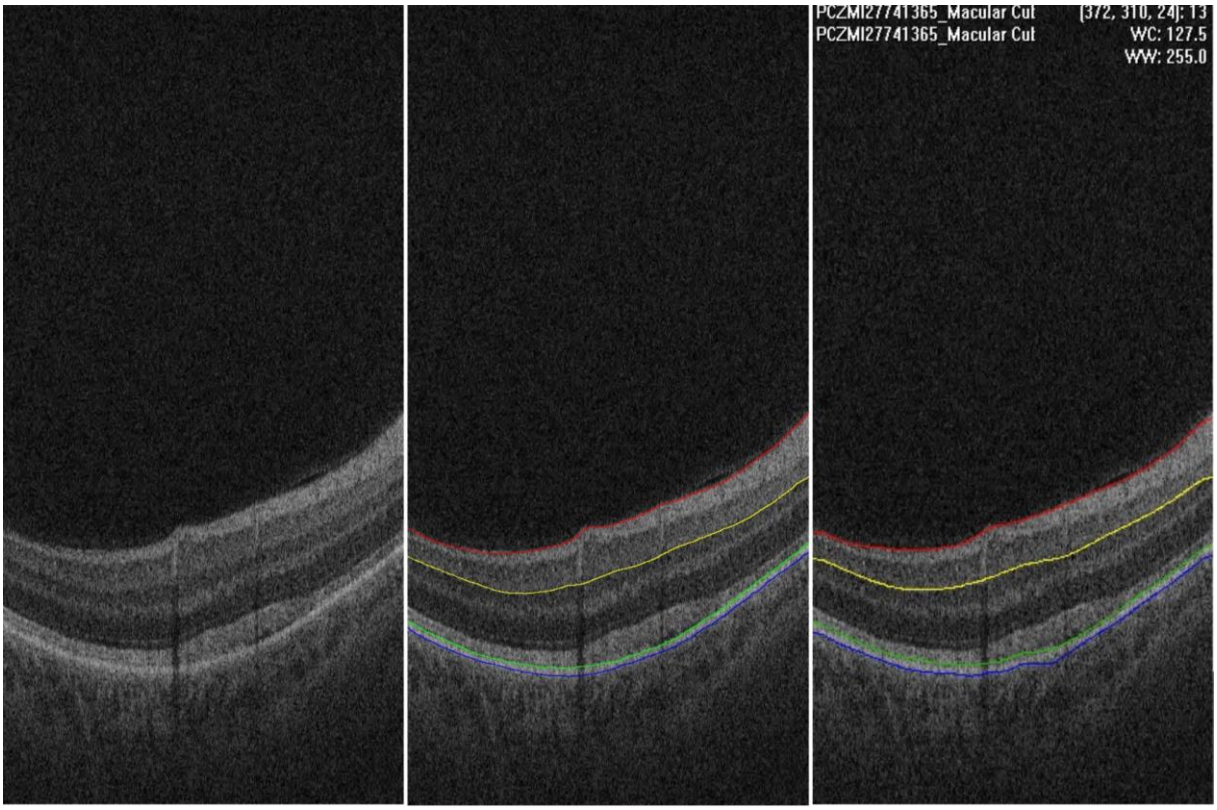

Fig. 3.2 Example of B-scan without much deformation. From left to right: raw image, reference segmentation from the AROI database, segmentation performed with OCT Explorer.

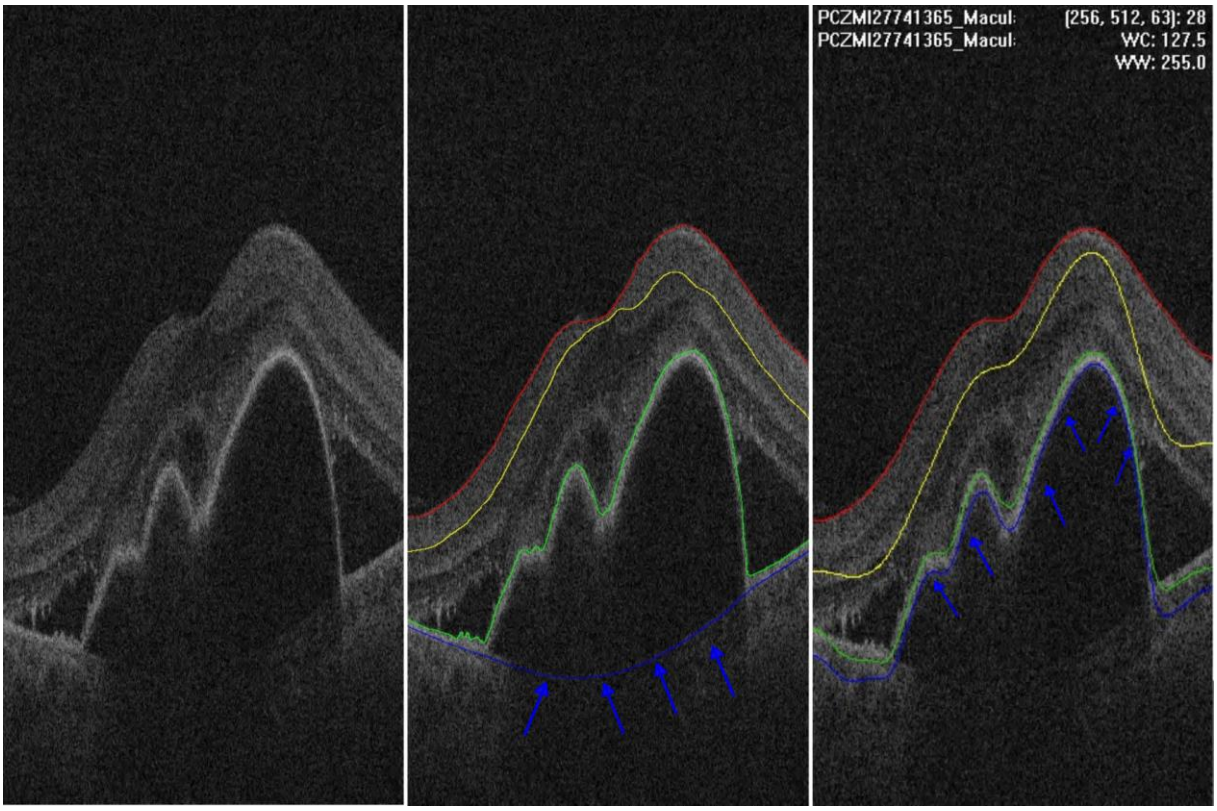

Fig. 3.3 Example with major pathological changes present. From left to right: raw image, reference segmentation from the AROI database, segmentation performed with OCT Explorer

More recently, retinal layers segmentation is performed by machine learning (ML), mostly using deep learning (DL) [74]–[79]. It is predominantly an outcome of introducing U-net architecture [21] in 2015 which accomplished good results in a small data regime. The U-net architecture consists of an encoder (contractive path), decoder (expanding path), and skip connections which enables simultaneous capturing of context and localization. It will be described in more detail in chapter 3.2.1.1.

One of the earlier and more influential paper is by Apostolopoulos et al. (2017) [74]. Authors inspired by U-net and ResNet [80] introduce BRU-net (Branch Residual U-Network) shown in Fig. 3.4. Original U-net architecture is modified with building blocks based on dilated convolutions to increase the effective receptive field. Blocks in encoder and decoder differ in max-pooling operation (BlockD) and up-sampling operation (BlockU). Also, the original U-net architecture is enhanced with residual connections (similar to ResUnet) and with batch normalization [81]. For training were used Adam optimizer [82] and mean square error (MSE) loss.

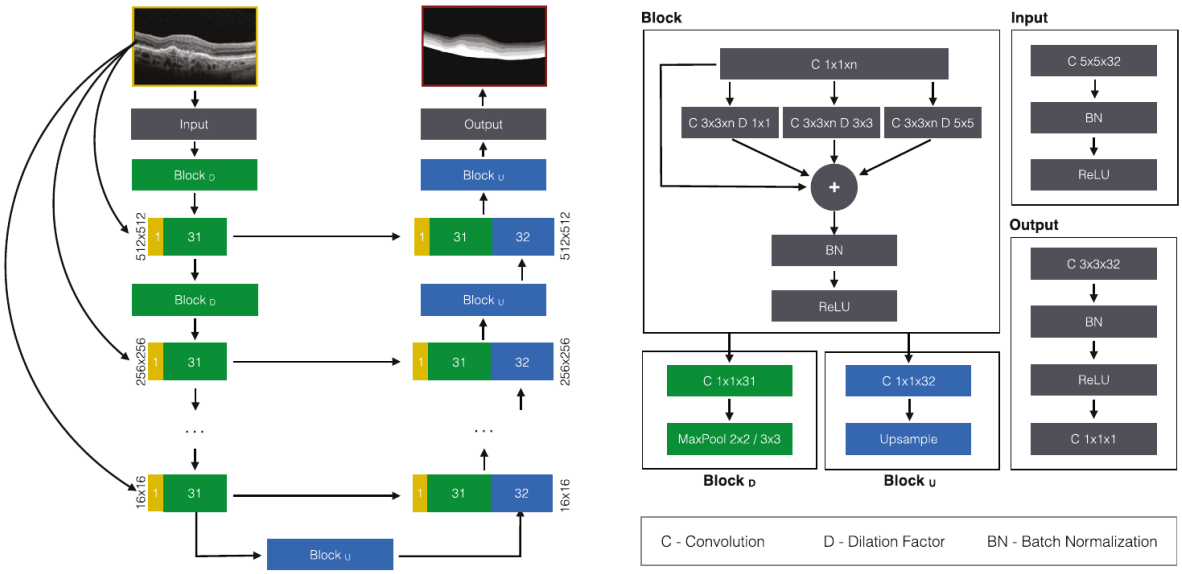

Fig. 3.4 BRU-net (Branch Residual U-Network). [74]

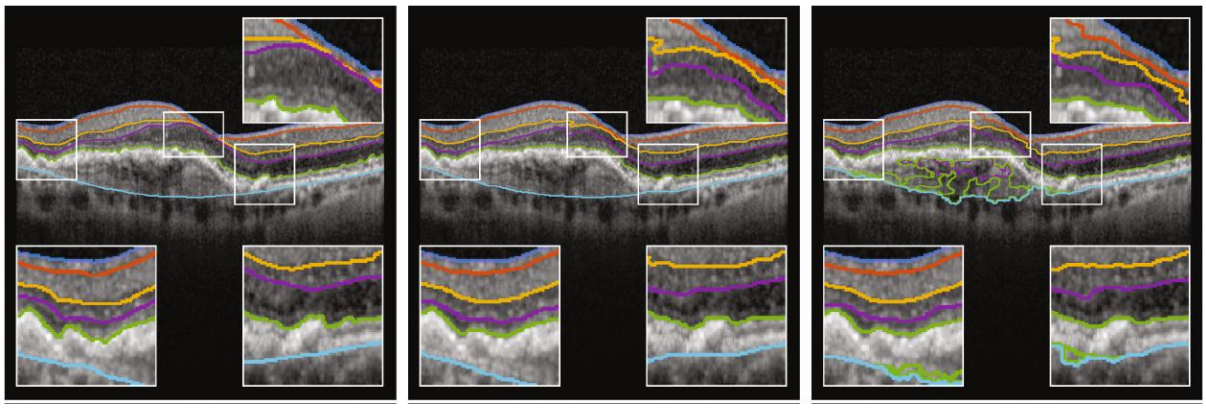

Fig. 3.5 From left to right: ground truth, segmentation done with BRU-net, and with U-net. [74]

Database on which results are evaluated consists of 20 macular OCT volumes (obtained from a Heidelberg Spectralis OCT device) from patients suffering from AMD. Unfortunately, the database is not publicly available. Segmentation of six boundaries (out of eleven; more details about retinal structure were provided in chapter 2.2.2.) were performed: internal limiting membrane (ILM), retinal nerve fiber layer (RNFL), ganglion cell layer (GCL), inner nuclear layer (INL), outer plexiform layer (OPL), and Bruch's membrane (BM). Fig. 3.5 shows some segmentation examples done with U-net and BRU-net. It is seen that segmentation performed by BRU-net is better in determining Bruch's membrane which is highly deformed due to pathology.

One of the major problems in segmenting retinal structures is the preservation of topology. Another paper I would like to point out is by He et al. (2018) [77] in which a cascade of two U-net architectures is proposed for the purpose of preserving the topology. Architecture (Fig. 3.6) consists of the first network marked as S-net which performs segmentation and the second network in cascade marked as R-net (regression network) which takes the topologically unconstrained S-net results and outputs layer thickness for each layer and each position.

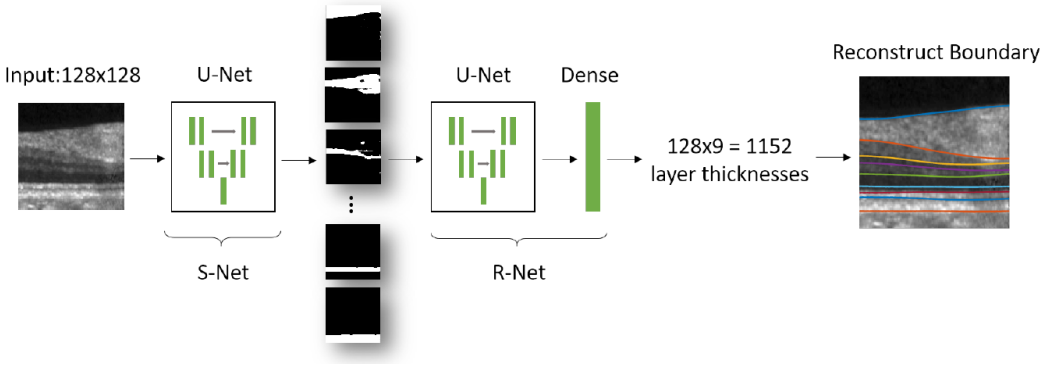

Fig. 3.6 Proposed cascaded framework for preserving topology. [77]

Networks are trained independently, S-net is trained with cross-entropy loss function, and R-net with mean squared loss function. Database on which results are evaluated consists of 20 macular OCT volumes (obtained from a Heidelberg Spectralis OCT device) from patients suffering from multiple sclerosis and 10 macular OCT volumes from healthy patients. Unfortunately, the database is not publicly available. Segmentation of nine layers was performed. Fig. 3.7 shows some segmentation examples. It is seen that segmentation after the second network in cascade is much better in preserving topology (which results in enhanced accuracy). Results that are shown are not from very deformed structure of the retina, without fluids and other pathological biomarkers present, so it remains questionable how good the results would remain in the case of the presence of far greater retinal deformity (as is the case with nAMD), but it can still be surmised that it is one of the effective ways to preserve the topology. Also, it should be mentioned that OCT volumes acquired from Heidelberg Spectralis OCT device have better quality (less noise) than OCT volumes acquired from Zeiss Cirrus device (as is the case for the AROI database).

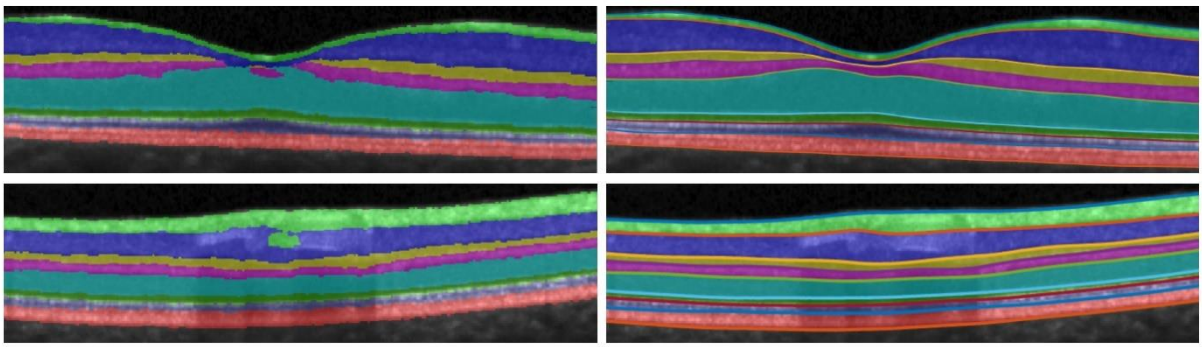

Fig. 3.7 From left to right: outputs from S-Net (with defects in topology) and outputs from R-Net with preserved topology for two examples (first and second row). [77]

3.1.2 Overview of methods for joint retinal layer and fluid segmentation

An approach in which layers and fluids are jointly segmented aims to take advantage of the interdependence of fluids and layers and thus achieve the best possible segmentation [45]–[50]. An overview of the more important studies will be given in follow-up.

One of the most frequently cited paper is from Roy et al. (2017) [46]. The proposed ReLayNet network architecture is shown in Fig. 3.8. It is a slightly modified U-net architecture. Similarly like in U-net architecture, each encoder block consists of a convolution layer, batch normalization (BN), ReLU activation function, and max pooling (MP). Each decoder block consists of unpooling layer, concatenation layer, convolution layer, batch normalization, and ReLU activation function. There are also skip connections typical for U-net architecture. The ReLayNet is trained by jointly optimizing two loss functions: weighted multi-class logistic loss and Dice loss. Data slicing approach is used wherein an OCT B-scan is sliced width-wise into a set of non-overlapping B-scan lines. Also, extensive data augmentation is used. [46]

Evaluation is done on the Duke SD-OCT publicly available dataset for DME patients [57] (already mentioned in chapter 2.1). Comparison with five other methods has been performed and ReLayNet achieved the best results. Segmentation of seven layers and all fluids was performed (without making a distinction between fluids). Segmentation results of the ReLayNet for two examples are shown in Fig. 3.9. The first example is OCT B-scan with no fluid present and the second with present fluids.

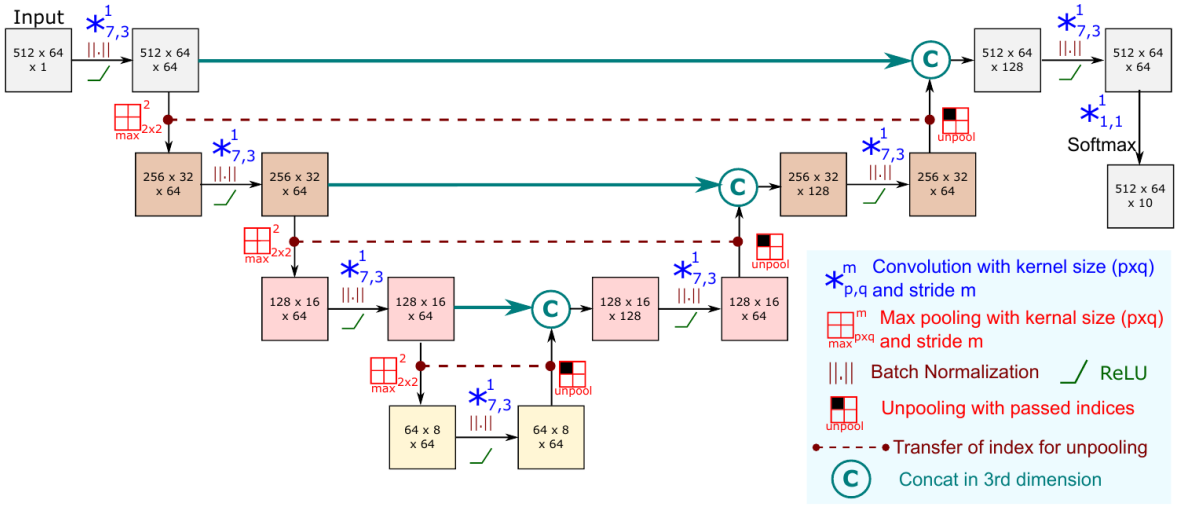

Fig. 3.8 ReLayNet architecture. [46]

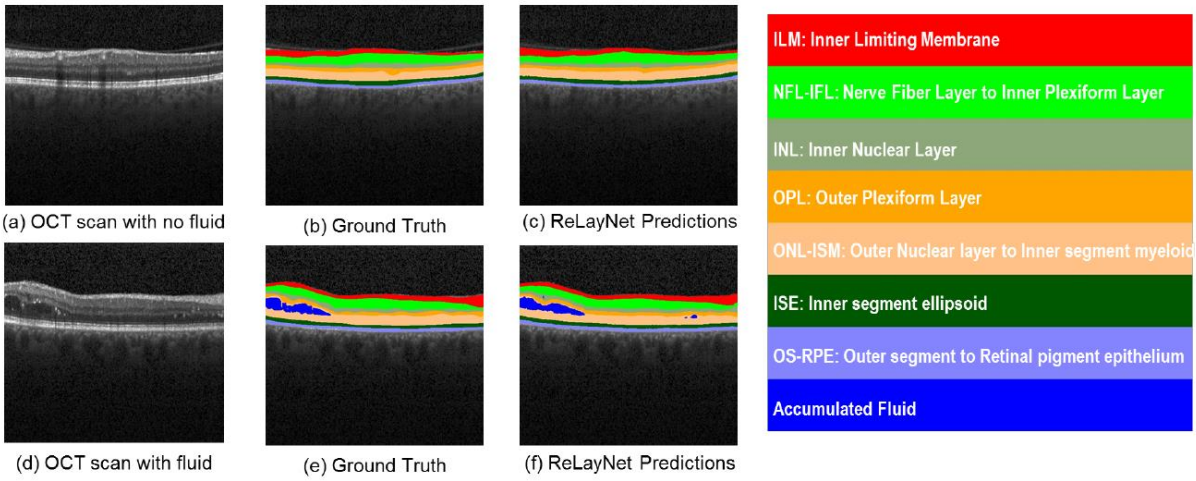

Fig. 3.9 Segmentation results of the ReLayNet. Example with no fluid present (first row) and with present fluids (second row). [46]

In the review paper [83], the authors give an overview of how the development of AI, or more specifically DL, has affected image analysis in ophthalmology. Results of different segmentation methods (Montouro et al. (2017) [45], Roy et al. (2017) [46], Lee et al. (2018) [84], Venhuizen et al. (2018) [9], Schlegl et al. (2017) [85]), some of which have been discussed in more detail before, are shown in Fig. 3.10.

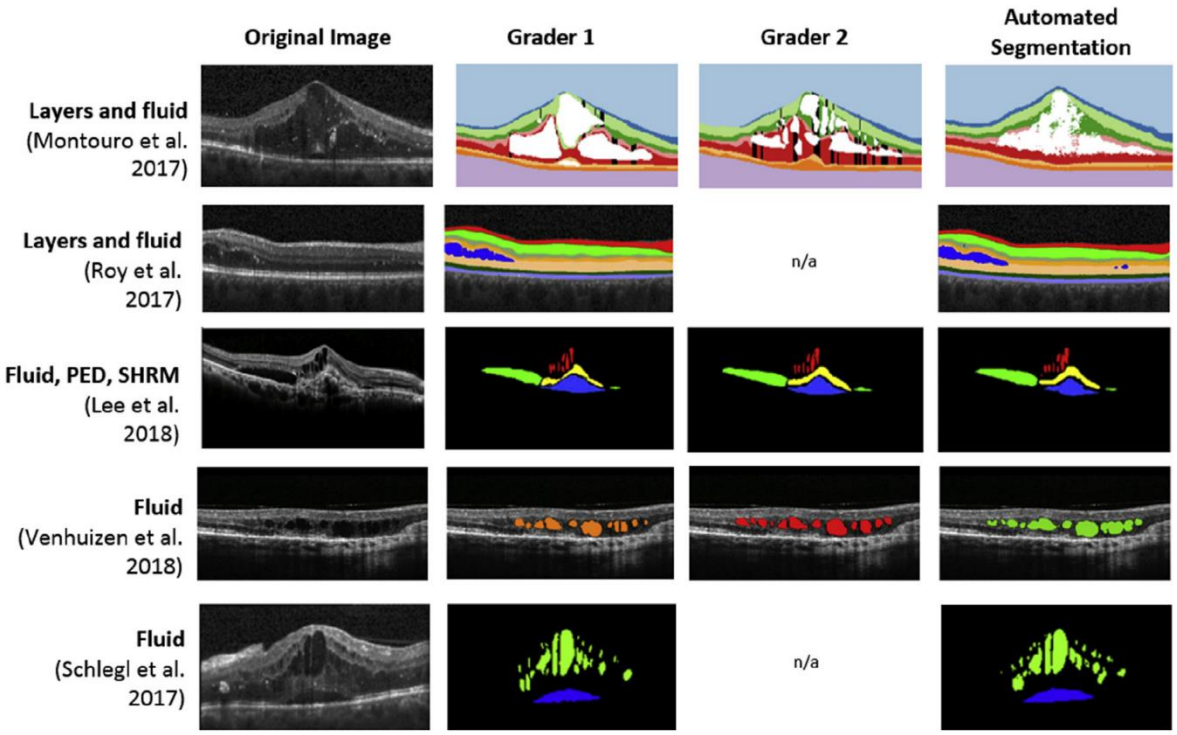

Fig. 3.10 Results of different segmentation methods. [83]

It can be concluded that although the development of DL has greatly contributed to the development of the field, there is still room for improvement and many ways in which AI can be introduced into clinical practice and thus modernize ophthalmology.

3.2 Model-centric approach towards automatic segmentation

In the first part, the model-centric approach will be presented, and in the second data-centric approach. The model-centric approach means that the data are constant while trying different architectures and within some deep neural net architecture tweaking parameters and hyperparameters. On the contrary, a data-centric approach means that with the same constant architecture we manipulate data (image pre-processing and post-processing, data augmentation, etc.). The final segmentation method could, and probably should consist of both approaches, but in such an orthogonal way, it is easier to see how the accuracy of the segmentation method depends on either data manipulation or model manipulation.

3.2.1 The baseline set of models for automatic segmentation

To set a baseline for deep learning (DL) methods results for automatic segmentation with standard U-net architecture [21] and two state-of-the-art architectures (U-net-like and U-net++) in medical image segmentation have been presented. Although many modifications of U-net architecture were proposed [49], [86]–[89] two state-of-the-art architectures were chosen to get further insights into challenges for automatic segmentation. The groundbreaking improvement in computer vision was achieved with ResNet architecture [80] and DenseNet architecture [90], where both reached breakthrough results in classification on the ImageNet dataset [91]. The logical step was to improve the U-net architecture in the ResNet and DenseNet style fashion. U-net-like architecture [51] and U-net++ [52] were chosen as two state-of-the-art architectures. The former combines the good sides of U-net and ResNet architectures and the latter is inspired by DenseNet architecture. A recent paper by Isensee et al. (2020) [92], where they proposed a nnU-net (out-of-the-box tool) for biomedical image segmentation that uses U-net-like architecture was another argument for this choice.

3.2.1.1 U-net architecture

Since its introduction in 2015, the U-net architecture [21] and its various modifications became widely popular in medical image segmentation [46], [93]–[95]. The U-net architecture consists of an encoder (contractive path), decoder (expanding path), and skip connections which enables simultaneous capturing of context and localization as it is shown in Fig. 3.11. The encoder blocks contain two convolutional filters (3x3) followed by batch normalization (BN) [49] and a ReLU activation function. Down-sampling is accomplished by max-pooling (MP) which reduces image size by half. Blocks in the decoder similarly consist of two transpose convolutions (3x3), BN, activation function (ReLU), and up-sampling (which doubles the size of the image – opposite to max-pooling). Skip connections are used to skip features from the contracting path to the expanding path.

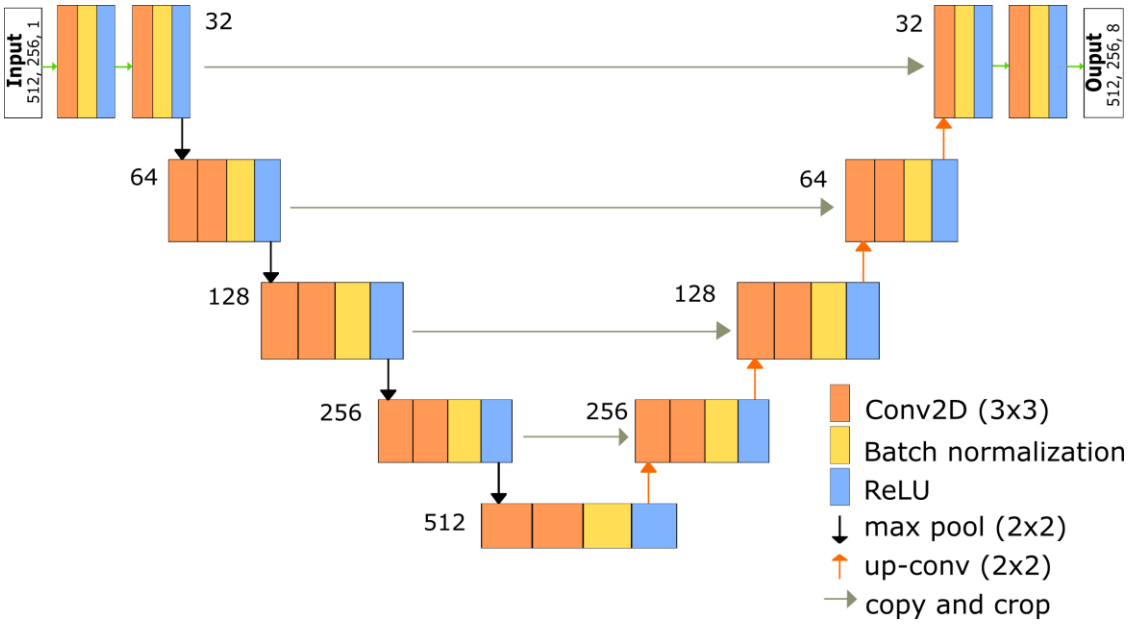

Fig. 3.11 The U-net architecture. [21]

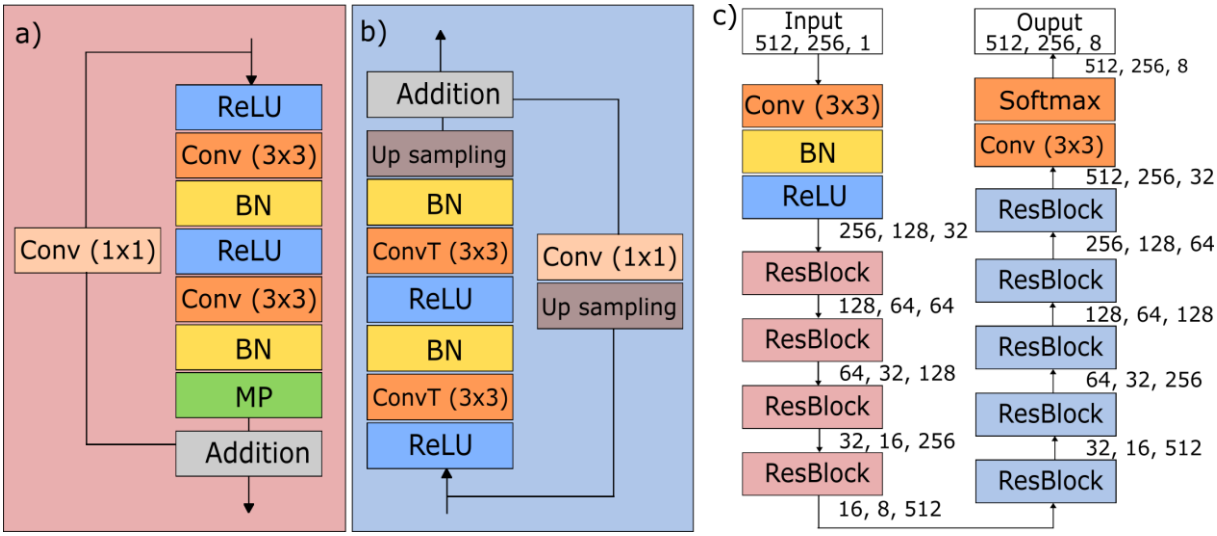

Fig. 3.12 a) A residual block in the encoder b) A residual block in the decoder. c) The used U-net-like architecture. [10]

3.2.1.2 U-net-like architecture

The U-net-like architecture [48] (Fig. 3.12) combines the good sides of U-net and ResNet architectures. In the down-sampling path, residual blocks contain convolutional filters (3x3) followed by batch normalization (BN) [49] and a ReLU activation function. Down-sampling is accomplished by max-pooling (MP) which reduces image size by half. Skip connection in the residual block is not just an identity connection but it contains a

convolutional filter (1x1) with strides equal to 2 and in that way is achieved down-sampling (image size is reduced by half, same as with MP in the main branch of the residual block). At the end of the residual block, the outputs from the main branch (also called a layer) and the skip connection are summarized. Residual blocks in the decoder are implemented in a similar manner: in a layer (the main branch) there are transpose convolutions (3x3), BN, activation function (ReLU), and up-sampling (which double the size of the image). In the skip connection, there is a convolution (1x1) and up-sampling.

3.2.1.3 U-net++ architecture (a nested U-net architecture for medical image segmentation)

The U-net++ architecture (a nested U-net architecture for medical image segmentation) is shown in Fig. 3.13. It is seen how U-net architecture is enhanced with dense blocks and convolution layers between the encoder and decoder. The purpose of modified skip pathways is to reduce the semantic gap between the feature maps of the encoder and decoder before merging. Each circle represents a set of convolution operations. The shaded part shows the original U-net architecture, while the rest (middle part) shows the difference from the original architecture: dense convolution blocks on the skip pathways and deep supervision. [52]

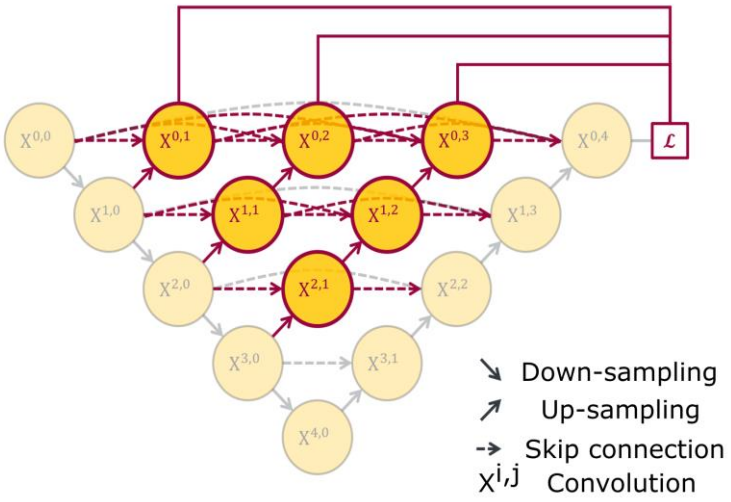

Fig. 3.13 U-net++: A nested U-net architecture. [52]

3.2.2 Training

Original images size 1024x512 pixels were resized to 512x256 pixels. Categorical cross-entropy loss [96] was used to train all models. The batch size was set to 4 (worse results were obtained with a larger batch size). The AdaBound optimizer [97] was used (as it combines advantages of Adaptive Moment Estimation (Adam) [82] and Stochastic Gradient Descent (SGD) [98]). Number of trainable parameters for standard U-net, U-net-like, and U-net++ architectures are 7764744, 8230536, and 9041832, respectively. Early stopping was used to prevent overfitting.

K-fold cross-validation was used where each fold contains images from 4 patients (1^{st} fold contains images from patients 1 to 4, 2^{nd} fold contains images from patients 5 to 8, and so on). K equals 6 in this procedure since, in that way, the test set share is approximately 15% as is a recommendation and common practice in a small data regime. It is not recommended to split the sets of images from the same patient across training, validation, and test set as adjacent B-scans are similar and that would lead to overestimated validation of the method.

The models were trained on Google Colab [99] with a GPU. The models were implemented in Python, using the Keras library with the TensorFlow backend.

Code is available on GitHub: https://github.com/mmelinscak/OCT-images-segmentation

3.2.3 Model prediction errors

Some examples of the automatic segmentation results are shown in Fig. 3.14. It is visible that results are good in case there are no significant pathological changes and deformation in the retinal structure. In case there are pathological changes, segmentation predictions are deficient in preserving the topology. Also, fluids segmentation should be enhanced. The main cause is the low representation of pixels belonging to these classes in the total number of pixels, especially when it comes to IRF (not present in all patients nor in all B-scans; in addition, it is regularly smaller than SRF and PED).

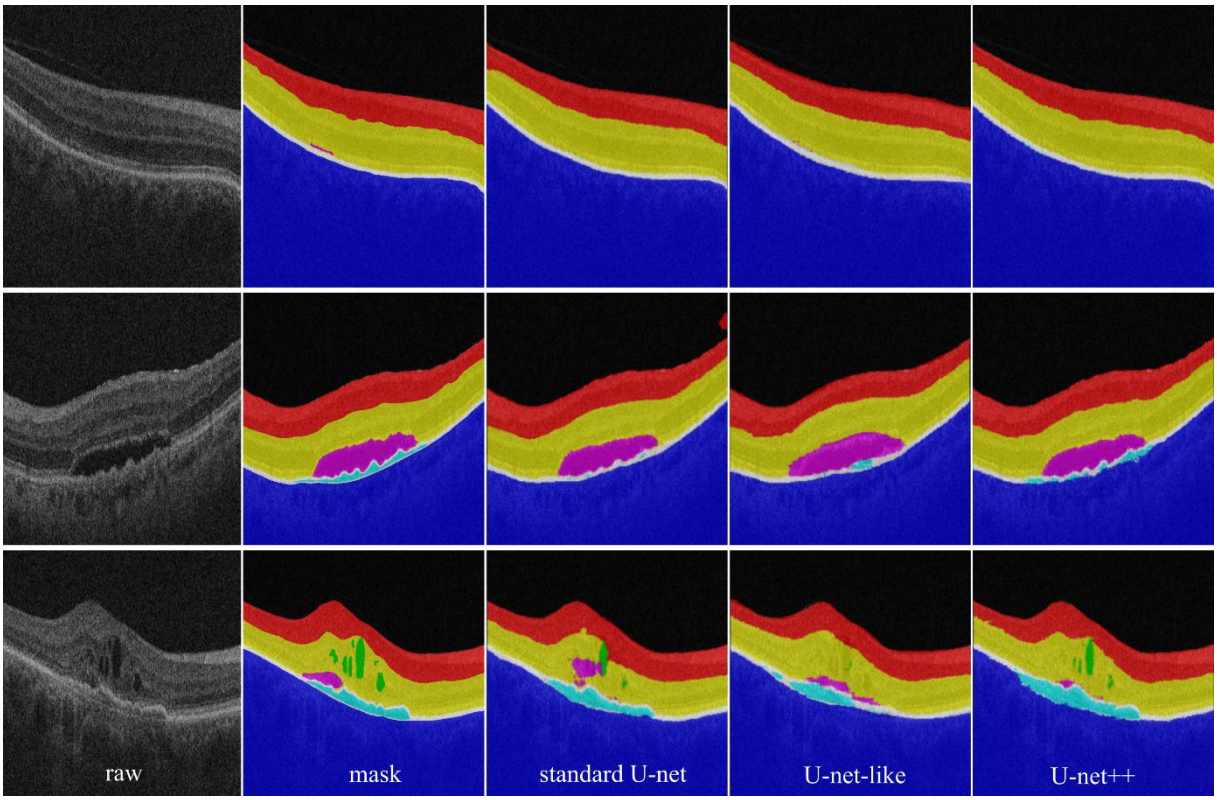

Fig. 3.14 Three examples of the segmentation results. First row: a case with less pronounced pathological changes. Second row: a case with more pathological changes (PED and SRF are present). Third row: a case with extensive pathological changes (PED, SRF, and IRF are present and there is a large distortion of layers). From left to right: raw image, expert annotation (mask), the prediction from the standard U-net architecture, the prediction from the U-net-like architecture, and the prediction from the U-net++ architecture. Images are cropped and only the ROI is visible.

3.2.4 Comparison of inter-observer error and model prediction error

The Dice score is used to evaluate results as it is a similarity measure often used as a metric in the segmentation of medical images. It is calculated according to Eq. (1). In Table 3.1 there are reported Dice scores (mean and standard deviation) for each class and for the inter- and intra-observer error, as well as the prediction errors from the U-net, U-net-like, and U-net++ models. The same results are shown in Fig. 3.15 with bar graphs. The Dice score is calculated for each patient in the test fold (that means four patients) but only for those B-scans with a reference segmentation. In case of fluids (either SRF or IRF) were not present on a single B-scan of the patient, the Dice score is exempted when calculating the mean and standard deviation of Dice score for that class and that patient. I found it more appropriate than setting it to zero or one as it would lead to overestimating or underestimating the metric value. In the case of patients in whom fluids were present on some of the B-scans but not all,

for B-scans where they were not present the value for Dice score was set to zero (the Dice score is not defined in the case of zero division which is a common situation in the absence of the class). It can be observed that for the inter-observer case and in all cases of the automatic segmentation the biggest errors occur in class 3 (surface between RPE and BM) and in classes that represent fluids (PED, SRF, and IRF). One of the factors that contribute to the complexities of automatic segmentation is significant class imbalance: the background (area above ILM and under BM) occupies as much as 83.26% in the total number of pixels, while IRF occupies only 0.12%; surface between RPE and BM occupies 1.07%, SRF 1.05%, PED 1.5%. Further on, out of a total of 1136 scans, PED, SRF, and IRF are present in 1014 (89.26%), 648 (57.04%), and 229 (20.16%) B-scans, respectively.

Table 3.1 The Dice score (mean and standard deviation) in the inter-observer case, the intra-observer case, for standard U-net model, U-net-like model, and U-net++ model. [10]

	Above ILM	ILM - IPL/INL	IPL/INL - RPE	RPE - BM	Under BM	PED	SRF	IRF
Inter-observer	0.982 (0.072)	0.950 (0.111)	0.948 (0.112)	0.699 (0.129)	0.989 (0.114)	0.860 (0.301)	0.876 (0.366)	0.735 (0.280)
Intra-observer	0.998 (0.003)	0.973 (0.008)	0.970 (0.117)	0.778 (0.092)	0.998 (0.001)	0.912 (0.242)	0.924 (0.331)	0.844 (0.140)
Standard U-net	0.995 (0.011)	0.950 (0.028)	0.923 (0.083)	0.669 (0.129)	0.988 (0.016)	0.638 (0.173)	0.513 (0.294)	0.480 (0.252)
U-net-like	0.995 (0.004)	0.899 (0.040)	0.890 (0.066)	0.476 (0.132)	0.988 (0.014)	0.533 (0.139)	0.365 (0.291)	0.040 (0.061)
U-net++	0.992 (0.011)	0.944 (0.032)	0.924 (0.064)	0.641 (0.133)	0.986 (0.017)	0.622 (0.159)	0.465 (0.297)	0.432 (0.265)

This research suggests that more complex architectures result in only slightly enhanced outcomes, no enhancement at all, or worse outcomes. For classes that represent fluids, the results are noticeably worse, compared to human error. Better results are obtained with U-net++ architecture than with U-net-like architecture, probably because U-net-like architecture lacks skip connections between encoder and decoder (skip connections only exist in each residual block within encoder/decoder) while there are dense blocks and convolutional layers between encoder and decoder in U-net++ architecture. Preliminary, it could be concluded that better segmentation accuracy cannot be obtained only with more complex architectures, but

rather using some of the efficient techniques in case of distinct class imbalance and the need for preserving the topology.

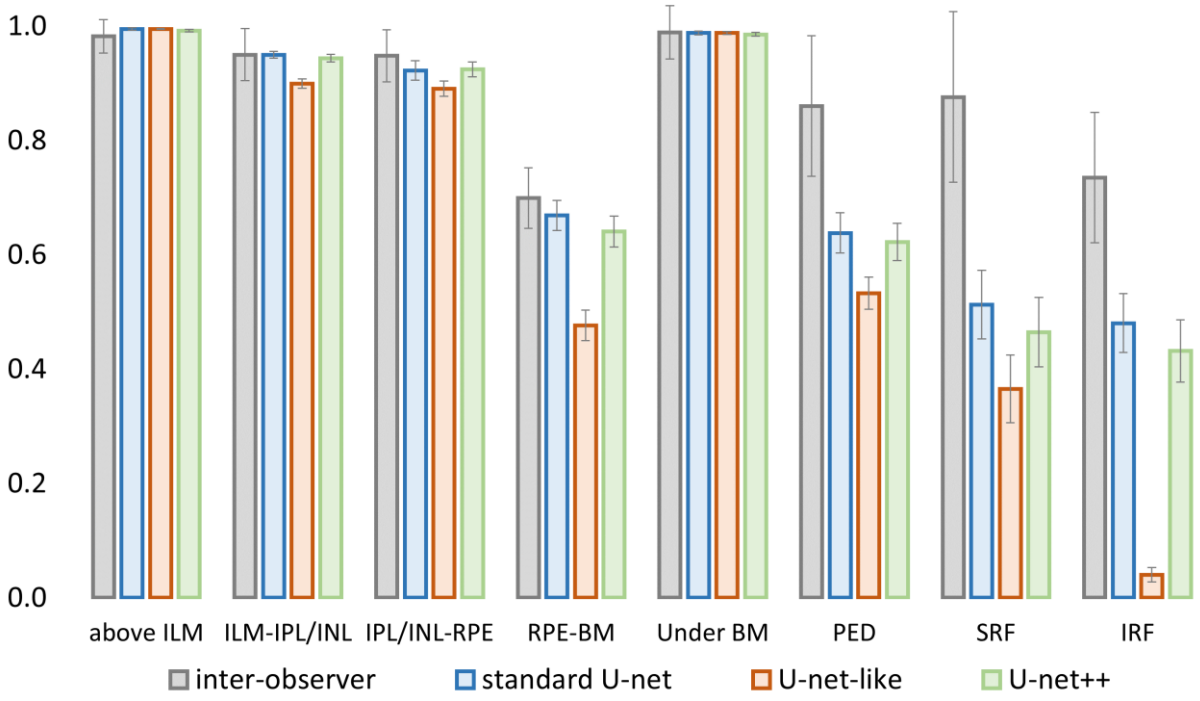

Fig. 3.15 The Dice scores (mean and standard error of the mean) for inter-observer variability, for standard U-net model, U-net-like model, and U-net++ model. [10]

Fig. 3.16 shows the confusion matrix for inter-observer error, and for model (U-net, U-net-like, and U-net++) prediction error. In the case of automatic segmentation, it is observed that IRF is often misclassified as the surface between IPL/INL and RPE (class 2) as it is in that area. In case the IRF is smaller or does not differ significantly in intensity and texture from the surface between IPL/INL and RPE, the model does not recognize it as a separate class. Also, it is observed that in a similar way SRF is often misclassified as PED or surface between IPL/INL and RPE. PED is mostly misclassified as an area under BM, and with further inception of individual predictions, it is apparent that it happens when Bruch's membrane is not clearly visible (due to geographic atrophy or some other pathological changes). These results of automatic segmentation for three models (standard U-net architecture and two state-of-the-art architectures) can serve as a baseline for the further development of deep learning models.

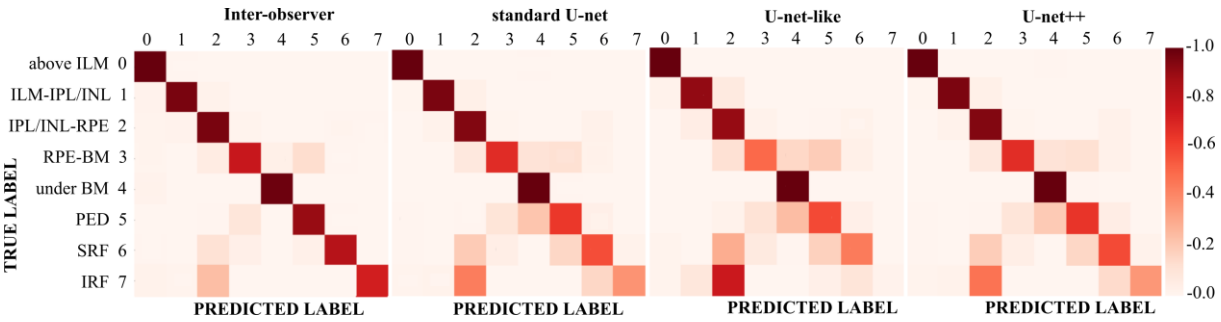

Fig. 3.16 The confusion matrix for inter-observer error, for U-net model, U-net-like model, and U-net++ model prediction error. [10]

3.3 Data-centric approach towards automatic segmentation

As stated earlier, in this section will be examined how various data manipulations affect segmentation accuracy.

3.3.1 Data augmentation

Data augmentation is a common "recipe" in small data regimes. It is a way to artificially increase the number of images. Images are transformed and in such a way are produced images that are not present in the dataset. In a survey from Shorten et al. [100] an extensive overview of the used data augmentation is given.

Natural images collected from most social media have human bias and are usually centered so it is meaningful to crop them as for real-world application it will not be so perfect. For instance, it is meaningful to change the color of the tree where the original image was taken in spring (green leaves), to make it look like in autumn (yellow-red leaves). In a real-world scenario, cats can be behind something and therefore covered, the sign on the road can only be partially visible, etc.

Unfortunately, there are not many realistic transformations that can be done on OCT images as they are acquired in a controlled situation and by following a standardized protocol. They always cover the same region centered at the fovea, images acquired from some type of OCT device are always of the same size, all images are noisy, grayscale, and so on. That is, there are not many transformations that can be applied after which images remain realistic since they are uniform in many respects. Fig. 3.17 shows some examples of transformations

that lead to still realistic images, e.g., horizontal flipping which mimics converting left eye to right and vice versa; rotating for some small number of degrees remains images in the domain of realistic scenarios.

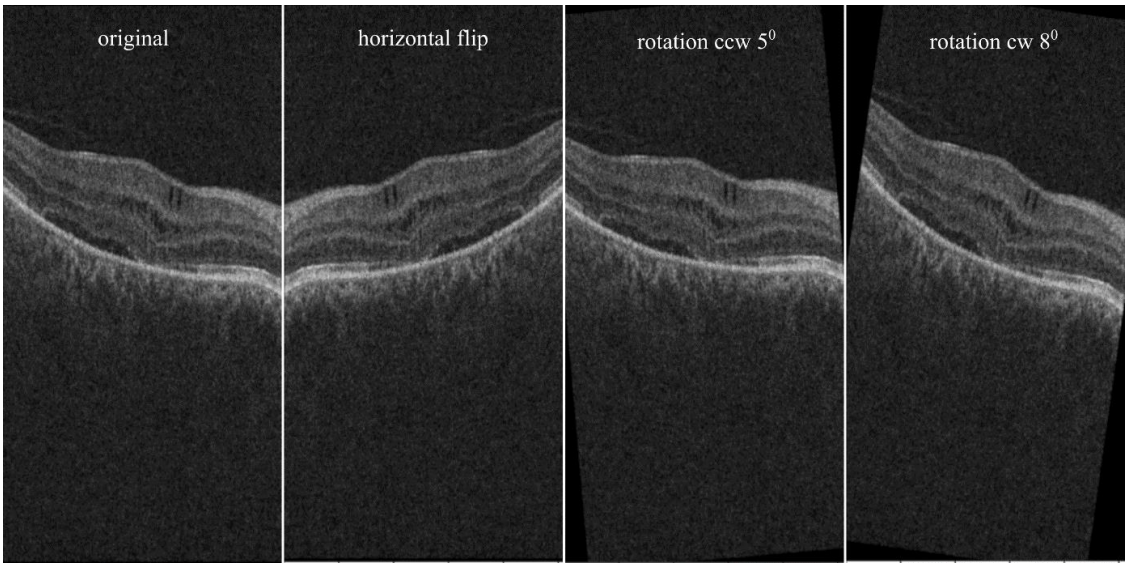

Fig. 3.17 Realistic transformations of OCT images.

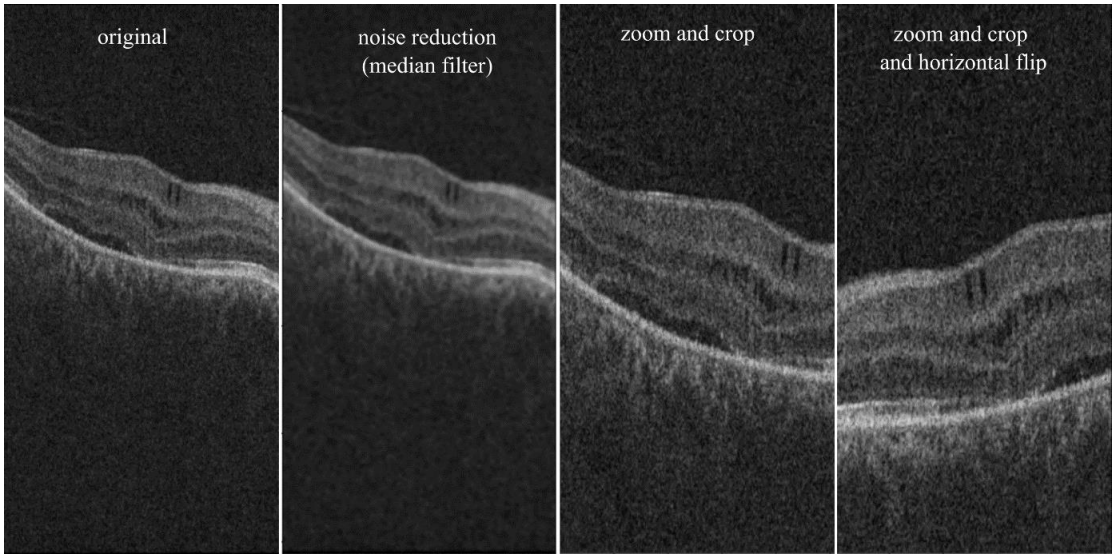

Fig. 3.18 Unrealistically transformed retinal OCT images.

It is questionable if unrealistically transformed images can still help in training the neural network. In that way, the network is trained on a dataset that could be considered from a different distribution than the test dataset (data augmentation is not performed on the test set). Fig. 3.18 shows some examples of transformed images that could not be present in the original dataset. For instance, cropping and zooming give images such as will never appear in

the test dataset. To see if such images still can enhance the accuracy of segmentation, I tried training with both realistic and unrealistic transformations.

3.3.1.1 U-net model

Firstly, the training with applied data augmentation was done on the example of the standard U-net model. Adding images that are not in the domain of realistic images decreased the performance of the model as could be expected. Unfortunately, training the model with only realistic data augmentation did not improve results much or in some cases did not improve at all. It could be explained that only minor data augmentation (with horizontal flipping and rotating) cannot greatly affect the results. Also, convolutions are equivariant to translation so they should capture same pattern wherever it is in the image space.

Table 3.2 The Dice score (mean and standard deviation) for standard U-net model and standard U-net model with applied data augmentation.

	Above ILM	ILM - IPL/INL	IPL/INL - RPE	RPE - BM	Under BM	PED	SRF	IRF
Standard U-net	0.995 (0.011)	0.950 (0.028)	0.923 (0.083)	0.669 (0.129)	0.988 (0.016)	0.638 (0.173)	**0.513** (0.294)	**0.480** (0.252)
Standard U-net + DA	0.995 (0.007)	**0.953** (0.017)	**0.935** (0.054)	**0.678** (0.125)	0.988 (0.018)	**0.653** (0.147)	0.512 (0.290)	0.378 (0.233)

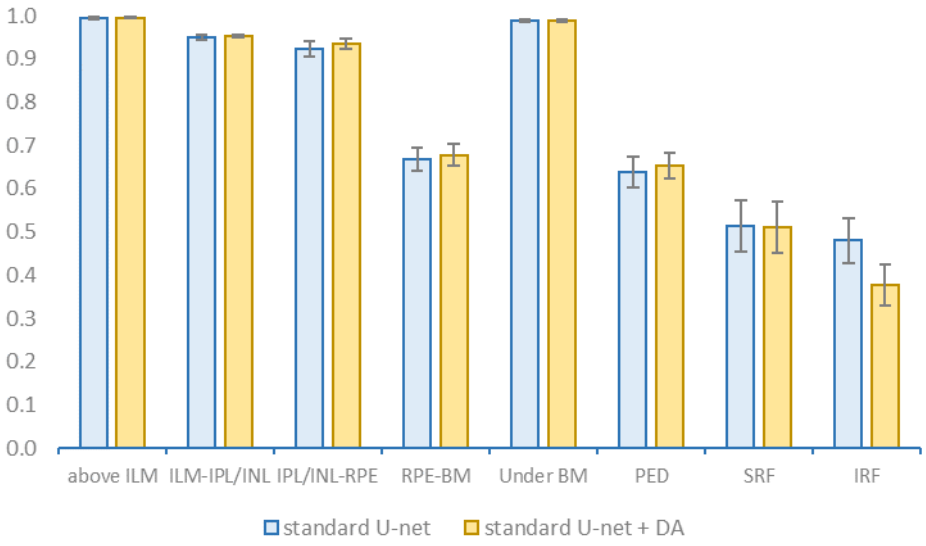

Fig. 3.19 Dice scores (mean and standard error of the mean) for standard U-net model and standard U-net model with applied data augmentation.

In Table 3.2 there are reported Dice scores (mean and standard deviation) for each class and for the prediction errors from the standard U-net model and, standard U-net model with applied data augmentation. The same results are shown in Fig. 3.19 with bar graphs.

3.3.1.2 Modified U-net-like model

Not to jump to conclusions and to generalize only on one example of the model, the same experiment was done with a U-net-like model. First, the U-net-like model was modified by adding skip connections. In chapter 3.2.1.2 was surmised that the inferior performance of U-net-like architecture is the result of lacking skip connections between encoder and decoder. To see if adding skip connections will enhance results, firstly, training was performed with a modified U-net-like model without data augmentation. Adding skip connections did improve the performance of the model, so further experiments with data augmentation are done with a modified U-net-like model. Unlike the case with the standard U-net model where data augmentation did not lead to enhancement of the model performance, in the case of a modified U-net-like model, after data augmentation (with only realistic transformations) model performance was improved.

In Table 3.3 there are reported Dice scores (mean and standard deviation) for each class and for the prediction errors from the U-net-like model, modified U-net-like model, and modified U-net-like model with applied data augmentation. The same results are shown in Fig. 3.20 with bar graphs. It can be observed that better results are achieved with modified U-net-like architecture, and that data augmentation did improve results, and overall performance gain more enhancement than in the case of the standard U-net model.

Table 3.3 The Dice score (mean and standard deviation) for the U-net-like model, modified U-net-like model, and modified U-net-like model with applied data augmentation.

	Above ILM	ILM - IPL/INL	IPL/INL - RPE	RPE - BM	Under BM	PED	SRF	IRF
U-net-like	0.995 (0.004)	0.899 (0.040)	0.890 (0.066)	0.476 (0.132)	0.988 (0.014)	0.533 (0.139)	0.365 (0.291)	0.040 (0.061)
Modified U-net-like	0.997 (0.003)	0.926 (0.034)	0.909 (0.069)	0.586 (0.145)	0.988 (0.017)	0.561 (0.152)	0.424 (0.316)	0.145 (0.122)
Modified U-net-like + DA	**0.997** (0.005)	**0.947** (0.028)	**0.928** (0.059)	**0.649** (0.124)	**0.991** (0.013)	**0.662** (0.123)	**0.438** (0.319)	**0.322** (0.252)

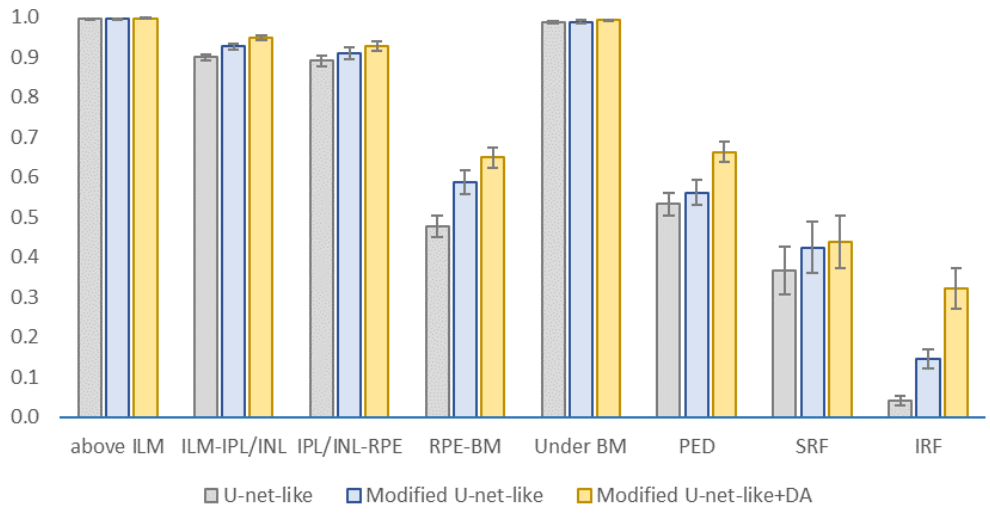

Fig. 3.20 Dice scores (mean and standard error of the mean) for the U-net-like model, modified U-net-like model, and modified U-net-like model with applied data augmentation.

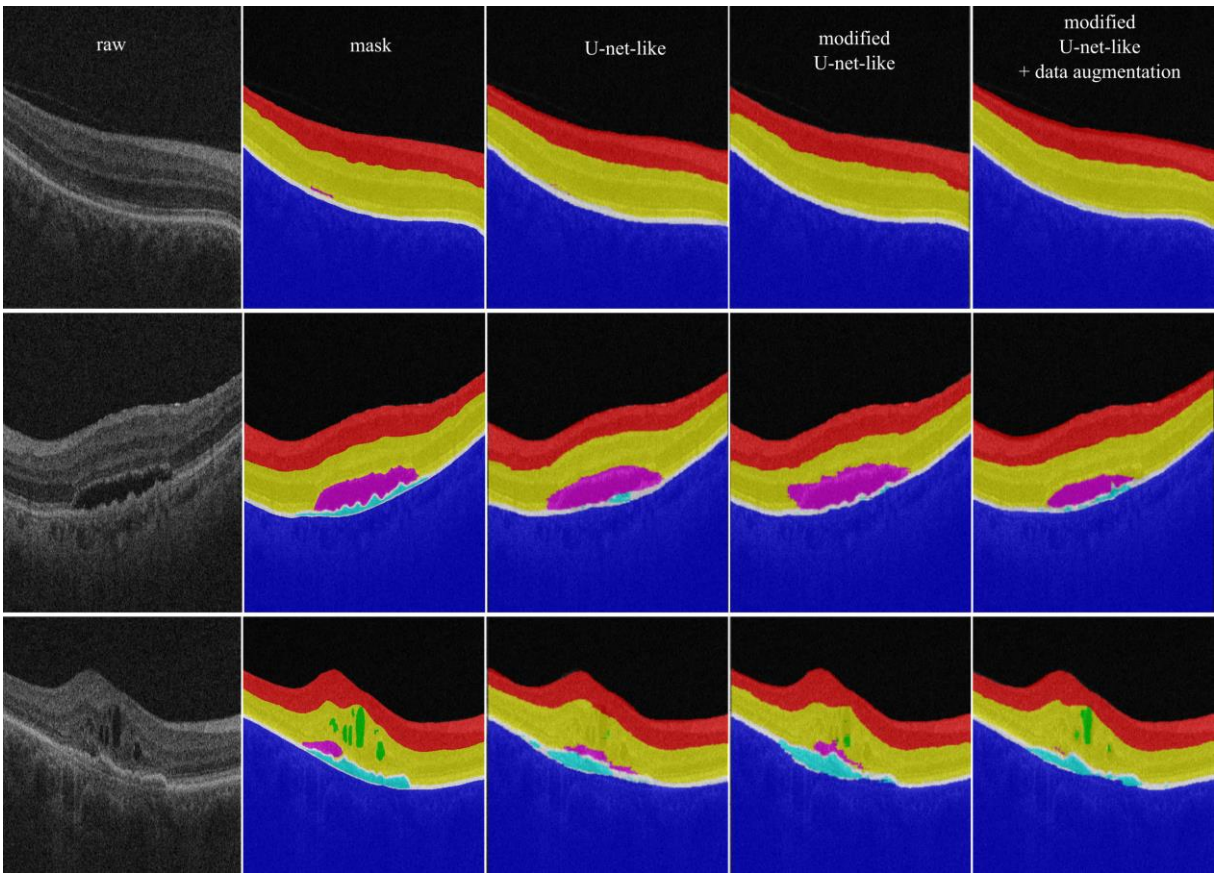

Fig. 3.21 Three examples of the segmentation results. First row: a case with less pronounced pathological changes. Second row: a case with more pathological changes (PED and SRF are present). Third row: a case with extensive pathological changes (PED, SRF, and IRF are present and there is a large distortion of layers). From left to right: raw image, expert annotation (mask), the prediction from the U-net-like architecture, the prediction from modified U-net-like architecture, and the prediction from modified U-net-like architecture with applied data augmentation. Images are cropped and only the ROI is visible.

Some examples of the automatic segmentation results are shown in Fig. 3.21. It is visible that there is not much difference in segmentation results in case there are no significant pathological changes and deformation in retinal structure. In case there are pathological changes, segmentation predictions are enhanced with data augmentation (third case in which IRF is detected with a modified U-net-like model, and even more with applying data augmentation).

Although my intention was to make analysis in as much orthogonal way as possible, that is to see how manipulating with just each parameter either on the data (e.g., data augmentation) or architecture results in overall performance, fully orthogonal analysis is not possible for neural networks models as there are too many interdependencies between data, model, and all parameters. All results in which only data are manipulated, and the model is constant and vice versa should be taken approximately and as a guide to see what leads to better performance and what decreases performance.

3.3.2 Image post-processing

Given how challenging tasks are segmentation of retinal structures and pathological biomarkers, various methods of image post-processing are often used as a final part in order to improve the accuracy of segmentation. Results of the more commonly used methods will be given below.

3.3.2.1 Median filter

The median filter is often used to remove noise from an image. The main idea behind applying the median filter as a post-processing step is to remove outlier values. By applying the median filter, each pixel is replaced with the median of neighboring pixels. The basic problem associated with noise removal algorithms is their intrinsic consequence of reducing image resolution. If kernel size is too big some details could be lost, so it is always a trade-off between smoothing the image by removing the noise and losing some details.

The training was done on the example of the standard U-net model with all the same parameters as in previous chapters. In Table 3.4 there are reported Dice scores (mean and standard deviation) for each class and for the prediction errors from the standard U-net model

with no post-processing step, and with applied median filter with kernel size 4 and 8. It can be observed that there is no improvement in segmentation accuracy.

Table 3.4 The Dice score (mean and standard deviation) for standard U-net model with no post-processing step, and with applied median filter with kernel size 4 and 8.

	Above ILM	ILM - IPL/INL	IPL/INL - RPE	RPE - BM	Under BM	PED	SRF	IRF
No post-processing	0.995 (0.011)	**0.950** (0.028)	0.923 (0.083)	**0.669** (0.129)	**0.988** (0.016)	**0.638** (0.173)	**0.513** (0.294)	**0.480** (0.252)
Median filter size=4	0.994 (0.006)	0.942 (0.026)	**0.930** (0.051)	0.660 (0.129)	0.987 (0.018)	0.613 (0.183)	0.457 (0.282)	0.358 (0.214)
Median filter size=8	**0.995** (0.005)	0.941 (0.026)	0.929 (0.051)	0.634 (0.134)	0.986 (0.018)	0.609 (0.167)	0.459 (0.287)	0.299 (0.225)

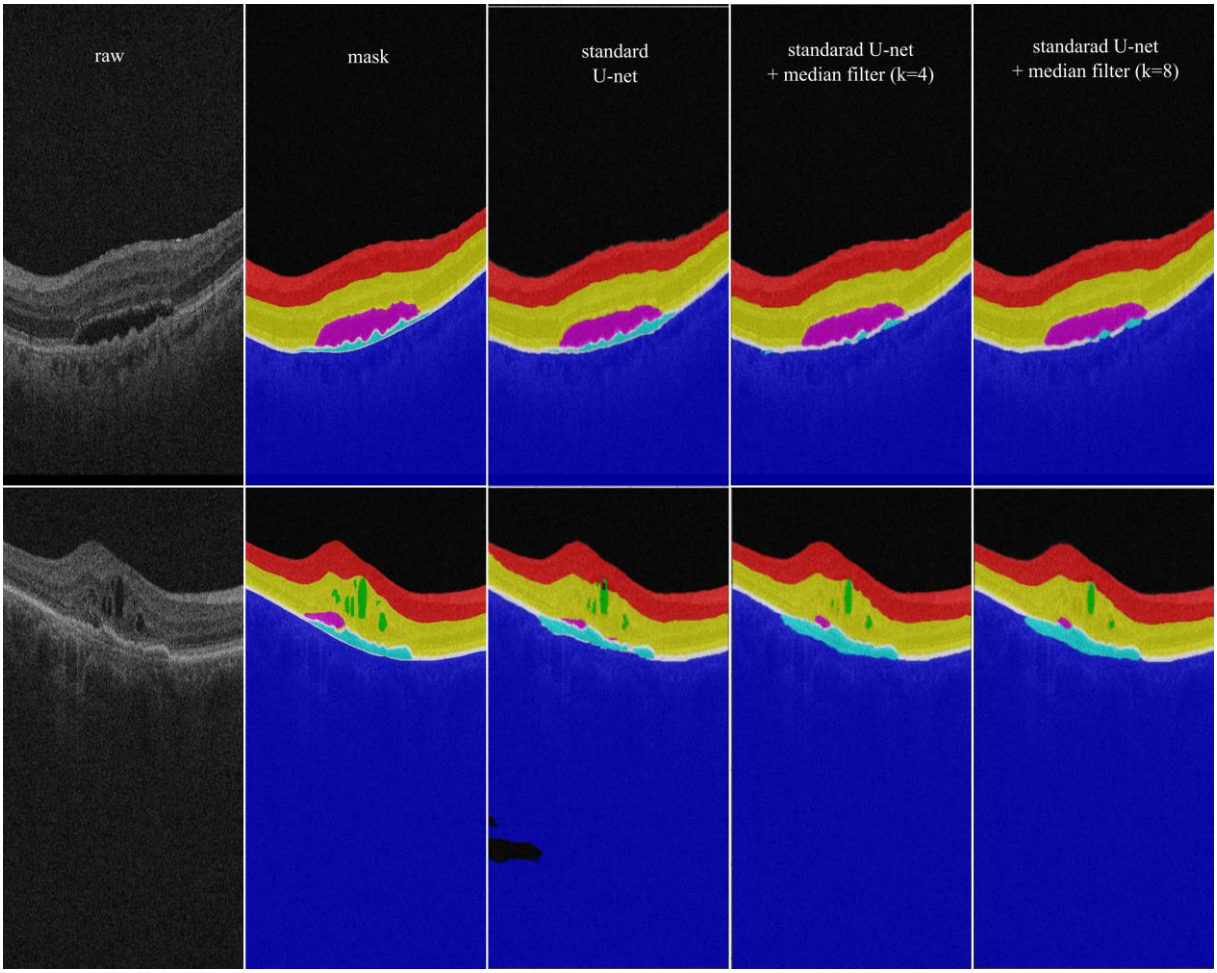

Fig. 3.22 Examples for one raw image and ground truth mask, followed by model predictions without post-processing and with applying a median filter with kernel size equals 4 and 8.

Fig. 3.22 shows examples for one raw image and ground truth mask, followed by model predictions without post-processing and with applying a median filter with kernel size equals 4 and 8. It can be observed that in the first case the lines and shapes are less serrated, but some of the fluid (PED) is lost. In the second case, it can be observed that there are some enhancements in segmentation, but also some drawbacks: by applying median filter misclassified area under BM is classified correctly; the shape of IRF is cleaner, but part of it is lost (with bigger size of the kernel, it is even more pronounced); part of SRF which was FP is removed with a median filter, but part of correctly classified SRF is also removed.

3.3.2.2 Mathematical morphology

Mathematical morphology is a technique for the processing of geometrical structures, based on set theory, lattice theory, and topology. The basic morphological operators are erosion, dilation, opening, and closing. These transformations involve the interaction between an image and a structuring set. I will not go into all mathematical details, rather show few examples. Intuitively, dilation magnifies an image object and erosion decreases it. Erosion and dilation can be used in parallel and series. Opening generally smooths a contour in an image. Closing tends to narrow smooth sections of contours, eliminating small holes, and filling gaps in contours. [101]

The training was done on the example of the standard U-net model with all the same parameters as in previous chapters. Fig. 3.23 shows an example of prediction with applied mathematical morphology operations. It can be observed that predictions have some FP classified IRF (marked in green). Also, SRF is poorly classified, and there are few yellow spots in the red area (ILM-IPL/INL misclassified as IPL/INL-BM) After applying erosion, misclassified IRF and IPL/INL-BM disappear. PED is shrinking and is replaced with RPE-BM class. When applying dilation, SRF, IRF, area IPL/INL-BM, and PED are expanded. With the opening, the shapes are more regular and smoother, like after applying the median filter. With the closing, there is not much change to the original prediction.

Fig. 3.24 shows another example and similar behavior of applying morphological operations can be noticed.

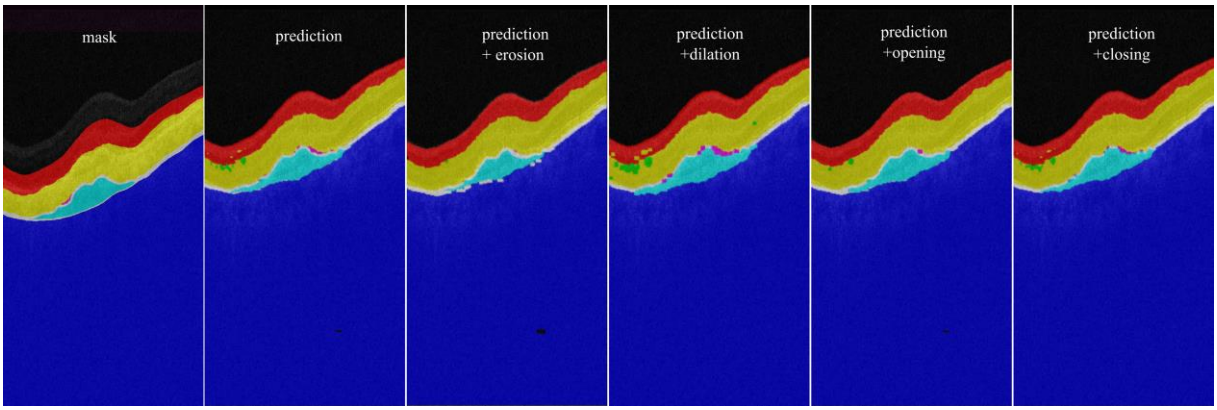

Fig. 3.23 Examples for one raw image and ground truth mask, followed by model predictions without post-processing and with applying mathematical morphology operations: erosion, dilation, opening, and closing.

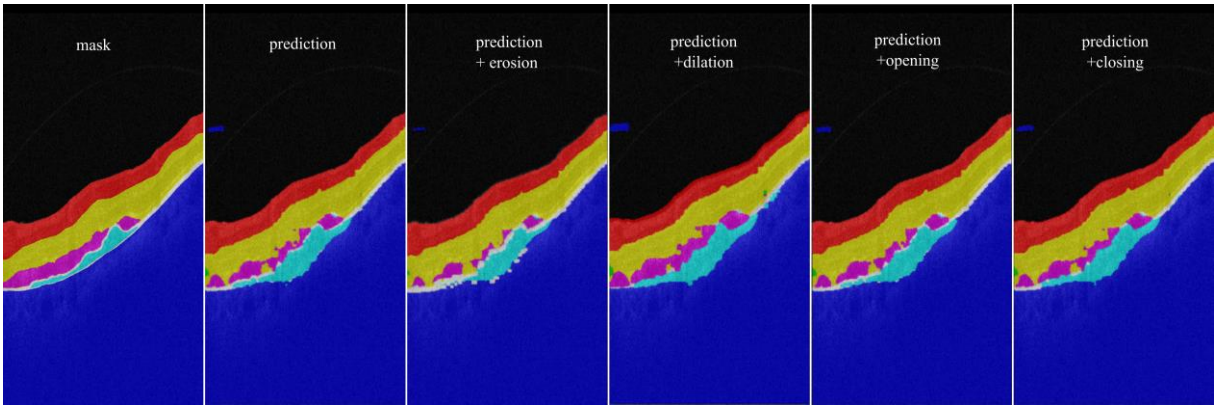

Fig. 3.24 Examples for one raw image and ground truth mask, followed by model predictions without post-processing and with applying mathematical morphology operations: erosion, dilation, opening, and closing.

3.3.2.3 Conditional random fields

Conditional random fields (CRF) are a type of discriminative undirected probabilistic graphical model – the prediction is modeled as a graphical model, which implements dependencies between the predictions. They are often used as post-processing steps in image segmentation tasks. I opted for them as CRF take context into account and therefore should enhance segmentations considering topology preserving.

In this study, CRF was implemented as fully connected CRF with Gaussian edge potentials (pairwise edge potentials are defined by a linear combination of Gaussian kernels) [102]. In [102] is described how CRF could be implemented as fully connected, which means that they are defined on the entire set of image pixels and still effective due to efficient approximate inference algorithm.

A fully connected CRF learns pairwise potentials on all pairs of pixels in the image – the pairwise edge potentials are determined by a linear combination of Gaussian kernels. Equations and all the details of how unary potential and contrast-sensitive two-kernel potentials are defined can be seen in the original paper [102]. Two-kernel potentials are defined as the sum of appearance kernel and smoothness kernel. The appearance kernel takes into consideration that adjacent pixels with similar intensity probably belong to the same class. The levels of nearness and similarity could be regulated by two parameters. The smoothness kernel removes small, isolated regions.

In Table 3.5 there are reported Dice scores (mean and standard deviation) for each class and for the prediction errors from the standard U-net model with no post-processing step, and with applied conditional random fields. The number of the iterations of the inference algorithm was set to 5. It can be observed that all values for the Dice score are increased except for IRF and RPE-BM classes. By setting different parameters, which would emphasize more difference in pixel intensity of nearby pixels, different results would be obtained. By setting relatively large parameters in the smoothness kernel some TP segmented pixels belonging to IRF are also removed.

Table 3.5 The Dice score (mean and standard deviation) for standard U-net model with no post-processing step, and with applied CRF.

	Above ILM	ILM - IPL/INL	IPL/INL - RPE	RPE - BM	Under BM	PED	SRF	IRF
No post-processing	0.995 (0.011)	0.950 (0.028)	0.923 (0.083)	**0.669** (0.129)	0.988 (0.016)	0.638 (0.173)	0.513 (0.294)	**0.480** (0.252)
CRF	**0.996** (0.004)	**0.953** (0.023)	**0.940** (0.041)	0.667 (0.109)	**0.990** (0.009)	**0.650** (0.150)	**0.554** (0.251)	0.443 (0.241)

Fig. 3.25 shows examples for one raw image and ground truth mask, followed by model predictions without post-processing and with applied conditional random fields. It can be observed that in the first example, segmentation looks cleaner, shapes of PED and SRF are more coherent. Also, some false positive segmented IRF is removed. In the second example, except that some false positive IRF and SRF were removed, a large portion of the false positive segmented area under BM (background) is removed and properly segmented. Similarly, in the third case, a large portion of the false positive segmented area above ILM (background) is properly segmented after applying CRF. These are results of tuned

parameters in the appearance kernel and smoothness kernel. By setting different parameters, which would emphasize more difference in pixel intensity of nearby pixels than in localization, different results would be obtained.

The other way how CRF could be implemented is by adding a layer in form of a recurrent neural network (RNN) as the last layer in the convolutional neural network.

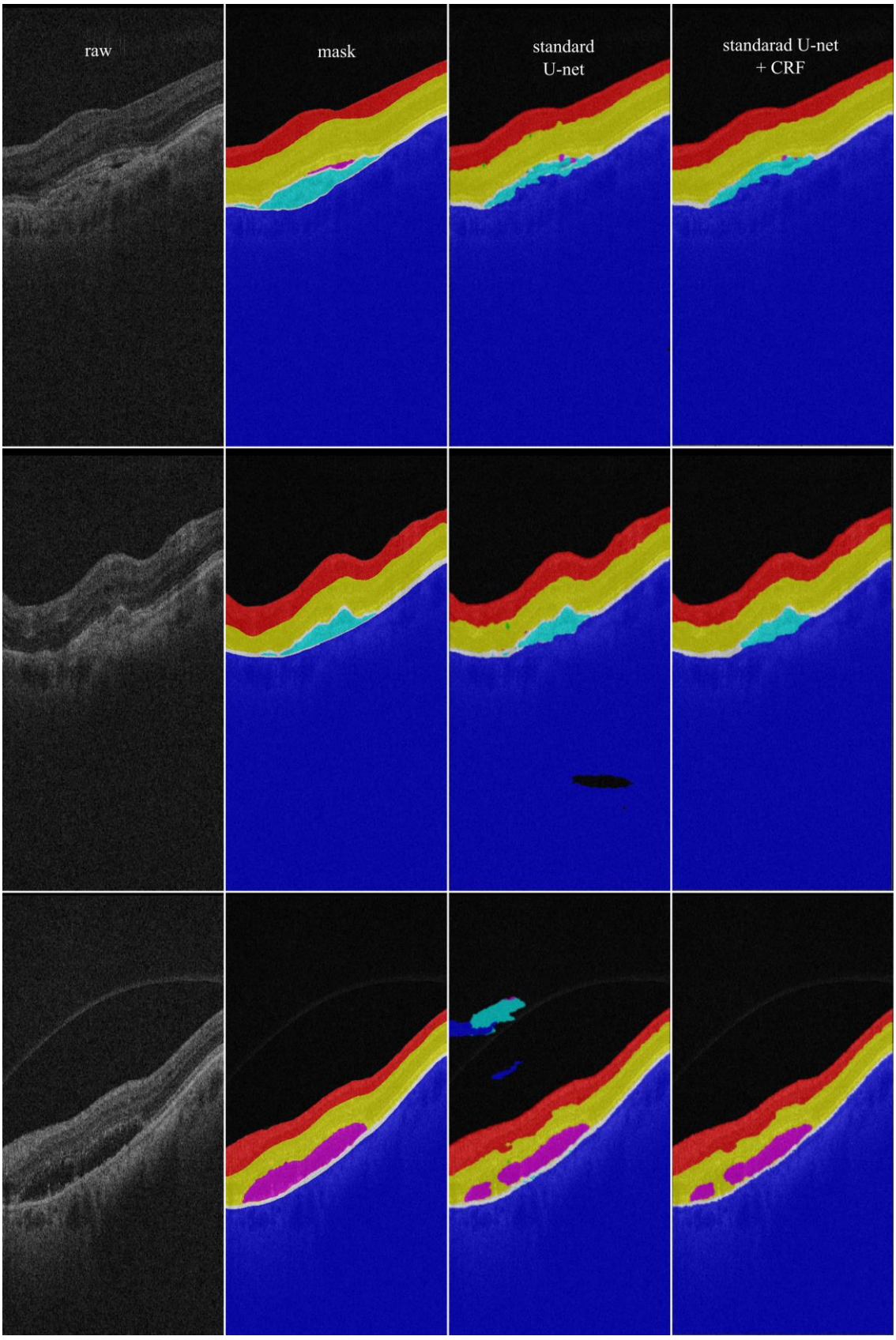

Fig. 3.25 Examples for one raw image and ground truth mask, followed by model predictions without post-processing and with applying conditional random fields.

3.4 Proposed methods for joint retinal layer and fluid segmentation

Table 3.6 shows the Dice score (mean and standard deviation) in the inter-observer case, for standard U-net model, standard U-net model with applied CRF, U-net++ model, U-net-like model, modified U-net-like model, modified U-net-like model with data augmentation, and modified U-net-like model with data augmentation and applied CRF.

Table 3.6 The Dice score (mean and standard deviation) in the inter-observer case, for standard U-net model, standard U-net model with applied CRF, U-net++ model, U-net-like model, modified U-net-like model, modified U-net-like model with data augmentation, and modified U-net-like model with data augmentation and applied CRF.

	Above ILM	ILM - IPL/INL	IPL/INL - RPE	RPE - BM	Under BM	PED	SRF	IRF
Inter-observer	0.982 (0.072)	0.950 (0.111)	0.948 (0.112)	0.699 (0.129)	0.989 (0.114)	0.860 (0.301)	0.876 (0.366)	0.735 (0.280)
Standard U-net	0.995 (0.011)	0.950 (0.028)	0.923 (0.083)	0.669 (0.129)	0.988 (0.016)	0.638 (0.173)	0.513 (0.294)	**0.480** (0.252)
Standard U-net + CRF	0.996 (0.004)	0.953 (0.023)	**0.940** (0.041)	0.667 (0.109)	0.990 (0.009)	0.650 (0.150)	**0.554** (0.251)	0.443 (0.241)
U-net++	0.992 (0.011)	0.944 (0.032)	0.924 (0.064)	0.641 (0.133)	0.986 (0.017)	0.622 (0.159)	0.465 (0.297)	0.432 (0.265)
U-net-like	0.995 (0.004)	0.899 (0.040)	0.890 (0.066)	0.476 (0.132)	0.988 (0.014)	0.533 (0.139)	0.365 (0.291)	0.040 (0.061)
Modified U-net-like	**0.997** (0.003)	0.926 (0.034)	0.909 (0.069)	0.586 (0.145)	0.988 (0.017)	0.561 (0.152)	0.424 (0.316)	0.145 (0.122)
Modified U-net-like + DA	0.997 (0.005)	0.947 (0.028)	0.928 (0.059)	0.649 (0.124)	0.991 (0.013)	0.662 (0.123)	0.438 (0.319)	0.322 (0.252)
Modified U-net-like + DA + CRF	**0.997** (0.003)	**0.955** (0.012)	0.930 (0.032)	**0.670** (0.113)	**0.992** (0.010)	**0.670** (0.140)	0.500 (0.132)	0.448 (0.241)

Only models in which some of the modifications contributed to better results were selected. For example, models that use a median filter as post-processing are not shown as the results after applying the median filter turn out to be worse. As can be seen from the results, no model gives the best results for all classes. Better results could be achieved with an ensemble of multiple methods (multiple models in parallel), but since the goal is to build the model that will be the part of an OCT device when faster processing time is required, the

question is how acceptable such an approach is – it depends on the OCT device and associated hardware.

Any tuning of the model parameters in the case of a small database without enough diversity (consisting of retinal OCT images of only 24 patients suffering from nAMD) would lead to overfitting to the database itself with loss of generalization to other (unseen) images. When developing an OCT device that would be acceptable for introduction into clinical practice, it is necessary to collect a database of images with sufficient diversity (meaning different age groups, different diseases, different ethnicities, etc.) and then it is necessary to adjust the model parameters to obtain better performance.

4 METHODS FOR RETINAL FLUID SEGMENTATION IN OPTICAL COHERENCE TOMOGRAPHY IMAGES

4.1 Overview of methods for retinal fluids segmentation

Despite numerous methods which have been developed to date, methods for automatic segmentation are only reliable on retinal images for patients with no disease or in the early stage of the disease. When nAMD is present, fluids significantly destroy and distort the structure of retinal layers, resulting in a complex interaction of fluids and layers. Therefore, automatic segmentation becomes much more complex and challenging.

Due to the complexity of retinal fluids segmentation, automatic methods that have been developed are usually a combination of various methods. Previously developed methods are: methods based on a combination of techniques for image analysis [103]–[105], methods based on graph theory [106], [107], and various techniques of machine learning where feature extraction is performed by the method of image analysis [108]–[110]. Due to the growing popularity of deep learning methods, DL algorithms are also applied to the segmentation of retinal fluids [84], [85], [111]. It is difficult to evaluate and compare different methods as there is a lack of publicly available databases of manually labelled images, and results depend significantly on the number of processed images, quality of images (related to the type of OCT device), present disease, etc.

As mentioned in chapter 2.1 in 2017 “The Retinal OCT fluid Detection and Segmentation Benchmark and Challenge” (RETOUCH) [55] was organized. The benchmarking resource database consisted of OCT volumes from 112 patients suffering from AMD and retinal vein occlusion (RVO). OCT volumes were acquired from three types of OCT devices: Topcon, Spectralis, and Cirrus. Three types of retinal fluids were annotated: SRF, PED, and IRF. [56] All applied methods (eight teams were participating) for automatic segmentation and detection of fluids were based on deep learning methods, most of them relied on U-net architecture. Differences in results mostly depend on whether they used data augmentation, layer segmentation, and post-processing. Five of them used data augmentation. Four out of eight teams used layer segmentation although layer annotations have not been provided, instead, they relied on some previously developed methods. All but one team used

post-processing (morphologic, median filter, random forest, graph-cut). In Table 4.1 are shown Dice scores for all three fluids and all eight teams. Dice scores for PED range from 0.63±0.27 (maximum marked in bold) to 0.77±0.23 (minimum marked in blue color). Dice scores for SRF range from 0.53±0.34 to 0.75±0.30. Dice scores for IRF range from 0.49±0.20 to 0.82±0.19. There are lots of B-scans with no fluids, and it is not described how Dice score is calculated in such cases: whether it is set to zero, to one, or such B-scans are disregarded from the calculation – given the proportion of B-scans without fluids, this significantly contributes to the result. Nevertheless, the results have shown that the fluids segmentation task is still challenging and hard to accomplish.

Table 4.1 The Dice score (mean and standard deviation) for all three fluids and for all eight teams that attended the challenge. The highest values are marked in bold, and the lowest values are in light blue. [56]

	PED	SRF	IRF
SFU	0.74 (0.24)	**0.75 (0.30)**	**0.82 (0.19)**
UMN	**0.77 (0.23)**	0.71 (0.33)	0.70 (0.20)
MABIC	0.70 (0.29)	0.66 (0.32)	0.78 (0.22)
RMIT	0.69 (0.25)	0.70 (0.31)	0.73 (0.20)
RetinAI	0.71 (0.29)	0.67 (0.33)	0.74 (0.19)
Helios	0.66 (0.26)	0.68 (0.30)	0.63 (0.19)
NJUST	0.63 (0.27)	0.53 (0.34)	0.57 (0.21)
UCF	0.63 (0.24)	0.54 (0.33)	0.49 (0.20)

I would also like to highlight a paper from Lee et al. (2018) [84] as they work with images from patients with nAMD and as they differentiate SRHM from SRF which makes fluid segmentation even more challenging. Dataset consists of 930 B-scans from 93 patients (central B-scans in the foveal area) with nAMD. Images were acquired with Spectralis OCT device (better quality of images than with images acquired from Cirrus Zeiss OCT device as is the case in the AROI database).

The architecture they used is based on U-net architecture with minor modifications (Fig. 4.1). Data augmentation is applied to compensate for the small number of images. Training is done with optimizing Dice loss as the problem is reduced to semantic segmentation where the Dice score is usually a metric for evaluation.

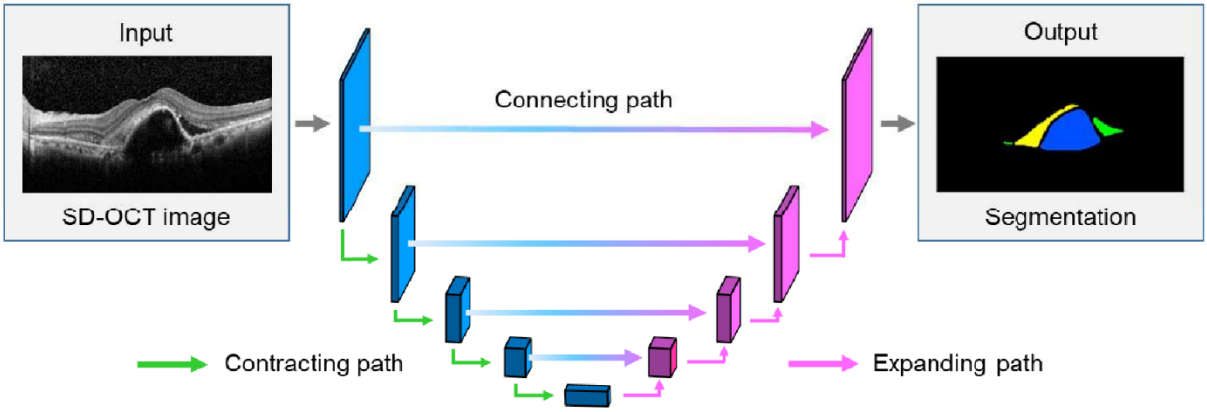

Fig. 4.1 CNN based on U-net. [84]

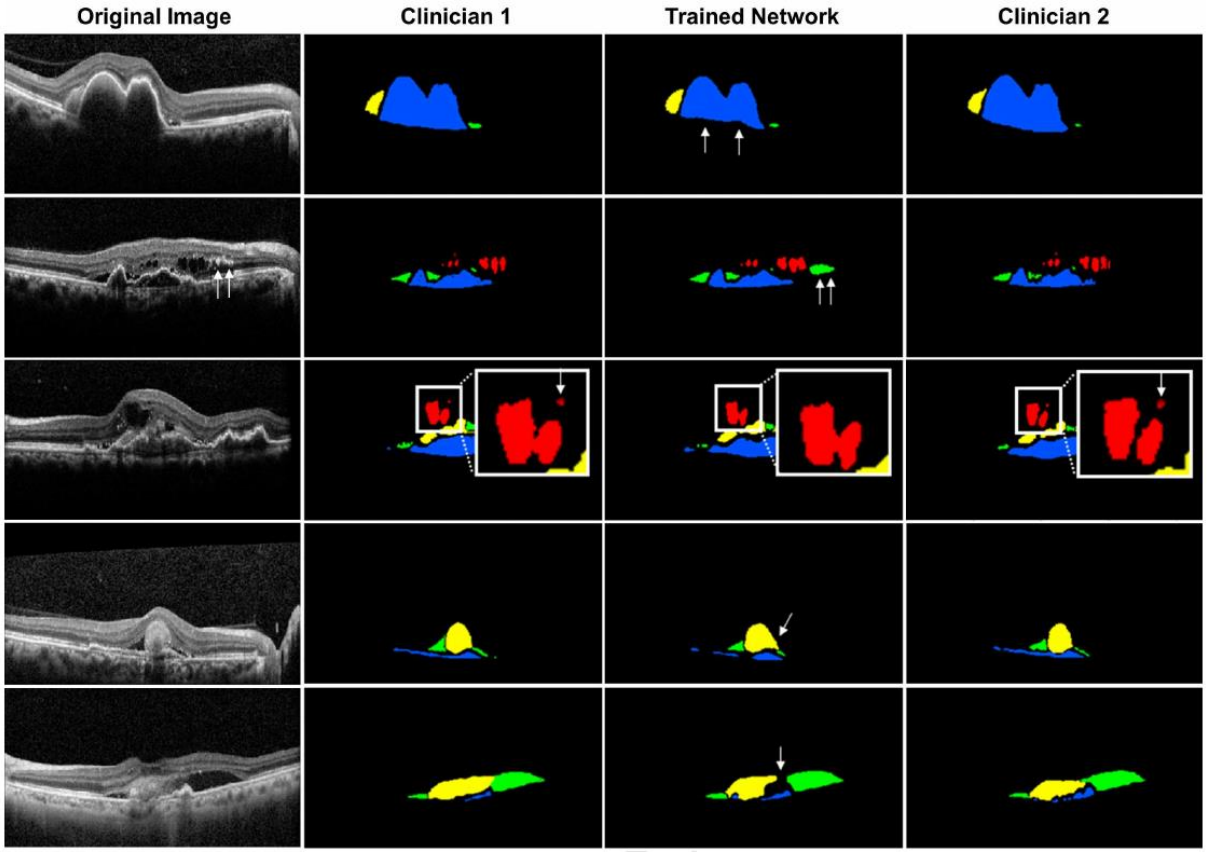

Fig. 4.2 Some examples of segmentation. IRF is marked with red color, SRF with green, PED with blue, and SHRM with yellow. The arrows indicate the areas in which the model errs during segmentation. [84]

Some examples of segmentation are shown in Fig. 4.2. IRF is marked with red color, SRF with green, PED with blue, and SHRM with yellow. The arrows indicate the areas in which the model makes a mistake during segmentation. Dice scores range from 0.78 to 0.82 which are quite high values. Unfortunately, neither code, nor database are publicly available, and therefore study could not be replicated.

4.2 Is joint segmentation of layers and fluids better and how much?

In the literature [45]–[50] is often stated that simultaneous segmentation of fluids and layers gives better results. As previously said, to detect IRF, SRF, and PED, the knowledge of their location within or outside specific retinal layers can be used to facilitate their detection and differentiation, or more specific: the IRF is a localized extracellular intraretinal fluid accumulation seen as a hyporeflective area located anywhere between ILM and ELM, the SRF is a subretinal fluid accumulation seen as a hyporeflective area between photoreceptor layer and RPE, and the PED is either a hyporeflective accumulation of fluid or hyperreflective accumulation of other material between RPE and BM. The segmentation of layers itself has its diagnostic value and although it is not only for the purpose of more accurate segmentation of fluids while deciding which layers to be segmented, one of the reasons was better localization of fluids.

Although it is intuitively clear that simultaneous segmentation should give better results, in this part of the research I wanted to examine how much joint segmentation is better in terms of obtaining more accurate fluid segmentation. To compare the segmentation of fluids in the case when only they are labeled (layers are considered as background) and, in the case, when joint segmentation of layers and fluids is done, both segmentations were done using the standard U-net architecture.

Original images size 1024x512 pixels were resized to 512x256 pixels. Categorical cross-entropy loss was used to train both models. The batch size was set to 4. The AdaBound optimizer [97] was used. K-fold cross-validation was used where each fold contains images from 4 patients. The models were trained on Google Colab [99] with a GPU. The models were implemented in Python, using the Keras library with the TensorFlow backend.

In Table 4.2 there are reported Dice scores (mean and standard deviation) for each fluid class in both cases. The same results are shown in Fig. 4.3 with bar graphs. It can be observed that higher values are obtained in case joint segmentation of layers and fluids were done.

Examples of annotated fluids (PED, SRF, and IRF) in both cases (joint segmentation of layers and fluids, and only fluids segmentation) are shown in Fig. 4.4 for situations with less

pronounced pathological changes and in Fig. 4.5 for situations with more pathological changes. In the case of minor pathological changes, no large difference in segmentation is noticed. When there are major pathological changes, simultaneous segmentation of fluids and layers contributes more to better segmentation of the fluids themselves.

Table 4.2 The Dice score (mean and standard deviation) in the case in which joint segmentation of fluids and layers was performed (results are shown only for fluids), and in the case in which only segmentation of fluids was performed.

	PED	SRF	IRF
Joint segmentation	0.638 (0.173)	0.513 (0.294)	0.480 (0.252)
Just fluids	0.594 (0.166)	0.439 (0.291)	0.414 (0.221)

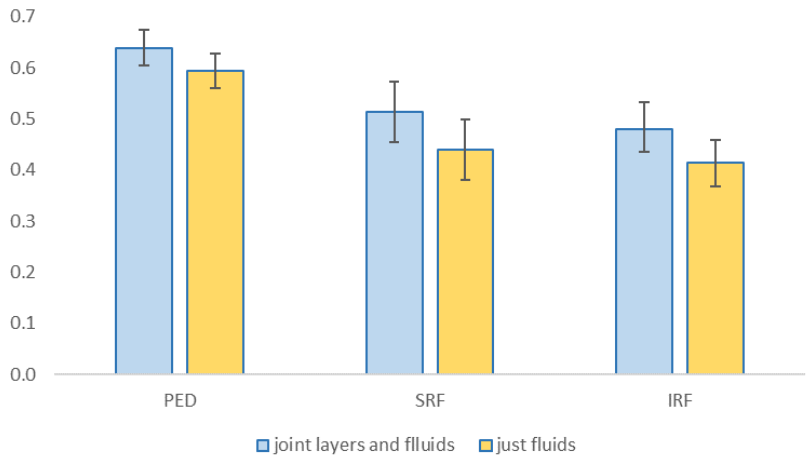

Fig. 4.3 The Dice scores (mean and standard error of the mean) for fluids segmentation in case joint segmentation were performed and only fluids segmentation was performed.

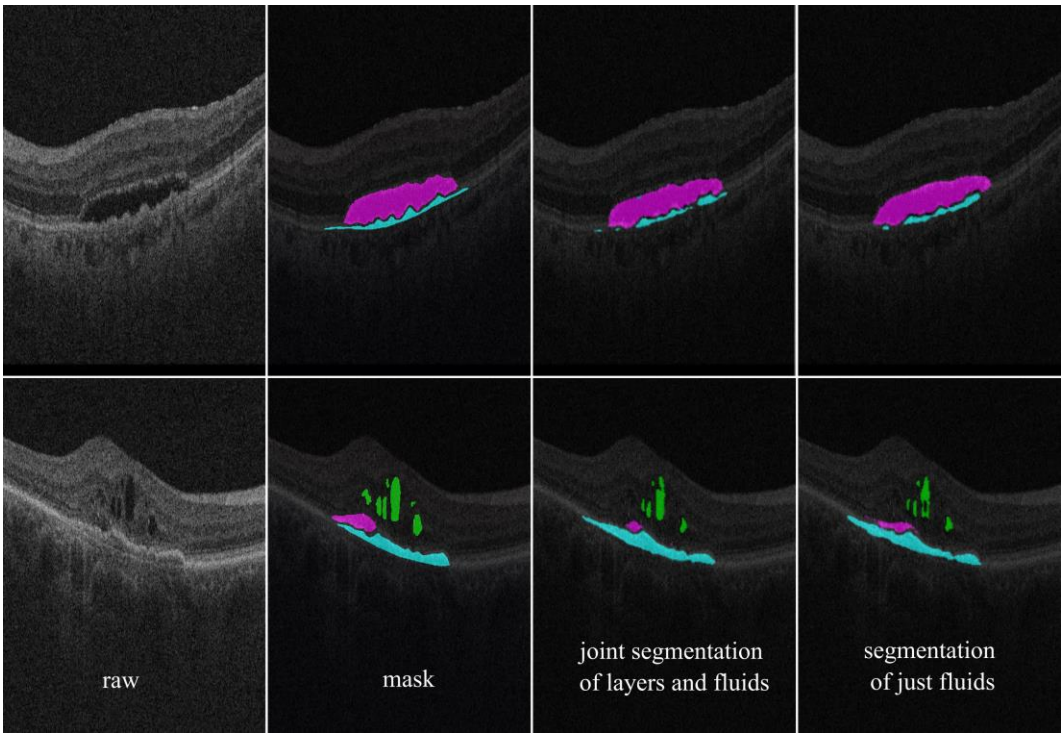

Fig. 4.4 A case with less pronounced pathological changes: examples of annotated fluids (PED, SRF, and IRF) in case joint segmentation of layers and fluids were performed and only fluids segmentation was performed. Images are cropped and only the ROI is visible.

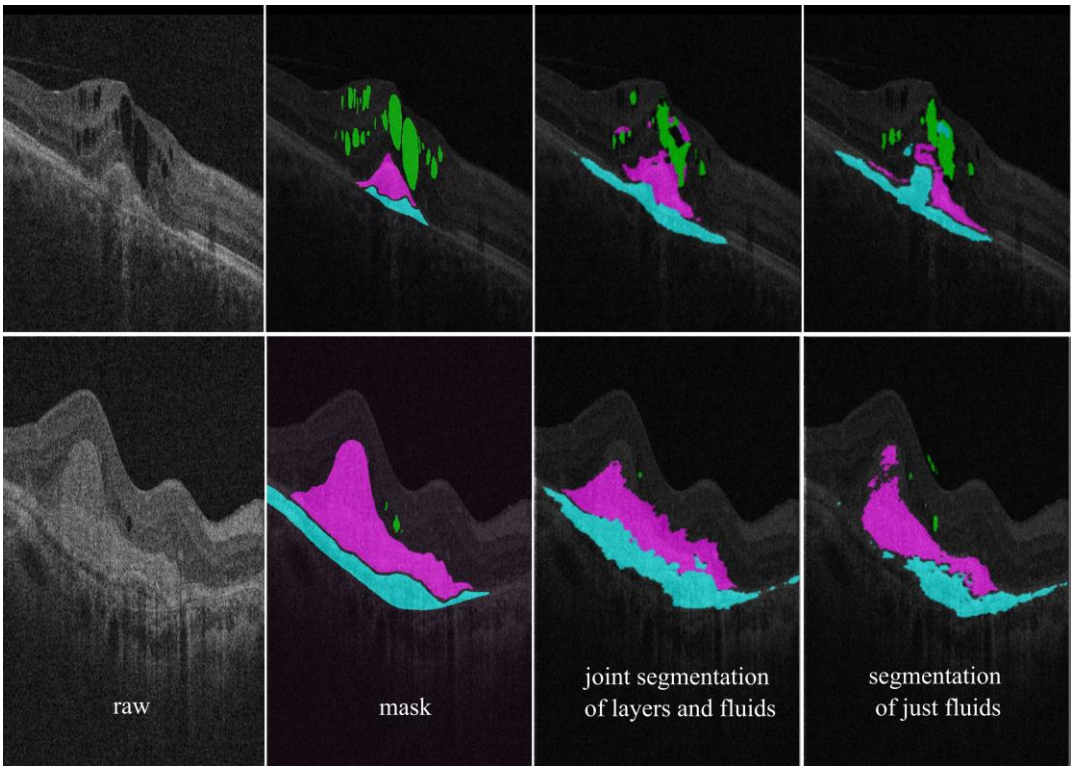

Fig. 4.5 A case with more pronounced pathological changes: examples of annotated fluids (PED, SRF, and IRF) in case joint segmentation of layers and fluids were performed and only fluids segmentation was performed. Images are cropped and only the ROI is visible.

4.3 Method for retinal fluids segmentation with transfer learning

4.3.1 Transfer learning for medical images

In recent years, the segmentation of medical and thus ophthalmic images is performed almost exclusively using deep learning methods. One of the preconditions for the success of deep learning algorithms is the large amount of data available. But in medicine, this is rarely the case. In the case of supervised learning, images manually annotated by an expert are required, which is very time-consuming. Even more, image labeling is not part of clinical practice, and labeling is done solely to develop machine learning methods. As images from different OCT devices and different device generations vary, it is required that the segmentation algorithms be as robust as possible. This means that in case of migration to a new device it is not necessary to re-annotate a larger number of images, and existing algorithms will be upgraded with a smaller number of new images.

Although transfer learning is often applied in medicine, by using the known architectures of neural networks (Xception [112], Inception V3 [113], ResNet50 [114], MobileNet [115], VGG16, and VGG19 [116]) pretrained on the ImageNet dataset, in ophthalmology this is rarely the case [23]. ImageNet is a database with over 14 million natural images. Intuitively, it is not clear how would transfer learning help in the case of medical images which are often grayscale and dissimilar to natural images. Several papers [24]–[26] examine the adequacy of such a method since the ImageNet dataset is a database of natural images that bears no resemblance to medical ones. In a recent paper [27], a hybrid architecture was proposed in which only the first few layers of networks such as the ResNet architecture (pretrained on ImageNet) would be used and then some lighter versions of CNN would be applied (and fine-tuned) in the end. Optimizing transfer learning (with model pretrained on ImageNet) is still an active area of research in the case of medical images.

4.3.2 How well do algorithms generalize or how transferable are the data?

The aim of this part of the research was twofold. First, to see how the performance of the algorithm depends on the amount of data. Since medical labeled images are difficult to collect, I found it useful to get an insight into the performance dependence of the algorithm on

the number of images. To begin with, the network was trained from scratch with a different amount of data to see how performance varies. In order to perform the segmentation of retinal structures and pathological biomarkers, the AROI database (described in chapter 2.2.) was used. And second, it was explored how many additional labeled images are needed and how much transfer learning can help in the case the network was trained on one dataset (source data) and then migrated to another (target data). As a source dataset, the image database from the RETOUCH challenge (The Retinal OCT Fluid Detection and Segmentation Benchmark and Challenge) was used [55]. Transfer learning was then performed so that the network was trained with part of the images from RETOUCH challenge and then fine-tuned with the images of the target database: again, with a different number of images.

The AROI database of manually annotated images was used and for convenience, let it be called the target dataset. The other dataset was from the RETOUCH challenge [55] – let it be called the source dataset. The database consists of images from three types of devices (Topcon, Spectralis, and Cirrus). For each type of device, all B-scans for 24 patients were labeled. Only part of the images from the Cirrus device (1000 randomly picked images) was used to train the network. Unlike the target dataset, where all B-scans contained intraretinal fluids, many B-scans did not contain pathological changes (whole B-scans belonged to the background class). It resulted in an even greater disproportion of background pixels and pixels belonging to fluids. The database included images of patients with AMD, retinal vein occlusion (RVO), and diabetic macular edema (DME). More details can be found on the website [55].

Segmentation was done using the U-net architecture [21] as it is almost the standard in the segmentation of medical images. To make it easier to see the dependence of the algorithm's performance on the amount of data, the problem was reduced to binary segmentation (all three fluids belonged to the same class). Data augmentation, which is usually used in the case of a small amount of data, had not been performed, in order to better isolate the dependence of network performance on the amount of data.

In the first case, the target dataset was divided into a training set (1000 images) and a test set (136 images: patient03, patient21, patient22, patient23, patient24). Furthermore, the training set was divided into three groups with a different number of images: 256, 512, and 1000. The network was trained from scratch. Dice loss was used to train the model, since the

dice score is usually used as a metric for evaluation, and the goal was to maximize it. The model was also trained with binary cross-entropy loss, but worse results were obtained. The batch size was set to 4. Adam optimizer [82] with a learning rate 1e-5 was used. The model had 7759521 trainable parameters. The training was performed during 70 epochs, without early stopping. Dropout was used to prevent overfitting. The validation dataset was 10% of the training dataset.

In the second case, the network training was performed with the source dataset. The same hyperparameters were used. The network had only been trained for 20 epochs, as the purpose was only to have a pretrained network. Fine-tuning of all layers with the same three data groups from the target dataset (256, 512, and 1000) was then performed. The training was executed during 50 epochs with all the same parameters as in the first case.

In the first case, when the model is trained from scratch, as expected, the performance is better due to the larger number of images (it is well known that deep learning algorithms are "data-hungry"). Obtained dice scores (mean ± SEM) on a test dataset (224 images) for 256, 512, and 1000 images are: 0.579±0.272, 0.632±0.279, and 0.731±0.268, respectively.

Fig. 4.6 shows examples for one raw image and ground truth mask, followed by model predictions on the test set in case the model is trained with 256, 512, and 1000 training images. As it is as expected, segmentations are better with a larger amount of data. With an experiment like this, we can get a better insight into how much data we need, since there is always a trade-off between the amount of manually annotated data, we can/want to collect and the performance of algorithms. The second row shows segmentation predictions for the second case (with the model pretrained on the source dataset and then fine-tuned on the target dataset). In this case, obtained dice scores (mean ± SEM) on the test dataset were 0.612±0.284, 0.677±0.272, and 0.826±0.270 respectively. It is evident that predictions after transfer learning have more "serrated" shapes while predictions from a model trained from scratch are more oval. As shape types are one of the key features in distinguishing different kinds of intraretinal fluids, this is a promising result.

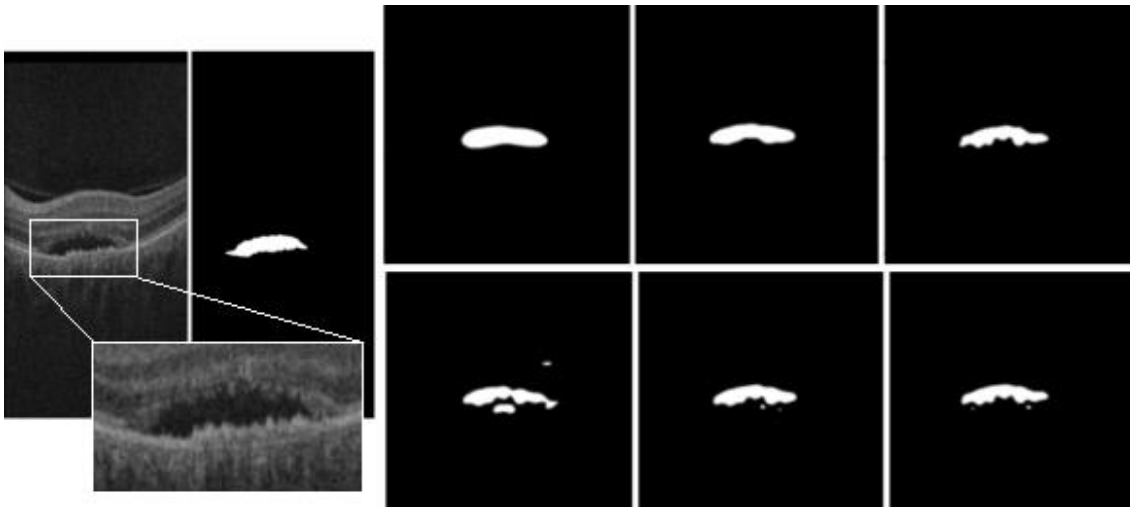

Fig. 4.6 First row (from left to right): raw image from test dataset, ground truth (manual annotation), prediction from a network trained with 256 images, 512 images, and 1000 images. Second row (from left to right): predictions after training the network with 1000 images from source dataset and fine-tuning pretrained network with 256 images, 512 images, 1000 images (from left to right) from target data. [117]

Fig. 4.7 shows a graph with mean dice scores for all cases. It is visible that the mean value of the dice score is higher in the second case (with transfer learning).

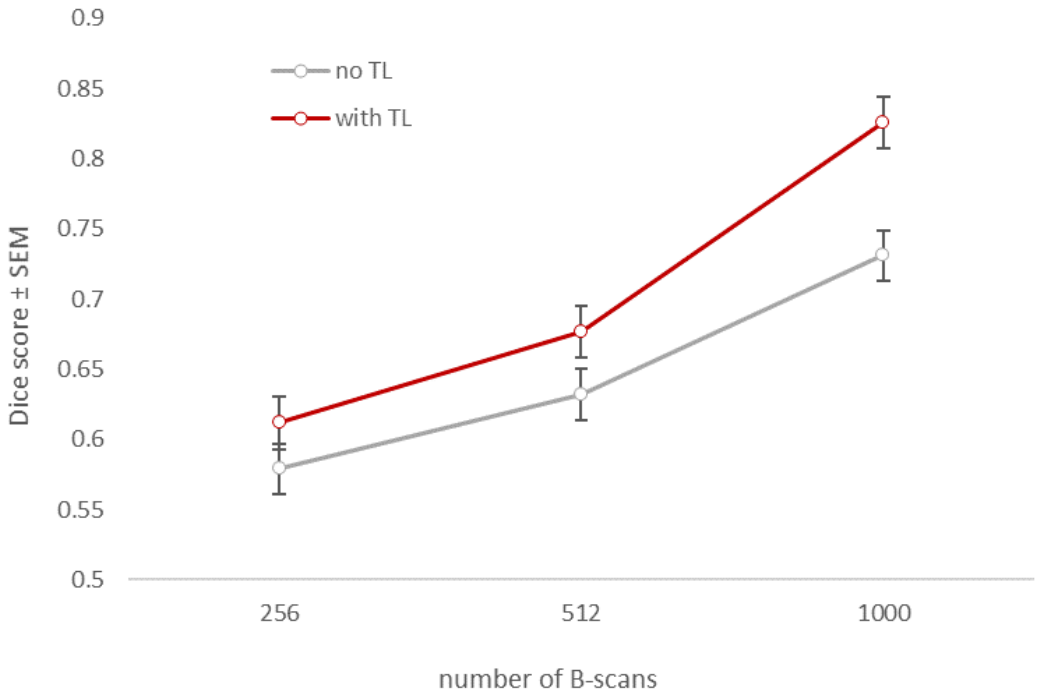

Fig. 4.7 Mean dice score with error bars (standard error of the mean – SEM) for training model from scratch (no TL) and in case of fine-tuning the pretrained model (with TL). [117]

Although the model trained with 1000 images from the source dataset achieved a high value for the Dice score on the training test (0.717), the value of the dice score is very low for the test set (0.348). This means that in the case of migration to a new OCT device, it would still be necessary to additionally train the model, but with a smaller number of images than in the case of training the model from scratch. The reason is probably that both datasets cannot be considered to fit the same distributions despite the same type of OCT device being used in both cases. One of the reasons is the differences in annotations (we have seen that there is quite a significant inconsistency between expert opinions as is described in chapter 2.2.3). Another reason may be that the source dataset contains images of patients with three types of disease, and the target dataset only contains patients with nAMD. Therefore, we have a case of a model trained on data from one distribution and tested on a different image distribution.

By visualizing convolutional filters, it becomes visible what the network is actually learning. Fig. 4.8 shows a visualization of convolution layers in the case of a training model from scratch with 256 images (first row). Then for the pretrained model (second row), and after fine-tuning the pretrained model with 256 images (third row). The first three layers belong to the encoder and the last three layers to the decoder. It is visible that in the third row, convolution layers appear to be a combination of convolution layers shown in the first two rows, as might be expected. It is evident that convolution layers are clearer and more complete after fine-tuning with additional 256 target images than after only training with 1000 source images.

It has been demonstrated that the initial layers, which are usually similar to Gabor filters, are not clear even after transfer learning when the source database is similar. It can be concluded that using images from different distributions leads to better results than increasing the number of images from the same distribution. However, the generality of this finding requires further investigation.

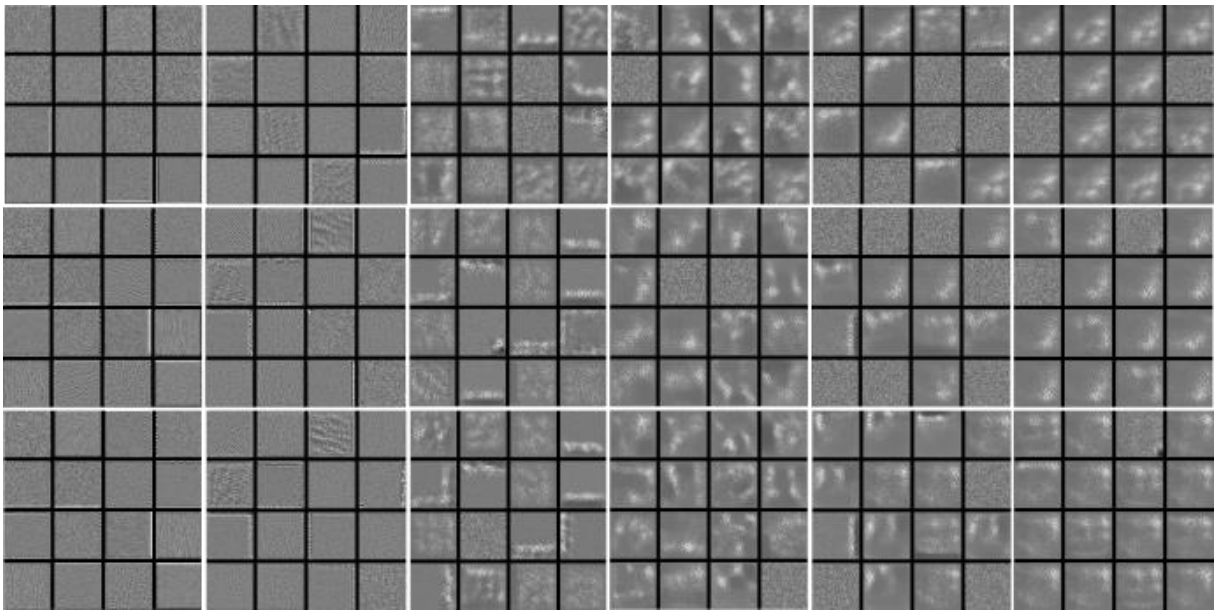

Fig. 4.8 Visualization of convolution layers 2, 4, 7, 10, 13, 16 (from left to right). Layers 2, 4, and 7 belong to the encoder part of U-net architecture, and 10, 13, and 16 to the decoder part. First row: network trained from a scratch with 256 images (target data). Second row: network trained with 1000 images (source data). Third row: network fine-tuned with target data (256 images) after learning with source data (1000 images). [117]

4.4 Fluid segmentation – should we collect more data with just fluids labeled or perform joint layer and fluid segmentation on a smaller amount of data?

In chapter 4.2 it was shown that the joint segmentation of fluids and layers gives better results than fluids alone. In case only fluids segmentation is required (as it is for some diagnostic purposes), the question arises whether it is better to collect as many labeled B-scans with labeled fluids as possible, or it is better to limit oneself to a smaller number of patients but to do labeling of layers and fluids at the same time. For this purpose, transfer learning was performed, where the source dataset is again the image database from the RETOUCH challenge, with all B-scans being taken this time (24 patients or 3072 B-scans). The target database was the AROI database (1336 B-scans with annotated fluids, layers were considered as background). After training the network on the source dataset, fine-tuning was done on the target dataset.

The training was performed with all the same parameters as in chapter 4.2: the U-net architecture was used for segmentation. Original images size 1024x512 pixels were resized to 512x256 pixels. Categorical cross-entropy loss was used to train both models. The batch size

was set to 4. The AdaBound optimizer [97] was used. K-fold cross-validation was used where each fold contains images from 4 patients. The models were trained on Google Colab [99] with a GPU. The models were implemented in Python, using the Keras library with the TensorFlow backend.

In Table 4.3 there are reported Dice scores (mean and standard deviation) for each fluid class and for both cases, i.e. in the case when the network was trained only on the AROI database (marked with 'No TL') and in the case when it was first trained on the RETOUCH database and then fine-tuning was done on the AROI database (marked as 'With TL'). Additionally, in the third row are presented results in case simultaneous segmentation of fluids and layers were done. It can be observed that higher values are obtained in case joint segmentation of layers and fluids is done than in case transfer learning was performed. It would be too ambitious to conclude that is a general rule but to be very careful and thoughtful when deciding which data to collect. The same results are shown in Fig. 4.9 with bar graphs.

Examples of annotated fluids (PED, SRF, and IRF) in both cases (with transfer learning and without transfer learning) are shown in Fig. 4.10 for situations with less pronounced pathological changes and in Fig. 4.11 for situations with more pathological changes. In the case of minor pathological changes, no large difference in segmentation is seen. When there are major pathological changes, transfer learning contributes more to better segmentation of the fluids. It can be concluded that transfer learning functions similar to joint segmentation of fluids and layers in terms of improving the accuracy of the segmentation of the fluids themselves.

Table 4.3 The Dice score (mean and standard deviation) in the case with no TL, with TL and in which joint segmentation of fluids and layers was performed (results are shown only for fluids).

	PED	SRF	IRF
No TL	0.594 (0.166)	0.439 (0.291)	0.414 (0.221)
with TL	0.622 (0.173)	0.508 (0.273)	0.448 (0.246)
Joint segmentation	**0.638** (0.173)	**0.513** (0.294)	**0.480** (0.252)

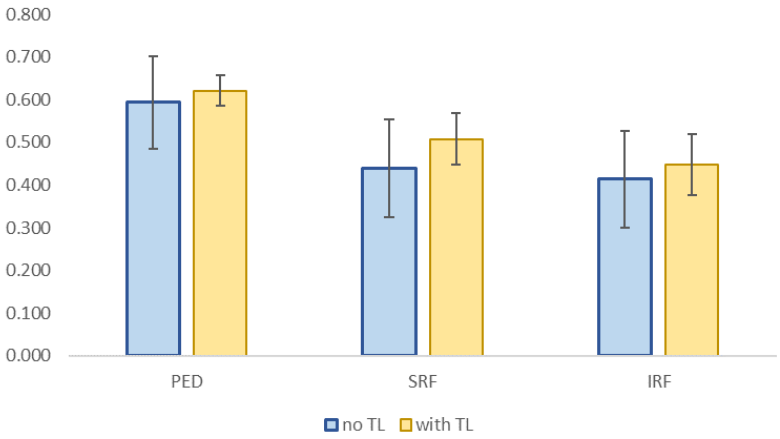

Fig. 4.9 The Dice scores (mean and standard error of the mean) for fluids segmentation in case with no TL and with TL.

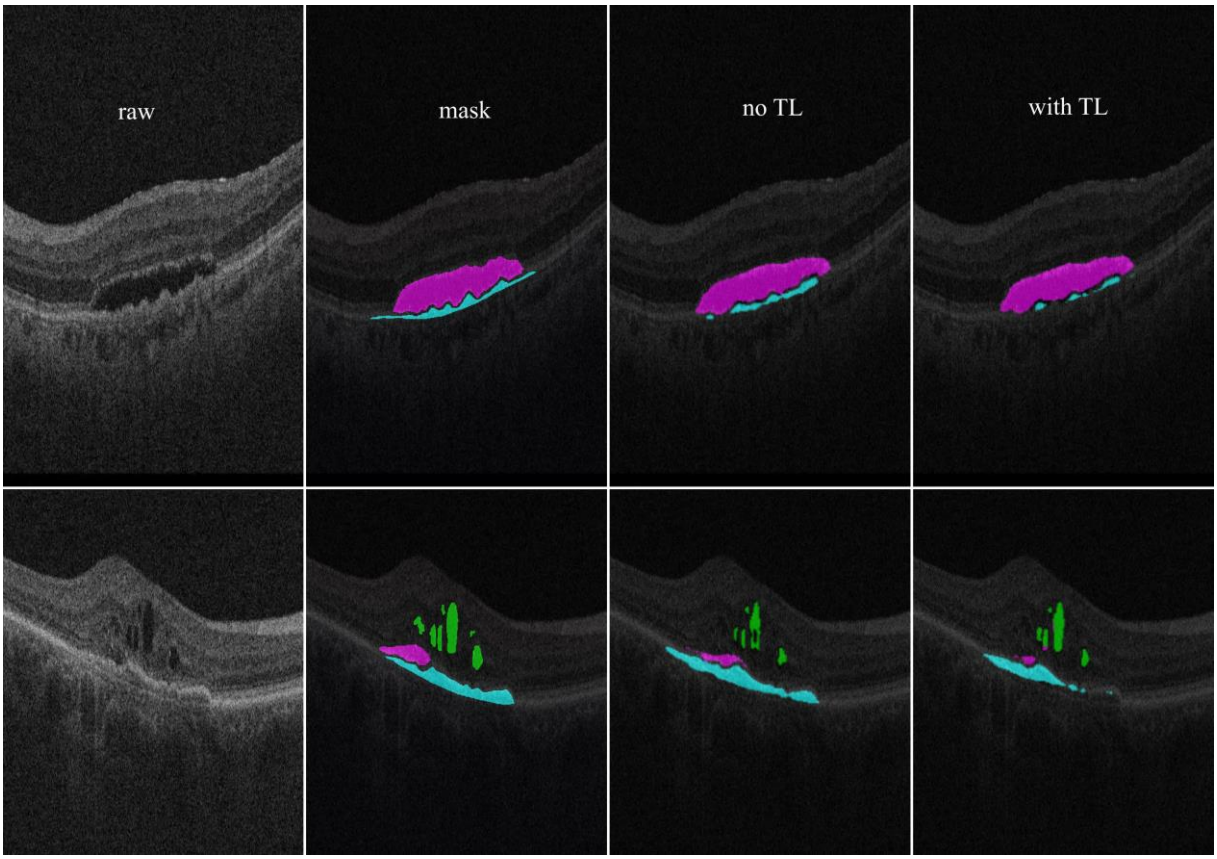

Fig. 4.10 A case with less pronounced pathological changes: examples of annotated fluids (PED, SRF, and IRF) in case with no TL and with TL.

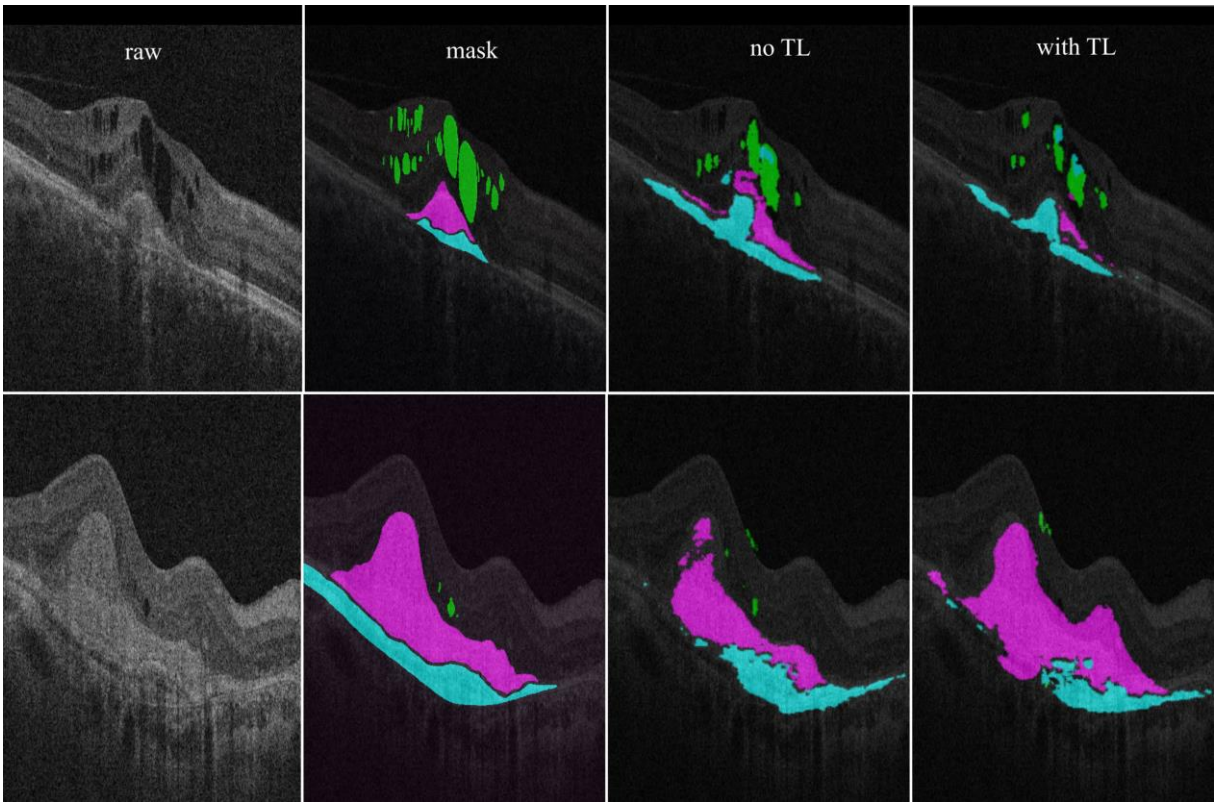

Fig. 4.11 A case with more pathological changes: examples of annotated fluids (PED, SRF, and IRF) in case with no TL and with TL. When the structure of the retina is significantly disturbed, segmentation with TL gives visibly better results. First row: although both segmentations are quite poor, segmentation with TL is more consistent, it is especially visible around PED and SRF. Second row: in the case with TL segmentation of SRF is visibly better and whole segmentation is more coherent.

5 CONCLUSION

Given the lack of publicly available databases of retinal OCT images, within this thesis work, the materials for the AROI database were collected in cooperation with the Department of Ophthalmology (Sestre milosrdnice University Hospital Center, Zagreb). The AROI database contains a large sample of 1136 B-scans with exhaustive annotations (both fluids and retinal layers were annotated by an expert ophthalmologist). As images are collected from patients suffering from nAMD where pathologic biomarkers and large distortion of the retinal structure are present, automatic segmentation of such images presents a significant challenge. In addition, the images have a significant amount of noise (quality is worse than in images acquired from some other OCT devices), there is a huge class imbalance, and 1k images are still a small data regime – all this makes automatic segmentation extremely difficult and challenging.

We have limited ourselves to collecting images from only one type of OCT device and from patients suffering from one type of disease. However, developing algorithms on as large and as diverse databases as possible would provide for more robust algorithms that could be implemented in commercial OCT device software. Hopefully, open access will become a common practice for the majority of research groups in the future and online image collections and repositories will contribute to building a single database covering various diseases, various types of devices, and various retinal structures annotated from different experts.

To enable the validation of algorithms for automatic segmentation, intra- and inter-observer errors were calculated. However, it is still not clear what level of segmentation accuracy we need as manual segmentation is rarely performed in clinical practice. There is probably no universal rule, and the required accuracy of segmentation will depend on the purpose – whether it is a diagnosis or prediction of the outcome of anti-VEGF therapy or prediction of another eye disease.

After building the database, and establishing the framework for evaluation, automatic segmentation of retinal layers and fluids was explored in the second part of the research, in as systematic way as possible. First, I investigated how different architectures affected

segmentation accuracy. These results of automatic segmentation for three models (standard U-net architecture and two state-of-the-art architectures) can serve as a baseline for the further development of deep learning models. It is shown that enhanced architectures do not contribute much or almost at all to segmentation accuracy. This can be explained by the fact that more advanced architectures still do not solve the biggest challenges: huge class imbalance, poor quality of images, and most of all, the necessity of expert ophthalmic knowledge for proper segmentation.

Furthermore, it was investigated how data manipulation affects segmentation performance and how automatic segmentation can be improved with data augmentation, transfer learning, and post-processing. It is shown that all these methods can boost performance but within limits. Exploration was performed in an orthogonal way, which means that there has always been one change over the data. Although an entirely orthogonal way is not possible, as there are too many interdependencies in each DL model, all these results still can serve as guidelines on how to improve results and what works and what does not.

The question that arises is why state-of-the-art OCT devices still lack abilities for good automatic segmentation despite numerous published papers in recent years which report very good results. From my perspective, the answer can be broken down into several aspects. First, most published papers are evaluated on publicly unavailable data. If data and code are not openly available, results are not repeatable/reproducible. All models that are evaluated on some small dataset without enough diversity will not be applicable in clinical practice as they will not generalize well. As a result, we have a field crowded with published papers, with no possibility for a real comparison of models, with a reproducibility crisis, and with a lack of open-access datasets.

Large amounts of data and raising computing power are the two main pillars on which advances in computer vision rest. Large databases like ImageNet were one of the main prerequisites for continually developing new and enhanced models and thus enabling advances and breakthroughs in computer vision. Unfortunately, in ophthalmology, one of the two main pillars is missing, and we are faced with something like the curse of small data. That is, at least as long as most DL algorithms rely on supervised, or semi-supervised methods. One way would be to bypass a large amount of data by adding ophthalmic knowledge as

automatic segmentation models often make mistakes where pattern recognition is not sufficient but also additional knowledge is needed. For a long time, my intuition was that this was the only way, but it was probably just a human bias – "machines should do it in a human way". Given the complexity of ophthalmic knowledge, and that it can be hardly presented by symbolic logic, and that there is no good way to incorporate symbolic logic with DL models, it does not seem a very probable solution. The other solution that seems more probable is by generating synthetic data which can be obtained by simulating the retina and generating images by simulating the mechanisms of the disease and the way in which they deform the retina, and simultaneously adding pathological structures. Nowadays, it is usually the case that much of the training of the neural networks is done with simulated data or in simulated environments – the biggest advances in AI that we have had the opportunity to witness so far are the result of simulated data and training neural networks in simulated (virtual) reality. Unfortunately, that part was not feasible within this thesis work, but hopefully, it will be done in the future.

BIBLIOGRAPHY

[1] R. D. Jager, W. F. Mieler, and J. W. Miller, "Age-related macular degeneration," *New England Journal of Medicine*, vol. 358, no. 24, pp. 2606–2617, 2008.

[2] W. L. Wong *et al.*, "Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis," *The Lancet Global Health*, vol. 2, no. 2, pp. e106–e116, Feb. 2014, doi: 10.1016/S2214-109X(13)70145-1.

[3] R. Klein, B. E. K. Klein, and K. L. P. Linton, "Prevalence of Age-related Maculopathy," *Ophthalmology*, vol. 99, no. 6, pp. 933–943, Jun. 1992, doi: 10.1016/S0161-6420(92)31871-8.

[4] H. R. Coleman, C.-C. Chan, F. L. Ferris, and E. Y. Chew, "Age-related macular degeneration," *The Lancet*, vol. 372, no. 9652, pp. 1835–1845, Nov. 2008, doi: 10.1016/S0140-6736(08)61759-6.

[5] P. Kaszubski, T. Ben Ami, C. Saade, and R. T. Smith, "Geographic Atrophy and Choroidal Neovascularization in the Same Eye: A Review," *Ophthalmic Res*, vol. 55, no. 4, pp. 185–193, 2016, doi: 10.1159/000443209.

[6] H. R. Patel and D. Eichenbaum, "Geographic atrophy: clinical impact and emerging treatments," *Ophthalmic Surg Lasers Imaging Retina*, vol. 46, no. 1, pp. 8–13, Jan. 2015, doi: 10.3928/23258160-20150101-01.

[7] D. Brown *et al.*, "Current Best Clinical Practices—Management of Neovascular AMD," *Journal of VitreoRetinal Diseases*, vol. 1, no. 5, pp. 294–297, Sep. 2017, doi: 10.1177/2474126417725946.

[8] J. M. Schmitt, "Optical coherence tomography (OCT): a review," *Selected Topics in Quantum Electronics, IEEE Journal of*, vol. 5, no. 4, pp. 1205–1215, 1999.

[9] F. G. Venhuizen *et al.*, "Automated Staging of Age-Related Macular Degeneration Using Optical Coherence Tomography," *Invest. Ophthalmol. Vis. Sci.*, vol. 58, no. 4, pp. 2318–2328, Apr. 2017, doi: 10.1167/iovs.16-20541.

[10] M. Melinščak, M. Radmilović, Z. Vatavuk, and S. Lončarić, "Annotated retinal optical coherence tomography images (AROI) database for joint retinal layer and fluid segmentation," *Automatika*, vol. 62, no. 3, pp. 375–385, Jul. 2021, doi: 10.1080/00051144.2021.1973298.

[11] C. E. Mendoza-Santiesteban, I. Gabilondo, J. A. Palma, L. Norcliffe-Kaufmann, and H. Kaufmann, "The Retina in Multiple System Atrophy: Systematic Review and Meta-Analysis," *Front. Neurol.*, vol. 8, p. 206, May 2017, doi: 10.3389/fneur.2017.00206.

[12] D. Huang *et al.*, "Optical coherence tomography," *Science*, vol. 254, no. 5035, pp. 1178–1181, Nov. 1991, doi: 10.1126/science.1957169.

[13] A. F. Fercher, “Optical coherence tomography,” *Journal of Biomedical Optics*, vol. 1, no. 2, pp. 157–173, 1996.

[14] W. Drexler and J. Fujimoto, “State-of-the-art retinal optical coherence tomography,” *Progress in Retinal and Eye Research*, vol. 27, no. 1, pp. 45–88, Jan. 2008, doi: 10.1016/j.preteyeres.2007.07.005.

[15] A. Mishra, A. Wong, K. Bizheva, and D. A. Clausi, “Intra-retinal layer segmentation in optical coherence tomography images,” *Opt. Express*, vol. 17, no. 26, pp. 23719–23728, Dec. 2009, doi: 10.1364/OE.17.023719.

[16] R. Kafieh, H. Rabbani, and S. Kermani, “A Review of Algorithms for Segmentation of Optical Coherence Tomography from Retina,” *J Med Signals Sens*, vol. 3, no. 1, pp. 45–60, 2013.

[17] M. E. J. van Velthoven, D. J. Faber, F. D. Verbraak, T. G. van Leeuwen, and M. D. de Smet, “Recent developments in optical coherence tomography for imaging the retina,” *Progress in Retinal and Eye Research*, vol. 26, no. 1, pp. 57–77, Jan. 2007, doi: 10.1016/j.preteyeres.2006.10.002.

[18] M. D. Abràmoff, M. K. Garvin, and M. Sonka, “Retinal Imaging and Image Analysis,” *IEEE Reviews in Biomedical Engineering*, vol. 3, pp. 169–208, 2010, doi: 10.1109/RBME.2010.2084567.

[19] M. Choma, M. Sarunic, C. Yang, and J. Izatt, “Sensitivity advantage of swept source and Fourier domain optical coherence tomography,” *Optics Express*, vol. 11, no. 18, pp. 2183–2189, 2003.

[20] K. M. Lee, “Segmentations of the intraretinal surfaces, optic disc and retinal blood vessels in 3D-OCT scans,” *Theses and Dissertations*, p. 247, 2009.

[21] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” *arXiv:1505.04597 [cs]*, May 2015, [Online]. Available: http://arxiv.org/abs/1505.04597

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” p. 8.

[23] G. Litjens *et al.*, “A survey on deep learning in medical image analysis,” *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[24] K. He, R. Girshick, and P. Dollár, “Rethinking ImageNet Pre-training,” *arXiv:1811.08883 [cs]*, Nov. 2018, Accessed: Jan. 07, 2020. [Online]. Available: http://arxiv.org/abs/1811.08883

[25] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?,” *arXiv:1411.1792 [cs]*, Nov. 2014, Accessed: Jan. 07, 2020. [Online]. Available: http://arxiv.org/abs/1411.1792

[26] J. Lee, P. Sattigeri, and G. Wornell, “Learning New Tricks From Old Dogs: Multi-Source Transfer Learning From Pre-Trained Networks,” p. 11.

[27] M. Raghu, J. Kleinberg, C. Zhang, and S. Bengio, “Transfusion: Understanding Transfer Learning for Medical Imaging,” p. 11.

[28] R. F. Spaide *et al.*, “Consensus Nomenclature for Reporting Neovascular Age-Related Macular Degeneration Data,” *Ophthalmology*, vol. 127, no. 5, pp. 616–636, May 2020, doi: 10.1016/j.ophtha.2019.11.004.

[29] U. Schmidt-Erfurth and S. M. Waldstein, “A paradigm shift in imaging biomarkers in neovascular age-related macular degeneration,” *Progress in Retinal and Eye Research*, vol. 50, pp. 1–24, Jan. 2016, doi: 10.1016/j.preteyeres.2015.07.007.

[30] M. Ritter *et al.*, “Intraretinal cysts are the most relevant prognostic biomarker in neovascular age-related macular degeneration independent of the therapeutic strategy,” *Br J Ophthalmol*, vol. 98, no. 12, pp. 1629–1635, Dec. 2014, doi: 10.1136/bjophthalmol-2014-305186.

[31] T. Ristau *et al.*, “Relationship between Visual Acuity and Spectral Domain Optical Coherence Tomography Retinal Parameters in Neovascular Age-Related Macular Degeneration,” *Ophthalmologica*, vol. 231, no. 1, pp. 37–44, Oct. 2013, doi: 10.1159/000354551.

[32] A. S. Willoughby *et al.*, “Subretinal Hyperreflective Material in the Comparison of Age-Related Macular Degeneration Treatments Trials,” *Ophthalmology*, vol. 122, no. 9, pp. 1846-1853.e5, Sep. 2015, doi: 10.1016/j.ophtha.2015.05.042.

[33] S. M. Waldstein *et al.*, “Correlation of 3-Dimensionally Quantified Intraretinal and Subretinal Fluid With Visual Acuity in Neovascular Age-Related Macular Degeneration,” *JAMA Ophthalmol*, vol. 134, no. 2, p. 182, Feb. 2016, doi: 10.1001/jamaophthalmol.2015.4948.

[34] H. Lee, A. Jo, and H. C. Kim, “Three-Dimensional Analysis of Morphologic Changes and Visual Outcomes in Neovascular Age-Related Macular Degeneration,” *Invest. Ophthalmol. Vis. Sci.*, vol. 58, no. 2, p. 1337, Feb. 2017, doi: 10.1167/iovs.16-20637.

[35] H. Bogunović *et al.*, “Prediction of Anti-VEGF Treatment Requirements in Neovascular AMD Using a Machine Learning Approach,” *Invest. Ophthalmol. Vis. Sci.*, vol. 58, no. 7, pp. 3240–3248, Jun. 2017, doi: 10.1167/iovs.16-21053.

[36] A. P. Adamis, C. J. Brittain, A. Dandekar, and J. J. Hopkins, “Building on the success of anti-vascular endothelial growth factor therapy: a vision for the next decade,” *Eye*, Jun. 2020, doi: 10.1038/s41433-020-0895-z.

[37] A. Sheyman and A. A. Fawzi, Eds., *Retinal Vascular Disease*. Singapore: Springer Singapore, 2020. doi: 10.1007/978-981-15-4075-2.

[38] Y. Peng *et al.*, “Predicting risk of late age-related macular degeneration using deep learning,” *arXiv:2007.09550 [cs, eess]*, Jul. 2020, Accessed: Jul. 25, 2020. [Online]. Available: http://arxiv.org/abs/2007.09550

[39] J. Yim *et al.*, “Predicting conversion to wet age-related macular degeneration using deep learning,” *Nat Med*, May 2020, doi: 10.1038/s41591-020-0867-7.

[40] J. Siedlecki *et al.*, “Impact of Sub-Retinal Fluid on the Long-Term Incidence of Macular Atrophy in Neovascular Age-related Macular Degeneration under Treat & Extend Anti-Vascular Endothelial Growth Factor Inhibitors,” *Sci Rep*, vol. 10, no. 1, p. 8036, Dec. 2020, doi: 10.1038/s41598-020-64901-9.

[41] K. T. Kim, H. Lee, J. Y. Kim, S. Lee, J. B. Chae, and D. Y. Kim, “Long-Term Visual/Anatomic Outcome in Patients with Fovea-Involving Fibrovascular Pigment Epithelium Detachment Presenting Choroidal Neovascularization on Optical Coherence Tomography Angiography,” *JCM*, vol. 9, no. 6, p. 1863, Jun. 2020, doi: 10.3390/jcm9061863.

[42] U. Schmidt-Erfurth, W.-D. Vogl, L. M. Jampol, and H. Bogunović, “Application of Automated Quantification of Fluid Volumes to Anti–VEGF Therapy of Neovascular Age-Related Macular Degeneration,” *Ophthalmology*, p. S0161642020302682, Mar. 2020, doi: 10.1016/j.ophtha.2020.03.010.

[43] S. Sadda *et al.*, “Errors in Retinal Thickness Measurements Obtained by Optical Coherence Tomography,” *Ophthalmology*, vol. 113, no. 2, pp. 285–293, Feb. 2006, doi: 10.1016/j.ophtha.2005.10.005.

[44] S. M. Waldstein, B. S. Gerendas, A. Montuoro, C. Simader, and U. Schmidt-Erfurth, “Quantitative comparison of macular segmentation performance using identical retinal regions across multiple spectral-domain optical coherence tomography instruments,” *Br J Ophthalmol*, vol. 99, no. 6, pp. 794–800, Jun. 2015, doi: 10.1136/bjophthalmol-2014-305573.

[45] A. Montuoro, S. M. Waldstein, B. S. Gerendas, U. Schmidt-Erfurth, and H. Bogunović, “Joint retinal layer and fluid segmentation in OCT scans of eyes with severe macular edema using unsupervised representation and auto-context,” *Biomedical Optics Express*, vol. 8, no. 3, p. 1874, Mar. 2017, doi: 10.1364/BOE.8.001874.

[46] A. G. Roy *et al.*, “ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks,” *Biomedical Optics Express*, vol. 8, no. 8, p. 3627, Aug. 2017, doi: 10.1364/BOE.8.003627.

[47] H. Wei and P. Peng, “The Segmentation of Retinal Layer and Fluid in SD-OCT Images Using Mutex Dice Loss Based Fully Convolutional Networks,” *IEEE Access*, vol. 8, pp. 60929–60939, 2020, doi: 10.1109/ACCESS.2020.2983818.

[48] J. De Fauw *et al.*, “Clinically applicable deep learning for diagnosis and referral in retinal disease,” *Nat Med*, vol. 24, no. 9, pp. 1342–1350, Sep. 2018, doi: 10.1038/s41591-018-0107-6.

[49] J. Wang *et al.*, “S-D Net: Joint Segmentation and Diagnosis Revealing the Diagnostic Significance of Using Entire RNFL Thickness in Glaucoma,” p. 10.

[50] Z. Mishra, A. Ganegoda, J. Selicha, Z. Wang, S. R. Sadda, and Z. Hu, “Automated Retinal Layer Segmentation Using Graph-based Algorithm Incorporating Deep-learning-derived Information,” *Sci Rep*, vol. 10, no. 1, p. 9541, Dec. 2020, doi: 10.1038/s41598-020-66355-5.

[51] K. Team, “Keras documentation: Image segmentation with a U-Net-like architecture.” https://keras.io/examples/vision/oxford_pets_image_segmentation/ (accessed Sep. 14, 2020).

[52] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: A Nested U-Net Architecture for Medical Image Segmentation,” *arXiv:1807.10165 [cs, eess, stat]*, Jul. 2018, Accessed: Mar. 30, 2020. [Online]. Available: http://arxiv.org/abs/1807.10165

[53] M. W. M. Wintergerst *et al.*, “Replication and Refinement of an Algorithm for Automated Drusen Segmentation on Optical Coherence Tomography,” *Sci Rep*, vol. 10, no. 1, p. 7395, Dec. 2020, doi: 10.1038/s41598-020-63924-6.

[54] Q. Chen *et al.*, “Automated drusen segmentation and quantification in SD-OCT images,” *Medical Image Analysis*, vol. 17, no. 8, pp. 1058–1072, Dec. 2013, doi: 10.1016/j.media.2013.06.003.

[55] “RETOUCH - MICCAI 2017 Workshop,” Dec. 13, 2017. https://retouch.grand-challenge.org/workshop/ (accessed Dec. 13, 2017).

[56] H. Bogunovic *et al.*, “RETOUCH -The Retinal OCT Fluid Detection and Segmentation Benchmark and Challenge,” *IEEE Transactions on Medical Imaging*, pp. 1–1, 2019, doi: 10.1109/TMI.2019.2901398.

[57] S. J. Chiu, M. J. Allingham, P. S. Mettu, S. W. Cousins, J. A. Izatt, and S. Farsiu, “Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema,” *Biomedical Optics Express*, vol. 6, no. 4, p. 1172, Apr. 2015, doi: 10.1364/BOE.6.001172.

[58] S. M. Khan *et al.*, “A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability,” *The Lancet Digital Health*, p. S2589750020302405, Oct. 2020, doi: 10.1016/S2589-7500(20)30240-5.

[59] M. Melinščak, M. Radmilović, Z. Vatavuk, and S. Lončarić, “AROI: Annotated Retinal OCT Images database,” in *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, Opatija, Sep. 2021, p. accepted for publication.

[60] “GIMP,” *GIMP*. https://www.gimp.org/ (accessed Jan. 23, 2021).

[61] “OCT image database - Image Processing Group.” https://ipg.fer.hr/ipg/resources/oct_image_database# (accessed Jan. 28, 2021).

[62] M. Melinscak, “Retinal OCT image segmentation,” Feb. 2021, doi: 10.17605/OSF.IO/5WYR3.

[63] G. Staurenghi, S. Sadda, U. Chakravarthy, and R. F. Spaide, “Proposed Lexicon for Anatomic Landmarks in Normal Posterior Segment Spectral-Domain Optical Coherence Tomography,” *Ophthalmology*, vol. 121, no. 8, pp. 1572–1578, Aug. 2014, doi: 10.1016/j.ophtha.2014.02.023.

[64] A. Daruich *et al.*, “Mechanisms of macular edema: Beyond the surface,” *Progress in Retinal and Eye Research*, vol. 63, pp. 20–68, Mar. 2018, doi: 10.1016/j.preteyeres.2017.10.006.

[65] M. D. Abràmoff, M. K. Garvin, and M. Sonka, “Retinal Imaging and Image Analysis,” *IEEE Rev Biomed Eng*, vol. 3, pp. 169–208, Jan. 2010, doi: 10.1109/RBME.2010.2084567.

[66] Z. B. Popović and J. D. Thomas, “Assessing observer variability: a user’s guide,” *Cardiovasc. Diagn. Ther.*, vol. 7, no. 3, pp. 317–324, Jun. 2017, doi: 10.21037/cdt.2017.03.12.

[67] A. Yazdanpanah, G. Hamarneh, B. R. Smith, and M. V. Sarunic, “Segmentation of Intra-Retinal Layers From Optical Coherence Tomography Images Using an Active Contour Approach,” *IEEE Transactions on Medical Imaging*, vol. 30, no. 2, pp. 484–496, Feb. 2011, doi: 10.1109/TMI.2010.2087390.

[68] M. K. Garvin, M. D. Abramoff, R. . Kardon, S. R. Russell, X. Wu, and M. Sonka, “Intraretinal Layer Segmentation of Macular Optical Coherence Tomography Images Using Optimal 3-D Graph Search,” *IEEE Transactions on Medical Imaging*, vol. 27, no. 10, pp. 1495–1505, Oct. 2008, doi: 10.1109/TMI.2008.923966.

[69] S. J. Chiu, X. T. Li, P. Nicholas, C. A. Toth, J. A. Izatt, and S. Farsiu, “Automatic segmentation of seven retinal layers in SDOCT images congruent with expert manual segmentation,” *Optics express*, vol. 18, no. 18, pp. 19413–19428, 2010.

[70] R. Kafieh, H. Rabbani, M. Abramoff, and M. Sonka, “Intra-retinal layer segmentation of optical coherence tomography using diffusion map,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 1080–1084. doi: 10.1109/ICASSP.2013.6637816.

[71] M. K. Garvin, M. D. Abramoff, Xiaodong Wu, S. R. Russell, T. L. Burns, and M. Sonka, “Automated 3-D Intraretinal Layer Segmentation of Macular Spectral-Domain Optical Coherence Tomography Images,” *IEEE Transactions on Medical Imaging*, vol. 28, no. 9, pp. 1436–1447, Sep. 2009, doi: 10.1109/TMI.2009.2016958.

[72] “Ophthalmic Image Analysis | The Iowa Institute for Biomedical Imaging.” https://www.iibi.uiowa.edu/ophthalmic-analysis (accessed Jul. 25, 2021).

[73] M. Sonka and M. D. Abràmoff, “Quantitative analysis of retinal OCT,” *Medical Image Analysis*, vol. 33, pp. 165–169, Oct. 2016, doi: 10.1016/j.media.2016.06.001.

[74] S. Apostolopoulos, S. De Zanet, C. Ciller, S. Wolf, and R. Sznitman, “Pathological OCT Retinal Layer Segmentation Using Branch Residual U-Shape Networks,” in *Medical Image Computing and Computer-Assisted Intervention − MICCAI 2017*, vol. 10435, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds. Cham: Springer International Publishing, 2017, pp. 294–301. doi: 10.1007/978-3-319-66179-7_34.

[75] A. Ben-Cohen *et al.*, “Retinal layers segmentation using Fully Convolutional Network in OCT images,” p. 8.

[76] M. Pekala, N. Joshi, T. Y. A. Liu, N. M. Bressler, D. C. DeBuc, and P. Burlina, “Deep learning based retinal OCT segmentation,” *Computers in Biology and Medicine*, vol. 114, p. 103445, Nov. 2019, doi: 10.1016/j.compbiomed.2019.103445.

[77] Y. He *et al.*, “Topology guaranteed segmentation of the human retina from OCT using convolutional neural networks,” *arXiv:1803.05120 [cs]*, Mar. 2018, Accessed: Jul. 04, 2018. [Online]. Available: http://arxiv.org/abs/1803.05120

[78] H. Xie *et al.*, “Globally Optimal Segmentation of Mutually Interacting Surfaces using Deep Learning,” *arXiv:2007.01259 [cs]*, Jul. 2020, Accessed: Jul. 10, 2020. [Online]. Available: http://arxiv.org/abs/2007.01259

[79] S. Borkovkina, A. Camino, W. Janpongsri, M. V. Sarunic, and Y. Jian, “Real-time retinal layer segmentation of OCT volumes with GPU accelerated inferencing using a compressed, low-latency neural network,” *Biomed. Opt. Express*, vol. 11, no. 7, p. 3968, Jul. 2020, doi: 10.1364/BOE.395279.

[80] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[81] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” p. 11.

[82] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *arXiv:1412.6980 [cs]*, Jan. 2017, Accessed: Jul. 06, 2021. [Online]. Available: http://arxiv.org/abs/1412.6980

[83] U. Schmidt-Erfurth, A. Sadeghipour, B. S. Gerendas, S. M. Waldstein, and H. Bogunović, “Artificial intelligence in retina,” *Progress in Retinal and Eye Research*, Aug. 2018, doi: 10.1016/j.preteyeres.2018.07.004.

[84] H. Lee, K. E. Kang, H. Chung, and H. C. Kim, “Automated Segmentation of Lesions Including Subretinal Hyperreflective Material in Neovascular Age-related Macular Degeneration,” *American Journal of Ophthalmology*, vol. 191, pp. 64–75, Jul. 2018, doi: 10.1016/j.ajo.2018.04.007.

[85] T. Schlegl *et al.*, “Fully Automated Detection and Quantification of Macular Fluid in OCT Using Deep Learning,” *Ophthalmology*, vol. 125, no. 4, pp. 549–558, Apr. 2018, doi: 10.1016/j.ophtha.2017.10.031.

[86] Z. Gu *et al.*, “CE-Net: Context Encoder Network for 2D Medical Image Segmentation,” *IEEE Trans. Med. Imaging*, vol. 38, no. 10, pp. 2281–2292, Oct. 2019, doi: 10.1109/TMI.2019.2903562.

[87] W. Liu, Y. Sun, and Q. Ji, “MDAN-UNet: Multi-Scale and Dual Attention Enhanced Nested U-Net Architecture for Segmentation of Optical Coherence Tomography Images,” *Algorithms*, vol. 13, no. 3, p. 60, Mar. 2020, doi: 10.3390/a13030060.

[88] G. Lin, A. Milan, C. Shen, and I. Reid, “RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation,” *arXiv:1611.06612 [cs]*, Nov. 2016, Accessed: Mar. 22, 2021. [Online]. Available: http://arxiv.org/abs/1611.06612

[89] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid Scene Parsing Network,” *arXiv:1612.01105 [cs]*, Apr. 2017, Accessed: Jul. 11, 2020. [Online]. Available: http://arxiv.org/abs/1612.01105

[90] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” *arXiv:1608.06993 [cs]*, Jan. 2018, Accessed: Jul. 19, 2020. [Online]. Available: http://arxiv.org/abs/1608.06993

[91] “ImageNet.” http://www.image-net.org/ (accessed Sep. 14, 2020).

[92] F. Isensee, “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nature Methods*, p. 14.

[93] B. Norman, V. Pedoia, and S. Majumdar, “Use of 2D U-Net Convolutional Neural Networks for Automated Cartilage and Meniscus Segmentation of Knee MR Imaging Data to Determine Relaxometry and Morphometry,” *Radiology*, vol. 288, no. 1, pp. 177–185, Jul. 2018, doi: 10.1148/radiol.2018172322.

[94] A. Sevastopolsky, “Optic Disc and Cup Segmentation Methods for Glaucoma Detection with Modification of U-Net Convolutional Neural Network,” *Pattern Recognit. Image Anal.*, vol. 27, no. 3, pp. 618–624, Jul. 2017, doi: 10.1134/S1054661817030269.

[95] B. Ait Skourt, A. El Hassani, and A. Majda, “Lung CT Image Segmentation Using Deep Neural Networks,” *Procedia Computer Science*, vol. 127, pp. 109–113, Jan. 2018, doi: 10.1016/j.procs.2018.01.104.

[96] Z. Zhang and M. Sabuncu, “Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels,” p. 11.

[97] L. Luo, Y. Xiong, Y. Liu, and X. Sun, “Adaptive Gradient Methods with Dynamic Bound of Learning Rate,” *arXiv:1902.09843 [cs, stat]*, Feb. 2019, Accessed: Feb. 04, 2021. [Online]. Available: http://arxiv.org/abs/1902.09843

[98] L. Bottou, “Large-Scale Machine Learning with Stochastic Gradient Descent,” in *Proceedings of COMPSTAT’2010*, Heidelberg, 2010, pp. 177–186. doi: 10.1007/978-3-7908-2604-3_16.

[99] “Google Colaboratory.” https://colab.research.google.com/notebooks/intro.ipynb (accessed May 17, 2020).

[100] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” *Journal of Big Data*, vol. 6, no. 1, p. 60, Jul. 2019, doi: 10.1186/s40537-019-0197-0.

[101] R. M. Haralick, S. R. Sternberg, and X. Zhuang, “Image Analysis Using Mathematical Morphology,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-9, no. 4, pp. 532–550, Jul. 1987, doi: 10.1109/TPAMI.1987.4767941.

[102] P. Krähenbühl and V. Koltun, “Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials,” *arXiv:1210.5644 [cs]*, Oct. 2012, Accessed: Aug. 11, 2021. [Online]. Available: http://arxiv.org/abs/1210.5644

[103] G. R. Wilkins, O. M. Houghton, and A. L. Oldenburg, “Automated segmentation of intraretinal cystoid fluid in optical coherence tomography,” *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 1109–1114, 2012.

[104] M. Pilch *et al.*, “Automated segmentation of pathological cavities in optical coherence tomography scans,” *Invest. Ophthalmol. Vis. Sci.*, vol. 54, no. 6, pp. 4385–4393, Jun. 2013, doi: 10.1167/iovs.12-11396.

[105] M. Sahoo, S. Pal, and M. Mitra, “Automatic segmentation of accumulated fluid inside the retinal layers from optical coherence tomography images,” *Measurement*, vol. 101, no. Supplement C, pp. 138–144, Apr. 2017, doi: 10.1016/j.measurement.2017.01.027.

[106] X. Chen, M. Niemeijer, L. Zhang, K. Lee, M. D. Abràmoff, and M. Sonka, “Three-dimensional segmentation of fluid-associated abnormalities in retinal OCT: probability constrained graph-search-graph-cut,” *IEEE transactions on medical imaging*, vol. 31, no. 8, pp. 1521–1531, 2012.

[107] W. Ding *et al.*, “Automatic detection of subretinal fluid and sub-retinal pigment epithelium fluid in optical coherence tomography images,” *Conf Proc IEEE Eng Med Biol Soc*, vol. 2013, pp. 7388–7391, 2013, doi: 10.1109/EMBC.2013.6611265.

[108] G. Quellec, K. Lee, M. Dolejsi, M. K. Garvin, M. D. Abràmoff, and M. Sonka, “Three-dimensional analysis of retinal layer texture: identification of fluid-filled regions in SD-OCT of the macula,” *Medical Imaging, IEEE Transactions on*, vol. 29, no. 6, pp. 1321–1330, 2010.

[109] A. Lang *et al.*, “Automatic segmentation of microcystic macular edema in OCT,” *Biomed Opt Express*, vol. 6, no. 1, pp. 155–169, Jan. 2015, doi: 10.1364/BOE.6.000155.

[110] D. Lu, W. Ding, A. Merkur, M. V. Sarunic, and M. F. Beg, “Multiple instance learning for age-related macular degeneration diagnosis in optical coherence tomography images,” in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, Apr. 2017, pp. 139–142. doi: 10.1109/ISBI.2017.7950487.

[111] D. Lu, M. Heisler, S. Lee, G. Ding, M. V. Sarunic, and M. F. Beg, “Retinal Fluid Segmentation and Detection in Optical Coherence Tomography Images using Fully Convolutional Neural Network,” *arXiv:1710.04778 [cs]*, Oct. 2017, [Online]. Available: http://arxiv.org/abs/1710.04778

[112] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 1800–1807. doi: 10.1109/CVPR.2017.195.

[113] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” *arXiv:1512.00567 [cs]*, Dec. 2015, Accessed: Jul. 07, 2021. [Online]. Available: http://arxiv.org/abs/1512.00567

[114] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Dec. 2015, Accessed: Sep. 14, 2020. [Online]. Available: https://arxiv.org/abs/1512.03385v1

[115] A. G. Howard *et al.*, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” *arXiv:1704.04861 [cs]*, Apr. 2017, Accessed: Jul. 07, 2021. [Online]. Available: http://arxiv.org/abs/1704.04861

[116] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv:1409.1556 [cs]*, Apr. 2015, Accessed: Jul. 07, 2021. [Online]. Available: http://arxiv.org/abs/1409.1556

[117] M. Melinščak and S. Lončarić, “Retinal OCT Image Segmentation: How Well do Algorithms Generalize or How Transferable are the Data?,” in *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, Sep. 2020, pp. 192–196. doi: 10.23919/MIPRO48935.2020.9245336.

[118] F. Chollet, *Deep learning with Python*. Shelter Island, New York: Manning Publications Co, 2018.

BIOGRAPHY

Martina Melinščak was born in 1975 in Brežice. She graduated from the XV Gymnasium (MIOC) in Zagreb. In 2000, she graduated from the Faculty of Electrical Engineering and Computing in Zagreb at the Department of Electronic Systems and Information Processing with the thesis "Determination of diastolic blood pressure by computational oscillometric method" under the supervision of professor Ante Šantić. In the same year, she enrolled in postgraduate studies. She received her master's degree in May 2007 with the paper "Analysis of the characteristics of voltage pulse plethysmography" under the supervision of professor Ante Šantić. In 2013, she enrolled in a doctoral study at the Faculty of Electrical Engineering and Computing.

From 2000 to 2005, she worked at Siemens as a software developer, then at the Croatian History Museum as an IT specialist, and at Libusoft Cicom as a programmer. Since January 2006, she has been employed at the Karlovac University of Applied Sciences as an assistant (later as a lecturer) for courses in the field of electrical engineering and computer science.

List of publications:

M. Melinščak, M. Radmilović, Z. Vatavuk, and S. Lončarić, "Annotated retinal optical coherence tomography images (AROI) database for joint retinal layer and fluid segmentation," *Automatika*, vol. 62, no. 3, pp. 375–385, Jul. 2021, doi: 10.1080/00051144.2021.1973298.

M. Melinscak, P. Prentasic, and S. Loncaric, "Retinal Vessel Segmentation using Deep Neural Networks:," in *Proceedings of the 10th International Conference on Computer Vision Theory and Applications*, Berlin, Germany, 2015, pp. 577–582. doi: 10.5220/0005313005770582.

M. Melinščak and S. Lončarić, "Retinal OCT Image Segmentation: How Well do Algorithms Generalize or How Transferable are the Data?," in *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, Sep. 2020, pp. 192–196. doi: 10.23919/MIPRO48935.2020.9245336.

M. Melinščak, M. Radmilović, Z. Vatavuk, and S. Lončarić, "AROI: Annotated Retinal OCT Images database," in *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, Sep. 2021, pp. 400-405.

ŽIVOTOPIS

Martina Melinščak rođena je 1975. godine u Brežicama. Završila je XV gimnaziju (MIOC) u Zagrebu. Godine 2000. diplomirala je na Fakultetu elektrotehnike i računarstva u Zagrebu na Zavodu za elektroničke sustave i obradbu informacija s radom „Određivanje dijastoličkog krvnog tlaka računskom oscilometrijskom metodom“ pod mentorstvom profesora Ante Šantića. Iste godine upisala je poslijediplomski studij. Magistrirala je u svibnju 2007. s radom „Analiza karakteristika naponske impulsne pletizmografije“ pod mentorstvom profesora Ante Šantića. Godine 2013. upisala je doktorski studij na Fakultetu elektrotehnike i računarstva.

Od 2000. do 2005. godine radila je u Siemensu na razvoju softvera, zatim u Hrvatskom povijesnom muzeju kao informatičar, te u Libusoft Cicomu kao programer. Od siječnja 2006. godine zaposlena je na Veleučilištu u Karlovcu kao asistent (kasnije kao predavač) za kolegije iz područja elektrotehnike i računarstva.