Multimodal information fusion for the diagnosis of diabetic retinopathy

Yihao Li

COLLEGE DOCTORAL BRETAGNE SCIENCES DE LA VIE ET DE LA SANTE

Université de Bretagne Occidentale

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE BRETAGNE OCCIDENTALE

ÉCOLE DOCTORALE N° 637
Sciences de la Vie et de la Santé
Spécialité : *Analyse et Traitement de l'Information et des Images Médicales*

Par
Yihao LI

Multimodal information fusion for the diagnosis of diabetic retinopathy

Fusion d'informations multimodales pour le diagnostic de la rétinopathie diabétique

Thèse présentée et soutenue à Brest, le 07/12/2023
Unité de recherche : LaTIM UMR 1101, Inserm

Rapporteurs avant soutenance :

Florence ROSSANT	Professeur, ISEP Paris
Saïd MAHMOUDI	Professeur, Université de Mons (Belgique)

Composition du Jury :

Président :	Saïd MAHMOUDI	Professeur, Université de Mons (Belgique)
Examinateurs :	Florence ROSSANT	Professeur, ISEP Paris
	Saïd MAHMOUDI	Professeur, Université de Mons (Belgique)
	Mostafa EL HABIB DAHO	Maître de Conférences (HDR) en disponibilité, Université de Tlemcen (Algérie) et Ingénieur de Recherche, UBO
Dir. de thèse :	Gwenolé QUELLEC	Directeur de Recherche, INSERM

Invité(s) :

Pierre-Henri CONZE	Maître de Conférences, IMT Atlantique

Acknowledgement

I would like to express my sincere gratitude to all those who have supported and guided me throughout the journey of completing this thesis. Without their unwavering assistance, this achievement would not have been possible.

First and foremost, I am deeply indebted to my thesis director, Pr. Gwenolé Quellec, for his invaluable guidance, insightful feedback, and continuous encouragement. His expertise and dedication played a pivotal role in shaping the direction of this research.

I extend my heartfelt thanks to the co-supervisors of my thesis, Dr. Mathieu Lamard, Dr. Mostafa El Habib Daho and Dr. Pierre-Henri Conze, whose teachings and mentorship provided me with a strong foundation of knowledge and critical thinking skills. The valuable contributions and discussions they provided helped me challenge my ideas and expand my horizons.

I am grateful to EviRed Project for their financial support, which allowed me to carry out this research and access necessary resources. Their investment in my academic pursuits is greatly appreciated.

I would like also to reveal my deepest appreciation for the jury members Pr. Florence ROSSANT and Pr. Saïd MAHMOUDI, who examined my thesis work and provided insightful suggestions.

I want to acknowledge the participants of my study, Dr. Hassan Alhajj and Mr. Rachid Zeghlache, whose willingness to share their time and insights. Their contributions are integral to the outcomes of this work.

My sincere appreciation goes to my friends and family for their unwavering support, patience, and understanding during this demanding period. Their encouragement and belief in me were constant sources of motivation.

In conclusion, I am deeply grateful to everyone who played a part in shaping this thesis. Your support has been indispensable, and I am honored to have had the opportunity to learn and grow with you by my side.

Table of Contents

List of Figures

List of Tables

INTRODUCTION

"The process of scientific discovery is, in effect, a continual flight from wonder."

— *Albert Einstein*

0.1 Context

Diabetes is a chronic disease characterized by an excess of sugar in the blood, affecting 422 million people worldwide, including 3.3 million in France [1]. One of the most common complications of diabetes is diabetic retinopathy (DR), the leading cause of blindness in the working population of developed countries. However, there is an important obstacle to combating DR: the existing classification system, based on old imaging technology, namely Color Fundus Photography (CFP), is insufficient to finely predict the evolution of the disease. Recent advances in imaging techniques, such as Ultra-Wide-Field CFP (UWF-CFP), Optical Coherence Tomography (OCT), and Optical Coherence Tomography Angiography (OCTA), have made it possible to provide more detailed and comprehensive fundus information to aid in the diagnosis of DR. There is, however, an expanding amount of data produced by these new imaging modalities, which requires high levels of expertise from humans. It will be difficult for most ophthalmologists to develop a predictive clinical score based on all these factors. The "Évaluation Intelligente de la Rétinopathie diabétique" (EviRed) project is thus developing an expert system that answers a medical need: replacing the current classification of diabetic retinopathy based on up-to-date imaging modalities and other medical data of the patient. It is expected to better predict the evolution of the disease and guarantee timely treatments.

0.2 Motivation

As part of the EviRed project, this thesis investigates the use of artificial intelligence in order to properly integrate this massive amount of data. It should deliver a better diagnosis, accurate prediction, and better decision-making by ophthalmologists during the follow-up of DR cases. Specifically, the goal was to design deep learning network architectures that combine the advantages of different imaging modalities in order to enhance diagnostic performance. As part of this thesis, we examined the following fusion scenarios involving 3D and 2D data:

1. Joint analysis of multi-modal information in OCTA.
2. Joint analysis of different specifications of OCTA acquisitions.
3. Joint analysis of OCTA and UWF-CFP.

These joint analyses tested and compared different state-of-the-art multimodal information fusion methods, and new architectures for DR diagnostics were proposed. The performance of the fusion of these modalities is clinically validated, and our experimental results also demonstrated the effectiveness of our multimodal fusion approach, resulting in significant improvements in joint analysis compared to unimodal images for different diagnostic tasks involving DR pathology. Our designed algorithm will be part of the EviRed system to provide more accurate information for future clinical DR diagnostic work. Ophthalmology is on the verge of a revolution in terms of screening, diagnosis, and management of pathologies. We hope this work will contribute to this revolution.

0.3 Thesis outline

The organization of this manuscript is as follows:

— **Chapter 1** begins with a discussion of the clinical background and development of diabetic retinopathy. Next, several modality data currently used in the clinic are analyzed. Following that, we examined the screening and treatment system relevant to DR. Finally, we introduced the EviRed project, which aims to establish an automated DR diagnostic system, thus introducing the objective of our study: Diagnostics of DR using multimodal information fusion.

— **Chapter 2** is an exploration of state-of-the-art multimodal fusion methods based on deep learning. Based on the analysis of a large number of literature reviews

in the field of DR and other fields of medicine, we have proposed five different architectures for multimodal fusion.

— **Chapter 3** describes the data used in the study. With the advancement of the EviRed project, our dataset has evolved and become complete. Data used at different stages of the thesis and their multimodal fusion objectives are described.

— **Chapter 4** presents our initial explorations of multimodal fusion using data from the early stages of the EviRed project. The performance of different fusion methods on the three modalities in the OCTA images is compared.

— **Chapter 5** describes our exploration of the unlabeled data from the EviRed project. We have tested both self-supervised and semi-supervised learning methods in order to enhance the diagnostic performance of the fusion model using the unlabeled data.

— **Chapter 6** contains the fusion tests of different OCTA acquisitions after we received the data annotations. A hybrid fusion framework utilizing high-resolution and ultra-widefield OCTA was proposed to assess DR severity automatically.

— **Chapter 7** presents an initial exploration of the fusion of Ultra-WideField Color Fundus Photograph and OCTA. Using Manifold Mixup and Squeeze-and-Excitation blocks, our model generates a compelling outcome through a feature-level fusion strategy.

— **Conclusion** We end with a conclusion and discuss future works in Chapter Conclusion.

Chapter 1

Diabetic retinopathy and thesis context

> "Wherever the art of Medicine is loved, there is also a love of Humanity."
>
> — *Hippocrates*

Diabetic retinopathy (DR) is a progressive eye condition characterized by damage to the retina caused by diabetes, potentially leading to vision impairment or even blindness if left untreated. This chapter briefly overviews the clinical background of DR and its development. There has

been an increase in the use of emerging medical imaging techniques for diagnosing DR in recent years. The existing diagnostic systems are difficult to adapt to emerging imaging technologies. As a result, we presented the EviRed project, which aims to develop an automated DR diagnostic system, thus introducing the goal of our study: Diagnostics of DR using multimodal information fusion.

1.1 Development and impact of diabetic retinopathy

The disease of diabetes mellitus is a heterogeneous group of disorders characterized by hyperglycemia resulting from an absolute or relative decrease in insulin levels [2]. Depending on the etiopathology, diabetes mellitus can be categorized into Type 1 and Type 2 diabetes [3]. In people with Type 1 diabetes (T1D), beta cells in the endocrine pancreas are destroyed as a result of an autoimmune process resulting in a severe insulin deficiency [4]. Type 2 diabetes (T2D) is characterized by impaired insulin action, which is mainly caused by a combination of factors and, in its later stages, leads to decreased insulin production [5]. It is estimated that more than 90% of those contributing to the steep rise in disease incidence have T2D, which can be attributed to the consumption of calorie-dense foods with low nutritional value, inactivity, and an increasing prevalence of obesity [6]. In recent years, T2D has reached epidemic proportions, primarily in developing nations, due to the adoption of American-style dietary habits [7]. Throughout the world, diabetes mellitus is on the rise, with an estimated 382 million people diagnosed in 2013, rising to 592 million by 2030 [8–10].

There is an increasing burden posed by diabetes complications in advanced nations and also in developing nations, and many of these com-

plications result from the vascular complications of diabetes [7]. The most common complication of diabetes is diabetic retinopathy, which is the leading cause of blindness in people of working age throughout the world [11]. It is estimated that about one-third of diabetic patients worldwide will develop diabetic retinopathy, with increased risk associated with longer disease duration, high hemoglobin A1C (HbA1c), and hypertension [12]. The number of diabetic retinopathy patients is expected to increase by 191 million by 2030 from 127 million in 2010 [13]. Besides the implications for the patient personally and financially, diabetic retinopathy has a significant impact on society as well. Healthcare costs for patients with diabetic retinopathy are almost double those of those without the disease [14]. As a result, the early diagnosis and treatment of DR are of great value to diabetic patients and to society in general.

1.2 Classification of diabetic retinopathy

The screening of the retina is essential for diabetic patients in order to detect and treat diabetic retinopathy at an early stage in order to avoid the risk of blindness [15]. Color Fundus Photographs (CFP) are one of the most cost-effective screening tools for diabetic retinopathy [16]. CFP have the advantage of clearly illustrating the optic disc, optic cup, and blood vessels [17]. The Early Treatment Diabetic Retinopathy Study (ETDRS) established the gold standard method for assessing DR severity [18, 19]. This method relies on the examination of seven standard retinal fields on 30° stereoscopic color fundus photographs. Despite its accuracy and reproducibility, this technique is labor-intensive and requires skilled photographers and skilled readers, as well as sophisticated photography equipment, film processing, and archiving [20]. The turnaround time between data ac-

quisition and interpretation in clinical trials can be several weeks. From the patient's perspective, it can be time-consuming and uncomfortable. Currently, it is recommended to perform two 45° retinopathographies of each eye, the first centered on the macula and the second centered on the papilla, allowing analysis of the mid-nasal periphery [21]. These recommendations supersede those of the ETDRS. Nowadays, the retinograph can produce high-quality images without pupillary dilation, allowing patients to save time and experience greater comfort during the procedure.

Diagnosis of DR is based on the clinical manifestations of vascular abnormalities in the retina. Diabetic retinopathy consists of two major types: non-proliferative diabetic retinopathy (NPDR) and proliferative diabetic retinopathy (PDR) [22]. NPDR refers to the early stages of DR in which increased permeability and capillary occlusion are the primary signs observed in the retinal circulation. Even if the patient is asymptomatic, fundus photography can detect retinal pathologies such as microaneurysms, hemorrhages, and hard exudates during this stage [23]. DR which has progressed to PDR, is characterized by neovascularization. When the new abnormal vessels bleed into the vitreous during this stage (vitreous hemorrhage) or when tractional retinal detachment occurs, the patient may experience severe vision impairment [24].

According to the Early Treatment Diabetic Retinopathy Study [25], DR is classified based on the analysis of stereoscopic pairs of color photographs taken at seven-field fundus photography [26]. The appearance of different types of lesions on a retinal image indicates the presence of DR [23]. It includes microaneurysms (MA), haemorrhages (HM), and soft and hard exudates (EX) [19, 27]. According to the description of the lesions visible on the fundus, different classifications have been made in order to determine the severity, the prognosis, and the treatment for the lesions.

- The earliest indication of DR is microaneurysms, which appear as small red round dots on the retina [28]. The size is less than 125 micrometers in size, and the margins are sharp [29].
- The retinal hemorrhages are large spots with irregular margins, larger than 125 micrometers [30].
- The hard exudates appear as bright yellow spots on the retina due to the leakage of plasma. Their margins are sharp and found in the retina's outer layers [28].
- Soft exudates appear as spots of white on the retina as a result of swelling of the nerve fibers. It has an oval or round shape [28].

In addition to these lesions, some more serious lesions may occur in the advanced stages of DR, such as venous reduplication (VR), neovascularization (NV), and venous loops (VL). Based on the presence of the lesions, there are five stages of DR: no DR, mild NPDR, moderate NPDR, severe NPDR, and PDR. Two more recent classifications are currently used in France [21]: the ALFEDIAM classification (Association de langue française pour l'étude du diabète et des maladies métaboliques) [31] and the international classification of the American Academy of Ophthalmology (AAO) [32] represented Tab. 1.1. Fig. 1.1 shows the sample of color fundus photography of DR stages.

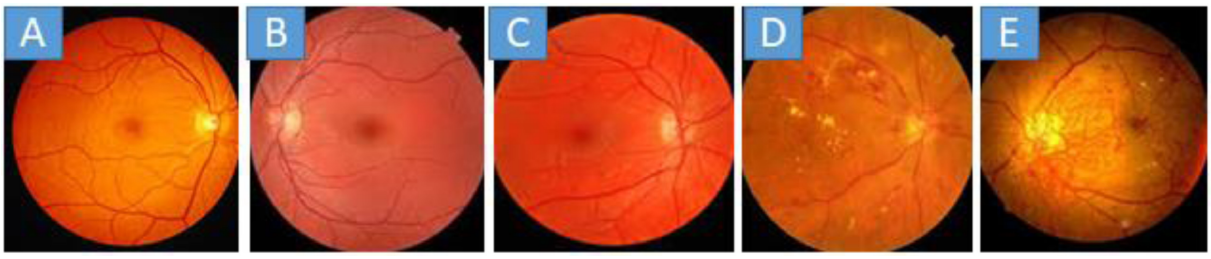


Figure 1.1 – The different DR stages in retinal fundus photographs: (a) No apparent retinopathy (b) Mild NPDR, (c) Moderate NPDR, (d) Severe NPDR, (e) PDR [33].

Table 1.1 – Diabetic Retinopathy Disease Severity Scale for AAO [32].

Proposed Disease Severity Level	Findings Observable on Dilated Ophthalmoscopy
No apparent retinopathy	No abnormalities.
Mild nonproliferative DR	Microaneurysms only.
Moderate nonproliferative DR	More than just microaneurysms but less than severe nonproliferative DR.
Severe nonproliferative DR	Any of the following: more than 20 intraretinal hemorrhages in each of 4 quadrants; definite venous beading in 2+ quadrants; Prominent intraretinal microvascular abnormalities in 1+ quadrant And no signs of proliferative retinopathy.
Proliferative DR	One or more of the following: neovascularization, vitreous/ preretinal hemorrhage.

1.3 Novel retinal imaging technologies for diabetic retinopathy

Several modern fundus imaging techniques for DR diagnosis have emerged as a result of progress in fundus photography and improvements in medical diagnostic systems: Ultra-Wide-Field imaging, Optical Coherence Tomography, Optical Coherence Tomography Angiography, etc.

1.3.1 Ultra-Wide-Field imaging

As a result of the relatively small retinal field covered by conventional CFP, its clinical application is limited. The technology of retinal imaging has evolved rapidly over the past decade. In recent years, the advent of ultra-wide-field (UWF) imaging, defined as a field-of-view of 100 degrees or

more by the Diabetic Retinopathy Clinical Research Network (DRCRnet), has allowed for the visualization of the far peripheral retina, areas that are beyond the field of view of traditional CFP [34]. By covering 3x more retinal surface than 45° pictures, UWF provides the best choice of new imaging technologies. About 30% to 40% of eyes have peripheral DR lesions that are located outside of the 45° photo fields [35]. In 10-15% of cases, the severity of the DR is underestimated by at least one step on 45° images. It has been noted that even problems that can threaten the vision, such as the development of new vessels, can occur mostly outside the 45°photos field in 53.9% of all cases [36]. The use of UWF imaging has made a significant contribution to the detection and management of DR [37]. According to current research, UWF imaging detects additional and more extensive PDR pathologies when compared to conventional CFP imaging [35].

1.3.2 Optical Coherence Tomography

The field of ophthalmology has seen remarkable advancements in retinal imaging technology, which now plays a crucial role in the clinical diagnosis of DR. As a major advancement, Optical Coherence Tomography (OCT) has been a game-changer since its introduction in 1991. OCT has transformed not only the evaluation of the retina but the entire field of ophthalmology [38]. The OCT technique is non-invasive, non-contact, and uses laser refraction to visualize anatomical structures in cross-section, similar to histology [39]. This is the only examination that is capable of obtaining cross-sections of the retina with a definition of five to ten microns [21]. It assists in the detection of vascular abnormalities, including microaneurysms, non-perfused areas, Intraretinal microvascular abnormalities (IRMA), macular edema, and preretinal neovascularization associated with diabetes [40].

1.3.3 Optical Coherence Tomography Angiography

Based on the OCT's foundations, Optical Coherence Tomography Angiography (OCTA) offers a non-invasive method for producing detailed and depth-resolved images of the chorioretinal microvasculature. The technique works by analyzing differences between two scans taken at the same location. Moving structures, such as red blood cells, generate a decorrelation signal. Thus, by detecting these signals, OCTA can highlight the retinal vascular networks, offering a rich picture of the retina's health [41]. Recently, swept-source technology has been used in OCTA, leading to the development of Swept-Source OCTA (SS-OCTA). This new approach, lauded for its non-invasive, safe, and repeatable imaging of retinal blood flow, has been the subject of numerous studies exploring its potential in diagnosing, screening, and monitoring DR [42–45]. The technological leap from SD-OCTA to SS-OCTA allowed imaging larger fields of view: most of the initial studies use SS-OCTA equipment that can capture a $12 \times 12\text{mm}^2$ area in a single scan (as opposed to typically $3 \times 3\text{mm}^2$ or $6 \times 6\text{mm}^2$ previously) [44, 46, 47]. This imaging area can be further expanded by stitching together multiple scans or adding dioptric lenses [42, 45, 48, 49], although these techniques may require longer acquisition times and are likely to introduce more artifacts [37]. Machines recently developed that can obtain 15×15 mm^2 or wider retinal blood flow images by a single scan have emerged to solve these problems well and provide a fast, reliable solution for DR diagnosis and screening [50, 51]. The introduction of such ultra-widefield SS-OCTA (UWF-SS-OCTA) has offered a broader view for assessing DR lesions [26].

1.3.4 Other technologies

Adaptive optics (AO) is a technique that reduces the effects of optical aberrations on optical systems to improve their performance [52]. The adaptive optics scanning laser ophthalmoscopy (AO-SLO) technique is a noninvasive, objective, and direct method of examining the retinal microvasculature [53]. Through adaptive optics, researchers are able to correct for ocular aberrations and obtain high-resolution retinal images of photoreceptors [54], blood flow [55], blood corpuscles [56], capillary networks [57], retinal wall [58], and retinal nerve fiber layers [59]. In light of the fact that retinal diabetic changes result from microcirculatory disturbances, AO-SLO, which can be used to observe blood corpuscles directly in the parafovea, may serve as a technique for assessing retinal hemorheology in capillary networks [53, 60].

Monitoring retinal blood flow is essential for understanding the pathophysiology of DR, which requires a temporal resolution beyond the capability of present-day OCT-A systems [61]. A new full-field imaging method, laser Doppler holography, was introduced in order to measure blood flow within the retina and choroid with as yet unrivaled temporal resolution [62]. Laser Doppler holography uses Doppler spectral broadening of light backscattered by the retina to create the angiographic contrast [61].

1.4 Screening and treatment of diabetic retinopathy

1.4.1 Screening of diabetic retinopathy

The screening process for diabetic retinopathy contributes to the early detection of advanced stages of the disease, which is very important for the selection of a treatment that is appropriate and for preventing further

vision loss [63, 64]. While laser photocoagulation reduces the risk of vision loss associated with diabetic retinopathy, the disease still contributes to a significant amount of blindness and visual impairment in most developed countries [65–68]. It is largely due to the fact that the diagnosis is often made too late for treatment to prevent complications from occurring [69]. Regular eye examinations are the only way to identify and treat patients before vision-threatening complications arise [70]. It has been recommended that an annual eye examination be conducted following the diagnosis of diabetes by the l'étude du diabète et des maladies métaboliques (Alfediam), Agence nationale d'accréditation et d'évaluation en santé (Anaes), Agence française de sécurité sanitaire des produits de santé (Afssaps) and Haute Autorité de santé (HAS) [1] [71–73]. However, in France, annual funduscopic examinations of all diabetic patients are not sufficiently carried out [69]. It has been confirmed by an Echantillon National Témoin Représentatif des Personnes Diabétiques (ENTRED) study [74]. In addition to the increase in diabetic patients, there is a decrease in the number of ophthalmologists, which contributes to the low number of annual eye examinations [75]. Based on a 2002 study by the French Ministry of Health, the patient-to-physician ratio is not expected to improve in the next 15 years [76].

The French DR screening process typically involves an ophthalmologist performing a funduscopic examination [69]. Fundus photography using a non-mydriatic camera, followed by review by an ophthalmologist, is an alternative method of evaluating DR, which is at least as sensitive as ophthalmoscopy [77–80]. Further, non-mydriatic cameras can capture high-quality digital fundus photographs without requiring pupil dilation and can be transmitted to remote experts via the Internet. The characteristics of diabetic retinopathy, combined with the advancements in data

1. `http://www.has-sante.fr`

transmission technology, provide an ideal platform for telemedicine [81]. Several countries have already established telemedical networks to screen for diabetic retinopathy with good results [82–87].

In certain regions of France, a telemedicine network has been established to improve DR screening. Founded in 2004, the Ophthalmology Diabetes Telemedicine (Ophdiat®) network aims to improve the ophthalmological screening of diabetic patients in the Île-de-France region (12,000 km^2; 11 millions inhabitants) [69]. This network consists of screening centers that are linked through a central server to ophthalmological reading centers, as shown in Fig. 1.2. DR screening was performed by trained orthoptists or nurses who were legally authorized to take retinal photographs [88]. Using a non-mydriatic funduscopic camera, three 45° non-stereoscopic retinal digital photographs were obtained for each eye without pupil dilation [69]. Images were originally stored on a conventional personal computer. A team of five certified ophthalmologists from the Reading Centre downloaded the stored images. The report was generated by readers and included a diagnosis of diabetes-related ocular disorders, a diagnosis of non-diabetic ocular disorders, and recommendations for further care. In cases of referable DR, according to the International DR classification [32], patients should consult an ophthalmologist. Those with normal examinations were invited to retest their eyes the following year. Each retinal image was graded in approximately five minutes. As a result of the evaluation, the evaluation report was uploaded from the server, printed at the screening center, and sent to the general practitioner and patient [69]. As a result, the Ophdiat® network represents a reliable screening program for DR that can be used in a variety of healthcare settings [89]. Both patients and physicians have well received the program.

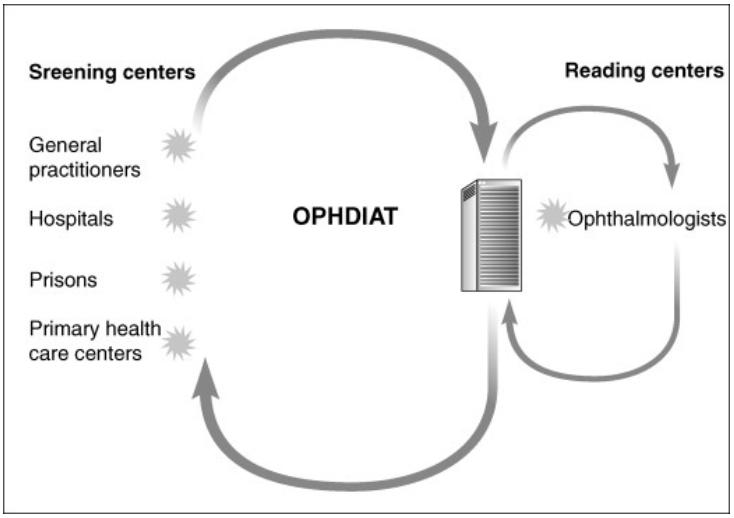


Figure 1.2 – An overview of the Ophdiat® network. The screening centers in Île-de-France are connected to an ophthalmological reading center via a central server [69].

1.4.2 Treatment of diabetic retinopathy

The management of diabetic retinopathy (DR) relies primarily on a good control of diabetes mellitus. However, when the severity of the vascular lesions warrants further treatment, laser photocoagulation or vitreoretinal surgery may need to be performed [90]. Currently, there are several treatment options available for different stages of DR: (1) For a long time, laser treatment has been considered an evidence-based treatment for DME and PDR [91–93]. Despite the fact that new therapies with anti-vascular endothelial growth factor and corticosteroids are revolutionizing the management of DME, laser photocoagulation remains the standard of care, both in PDR and DME [94, 95]. (2) Currently, intravitreal injections are a validated treatment option, which may increase the ocular therapeutic effects of many agents and reduce the incidence of systemic adverse events [96]. Several studies have demonstrated that different intravitreal agents are effective not only in preventing vision loss but also in restoring visual

acuity [97–99]. (3) According to the results of randomized clinical trials, intravitreal ranibizumab (IVR) is currently an approved therapeutic option for the treatment of DME [100, 101]. IVR has been evaluated both as a monotherapy and as a combination treatment with laser photocoagulation [102, 103]. (4) For patients who do not respond to laser photocoagulation in both DME and PDR, pars plana vitrectomy (PPV) is considered an option [104, 105].

1.4.3 Conclusion

As medical devices and medical systems develop, diabetic retinopathy screening and treatment systems are improving. Currently, the following solutions are available: (1) For diabetic patients, initial screening is performed using 45° CFP images. (2) Further diagnosis and follow-up using more advanced medical image imaging modalities (e.g., UWF, OCTA) for patients with pathology. (3) Treatment of the patient, if necessary, such as laser photocoagulation. Ophdiat®'s advent has undoubtedly improved the efficiency of initial patient diagnosis and screening. However, it remains inefficient for follow-up and further precise diagnosis of pathology. In one sense, new medical imaging equipment is still in the process of becoming popular, and hospitals require extra time and funding to upgrade their facilities. On the other hand, physicians need time to become familiar with the use of more advanced imaging technologies. The development of a comprehensive diagnostic and follow-up system is urgently needed as medical systems evolve and patients live longer.

1.5 EviRed Project

There is an important obstacle to combating DR with the existing classification system based on old imaging technology, which is insufficient to finely predict the evolution of the disease: in 50% of cases, ophthalmologists overestimate or underestimate the possibility of complications. Compared to standard CFP, ultrawide-field photography provides useful information regarding the periphery of the retina that is absent from standard photography. Optical coherence tomography, which produces cross-sectional images with a resolution of a few microns, serves as the gold standard in the diagnosis of diabetic macular edema. In addition, it has been enhanced with OCTA, which can demonstrate the retina's vasculature non-invasively. Widefield OCTA is capable of demonstrating areas of non-perfusion, which are a hallmark of DR and cannot be resolved by fluorescein angiography alone. There is, however, an expanding amount of data produced by these new imaging modalities, which requires high levels of expertise from humans. It will be difficult for most ophthalmologists to develop a clinical score based on these factors. That is why it is important to propose a thorough review of DR diagnosis by replacing the existing classification with an expert system based on artificial intelligence (AI) that combines newly available imaging techniques and patient information (age, gender, blood pressure, and glycemic control, state of the other eye, etc.) to provide diagnosis and prediction. Further, this system is able to predict the development of complications in the next 12 months (macular edema or proliferative DR). The Évaluation Intelligente de la Rétinopathie diabétique (EviRed) consortium (AP-HP, UBO, ZEISS, Evolucare Technologies, Université Paris Diderot, and ADCIS) developed a fundus photo RD screening system based on artificial intelligence (CE marking and market release

pending), and it is reinforced by the world's leading eye imaging company, ZEISS, which has extensive research and development experience.

The main objective of the EVIRED project is to develop and validate an AI-based expert system assisting the ophthalmologist by improving prediction of evolution and decision-making during diabetic retinopathy (DR) follow-up. This main objective will consist of the validation of the prognostic tool and the evaluation of how accurately the algorithm can predict progression to severe retinopathy (defined by the presence of proliferative DR and/or severe macular edema involving the center of the macula or the need for laser photocoagulation, vitrectomy, or intravitreal injection) in the following year. It will replace the current diagnosis of DR using a classification mainly based on outdated fundus photography captured with a 45°field and providing an insufficient prediction precision. It will use Artificial Intelligence (AI) trained on a large data set of images (provided by the best fundus imaging devices available today), with medical data of importance concerning the patient, to provide a better diagnosis and a prediction of evolution. It should facilitate a better diagnosis, accurate prediction, and better decision-making by ophthalmologists during the follow-up of DR cases. As a result, critical progress should be made in the management of DR.

The Work program for Evired is shown in Fig. 1.3. As part of WP1, EviRed developed a platform and annotation tools for a virtual reading center. It was administered by the AP-HP (Lariboisière) as the reference center for DR. A total of ten ophthalmology departments and thirteen diabetology departments (a total of 5000 diabetic patients) were planned to contribute to clinical studies (WP2). My laboratory, LaTIM, is expected to develop algorithms based on multimodal data; to support this effort, the retrospective extraction of 2000 images was initially planned (WP3). The

algorithms were trained on various imaging technologies using a prospective dataset of 1000 diabetic patients. For improved diagnosis and risk prediction, multimodal information fusion combined all images and patient data. It is planned to compare the algorithm with current practices using a series of 1000 images read by a human expert using the current classification system. In addition, algorithms were applied to images captured by different systems in order to assess their reliability on those systems as well. Together with international experts, the AP-HP will develop a consensus on how to integrate the system into clinical practices and daily decision-making (WP4). With the assistance of Sécurité Sociale, the consortium will also conduct studies on improving general care pathways for diabetic patients. The medico-economic research unit of AP-HP will also assess the impact of AP-HP's system on care pathways and costs. In WP5, industrial partners will industrialize the system in a variety of formats, including stand-alone boxes, software, or integrated into machines. Evolucare and ADCIS, with their networks in France and abroad, and ZEISS, with its global marketing force, will be offering the system globally. WP6 manages all partners and entities involved in EviRed and handles appropriate IP transfers and dissemination.

The added-value of EviRed is to develop a medical decision support tool for diagnosis. In other existing projects, retrospective data was used, and artificial intelligence was used as a substitute for humans, primarily during screening. This project is unique in that it brings together knowledge of diabetology and ophthalmology, R&D academic skills, a large cohort of patients, companies who have already proven they can work together, and a world leader in eye imaging. Additionally, EviRed will minimize unnecessary burdens and costs by improving care personalization and by enabling faster treatment, which should preserve vision to a greater extent.

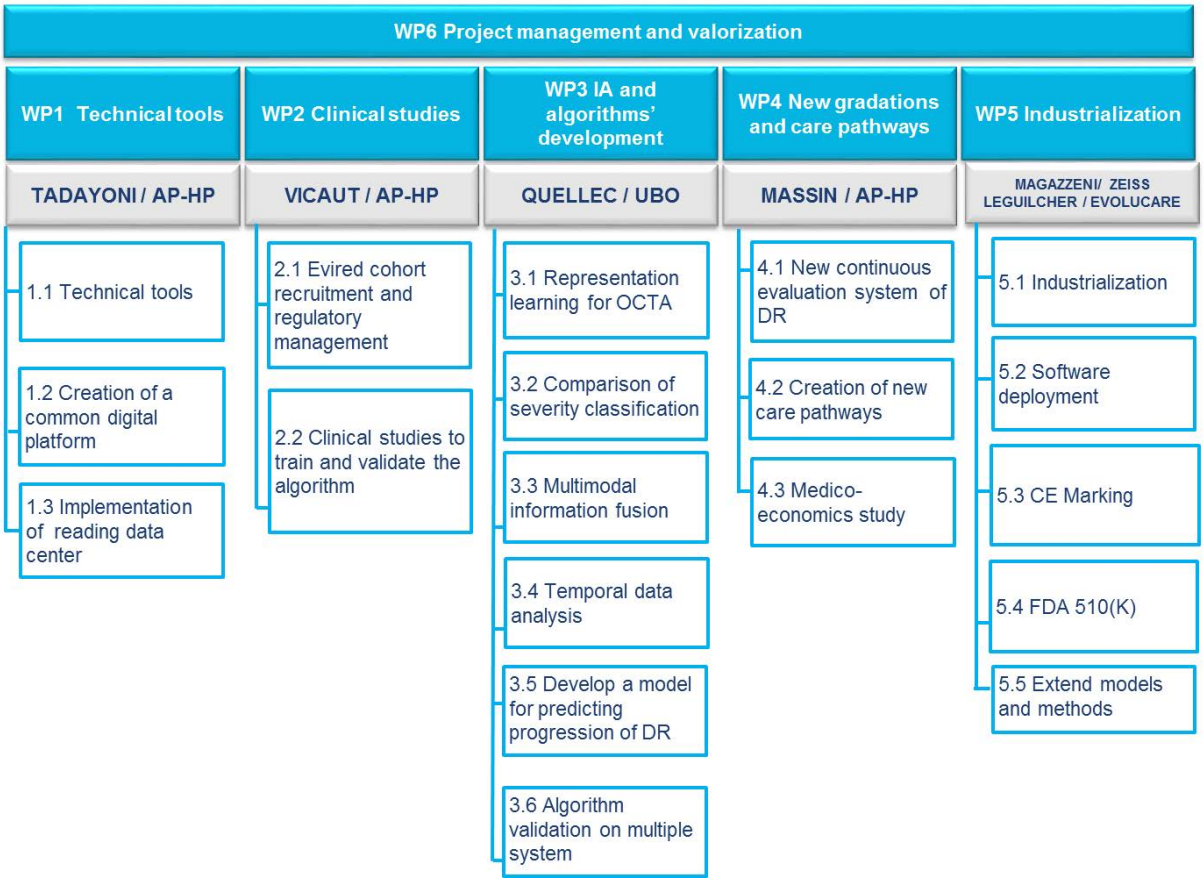


Figure 1.3 – Scientific program and structure of the proposal of Evired.

This would facilitate the diagnosis and make the procedure accessible to all ophthalmologists, thereby contributing to the spread of expertise.

1.6 Conclusion

This thesis was conducted as part of the Evired project, as part of WP3, and I intend to examine a fundamental aspect of artificial intelligence in this context: How can all images acquired during an eye examination, as well as information about a patient's clinical context be analyzed together to develop recommendations that are automatically generated? With the advent of new imaging technologies, the diagnosis of diabetic retinopathy has become much more accurate and comprehensive, and I hope to combine these advantages to arrive at an optimal diagnosis. The study of time-series analysis is being conducted by another PhD student in parallel. Jointly, the results of these two theses will address the main AI challenges of the Evired project. As part of my thesis, I will examine the following multimodal fusions:

— Joint analysis of multi-modal information in OCTA.
— Joint analysis of different specifications of OCTA acquisitions.
— Joint analysis of OCTA and UWF-CFP.

As a result, we will develop a new automated diagnostic model for diabetic retinopathy, which will utilize advanced imaging data for integrated analysis to obtain more accurate diagnostic results. In conjunction with new medical devices, the models we develop can assist ophthalmologists in reducing the amount of work and expertise required.

Chapter 2

State of the art literature review

"If you want the present to be different from the past, study the past."

— *Baruch Spinoza*

MULTIMODAL medical imaging plays a pivotal role in clinical diagnosis and research, as it combines information from various imaging modalities to provide a more comprehensive understanding of the underlying pathology. Deep learning-based multimodal fusion techniques have emerged as powerful tools for improving medical image classification. This chapter presents an extensive analysis of the developments in deep learning-based multimodal fusion for the diagnosis of diabetic retinopathy (DR) and other medical classification tasks. We examine the complementary relationships among the modalities for diagnostic DR and other common clinical modalities and discuss three major fusion schemes for multimodal classification networks: input fusion, intermediate fusion (subdivided into single-level fusion, hierarchical fusion, and attention-based fusion), and output fusion. By evaluating the performance of these fusion techniques, we provide insight into the suitability of different network architectures for the diagnosis of DR.

2.1 Image-based Computer-aided Diagnosis

2.1.1 Introduction

As part of the interface between medicine and computer science, computer-aided diagnosis (CAD) systems can be viewed as a cutting-edge expert and intelligence systems [106]. Diagnostic rules can be used in CAD systems to simulate the decision-making process of a skilled human practitioner in medicine. The development of CAD in medicine has been influenced by a number of factors [107]. In addition to the complexity of the medical diagnosis process itself, large amounts of complex clinical data pertinent to many diseases and conditions are also available. There is also a

large amount of diagnostic knowledge and advances in computer science (especially in the fields of AI and machine learning). The third artificial intelligence (AI) boom, as illustrated in Fig. 2.1, is gaining momentum. Particularly, the CAD field for medical images is undergoing a transformation due to the advent of ANN called deep learning, which is part of machine learning techniques [108].

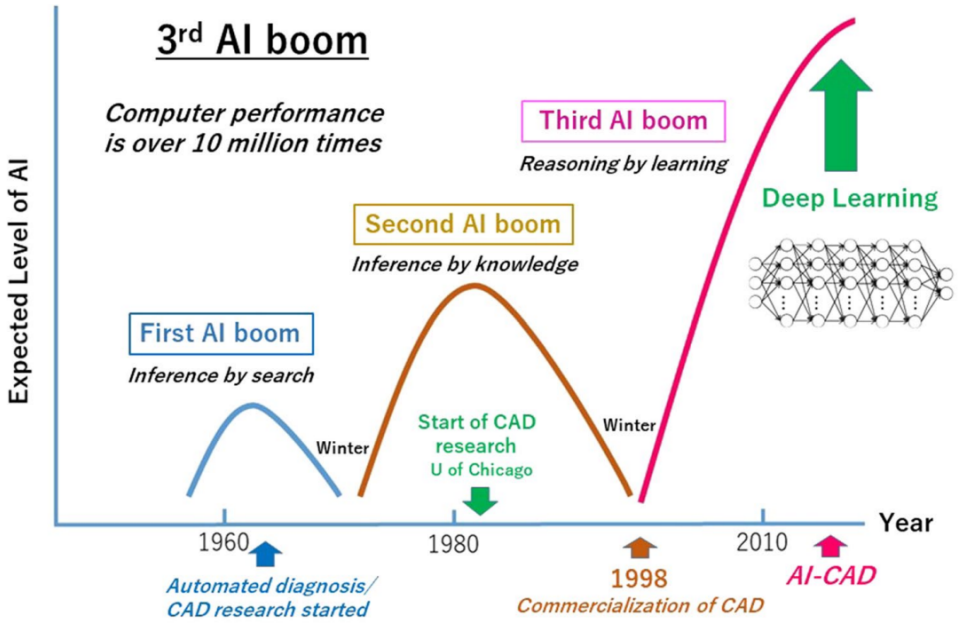


Figure 2.1 – There have been two AI booms in the past, and a third is currently underway. The progress of CAD research is closely related to the advancement of AI technology [108].

2.1.2 Deep Learning

A computer science field called AI is devoted to the creation of algorithms that solve problems that usually require human intelligence to solve [109]. The term machine learning refers to a subcategory of AI in which computers are able to learn without being explicitly programmed. As part of classic 'Machine Learning', human experts select imaging features that

appear to be most representative of the visual data and apply statistical techniques in order to classify the data based on these features [110]. The concept of 'Deep Learning', a subtype of machine learning, involves the use of representation learning without the use of feature selection [111]. It is instead the algorithm that determines which features are most effective for classifying the data by itself [109]. In the presence of sufficient training data, representation learning could potentially outperform hand-engineered features [112]. An illustration of the interrelationship between the three terms can be found in Fig. 2.2.

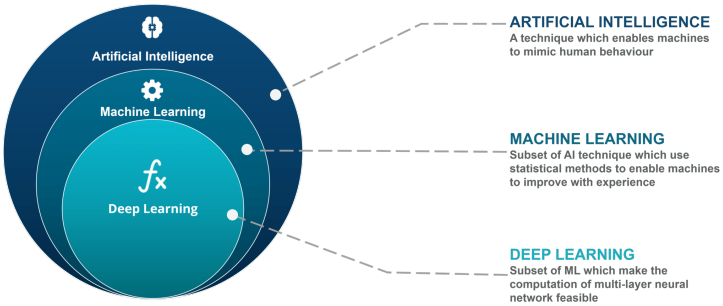


Figure 2.2 – Subsets of AI [113].

The majority of deep learning methods are based on artificial neural networks (ANN) [114]. The ANN is based loosely on the hypothesis that biological neural networks operate: data enters the dendrites of a neuron, something mathematical is done within the neuron, and the result is output through the axon [115]. The biological neural network is composed of the dendrites and axon terminals of many neurons that are interconnected in the cortex of the brain [116]. Fig. 2.3 illustrates an ANN example. The circles in the figure represent individual neurons, while the arrows connecting the neurons represent the weighting factors. Throughout the hierarchical structure, the left input layer is simulated to the right output layer. In its multilayered structure, a number of middle layers are interconnected.

There is a great deal of power in ANNs due to the large number of neurons that are arranged in interconnected deep hidden layers, which gives the network its computational power [108].

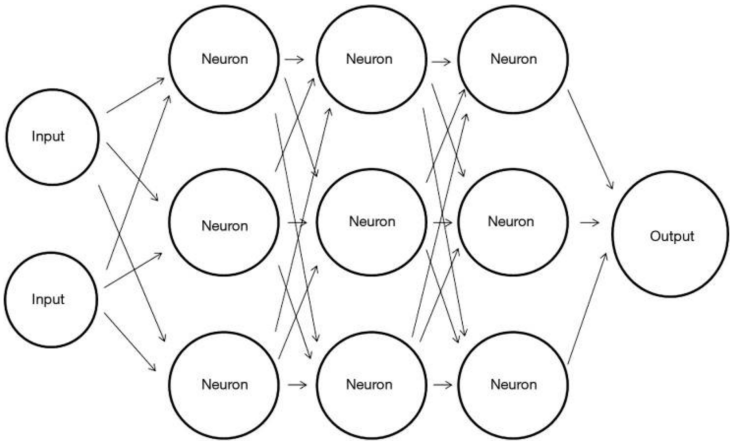


Figure 2.3 – Schematic representation of three hidden layers in a simple ANN [115].

Convolutional neural networks

The most commonly used network in medical image diagnosis is the convolutional neural network (CNN), a subtype of the ANN [117]. It was first proposed by LeCun et al. [118] in 1990 to classify digits and later used to recognize handwritten numbers on bank checks. However, the big breakthrough was achieved in 2012 with the ImageNet challenge [119].

Two key differences exist between ANNs and CNNs. First, CNNs perform convolution operations on images by sharing weights within the network [120]. This method eliminates the need for learning separate detectors for the same object occurring at different positions in an image, making the network equivariant with respect to translations [117]. In addition, it significantly reduces the number of parameters that need to be learned [115]. Secondly, CNNs typically incorporate pooling layers using a permu-

tation invariant function, such as max or mean, to aggregate pixel values of neighborhoods [120]. The convolutional layer can then be translated invariantly and have a larger receptive field [116]. Fully connected layers are usually added at the end of the convolutional stream of the network. The CNN structure is shown in Fig. 2.4, which includes three types of layers: convolution, pooling, and fully connected.

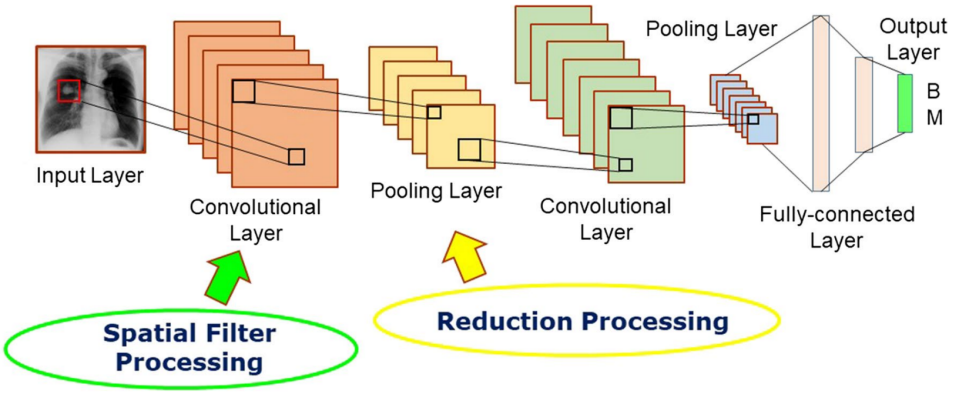


Figure 2.4 – An example of a CNN's classification of benign (B) and malignant (M) tumors on the chest radiograph. Processing of feature extraction takes place on convolutional layers and pooling layers, which are intermediate layers, and classifying processing occurs on all following fully connected layers. Convolutional layers act in a similar manner to spatial filtering in conventional image processing, while pooling layers function in a similar manner to reduction layers [108].

Transformers

In recent years, Transformers [121] have dominated the field of natural language processing (NLP), with applications in speech recognition [122], synthesis [123], text to speech [124], and natural language generation [125]. In addition to achieving unprecedented success on natural language tasks, Transformers have been successfully applied to numerous computer vision problems, achieving state-of-the-art results and leading researchers to reconsider the supremacy of convolutional neural networks as de facto op-

erators [126]. Taking advantage of these advances in computer vision, the medical imaging field has also seen a growing interest in Transformers that are capable of capturing global context as opposed to CNNs with local receptive fields [127]. During the past few years, the medical imaging community has witnessed an exponential increase in the use of Transformer-based techniques, especially after the introduction of Vision Transformer (ViT) [128].

Transformers are based on self-attention, which allows every element of a sequence to interact with every other and find out who they should pay more attention to [129]. As a result, they are better able to capture explicit long-range dependencies [130]. Other benefits of transformers include their ability to scale up more easily [131] and their resistance to corruption [132]. Further, their weak inductive bias enables them to perform better than CNNs when large-scale models and datasets are considered [131, 133, 134]. There has been a surge of interest in further developing Transformer-based models following encouraging results in several medical imaging applications [135–138].

An image classification model based on the basic architecture of conventional transformers is known as ViT. The structure of ViT is shown in Fig. 2.5. ViT converts input images into a series of patches, each encoded with a positional encoding method to provide spatial information [139]. When the patches are fed into the transformer along with the class token, the multi-head self-attention is calculated, and the learned embeddings of patches are output [129]. This image representation is derived from the state of the class token at the output of the ViT. Finally, the learned image representation is classified using a multi-layer perceptron (MLP) [128]. ViTs can also use feature maps from CNNs for relational mapping in addition to raw images [139].

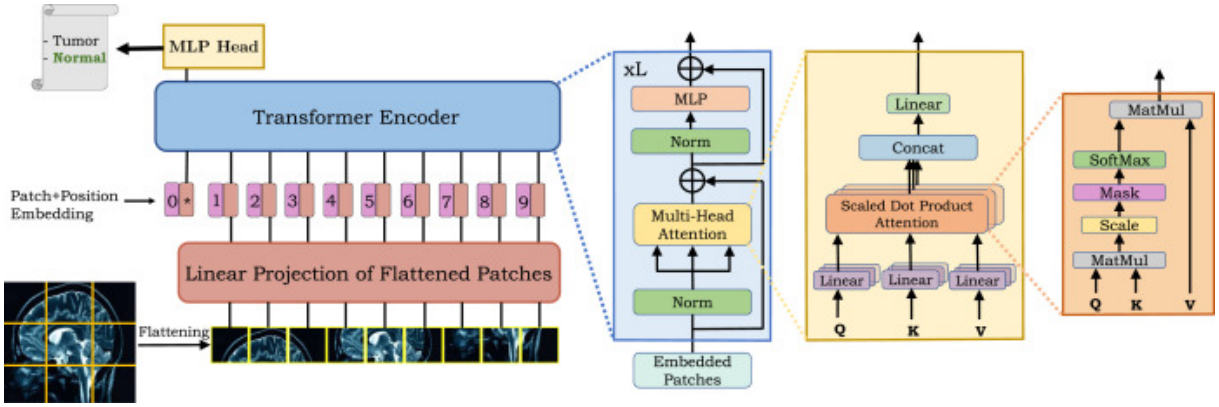


Figure 2.5 – The Vision Transformer architecture (left) and the encoder block (right). The vision transformer splits the input image into patches and projects them (after flattening) into a feature space where a transformer encoder processes them to produce a classification result [126].

2.1.3 Conclusions

CAD systems have made significant process recently. As their performance increase, they become more and more relevant and could potentially reduce the workload of clinicians in various use cases. For this reason, deep learning techniques are the first and most essential element of our research.

2.2 Information fusion techniques for multimodal medical image classification

Due to the fact that we are at the beginning of a new ophthalmic photography technology, there are currently very few multimodal fusion algorithms in which we are interested. In light of the current deep learning DR diagnosis papers, it is difficult to summarize and propose innovations for multimodal fusion methods. As a consequence, we have broadened our scope and proposed exploring multimodal methods in the entire medical image field, and after summarizing and learning systematic multimodal fu-

sion methods from them, we will use these methods in order to diagnose DR.

2.2.1 Introduction

Context

In recent years, the field of medical image analysis has attempted to apply deep learning-based methods to the classification of various diseases, notably related to the brain [140–142], breasts [143–145], prostate [146–148] and eyes [149, 150]. The ability to accurately classify and diagnose diseases from medical images has the potential to revolutionize healthcare by improving diagnostic accuracy, reducing human error, and enabling more personalized treatment planning. This has driven the need for robust and efficient methods for analyzing medical images from multiple imaging modalities.

With advances of medical image acquisition systems, many new imaging modalities have been used to diagnose patients [151–153], resulting in larger and more diverse datasets. An imaging modality alone does not usually provide all the information needed to ensure accurate clinical diagnosis. Therefore, clinicians increasingly base their diagnosis on images obtained from a variety of sources: a combination of abundant information can be used in clinical practice with more confidence. Following this trend and to improve diagnosis results, AI-based classification models are increasingly being developed by combining data from multiple modalities to take advantage of both redundancies and complementarities across modalities. By using multimodal approaches, medical images from a variety of modalities are combined to provide complementary information that can contribute to improving diagnostic results.

Traditional methods

Non-deep learning-based information fusion strategies, relying on traditional image processing and machine learning, have been reviewed in a previous survey [154]. We summarize hereafter the main developments and highlight the benefits of non-deep learning-based information fusion.

Input fusion is the most commonly used strategy among traditional methods. It involves the fusion of images from various modalities into structured data and fuses them into different categories depending on the fusion domain: spatial fusion [155–159], frequency fusion [160–165] and sparse representation [166–168]. In spatial fusion, multimodal images are combined at the pixel level, but this approach often leads to spectral degradation [169] and color distortion [170]. Frequency fusion, which involves transforming the input image into the frequency domain, is more complex and results in limited spatial resolution [171]. Sparse representation, on the other hand, can be sensitive to registration errors and lacks attention to details [170]. These limitations highlight the need for more advanced techniques, such as deep learning-based multimodal fusion methods, able to overcome the challenges faced by traditional methods.

Other strategies include intermediate and output fusion, which do not require registration of the input images. Intermediate fusion involves extracting features from different imaging modalities, concatenating them, and feeding them into a classifier, generally a support vector machine (SVM), for diagnosis [172–174]. This approach requires extensive testing and rich domain knowledge for feature extraction and selection. On the other hand, output fusion involves stacking the data results from unimodal models, such as SVM [175]. Traditional methods contain complex pre-processing steps and simple model structures, which often result in in-

Table 2.1 – A list of multimodal image datasets. The list is sorted by the number of publications on PubMed (Keywords: dataset name AND 'multimodal').

Dataset	Year	Modalities	Body Organ(s)	Medical Diagnosis
ADNI	2004	sMRI, fMRI, PET	Brain	Alzheimer's Disease
BraTS	2012	MRI (TI, T2, T1c, FLAIR)	Brain	Brain Tumor
TCIA	2014	CT, MRI, PET, US, etc.	Brain, Breast, Lung, Kidney, Head-Neck, Liver, Pancreas, etc.	Common Cancer Disease
OASIS	2007	MRI, PET	Brain	Alzheimer's Disease
SPC	2018	Dsc, Clinical Image, Metadata	Skin	Skin Lesion
TCGA	2006	Pathological data, Genomic data	Brain, Lung, etc.	Common Cancer Disease
ABIDE	2012	sMRI, fMRI	Brain	Autism Spectrum Disorder (ASD)
ADHD-200	2011	sMRI, fMRI	Brain	Attention Deficit Hyperactivity Disorder (ADHD)
COBRE	2012	sMRI, fMRI	Brain	Schizophrenia
GAMMA	2021	OCT, Fundus Image	Eye	Glaucoma
CPM-RadPath	2019	MRI (TI, T2, T1c, FLAIR)	Brain	Brain Tumor
ISIT-UMR	2019	White Light RGB, Narrow Band Imaging (NBI)	Digestive Tract	Gastrointestinal Lesions
MRNet	2018	MRI (TI, T2)	Knee	Knee Injuries
CTU-UHB	2014	CT (FHR, UC)	Uterus	Fetal Distress Diagnosis

formation loss during feature extraction, making it difficult to fully exploit the complementarities between different modalities.

Besides requiring domain knowledge, these traditional multimodal fusion approaches do not fully utilize the complementary between multimodal features. In contrast, deep learning network architectures offer complex models that can explore more possibilities for multimodal fusion. Furthermore, various end-to-end models significantly reduce the amount of domain knowledge required for diagnosis purposes. This has led to the exploration of alternative approaches based on deep learning to address the challenges faced by traditional methods.

Paper selection

In our initial literature search, a total of 14 public multimodal image datasets were found: these datasets are listed in Tab. 2.1. The 14 public multimodal datasets will be described in detail in Sect. 2.2.2, and our summary is provided in Tab. 7.2. The final list of papers analyzed was established as follows. For each of the 14 datasets, we searched on Pubmed all publications mentioning the dataset name plus ((multimodality) OR (multimodal) OR (multi-modal) OR (multiparametric) OR (multiparametric)). Next, the resulting 14 lists were merged. Finally, based on the abstracts, we manually filtered articles that discuss multimodal information fusion using deep learning methods. Unfortunately, any public multimodal datasets focusing on classification tasks are available for some organs (breast, lung, prostate, kidneys, larynx, heart, liver) mentioned in the multimodal medical image analysis literature. Therefore, to broaden the scope of the search, we also included 19 relevant articles targeting these organs that use private datasets. This resulted in a list of 90 publications.

Taxonomy

As discussed in Sect. 2.2.1 and other surveys [151, 176, 177], multimodal fusion methods are traditionally classified as *input fusion*, *intermediate fusion* or *output fusion*, based on the stage of information fusion in the classification pipeline, as in Fig. 2.6(a). Note that some publications refer to input fusion as early fusion, while intermediate fusion may be considered as feature-level fusion, and output fusion is equivalent to decision-level fusion or late fusion [176, 177]. We show that intermediate fusion is currently the most popular category. In order to provide a broader understanding of multimodal deep learning networks, we further divide intermediate fusion

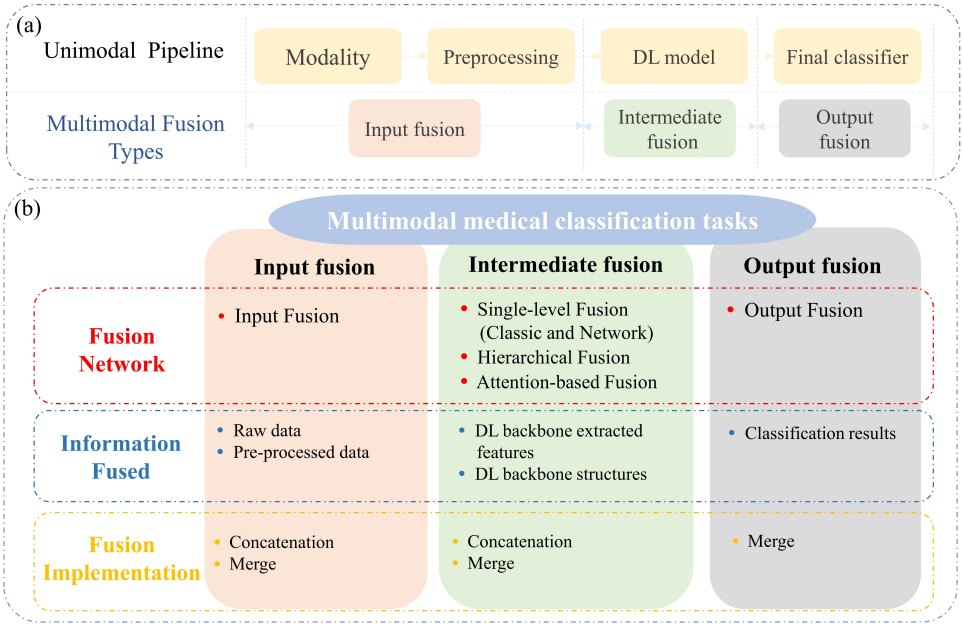


Figure 2.6 – (a) Unimodal classification task flow and different types of multimodal fusion based on the stages in which they perform information fusion. (b) Information fusion networks for the three types of multimodal fusion, inputs to information fusion, and the implementation of information fusion.

into *single-level fusion*, *hierarchical fusion* and *attention-based fusion*, as illustrated in Fig. 2.6(b). The proposed taxonomy is detailed and discussed in Sect. 2.2.4: it covers the majority of the current multimodal classification network architectures, providing insight into their stages and styles of information fusion.

2.2.2 Multimodal medical images

Imaging modalities

For medical diagnosis purposes, each imaging modality has its own characteristics and information. Different medical imaging modalities use different frequency bands of the electromagnetic spectrum in order to screen and diagnose different medical conditions in the human body [152]. There are different wavelengths and frequencies associated with each imaging

Table 2.2 – Typical imaging modalities and organs found in the multimodal medical image analysis literature.

Modalities	Body organs examined	Invasive/ Non-invasive	Description
Magnetic Resonance Image (MRI)	Brain, Prostate, Breast, etc.	Non-Invasive	In addition to high spatial resolution and exquisite soft tissue contrast, MRI can also display dynamic physiologic changes in three dimensions [179].
Positron Emission Tomography (PET)	Brain, Prostate, Breast, etc.	Invasive	The PET provides information about the organs' activity, as well as its sugar use as energy[180].
Computed Tomography (CT)	Lung, Bone, Oral, etc.	Non-Invasive (harmful)	CT is an excellent tool for detecting bone, joint, and soft tissue lesions that may affect bone, joints, or soft tissues [181].
Ultrasound (US)	Abdomen, Breast, etc.	Non-Invasive	In addition to showing the activity and function of certain organs in the body, US can also identify whether a tissue or organ contains fluid or gas [182].
Optical Coherence Tomography (OCT)	Eye, Heart	Non-Invasive	Biological tissues can be visualized in high-resolution with OCT scanning in two-dimensional or three-dimensional modes [38].
Dermatoscope (Dsc)	Skin	Non-Invasive	Dsc allows better visualization of subsurface structures and improved identification of skin diseases [183].

modality, as well as different characteristics (structure, function, etc.) [178]. Furthermore, medical imaging modalities can be classified as invasive or non-invasive. Invasive methods involve inserting an object into the body through an incision or needle injection in order to examine an organ, while non-invasive methods utilize some form of radiation or sound [152]. Table 2.2 shows some modalities that appear in multimodal medical image datasets.

Due to their complementary nature, there has been a significant focus on the following combinations of modalities targetting various diseases: (1) multi-parametric MRI (TI, T2, T1C, FLAIR) [184, 185], (2) MRI and PET [186, 187], (3) PET and CT [188], (4) multi-view ultrasound (US B-mode, US color Doppler) [143, 145], (5) Color Fundus Photographs (CFP) and Optical Coherence Tomography (OCT) [17], (6) Dsc and Clinical Im-

age [189, 190], and (7) combined diagnosis of Image Data and Clinical Data. The complementary relationships between these modal images will be briefly discussed.

Neurology and neurosurgery frequently use MRI. Different MRI images can be obtained by changing the factors affecting the magnetic resonance (MR) signal, and these different images are referred to as sequences. Depending on the sequence used, the behavior of tumors may vary, and it is essential to use multiple sequences to accurately determine tumor location and size [191]. T1-weighted (T1) and T2-weighted (T2) MRIs are the most common MRI sequences. Tomographic anatomical maps can be observed with the T1 sequence. The T2 sequence clearly shows the location and size of the lesion, but the puffy area around the tumor is blurred and difficult to discern [192]. To overcome this, the Fluid Attenuated Inversion Recovery (Flair) sequence is used. It provides better visualization of the area of puffiness around the tumor site, making it easier to detect the tumor's boundaries [193]. Furthermore, contrast-enhanced T1-weighted (T1c) sequences can be used to detect intra-tumor conditions and distinguish tumors from non-tumorigenic lesions [194]. T2 and Flair are suitable for detecting tumors with peritumoral edema, while T1 and T1c are suitable for detecting tumors without peritumoral edema [195].

Diffusion-weighted imaging (DWI) is another useful sequence designed to detect the random movements of water protons. Therefore, DWI sequence is a highly sensitive method for detecting acute strokes [196]. An increased apparent diffusion coefficient (ADC) value with lower signals of DWI images could reveal the fast diffusion of water molecules [197]. In addition to using multiple sequences, co-diagnosis using structural MRI (sMRI) and functional MRI (fMRI) is becoming increasingly popular [198, 199]. fMRI measures the small changes in blood flow that occur with brain

activity. This test can be used to determine which parts of the brain are performing critical functions and to determine the effects of strokes and other diseases on the brain [200].

The combination of PET and MRI, PET and CT has been recognized as a valuable method for screening and diagnosing various diseases [201–205]. The PET scan is preceded by the administration of a radioactive agent to the patient. This allows doctors to determine the metabolic processes in which the brain tissue is involved [180]. Compared to other imaging methods such as CT and MRI, PET has a high sensitivity and can detect lesions even if MRI/CT does not yet show abnormalities. PET also has high specificity, making it possible to determine whether a tumor is malignant based on its metabolism at the time of MRI/CT detection [206]. However, because PET scan lacks information about organ anatomy, they should be conducted in conjunction with CT/MRI scans [198]. In summary, the combination of PET and MRI/CT scans provides structural and functional information related to various diseases, improving the effectiveness of diagnosis. Fig. 2.7 shows the images of PET, CT, and MRI, as well as several sequences of MRI.

Availability, low cost, and safety make ultrasonography the most widely used clinical diagnostic tool. Conventional B-mode imaging is used to examine abnormal masses in tissues, while Color Doppler imaging shows the distribution of blood vessels within tissues [207]. The combined use of these two modalities is common in identifying cervical lymph nodes [208], diagnosing breast cancer [143, 145], and so forth [209–211].

In the diagnosis of ophthalmic diseases, CFP and OCT are the two most cost-effective methods [16]. These imaging modalities provide prominent biomarkers that can be used to identify glaucoma suspects, such as the vertical cup-to-disc ratio (vCDR) on fundus images and the retinal nerve

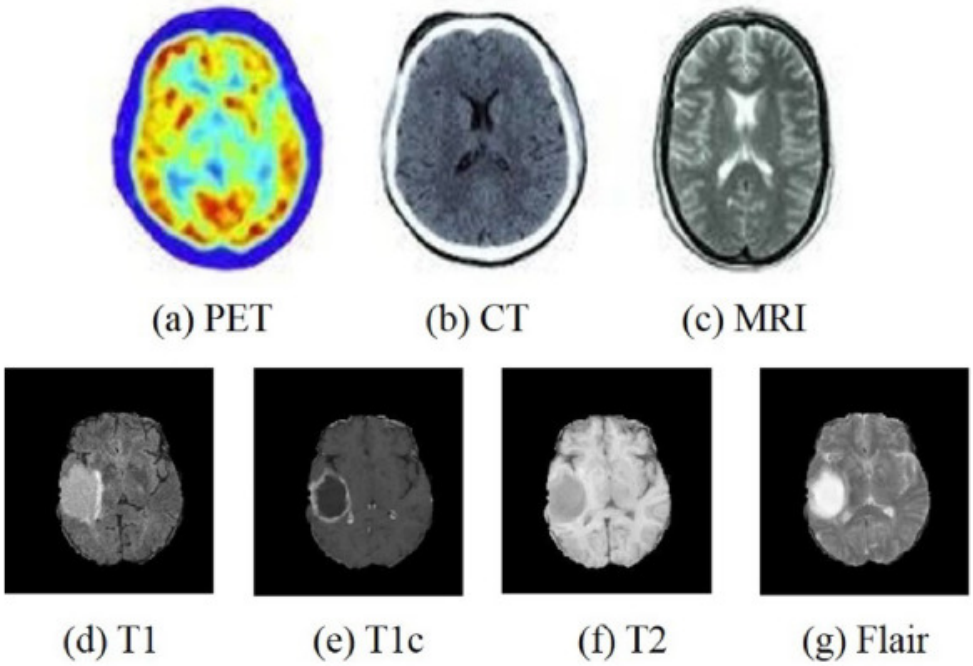


Figure 2.7 – (a)-(c) are the images of PET, CT, and MRI. (d)-(g) are the different sequences of MRI. Images from [195], with permission of the first author.

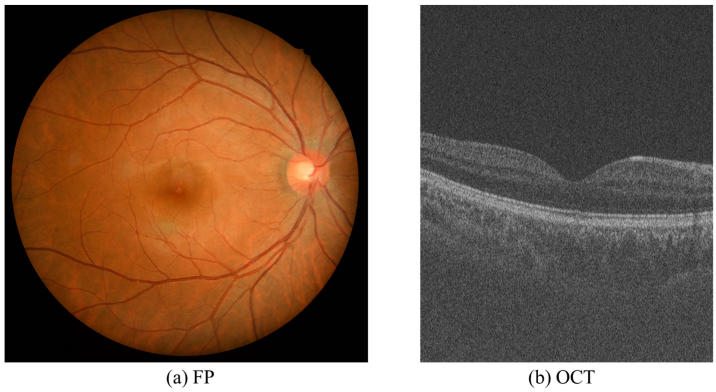
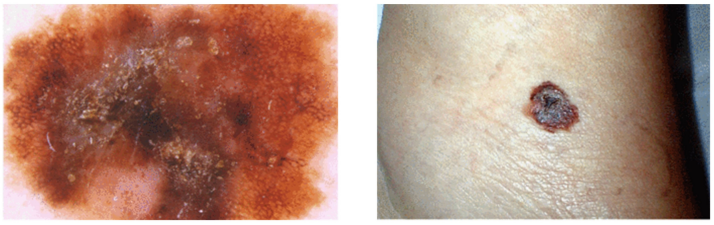


Figure 2.8 – Images of CFP and OCT from GAMMA challenge [17].

fiber layer thickness (RNFL) on an OCT image. A more accurate and reliable diagnosis is often achieved by taking both screenings in clinical practice [17]. Fig. 2.8 shows the images of CFP and OCT.

In the diagnosis of skin cancer, a combination of dermoscopic and clinical images is often used [189]. The clinical image is captured using a digital



(a) Dermoscopic images (b) Clinical images

Figure 2.9 – Dermoscopic and clinical images. Image from public datasets SPC [190].

camera and shows the visualized feature in different views and lighting conditions. On the other hand, dermoscopic images provide a clear view of the skin's subsurface structures and are obtained using a specific skin imaging technique in contact with the skin [212]. Fig. 2.9 shows the dermoscopic and clinical images.

In addition to multimodal image combinations, clinical information regarding the patient's medical history and symptoms can significantly contribute to the diagnosis of the disease. The textually recorded clinical data may contain implicit features that may improve the model's classification performance. Electronic Health Records (EHR) are commonly used to detect brain diseases by integrating image analysis features [213, 214]. Similarly, skin cancer detection also relies heavily on metadata [189, 215].

Multimodal image datasets

In the early stages of multimodal medical diagnosis, multimodal datasets are particularly valuable for testing various networks and developing fusion methods. However, the privacy and cost of medical images often make obtaining more comprehensive multimodal datasets challenging for researchers. Fortunately, there are several freely available multimodal datasets. These datasets provide information regarding the diagnosis of diseases at various locations in the body, as well as the analysis of various multimodal

combinations. These datasets are expected to contribute to the analysis of fusion methods and serve as a foundation for the future development of multimodal fusion methods.

Alzheimer's Disease Neuroimaging Initiative (ADNI) [1] is a multi-center longitudinal study to discover clinical, imaging, genetic, and biochemical biomarkers for Alzheimer's disease (AD). ADNI has three stages: ADNI 1 included 400 subjects diagnosed with mild cognitive impairment (MCI), 200 subjects with early AD, and 200 elderly control subjects; ADNI 2 added new participant groups: 150 elderly controls, 100 EMCI subjects, 150 late mild cognitive impairment (LMCI) subjects, and 150 mild AD patients; ADNI 3 added hundreds of new MCI subjects, mild AD subjects, and elderly controls. The MRI Brain Tumor Segmentation (BraTS) [2] challenge has been held since 2012 and currently includes classification tasks in addition to tumor segmentation. Each subject has four MRI modalities (T1, T1C, T2, and T2 FLAIR), human annotation of tumor segmentation, and tumor grade. The Cancer Imaging Archive (TCIA) [3] is a large-scale public database containing medical images of common tumors (lung cancer, prostate cancer, etc.) and corresponding clinical information (treatment protocol details, genetics, pathology, etc.). Open Access Series of Imaging Studies (OASIS) [4] seeks to make neuroimaging datasets freely accessible to the scientific community. OASIS-3 contains 755 cognitively normal adults and 622 individuals at various stages of cognitive decline ranging in age from 42-95 years. Seven-point Criteria Evaluation Database (SPC) [5] provides a database for evaluating computerized image-based prediction of the 7-point malignancy checklist for skin lesions. The dataset

1. https://adni.loni.usc.edu/
2. http://braintumorsegmentation.org/
3. https://www.cancerimagingarchive.net/
4. https://www.oasis-brains.org/
5. https://derm.cs.sfu.ca/Welcome.html

contains more than 2000 clinical and dermoscopy color images and structured metadata for training and evaluating CAD systems [190]. As part of the Cancer Genome Atlas (TCGA) [6], an internationally recognized cancer genomics project, more than 20000 primary cancer samples and matched normal samples were molecularly characterized [216]. The Autism Brain Imaging Data Exchange (ABIDE) [7] initiative now includes two large-scale collections, ABIDE I and ABIDE II, whose ultimate goal is to facilitate discovery science and comparative analysis across samples. ABIDE I contains 1112 datasets, including 539 from individuals with ASD and 573 from typical controls (ages 7-64 years, median 14.7 years across groups). ABIDE II contains 1114 datasets from 521 individuals with ASD and 593 controls (age range: 5-64 years). ADHD-200 Sample [8] is a grassroots initiative that aims to improve scientific understanding of the neural basis of ADHD through the implementation of open data sharing and discovery-based research methods. The Center for Biomedical Research Excellence (COBRE) [9] is providing raw anatomical and functional magnetic resonance imaging data from 72 patients with schizophrenia and 75 healthy controls (ages ranging from 18 to 65 in each group). The Glaucoma Grading from Multimodality Images (GAMMA) [10] Challenge is intended to facilitate the development of fundus and OCT-based glaucoma grading. GAMMA contains 2D fundus images and 3D OCT images of 300 patients. Computational Precision Medicine: Radiology-Pathology Challenge on Brain Tumor Classification 2019 (CPM-RadPath) [11] is a brain tumor classification challenge. Each patient contains multiple MRI sequences: T1, post-contrast

6. https://www.cancer.gov/ccg/research/genome-sequencing/tcga
7. http://fcon_1000.projects.nitrc.org/indi/abide/
8. http://fcon_1000.projects.nitrc.org/indi/adhd200/
9. http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html
10. https://aistudio.baidu.com/aistudio/competition/detail/90/0/introduction
11. https://www.med.upenn.edu/cbica/cpm-rad-path-2019/

T1-weighted (T1Gd), T2, and FLAIR. ISIT-UMR [12] is a dataset for the classification of gastrointestinal lesions in regular colonoscopy. The dataset consists of 76 polyps with white light and NBI videos from the same polyp. The MRNet [13] dataset consists of 1,370 knee MRI exams performed at Stanford University Medical Center between January 1, 2001, and December 31, 2012. There were 1104 (80.6%) abnormal exams in the dataset, with 319 anterior cruciate ligament (ACL) tears and 508 meniscal tears. CTU-UHB [14] is a database containing 552 cardiac tomography recordings from the Czech Technical University (CTU) in Prague and the University Hospital in Brno (UHB). As part of each CT, a fetal heart rate time series (FHR), as well as a uterine contraction (UC) signal, are recorded.

The previously mentioned datasets provide valuable resources for developing and testing multimodal fusion methods. They contain images of different medical modalities of the same patient, as well as images of different patients. Access to these datasets is available upon request and at no cost. We summarized the fusion methods presented in 41 articles that use ADNI, 11 articles that use TCIA, 4 articles that use BraTS, 3 articles that use COBRE, 3 articles that use ADHD-200, 2 articles that use CPM-RadPath, 2 articles that use ABIDE, 2 articles that use GAMMA, 2 articles that use SPC, 2 articles that use MRNet, 1 article that uses TCGA, 1 article that uses CTU-UHB, 1 article that uses OASIS, and 1 article that uses ISIT-UMR.

12. http://www.depeca.uah.es/colonoscopy_dataset/
13. https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002699
14. https://physionet.org/content/ctu-uhb-ctgdb/1.0.0/

Table 2.3 – Multimodal classification pipeline.

Stage	Description
Data preprocessing	The initial step of the classification task is to perform operations such as registration, denoising, and data augmentation on the raw data.
DL backbone	Extraction of high-dimensional features of data by the deep learning network structure.
Information fusion	Fusion of multimodal data/features by different methods.
Final classifier	The final stage of generating classification results from multimodal data.
Model evaluation	Different metrics are used to evaluate the performance of multimodal models.

2.2.3 Multimodal classification pipeline

Multimodal fusion research is still in its infancy, and the current definitions of fusion methods and processes are unclear. Many publications use the terms input fusion, intermediate fusion, and output fusion, but these terms do not necessarily have the same meaning. To provide a standardized framework for multimodal classification, we follow the five-stage pipeline proposed in [217], as shown in Tab. 2.3. These five stages can be used to summarize all medical multimodal classification tasks. This section provides clear definitions of each stage and presents the methods for implementing them. According to the order and structure of the information fusion stage and the deep learning (DL) backbone stage, we categorize the multimodal fusion methods into five methods in Section 4.2.3.

Pre-processing

Image pre-processing is crucial for multimodal medical classification tasks, as it enhances DL network efficiency and effectiveness in extracting features. Pre-processing techniques, such as image registration, cropping, denoising, resampling, intensity normalization, regions-of-interest (ROI)

extraction [218–220], and feature selection [143, 198, 221, 222], prepare the data for more accurate and efficient analysis by DL models.

To further improve the performance of these models, data augmentation techniques play an essential role in the pre-processing pipeline. For example, data augmentation helps prevent overfitting [223] using methods like random cropping, flipping, and rotation during training. In addition, increasing the training dataset's diversity improves the model's generalization capabilities.

Considering the large volumes of data generated by multimodal medical images, it is noteworthy that only a small fraction is relevant to diagnosing diseases. Therefore, feature selection emerges as a crucial pre-processing step, aiming to reduce data dimensionality while retaining pertinent information. Common feature selection methods include manual selection [224–226] and Principal Component Analysis (PCA) [227–229].

Another critical aspect of pre-processing in the context of multimodal medical images is image registration. By aligning multiple images from various modalities, image registration ensures the accurate matching of corresponding anatomical structures across image types. This alignment facilitates comprehensive data analysis and becomes particularly critical for input-level fusion, where combining complementary information from different modalities depends on proper alignment.

Information fusion

A key component of multimodal image classification is information fusion. Information fusion can be divided into input fusion, intermediate fusion, and output fusion based on the level at which information is fused. And there are two ways to achieve fusion [217], namely concatenation and merge. Concatenation involves the concatenation of data from different

Figure 2.10 – Two types of fusion. Orange and green: data of different modalities. Blue: the output fused data.

modalities into a single tensor for the next step. Merge involves complex calculations such as adding data from different modalities; the final result is a smaller amount of data. Fig. 2.10 illustrates the two types of fusion. Our study focuses on the fusion of different medical imaging modalities, and in the next section, we will examine the different fusion methods in greater detail.

Deep learning backbone

DL backbones are used to extract high-dimensional features of modalities during the modal classification process. Over recent years, several high-performing network architectures have emerged, including AlexNet [119], VGG [230], GoogLeNet [231], ResNet [232], DenseNet [233], AE [234–236], ViT [237], and others, providing state-of-the-art performance in classification. A summary of the common architectures for DL is presented in Tab. 2.4. DL has developed rapidly due to several factors, including the development of hardware devices like graphics processing units (GPUs) and tensor processing units (TPUs), which have greatly improved the training speed of DL networks. Additionally, publicly available datasets such as ImageNet [238] have facilitated the training and testing of various models. Furthermore, DL is capable of learning advanced features directly from data without requiring extensive expertise or prior experience, making it easily adaptable across various domains.

Table 2.4 – Some common architectures of deep neural networks. Different architectures are more suitable for different types of data.

Architecture	Description
Fully Connected Neural Network (FCNN)	FCNN are the most traditional deep neural networks. Every neuron in a layer is connected to every neuron in the layer below it [239].
Convolutional Neural Network (CNN)	CNN can model spatial structures, such as images or volumes. Convolutional kernels model local information by sliding over input data [239].
Autoencoders (AE)	By compressing and reconstructing the input data, AE learns low-dimensional encoding. There are different types of layers, such as convolutional and fully connected [240].
Transformer	Transformer is a model that uses a multi-headed attention mechanism. Feature extraction is solely based on attention [241].

In input fusion, a single backbone can extract features from fused modalities. However, in other fusion schemes, such as intermediate or output fusion, multiple DL backbones may be used to extract features from different modalities. In current multimodal fusion research, Convolutional Neural Networks (CNN) are the preferred choice of the majority of researchers due to their effectiveness in feature extraction from medical images. Many pre-trained models have already been tested on large datasets, making them suitable for use in medical imaging research. In the articles analyzed, CNNs were used in 65 articles, Fully Connected Neural Networks (FCNN) in 10 articles, Auto-Encoders (AE) in 8 articles, and Transformers in 6 articles.

Final classifier

Multimodal classification employs a final classifier to generate the classification results based on multimodal features or multiple independent

classification results, depending on the fusion scheme employed. In DL networks, the Fully Connected (FC) layer [188, 198, 199, 242] is often used as the final classifier. Other methods, such as SVM [227, 243], Random Forest [144], and Score Merge [244, 245] can also be used as final classifiers.

Evaluation metrics

Evaluation metrics for multimodal fusion tasks are similar to those used in unimodal classification tasks. Commonly used indicators for assessing the performance of multimodal fusion methods and DL networks in the context of medical classification tasks include True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These indicators can be used to calculate several performance metrics, such as sensitivity, specificity, accuracy, precision, and F1 score, among others. Additionally, the Area Under the Curve (AUC) and Kappa are commonly used metrics to evaluate medical classification tasks. The Receiver Operating Characteristic (ROC) curve is a graphical representation of a classifier's performance, and AUC quantifies the classifier's ability to distinguish between different classes, with a higher AUC indicating better discrimination.

- Accuracy (ACC) = $\frac{TP+TN}{TP+TN+FP+FN}$
- Sensitivity (SEN) = $\frac{TP}{TP+FN}$
- Specificity (SPEC) = $\frac{TN}{TN+FP}$
- F1 Score = $\frac{2\times TP}{2\times TP+FP+FN}$
- Positive Predictive Value (PPV) = $\frac{TP}{TP+FP}$
- Negative Predictive Value (NPV) = $\frac{TN}{TN+FN}$
- Area Under the receiver operating characteristic Curve (AUC)
- Cohen's Kappa (Kappa) = $\frac{p_0-p_e}{1-p_e}$

where p_0 is the accuracy and p_e is the sum of the products of the actual and predicted numbers corresponding to each category, divided by the square

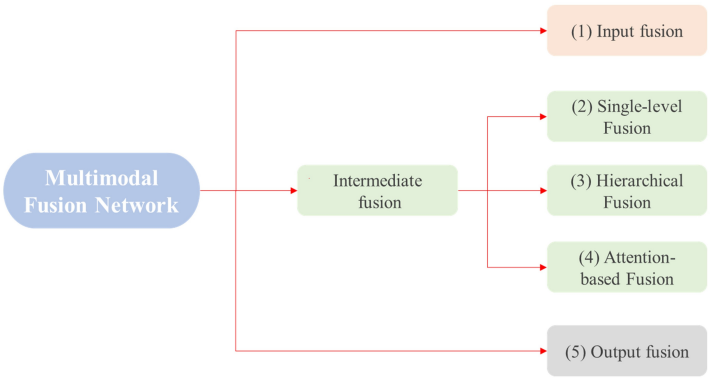


Figure 2.11 – Five types of multimodal fusion networks.

of the total number of samples.

2.2.4 Multimodal classification networks

Information fusion taxonomy for multimodal image classification

The positions of pre-processing and the final classifier are fixed during the process of multimodal classification. Based on the number and sequence of DL backbones and information fusion step, multimodal DL network architectures can be categorized into five types: input fusion, single-level fusion, hierarchical fusion, attention-based fusion, and output fusion, as shown in Fig. 2.11. As explained hereafter, single-level, hierarchical, and attention-based fusion are sub-categories of intermediate fusion. These categories describe how the network processes and combines the input modalities to produce classification results.

(1) **Input Fusion** can also be referred to as input-level fusion, where the information fusion phase precedes the DL backbone. Concatenation and Merge are two methods of information fusion. For the concatenation method, data of different modalities are used as different channels of the input. In the merge approach, data is fused at the pixel or voxel level, and the merged images are used as inputs for the DL classifier. The process

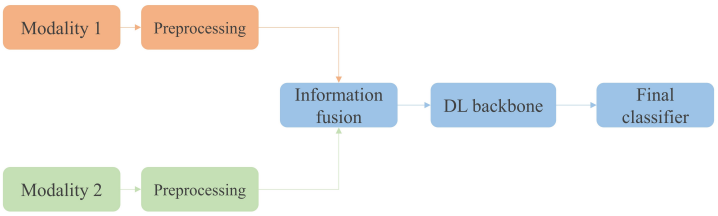


Figure 2.12 – Input fusion process diagram. Information fusion: Concatenation/Merge (Inputs).

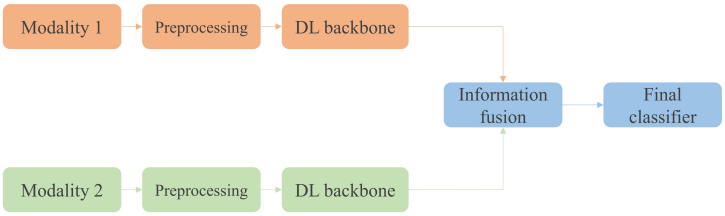


Figure 2.13 – Single-level fusion process diagram. Information fusion: Concatenation/Merge (Classic).

diagram for input fusion is shown in Fig. 2.12.

(2) **Single-level Fusion** involves information fusion after the DL backbone but before the final classifier. As part of a single-level fusion, the features extracted by the DL backbone are fused only once at some point before the classifier is applied. Depending on the network structure, it can be divided into two types: Classic Fusion and Network Fusion. In Classic Fusion, high-dimensional features are extracted from different modalities using different DL classifiers and then merged or concatenated. This is the most common network structure in intermediate fusion, so we call it *Classic*. Fig.2.13 illustrates the process diagram of Classic Fusion. In Network Fusion, the intermediate features of different modalities are first extracted using DL classifiers, followed by the extraction of high-level features of the fused modalities using additional DL backbones. Fig.2.14 shows the process diagram of Network Single-level Fusion.

(3) **Hierarchical Fusion** is an improvement over single-level fusion. In

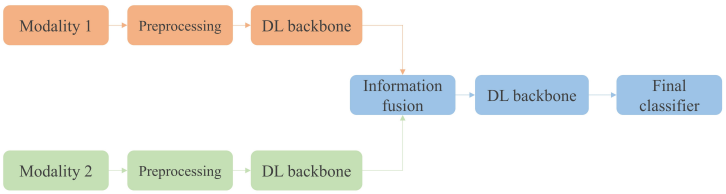


Figure 2.14 – Single-level fusion process diagram. Information fusion: Concatenation/Merge (Network).

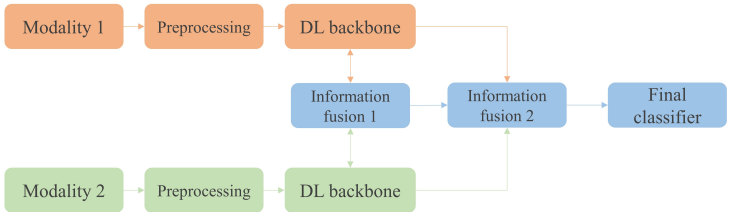


Figure 2.15 – Hierarchical fusion process diagram. Information fusion 1: Concatenation/Merge (Network). Information fusion 2: Concatenation/Merge (Classic).

this approach, the DL backbone extracts features from the data of different modalities, while features from each level are then fused at the network level by concatenation or merging. Additionally, further feature fusion is performed following the DL backbone. This allows for more complex feature combinations to be learned, improving classification accuracy. The process diagram for output fusion is shown in Fig. 2.15.

(4) The emergence of Transformers has led to the development of **Attention-based Fusion** as a new network architecture. Through its unique DL backbone, this architecture is able to extract features and implement feature fusion based on the attention relationship between different modalities. Fig. 2.16 illustrates the process diagram of attention-based fusion. A more detailed analysis of the network architecture will be presented in Sect. 2.2.4.

(5) **Output Fusion**, also known as decision-level fusion or late fusion, involves the use of DL backbones to extract high-dimensional features from different modalities of data. The extracted features are then used to generate separate classification results for each modality. These results are then

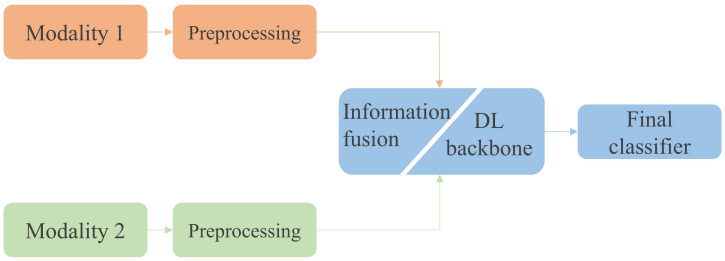


Figure 2.16 – Attention-based fusion process diagram.

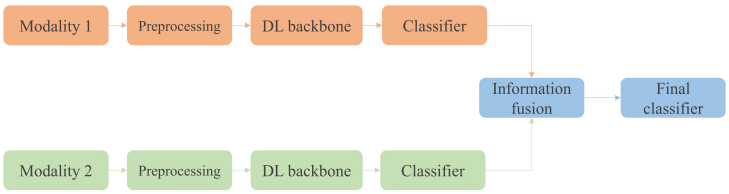


Figure 2.17 – Output fusion process diagram. Information fusion: Merge (Outputs).

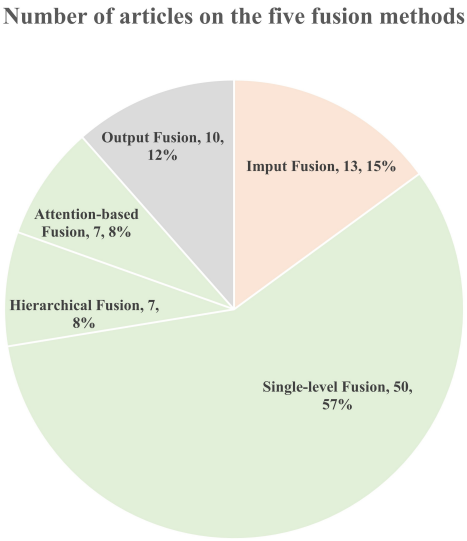


Figure 2.18 – The percentage of different fusion methods used in our study. Intermediate fusion accounts for 73% of all methods. Tags: multimodal classification networks, number of publications, percentage.

combined using a fusion technique, such as majority voting or averaging, to produce a final classification result. The process diagram for output fusion is depicted in Fig. 2.17.

Recent years have seen a growing trend toward the use of deep learning

networks in multimodal fusion research. Fig. 2.18 illustrates the distribution of five fusion methods in the scope of the study. In contrast to traditional methods, single-level fusion is the most commonly used method in DL multimodal fusion, followed by input and output fusion. Hierarchical fusion and attention-based fusion are also gaining attention and present great potential for research. These more recent fusion methods offer more complex ways of combining modalities, enabling deep learning networks to learn more powerful representations of multimodal data.

Input fusion networks

Input fusion combines data from multiple modalities into a single feature tensor fed into the deep neural network as an input. Input fusion typically involves the fusion of modalities with similar structures, making implementation relatively straightforward. Some modalities can be acquired together at the time of clinical photography (e.g., CT and PET). In many cases, these modalities have the same voxels and spacing after data processing, making obtaining registered multimodal data easy. Furthermore, the majority of input fusion tasks do not require re-modeling, only modifying the input part of the unimodal model to achieve multimodality. Fusion can be accomplished in three ways: concatenating or merging multimodal medical images, extracting high-dimensional features from multimodal images, and then fusing them.

(1) The registered multimodal data are fed into the DL classifier as input for different channels to obtain classification results, which is the most common input fusion approach. Fig. 2.19 illustrates this typical input fusion network architecture used in the research of [246–249]. [246] proposed a semi-automatic method for the classification of prostate cancer without feature selection. Several combinations of 3D volumes (e.g., ADC,

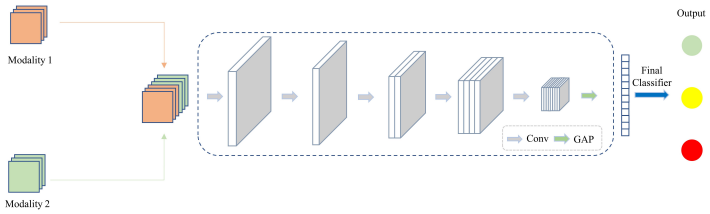


Figure 2.19 – Schematic diagram of the network architecture for input fusion. Information fusion method: Concatenation (Inputs).

DWI, and T2) are utilized as inputs of the CNN network. Each sequence is considered an input channel; the output is the classification of significant versus nonsignificant lesions. [247] employed MRI and PET to diagnose Alzheimer's disease. PET and MRI are used as two channels for the input of the CNN classification network, based on an ROI crop model to learn a classifier and fuse different features from MRI and PET. [248] concatenated T2, ADC and DWI for tumor foci classification using an end-to-end CNN network. In order to diagnose triple-negative breast cancer, [249] concatenated manually segmented multiparametric MRI images (PEI, DWI) into a CNN network. Despite the ease of implementing this fusion architecture, it has some limitations with regard to the modal data requirements. For instance, the registration performance of different modal data can influence the classification results. Moreover, this approach is not suitable for fusing heterogeneous data, such as 3D medical images and 1D clinical records, which have different characteristics and dimensions.

(2) The merging of images is another input fusion method in addition to concatenation. Various image modalities are fused at the pixel or voxel level in order to create a new fused image that is used for classification [140, 250]. [250] proposed an effective multimodal image fusion method for Alzheimer's disease diagnosis using MRI and PET. Through registration and mask coding, they were able to fuse gray matter (GM) and 18-

fluorodeoxyglucose positron emission tomography (FDG-PET) images to create a new imaging modality called "GM-PET". In the resultant composite image, the GM area is clearly highlighted, allowing AD diagnosis to be made while maintaining both the contour and metabolic characteristics of the subject's brain tissue. They then fed the fused images to the CNN for classification. The GM region cropped from the MRI image is mapped onto the PET image, resulting in the fusion of PET and MRI data in [140] research. In addition to providing anatomical and metabolic information about the brain, the fusion modality also allows the viewer to focus on the main features of the brain by reducing the visual noise. The benefit of fused images is that they contain a wealth of medical information, but the process of generating them often requires an extensive amount of prior medical knowledge.

(3) Some studies have performed input fusion after extracting features from multimodal images instead of performing a direct fusion of medical images [186, 227]. [227] used PCA to extract features from MRI, PET, and cerebrospinal fluid (CSF) and then concatenated these features into the Restricted Boltzmann Machine (RBM) network for the diagnosis of Alzheimer's disease. [186] manually extracted features from MRI and PET and then used stacked auto-encoder (SAE) to classify the concatenated multimodal features in order to diagnose Alzheimer's disease. The architecture of extracting features and combining them can solve the problem of multimodal heterogeneity. However, PCA-based or manual feature extraction requires prior knowledge and does not fully utilize image information.

In input fusion, fused data is used in single-branch feature extraction, and the network architecture design reduces network parameters and deployment difficulties significantly. Due to the fusion of the data at the input level, the complementary information from the different modalities is not

utilized to the fullest extent possible.

Single-level fusion networks

The single-level fusion process uses different DL backbones to extract features from different modalities separately, followed by an information fusion process before making the final decision. Based on the position of information fusion within the network architecture, it can be divided into classic fusion structures and network fusion structures.

(1) The most common single-level fusion architecture is to extract features from multimodal data by using different branches, then fuse these features and feed them to the final classifier [147, 185, 188, 215, 224, 234, 235, 243, 251–255]. A schematic diagram of its network architecture is shown in Fig. 2.20. After preprocessing the data, the architecture [224] extracted low-level 3D features from fMRI and sMRI to classify Attention Deficit Hyperactivity Disorder (ADHD) automatically. As soon as the features are concatenated, softmax classifiers are used to differentiate ADHD cases from typically developing children (TDC) cases. In order to diagnose breast cancer, [254] fused MRI (T1, T2) and clinical information. Two 3D ResNet-50 were used to extract features from contrast-enhanced T1 subtraction MR images and T2 MR images, while the FC layer provided clinical inputs. For the prediction of pathological complete response, the outputs of each 3D ResNet-50 and FC layer were concatenated, and the final FC layer with sigmoid activation function was used. Likewise, [215] employed ResNet and FC layers to extract features from DSC, Clinical Image, and metadata, then applied FC layers for skin lesion classification. Aside from these methods of concatenating modal features, complex computations can also be used to merge features. [252] used visual field (VF) and OCT for the diagnosis of glaucoma. VFNet and OCTNet were used

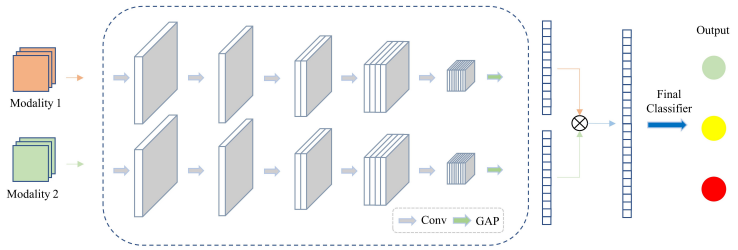


Figure 2.20 – Schematic diagram of the network architecture for classic single-level fusion. Information fusion method: Concatenation (Classic).

to extract features from the VF and OCT modes, respectively. A weighted average was used to obtain an aggregated representation from bimodal features using an attention module. Each modal feature was assigned a weight using a fully connected layer, followed by a sigmoid function to calculate a scalar value (0-1) indicating that feature's relative contribution to the aggregate representation. To aggregate all features, a global average pooling layer was also used. The results of glaucoma diagnosis were predicted using three fully connected layers and a softmax layer. For CT and PET modalities, [188] extracted features using CNN networks, merged the features using gated multimodal units (GMU), and classified lung cancer using FC layers. GMU, unlike the widely used connection operation, allows for the learning of intermediate representations of multimodality features by using hidden structures and gate controls, thereby enabling the prediction layer to assign weights more effectively to intrinsically associated features.

(2) Two stages can be described as the single-level fusion architecture for network fusion. The first stage involves extracting single-level features separately from different modalities using DL backbones, followed by the second stage of information fusion, which involves utilizing an additional DL backbone to extract high-level features from the fused features [225, 226, 256–259]. Lastly, the extracted high-level features are used in the final classification process. Fig. 2.21 illustrates a typical network fusion architec-

ture. [257] used cascaded CNN for the multimodal fusion of MRI and PET to diagnose Alzheimer's disease. They proposed a 2D CNN to combine the multimodality features and make the final classification. After 3D CNN output features are flattened to one dimension, the 1D feature vectors of MRI and PET are combined to produce a two-dimensional feature map for 2D CNN analysis. [226] developed a multimodal architecture for combining MRI, PET, and CSF features. Each modality's individual representation of high-level features is calculated using the stacked sparse extreme learning machine auto-encoder (sELM-AE). Another stacked sELM-AE is used to get the joint features from the high-level MRI, PET, and CSF features. The kernel-based extreme learning machine classifies the joint feature representation. With multimodality neuroimaging and genetic data, [260] proposed a three-stage deep feature learning DNN framework for Alzheimer's disease classification. Each modality's latent representation is learned in the first stage, then each pair of modalities' joint latent representations are learned in the second stage, and in the third stage, each pair of modalities' joint latent representations are used to create the classification model. [258] classified schizophrenia using sMRI, fMRI, and single nucleotide polymorphisms. The latent representations for the static functional network connectivity (sFNC), sMRI, and single nucleotide polymorphism (SNP) are learned using an autoencoder, multi-layered perceptron, and bi-directional long short-term memory (LSTM). The Multimodal Bottleneck Attention Module performs the fusion of the embeddings and then sends the combined embeddings to a variational autoencoder for encoding, followed by a SoftMax layer for classification.

The single-level fusion method is currently used to merge multiple medical modalities for classification tasks and can be applied to the fusion of different medical modalities. The method does not require a specific for-

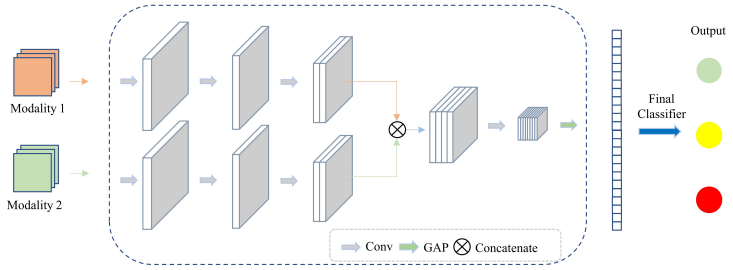


Figure 2.21 – Schematic diagram of the network architecture for single-level network fusion. Information fusion method: Merge (Network).

mat for the data as it extracts features from modalities using different branches and fuses data at a high-dimensional feature level. In this regard, single-level fusion is a suitable solution for unregistered or different dimensional data. Due to the fact that information fusion occurs only at the end of the network architecture, single-level fusion is not capable of analyzing low-dimensional features jointly.

Hierarchical fusion networks

Hierarchical fusion extends single-level fusion further in order to further exploit the complementary information between multimodal data. The hierarchical fusion process involves the fusion of different dimensional features and the classification of these jointly represented features through the process of fusion [229, 242, 261, 262]. There are two ways to implement hierarchical fusion: by using additional branches for multimodal feature fusion or by using fusion blocks for joining features from different modalities.

(1) The common hierarchical fusion architecture involves extracting different modalities via different branches and simultaneously combining multimodal features of different dimensions via another parallel branch. Finally, the high-dimensional features from the fusion branch and each modal branch are combined for classification. Fig. 2.22 shows a typical network

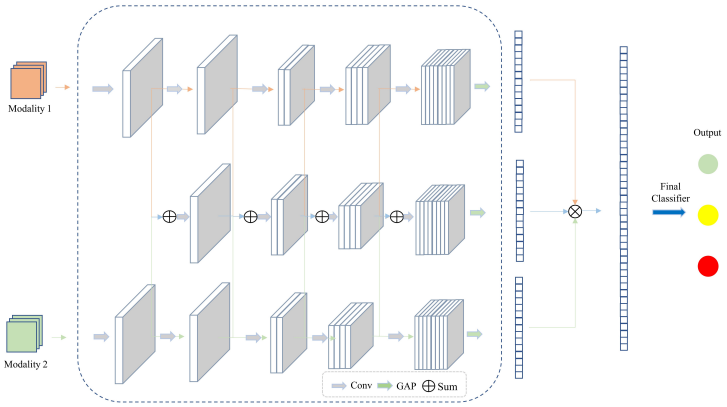


Figure 2.22 – Schematic diagram of the network architecture for hierarchical fusion. Information fusion method: Merge (Network) and Concatenation (Classic).

architecture for hierarchical fusion, [229, 262] utilized this network architecture. [229] utilized three sparse-response Deep Belief Network (DBN) branches to extract features from PET/MRI modalities, fuse them, and then employed an Extreme Learning Machine (ELM) to classify the fused features for brain diseases. [262] used a deep multi-modal fusion network (DMFNet) to fuse PET and MRI data for the diagnosis of Alzheimer's disease. Three branches are present in DMFNet, two of which extract features from the MRI and PET scans, respectively. A channel attention model is used to extract the features from each branch and merge the reweighted feature maps. In the third branch, fused features are further extracted.

(2) Hierarchical fusion can also be structured in another way by extracting features using different branches and fusing them in different dimensions by using fusion blocks, which are then returned to each modality branch for further fusion. The design of such a network structure can reduce the number of model parameters while fusing features at multiple levels. Fig. 2.23 illustrates a typical network architecture, [242, 263] utilized this network architecture. To classify brain diseases, [263] proposed a pathwise transfer deep convolution network that gradually learned and

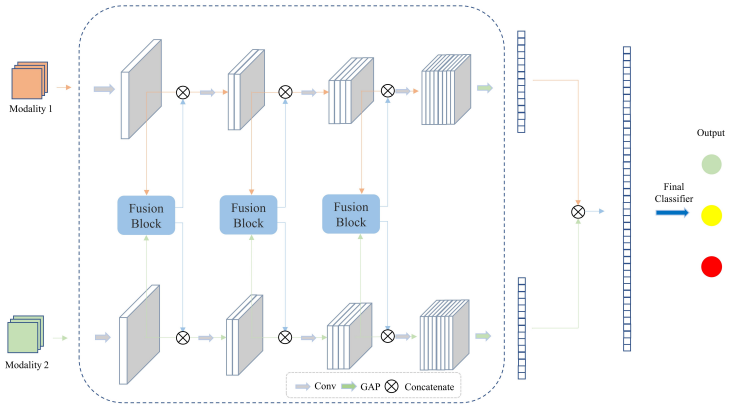


Figure 2.23 – Schematic diagram of another network architecture for hierarchical fusion. Information fusion method: Merge (Network) and Concatenation (Classic).

combined the multi-level and multimodal features of MRI and PET. The pathwise transfer blocks are designed to fully utilize complementary information from different imaging modalities. Pathwise transfer blocks are used to communicate information across PET and MRI, which helps to improve the classification model's performance. [242] proposed a multimodal MRI hierarchical-order multimodal interaction fusion network (HOMIF) to diagnose gliomas. There are two branch networks for each modality, several multimodal interaction modules with different scales and orderings, diverse learning constraints, and a predictive subnet in the framework. Each branch network has three CNN blocks with multiscale inputs and an arm with diverse high-order multimodal interaction (HOMI) modules to integrate and interact deeply with the multiscale features.

The multi-level feature fusion allows hierarchical fusion to explore more fully the complex and complementary information between modalities. Learning the synergy of multimodal data while maintaining the features of the modalities improves the model's classification performance [262]. However, as it involves the fusion of low-dimensional features, the registration of multimodal data may affect the classification performance of hierarchical

fusion.

Attention-based fusion networks

As attentional mechanisms [121] have been proposed and developed, more and more studies are beginning to incorporate attentional mechanisms into network architectures. Some of the network architectures mentioned above also included attention mechanisms in order to enhance the performance of the models. [262] added attention modules to reweight the modal features. [237] use a vision transformer (ViT) to extract the modal features and fuse them. These studies, however, only operate on unimodal modalities and do not utilize the attention mechanism for multimodal interactions. Recently, some studies have used the attention mechanism to extract and combine features [264–269]. This network architecture is called attention-based fusion, which is not related to any of the previous fusion architectures.

In the study of [264], they propose TransMed, which combines CNN and transformer to capture high-level cross-modalities and low-level features. First, TransMed sends the multimodal images to CNN, where they are processed as sequences, then transformers learn the relationships between them and predict the end result. TransMed is more efficient and accurate than existing multimodal fusion methods because it effectively models the global features of multimodal images.

Attention-based Hierarchical Multimodal Fusion (AHM-Fusion) is a novel fusion module [265] designed. The system includes both an early feature guidance module and a late feature fusion module, capturing deep interaction information between different multimodal features. In the early stage of feature aggregation, the early feature guidance module is used to capture multimodal interactions. To obtain classification results, late feature fusion

modules based on attention mechanisms are used. Through cascading double attention layers in the late feature fusion module, the deep interaction information is further captured. Then, they used a gating-based attention mechanism to decrease the impact of insignificant features in each modality.

[266] proposed a multimodal Medical Information Fusion (MMIF) framework that combines the Category Constrained-Parallel ViT framework (CCPViT) and the multimodal Representation Alignment Network (MRAN) as backbones, enabling the modeling of images and texts as unimodal features, as well as cross-modal features. CCPViT is proposed as a tool for learning key features of different modalities and for solving unaligned multimodal tasks. Then in MRAN, Cross-attention was used to cascade encoded images and decoded texts to explore deep-level interactive representations of cross-modal data, assisting with modal alignment and identifying abnormalities. MMIF is an image-text foundation modeling that could contribute to a much higher-precision classification model when compared with unimodal models.

Multimodal Mixing Transformer (3MT) was presented [267] as a novel technique to classify diseases. Based on neuroimaging data, gender, age, and the Mini-Mental State Examination (MMSE), They tested it for Alzheimer's Disease classification. Multimodal information is incorporated through a Cascaded Modality Transformers architecture with cross-attention. Different embedding layers are used to obtain Key (K) and Value (V) from imaging features and clinical data. K and V are then placed into a cross-attention layer with a latent code known as Query (Q). 3MT allows mixing unlimited modalities and formats and full data utilization.

Research is increasingly incorporating attention mechanisms, particularly Transformer structures, into multimodal classification tasks. While

performing cross-modal attention computation, a multi-level fusion of multimodal features is achieved. Furthermore, the Transformer structure is well-suited for joining modalities of different dimensions. Nevertheless, Transformer research in medical tasks is still in its infancy, and various studies tend to focus on solving particular problems, making it difficult to conclude a general multimodal classification architecture. A further important point to be noted is that while the success of Transformer is accompanied by pre-training on large datasets, the number of samples in medical datasets is often not sufficient to achieve the good training effect of Transformer. As a result, it is recommended that Transformer and CNN are used together in a hybrid fashion.

Output fusion networks

Fusion at the output level or fusion at the decision level can also be referred to as output fusion. In output fusion, for each modality that uses a separate DL backbone to extract features and make decisions, the results are merged into one final decision. Fig. 2.24 shows a typical network architecture for output fusion. The final Classifier of decision fusion can be achieved by simple operations [270, 271] such as voting, weighting, and averaging, or by classifiers [149, 187, 272, 273] such as SVM, extreme gradient boosting (XGBoost), adaptive boosting (AdaBoost), Decision Tree, and K-nearest neighbor (KNN). [270] used unweighted average, weighted average, weighted voting, and stacking to fusion the classification results from different modalities of US to identify breast tumors. [271] applied a linear weighted module to assemble the predicted probabilities of the pre-trained models based on the 4 MRI modalities for the classification of gliomas. In order to achieve the diagnosis of early glottic cancer, [273] used decision trees to combine the classification results from the sound data and the im-

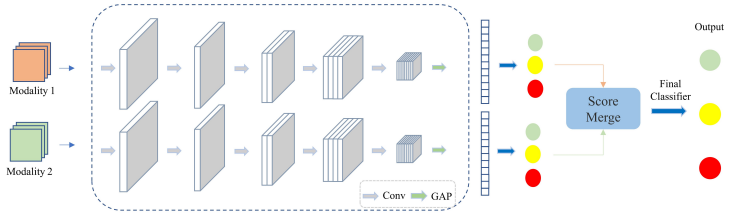


Figure 2.24 – Schematic diagram of the network architecture for output fusion. Information fusion method: Merge (Outputs).

age data. [272] used SVM, KNN, and linear discriminant analysis (LDA) to fuse the classification results of fMRI and sMRI to diagnose ADHD.

The output fusion process involves combining unimodal results from different modalities. As a result, it is relatively easy to implement and generally does not require additional training. It is, however, difficult to exploit the complementary information between different modalities because there is no feature fusion. Furthermore, output fusion may not improve classification performance if there are large differences in decisions between different modalities.

2.2.5 Discussion

Which fusion method is the best?

The choice of the fusion method is crucial when dealing with multimodal medical classification problems. We have quantitatively compared various fusion methods using the ADNI dataset to compare the performance of different fusion architectures. MRI and PET were used to diagnose Alzheimer's disease in all of these studies, and the number of subjects used in each study was relatively similar. There are three stages in the progression of Alzheimer's disease: normal cognition (NC), mild cognitive impairment (MCI), and Alzheimer's disease (AD). In spite of the fact that MCI does not significantly interfere with daily activities, a high risk of AD

progression has been consistently demonstrated in patients with MCI [274]. MCI subjects can be classified into MCI converters (cMCI) and MCI non-converters (ncMCI) to predict the transition risk of MCI. Tab. 2.5 shows the results of different fusion methods. When comparing these results, it should be noted that each paper relies on a different subset of patients.

Table 2.5 – Comparison of the results of different fusion methods on ADNI dataset. In the multi-classification task, 3 class is NC vs. MCI vs. AD, and 4 class is NC vs. ncMCI vs. cMCI vs. AD. Unit:%.

Research	Year	Fusion Methods	Dataset	NC vs. AD	NC vs MCI	Multi-classification
[186]	2015	Input Fusion	331 subjects: 77 NC, 102 ncMCI, 67 cMCI, 85 AD	ACC: 91.40 SEN: 92.32 SPE: 90.42	ACC: 82.10 SEN: 60.00 SPE: 92.32	4 Class ACC: 53.79 SEN: 52.14 SPE: 86.98
[250]	2021	Input Fusion	381 subjects: 126 NC, 160 MCI, 95 AD	ACC: 94.11 SEN: 93.33 SPE: 94.27	ACC: 85.00 SEN: 84.69 SPE: 85.60	3 Class ACC: 71.52 SEN: 55.67 SPE: 83.40
[140]	2022	Input Fusion	370 subjects: 130 NC, 129 MCI, 111 AD	ACC: 93.21 SEN: 91.43 SPE: 95.42	ACC: 86.52 SEN: 94.34 SPE: 81.64	3 Class ACC: 87.67
[243]	2014	Single-level Fusion	398 subjects: 101 NC, 128 ncMCI, 76 cMCI, 93 AD	ACC: 95.35 SEN: 94.65 SPE: 95.22	ACC: 85.67 SEN: 95.37 SPE: 65.87	-
[225]	2017	Single-level Fusion	202 subjects: 52 NC, 56 ncMCI, 43 cMCI, 51 AD	ACC: 97.13 SEN: 95.93 SPE: 98.53	ACC: 87.24 SEN: 97.91 SPE: 67.04	4 Class ACC: 57.00 SEN: 53.65 SPE: 85.05
[275]	2019	Single-level Fusion	392 subjects: 101 NC, 200 MCI, 91 AD	ACC: 98.47 SEN: 96.58 SPE: 95.39	ACC: 85.74 SEN: 90.11 SPE: 91.82	-
[276]	2022	Single-level Fusion	959 subjects: 264 NC, 273 ncMCI, 204 cMCI, 218 AD	ACC: 98.24 SEN: 98.82 SPE: 97.52	ACC: 94.59 SEN: 90.26 SPE: 96.98	-
[262]	2020	Hierarchical Fusion	500 subjects: 163 NC, 113 ncMCI, 105cMCI, 119 AD	ACC: 95.21 SEN: 93.56 SPE: 97.48	-	4 Class ACC: 86.15
[187]	2020	Output Fusion	398 subjects: 101 NC, 204 MCI, 93 AD	ACC: 99.27 SEN: 95.89 SPE: 98.72	ACC: 90.35 SEN: 88.36 SPE: 92.56	-

In general, we believe that deep multi-level fusion can better exploit

the synergy of multimodal data to produce better classification results. This is further supported by the results in Tab. 2.5. Compared with the input fusion, the single-level fusion has a more robust feature fusion, which improves the overall ACC of the middle fusion. Hierarchical fusion utilizing multi-level feature fusion did not improve significantly the performance of dichotomous classification but performed well for four class classifications. Generally, a complex model does not improve performance much when applied to a simple classification task. The more complex the network, the better it is at solving complex classification problems. When the number of categories for multi-category classification increases from two to four, the hierarchical fusion classification accuracy improves significantly. Last but not least, we note that the output fusion achieves excellent results on NC versus AD classification, thanks to the pre-training of different modal branches. With output fusion, DL backbones can be pre-trained on a large number of unimodal datasets and then fine-tuned on the multimodal datasets. Other datasets provide similar results. ABIDE data was combined with sMRI and fMRI to diagnose autism spectrum disorders. It was found that the hierarchical fusion [199] result was 87.2%, which was better than the input fusion [198] result of 65.5%. [258] used the COBRE dataset for the diagnosis of schizophrenia, and the accuracy of input fusion, output fusion, and single-level fusion was 70%, 78%, and 95%, respectively, when the same dataset was used.

It is difficult to determine a unified solution for a wide variety of multimodal fusion medical image fusion tasks. In spite of this, we can draw some preliminary conclusions from the above analysis. Modal registration is easier for medical modalities with similar structures, so input fusion, single-level fusion, and hierarchical fusion are all network structures worth investigating. Generally, single-level fusion and hierarchical fusion fuse deeper fea-

tures, which will improve the classification performance. When data have a wide range of structures or dimensions, single-level fusion and attention-based fusion are preferable solutions, as they are capable of handling a wide range of modal feature fusion scenarios. Lastly, if we have a large number of unimodal datasets for each modality in multimodal data, output fusion will perform well.

In addition to using a single multimodal fusion method, multiple fusion methods can be combined [189, 244, 245]. [189] achieved the classification of skin lesions using a combination of single-level fusion and output fusion. In order to improve the diagnosis of breast cancer, [245] fused multi-parametric MRI data at three levels: input, feature (intermediate), and decision (output). Combining different fusion methods can cumulate their advantages, allowing data from various perspectives to be fused and improving classification performance to some extent. It is one of the promising strategies that can be used when performing multimodal medical classification.

2.2.6 Conclusion

In this section, we conducted a comprehensive review of the development of deep learning-based multimodal medical classification tasks over the past few years. We examined the complementary relationships among several common clinical modalities and delved into five key architectures for deep learning multimodal classification networks: input fusion, single-level fusion, hierarchical fusion, attention-based fusion, and output fusion. Our study covered a wide range of multimodal fusion scenarios in medical classification, as well as the application domains for which different network architectures are most suitable.

2.3 Methodology of Automated DR diagnosis

Having reviewed and analyzed different fusion methods for multimodal medical classification tasks, our focus in this section was to review methods for DR diagnosis. We first examined the unimodal method used for different modalities and then briefly discussed the few multimodal fusion methods available. We hope that these methods related to DR diagnosis can provide references and insights for our thesis.

2.3.1 Unimodal diagnosis

Our research will start with deep learning-based unimodal ophthalmic imaging in order to provide ideas and solutions for future multimodal fusion research.

Color Fundus Photographs

CFP-based deep learning DR detection systems are relatively as the gold standard for early detection and diagnosis of DR. In addition, because of the ease and low cost of deployment of its shooting devices, many public datasets have emerged, e.g., EyePACS [84], APTOS [277], Messidor [278], DDR [279], E-Ophtha [280], etc. The emergence of these publicly available datasets has further advanced the development of deep learning techniques for CFP-based diabetic retinopathy diagnosis. In 2020, there has been a systematic review of deep learning algorithms for CFP-based detection of diabetic retinopathy [33]. The algorithms are essentially all CNN algorithms, with some combining CNNs with machine learning algorithms (CNN with random forest, decision tree, and support vector machine) [281–283]. The range of sensitivity and specificity of included studies was 30% to 100% and 70.7% to 98.5%, respectively. The area under the

receiver operating characteristic (AUROC) ranged from 95% to 99.3%. In addition, the accuracy of the DL model in detecting/classifying DR ranged from 75% to 97.28% [33].

Ultra-Wide-Field imaging

The development of ultra-wide-field imaging has revolutionized the assessment of diabetic retinopathy [284]. There have been several studies that have evaluated the utility of ultra-wide-field imaging in the diagnosis of diabetic retinopathy. According to [285], diabetic retinopathy severity grading was compared between Optomap ultra-wide-field images and ETDRS seven-standard field views. In spite of the fact that severity grades were identical in 85% of the images and within one severity level in 100% of the images, 19% of the images were assigned a higher retinopathy level in the ultra-wide-field view than in the ETDRS seven-field view. A study by [286] showed that Optomap ultra-wide-field images detected approximately 30% more peripheral neovascularization than standard two-field imaging in diabetic patients. Despite this, a limited number of studies have been conducted using only UWF images for deep learning-based diabetic retinopathy because the devices for capturing UWF are not widely available. With the Swin Transformer model, [287] proposed a hybrid preprocessing method for UWF that resulted in an average of classification accuracy(ACA) 0.72, Macro F1 0.7018, and Kappa 0.65 for the diagnostic task of diabetic retinopathy. With the help of the VGG-19 model, [288] was able to achieve accuracy, sensitivity, specificity, and Cohen's kappa score of 80% , 95%, 80% and 0.75, respectively, for DR diagnosis.

Optical Coherence Tomography

Images obtained with retinal optical coherence tomography provide valuable information regarding the health of the posterior eye, such as the retina and choroid, which can be used as a tool to identify the type and severity of diabetic retinopathy [289, 290]. The SD-OCT detects the disorganization of the inner retinal layers, which are also biomarkers of centrally involving diabetic macular edema (DME) with reduced vision [291]. Furthermore, the choroid of diabetic eyes with retinopathy was thinner than that of diabetic eyes without retinopathy [292]. The subtle changes in choroidal thickness on SS-OCT help distinguish early from more advanced stages of diabetes [293]. It has been reported that deep learning approaches have been applied for the analysis of OCT images to diagnose diabetic retinopathy [294]. A research article published in 2022 [295] presented a three-step system for diagnosing diabetic retinopathy (DR) utilizing optical coherence tomography (OCT) images as an example of diagnosing diabetic retinopathy (DR). As part of the process, retinal layers are segmented, 3D features are extracted, and backpropagation neural networks are used to classify the data. There is a 96.81% accuracy rate for the proposed system. Another example comes from [296], where an optical coherence tomography (OCT)-based deep learning CAD method is proposed to detect DR early through the use of structural 3D retinal scans. Three phases are involved in the development of the CAD system. In the first phase, the 3D-OCT was segmented into 12 layers. During the second phase, distinguishable features of higher-order reflectivity are extracted. The third stage involves classifying each layer based on the extracted features and applying a majority vote to the classification layer's output to obtain the global diagnosis. An accuracy of 96.88% was achieved by the proposed deep learning CAD

system. It should be noted, however, that there are few end-to-end OCT diagnostic methods that utilize deep learning.

Optical Coherence Tomography Angiography

I remind OCT-angiography is a motion-sensitive extension of OCT enabled by fast OCT acquisitions [297]. It was shown to provide quantitative information regarding blood flow within the retina and contrast with the retinal vasculature [298]. There is a clear benefit to OCTA for DR [299]: 1) Structure volume allows objective and quantitative assessment of diabetic macular edema, 2) flow Maximal Intensity Projections (MIP) allows quantification of retinal vascular plexus, non-perfusion and vessel density as well as the identification of damage; [300] lists the various biomarkers of DR and DME in OCTA acquisitions. CAD of DR using OCTA is an emerging area of research, motivated in part by the promise of useful biomarkers, as well as the challenge of integrating large amounts of data (i.e., 3-D ultra-widefield structural and flow images). To assist in the early detection, staging, and progression of DR, various quantitative metrics were automated [301]. Those metrics quantify retinal fluid volumes [302], retinal vasculature features (e.g., density, tortuosity) [303–305], avascular zones [302, 306], including the Foveal Avascular Zone (FAZ) [303, 307], and proliferative DR features such as neovascularization [308]. Based on a radiomics approach [305, 309, 310], these features were used to determine the severity of DR automatically. A number of methods have also been investigated to assess DR severity directly from OCTA images. With 2-D CNN, some authors classified 2-D en-face MIP images: [304, 310–313] classified one en-face flow MIP image (superficial plexus, deep plexus, or entire retina), [314] jointly classified two en-face flow MIPs (superficial and deep plexus) and their corresponding en-face structure MIPs, as well as two-

dimensional feature maps derived from feature segmentation and en-face flow MIPs jointly classified by [315] and [307]. Other authors have classified 3-D images using 3-D CNNs: [316] has classified a 3D image that includes two channels (structure and flow).

2.3.2 Multimodal diagnosis

It is true that the unimodal diagnosis method gives us ideas on how to handle different modal data, but for my thesis, we are more concerned with multimodal diagnosis, specifically the fusion method of multimodal information. There are already multimodal diagnostic studies, which often focus on the combined diagnosis of CFP images and OCT volumes [17, 317, 318]. [317] evaluated the feasibility and clinical utility of a deep learning-based dual-modality screening algorithm for DR and macular edema by combining fundus photos and OCT images in a community hospital. ResNet 101 [319] was used to classify DR stages on fundus images, and Faster R-CNN [320] was used to detect retinal abnormalities on OCT scans. The Glaucoma Grading from Multi-Modality Ages (GAMMA) Challenge was established by [17] to promote the development of fundus & OCT-based glaucoma grading. Across all 10 of the top solutions, a two-branch CNN model was used to extract and fuse features from the different modalities. [318] proposed a multimodal algorithm for the detection of glaucoma based on fundus photographs assessed with OCT. This multimodal model was combined with two image classifiers, a regression model, axial length, visual acuity, and the demographic numerical data of the participants. Finally, the classification results from different data are combined and analyzed by an integrated model. In spite of this, little research has been conducted on the multimodal fusion of emerging fundus photography techniques.

The acquisition of OCTA data can be divided into two volumes: the *structure volume* obtained by averaging successive 3-D scans, and the *flow volume*, which is determined by the amplitude of the local intensity variations across consecutive 3-D scans [321]. Structure and flow volume have been combined to provide depth-resolved, three-dimensional, micrometer-scale retinal images [38, 322–324]. Several studies have demonstrated the ability of combined structure volume and flow volume imaging to diagnose and detect DR pathology using quantitative measurements [325–327]. It is important to note that despite these advantages for the diagnosis of DR, a diagnostic platform based on combined structure volume and flow volume imaging will still require innovation before it can be translated to clinical practice [328]. The combined structure volume and flow volume data sets are large, and manual examination of these datasets can take a considerable amount of time. There may also be a lack of clinical infrastructure to meet these data analysis demands in underserved areas [329]. It is necessary to implement an automated CAD system in order to resolve these issues. The first objective of our research **Joint analysis of multi-modal information in OCTA** can be supported by a number of studies [328, 330, 331]. [328] proposed an automated diagnostic 3D CNN framework based on structure volume and flow volume that can be used to diagnose DR, AMD, and glaucoma. As part of the framework, a semi-sequential classifier is used, which consists of two parts with identical architectures, one of which diagnoses DR and AMD, and the other which diagnoses glaucoma. [330] proposed a framework for automating the classification of DR based on structure volume and flow volume data as inputs. First, inputs are resized to 160×224×224×2 pixels (two channels: structure and flow). This data is fed into a DR screening framework using a 3D CNN architecture. Two outputs are generated by the network: a non-referable DR

(nrDR) or referable DR (rDR) classification and a non-vision-threatening (nvtDR) or vision-threatening (vtDR) classification. Based on the results of the rDR and vtDR classifications, the multiclass DR classification is defined. [331] used ensemble learning techniques in conjunction with CNNs in order to classify referable DRs based on the structure volume and flow volume. VGG19-trained [230] networks performed better than those trained on deeper architectures. As a result of constructing ensemble networks based on four fine-tuned VGG19 architectures, accuracies were 92% and 90% for majority soft voting and stacking, respectively.

Unfortunately, there are no published DR diagnostic CAD systems based on deep learning for our second goal **Joint analysis of different specifications of OCTA acquisitions** and third goal **Joint analysis of OCTA and UWF-CFP** at this time. These multimodal fusions, however, have been demonstrated to be clinically valid in some clinical articles [37, 332–334]. In view of this, these directions are worth exploring and are the focus of research and innovation, as well as the focus of the thesis.

2.4 Conclusion

This summary of multimodal approaches to medical diagnostic tasks provides us with an overview of multimodal fusion techniques. Further, there is no doubt that we apply these methods to multimodal diagnosis in DR. Our first step will be to use input fusion, single-level fusion, and output fusion since these three methods have proven effective in a number of medical tasks and are relatively straightforward to implement, so we will utilize their results as a starting point. The emerging hierarchical fusion has shown superiority in complex classification, which is the focus of our study. In addition, although there is no stable and effective model

for attention-based fusion on multimodal diagnostic tasks, it is undergoing rapid development, and its performance on DR diagnosis should be examined. Lastly, the fusion of 2D and 3D data is a challenging task for our study. It will be necessary to consider how different fusion methods can be applied to different dimensions of data.

Chapter 3

Materials

"It is a capital mistake to theorize before one has data."

— *Sherlock Holmes*

This chapter provides a description of the materials collected and used for this thesis. The different protocols used to obtain the database and the acquisition methods are discussed.

3.1 Introduction

The thesis research work utilized three different types of datasets: the EviRed retrospective dataset, the EviRed prospective dataset, and the Sup-

plemental dataset. The EviRed prospective dataset follows the protocol of the EviRed study: it is the most comprehensive in terms of imaging modality available, in terms of follow-up, etc. The EviRed retrospective dataset is a smaller and less comprehensive collection of images intended to support the initial algorithm developments until enough patients are recruited in the EviRed study. Because this retrospective dataset was not large enough to demonstrate the generality of the initial algorithms, we also considered a Supplemental dataset targeting a different pathology and slightly different imaging modalities. It is important to note that the EviRed study began in December 2020, and the retrospective dataset was received in January 2021. Due to difficulties in collecting patient data and delays in ophthalmologists' annotation, we received the prospective EviRed dataset with annotations in November 2022, which resulted in some delays in developing and testing our fusion algorithm. The chronology of arrivals for the different stages of the EviRed dataset is shown in Fig.3.1.

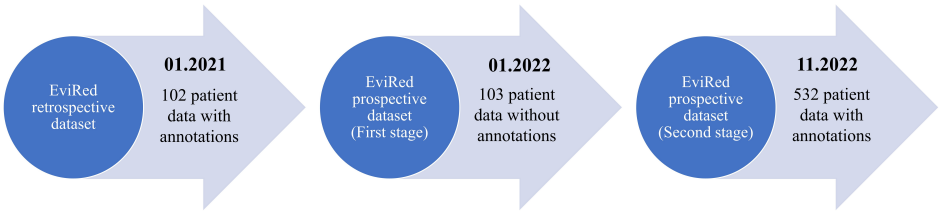


Figure 3.1 – Chronology of arrivals for different stages of the EviRed dataset.

3.2 EviRed retrospective dataset

The EviRed retrospective dataset consists of 102 patient data from two hospitals (Avicenne Hospital and Lariboisière Hospital) that are members of the Assistance publique – Hôpitaux de Paris (APHP). The Plex®Elite 9000 (Carl Zeiss Meditec Inc. Dublin, California, USA) is used to simul-

taneously acquire 3D structural OCT, 3D OCT angiography, and 2D line scanning ophthalmoscope (LSO) data for patients as Fig. 3.2. Scanning protocols included 3 × 3 mm^2, 6 × 6 mm^2, and 15 × 9 mm^2. The examination was conducted with patients' informed consent. The Declaration of Helsinki was followed during all procedures. The data is stored and transmitted via Nextcloud.

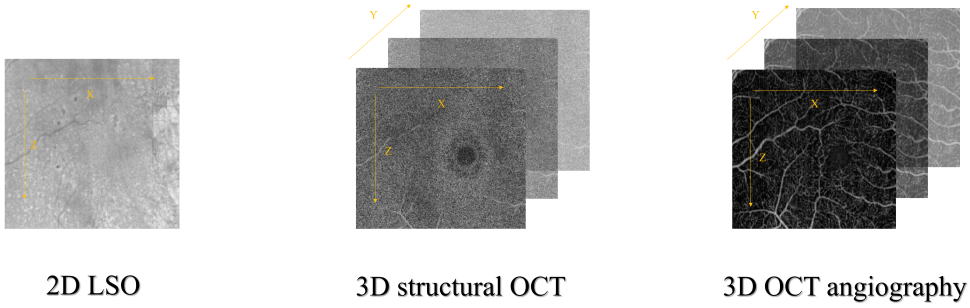


Figure 3.2 – Data from three imaging modalities in the EviRed retrospective dataset.

Annotations include pathological information and metadata relating to the patient's right and left eyes, including DR pathology, macular edema, date of birth, etc. According to the International Clinical Diabetic Retinopathy Disease Severity Scale (ICDR) scale, the DR severity level was graded by a retina specialist using fundus photographs: absence of diabetic retinopathy, mild nonproliferative diabetic retinopathy (NPDR), mild to moderate NPDR, moderate NPDR, moderate to severe NPDR, severe NPDR, proliferative diabetic retinopathy (PDR) and panretinal photocoagulation (PRP). The severity of 'mild to moderate NPDR' and 'moderate to severe NPDR' are non-standard severity labels: they underline the challenge of reliably assigning severity labels, even for an expert.

The EviRed raw data size is 300 × 1536 × 300 × 2 voxels for the 3 × 3 mm^2 SS-OCTA, 500 × 1536 × 500 × 2 voxels for the 6 × 6 mm^2 SS-OCTA and 834 × 1536 × 500 × 2 voxels for the 15 × 9 mm^2 SS-OCTA as Fig. 3.3.

The last channel presents the information of 3D structural OCT (Structure) and 3D OCT angiography (Flow), respectively. Furthermore, an additional LSO image of 512×664 was captured in addition to the LSO corresponding to the size of the en-face slice.

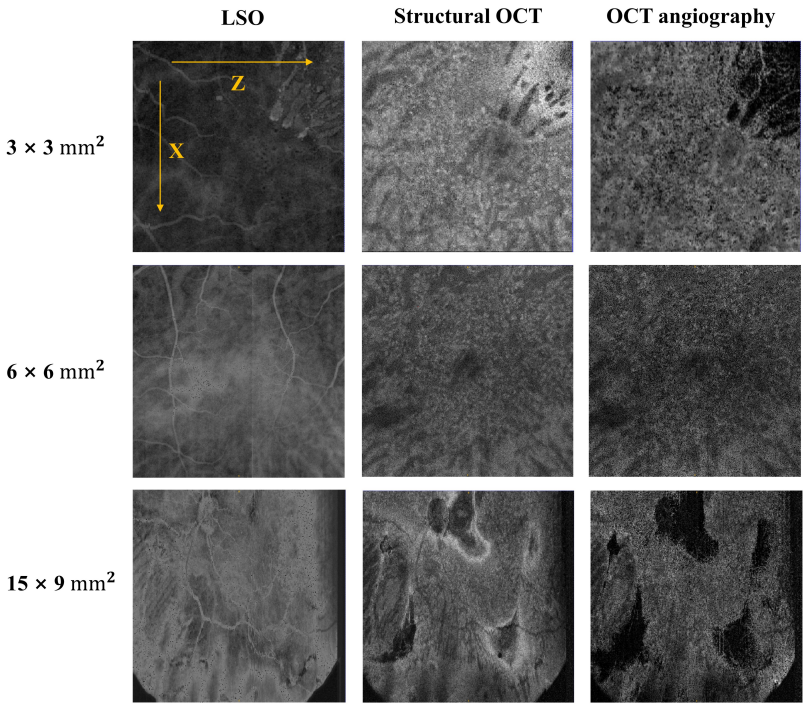


Figure 3.3 – 3D structural OCT and 3D OCT angiography en-face slices and LSO images from 3×3 mm^2 SS-OCTA, 6×6 mm^2 SS-OCTA and 15×9 mm^2 SS-OCTA.

3.3 EviRed prospective dataset

3.3.1 Introduction

For the EviRed prospective dataset, a cohort of up to 5000 diabetic patients is being recruited and followed for an average of 2 years. This cohort of patients is stratified to include 10% of patients with no DR or mild non-proliferative DR in both eyes, 10% of patients with complications in

both eyes, complications meaning proliferative DR (untreated or treated with panretinal photocoagulation PRP) or macular edema involving the center of the macula (untreated or previously treated with intravitreal injections), and 80% of patients with uncomplicated moderate to severe non-proliferative DR in at least one eye. Each year, general data as well as ophthalmological data and resource utilization data are collected. Retinal images and videos of both eyes are acquired using different imaging modalities, including widefield photography, OCT, and OCT angiography. All images and data are collected thanks to a common platform and centralized on a server. The EviRed cohort will be randomly split into two groups: one group of up to 4000 patients for building algorithms (training cohort) and one group of 1000 patients to evaluate them (evaluation cohort).

3.3.2 Data collection

Photography devices

The patients are followed according to usual clinical care, except that they have retinal imaging with two devices of different brands (Zeiss and/or another brand) instead of one for color fundus and OCT/OCT angiography. Zeiss is providing generously 6 * Swept Source OCT/OCTA Plex 9000 and 9 * CLARUS 5000 for the patient's analysis. Patients are followed with the same devices during the whole study.

— Ultra widefield fundus photography with CLARUS500 and/or other brands (Optos or Eidon).

— OCT and OCT Angiography exam with PLEX®Elite 9000 and/or OCT/OCTA device of another brand (Topcon, Spectralis, Optovue or Cirrus).

Required ethical and regulatory clearances

The recruitment of patients is being conducted in strict adherence to the Convention for the Protection of Human Rights and Fundamental Freedoms and EU regulations on ethical issues, including Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Directives 2001/20/EC, 2005/28/EC relating to implementation of good clinical practice in the conduct of clinical trials.

The participants of this RHU project also consider other international guidance, including:

— The Declaration of Helsinki by the World Medical Association.
— The International Ethical Guidelines for Biomedical Research Involving Human Subjects by the Council for International Organizations of Medical Sciences (CIOMS).
— The "Charter of Fundamental Rights" of the European Union (2000/C364/01).
— The Convention of the Council of Europe on "Human Rights and Biomedicine" (CETS164).
— The Declaration on Human Genetic Data adopted by UNESCO on October 16th 2003-The Universal Declaration on Bioethics and Human Rights adopted by UNESCO on October 19th 2005.

The AP-HP acts as the sponsor for all studies requiring promotion in the EviRed project.

Approval has been obtained from an Ethics Committee (CPP) as well as the authorization of the Regulatory Authority (ANSM). The study protocol was approved by the French South-West and Overseas Ethics Committee 4 on August 28, 2020 (Clinical Trial NCT04624737). The EviRed cohort is classified as category II of Loi Jardé since only interventions with

minimal risks and constraints were added to the current care by the study. According to the procedures of the sponsor (AP-HP), written consent by the patients is required to accept the acts added to the current practice by the research and to allow the use, for clinical research, of all her/his data collected during the study. As required by the law, a statement regarding the computerized filing of personal data collected for the research was submitted before the beginning of the research. The processing of the data collected in this research was authorized by the French Data Protection Authority (CNIL), which was a precondition for the beginning of the research.

The F-CRIN PARTNERS platform (PX) participated in drafting the protocol and information and consent forms, the according SOPs, and documents for the validation of the pharmacovigilance and vigilance of devices processes. DRCI or representative of the sponsor ensures approval by regulatory bodies and is in charge of notification and follow-up of the dossier of vigilance for CPP and ANSM. The EviRed consortium complies with Directive 95/46/EC on personal data protection (all patients are asked to specifically consent to use their data in addition to participating in follow-up).

Collection organization

The diabetic patients are being recruited in 14 Ophthalmology departments (recruitment centers) working with 18 Diabetology departments (non-recruitment centers). They will be seen yearly for 3 to 4 years. Each year, general data, as well as ophthalmological data, will be collected. A diagnosis will be made at the end of each visit, and usual care will be provided for each patient.

For the global management of the EviRed cohort, there are seven steps,

as represented in the following list:

1. Regulatory approvals and ethical follow up (completed)

   Before the clinical study is initiated, information or approval by the regulatory authorities, the ethics committees (EC), and any other competent authority (data protection, etc.) and registration in the clinical trial database (clinicaltrials.gov) is needed.

2. Development of the eCRF (completed)

   All information required by the protocol must be provided in the case report form and given by the investigator. Patient data will be collected and centralized anonymously on the electronic case report form (eCRF). The URC Lariboisière–St Louis platform (AP-HP) will design and develop an eCRF based on the paper CRF in collaboration with the study team. The anonymity of the subjects will be guaranteed by using a patient ID number on all documents necessary for research. Baseline, follow-up visits, and adverse events forms will be included in the eCRF. The study will be conducted using the CleanWeb® electronic data capture system, validated according to GCP guidelines. The principal investigators or sub-/co-investigators will fill in data in the electronic case report form (e-CRF) provided by the URC Lariboisière–St Louis (AP-HP). Data transmission to the web server will be performed by means of an Internet connection, and no specific software will need to be installed at the study site.

3. Centers selection and initiation (completed)

   There are (1) 14 Ophthalmology departments, all specialized in the management of retinal diseases, and (2) 18 Diabetology departments representing an active file of 55,000 diabetic patients. All the Ophthalmology departments have a consultation dedicated to the man-

agement of severe cases of DR, where ten to sixty diabetic patients are seen each week. In all Diabetology departments, an organization has been set up to screen for DR, using retinal photography graded by the ophthalmologist remotely. These organizations will thus allow the selection of diabetic patients.

4. EviRed Cohort recruitment

   Patients will be recruited in fourteen Ophthalmology departments, all specialized in the management of retinal diseases, working with 18 diabetology departments in which patients will be screened as well. There will be two sources of recruitment: the active cohort of diabetic patients followed in the Diabetology departments on one side and Ophthalmology consultations to which severe cases of diabetic retinopathy are directly referred on the other side. Fourteen Ophthalmology departments are listed:

   — Hôpital LARIBOISIERE
   — CHU DIJON
   — CHU BORDEAUX
   — CHU LYON CROIX ROUSSE
   — CHU AVICENNE
   — CHIC CRETEIL
   — CHU PITIE SALPETRIERE
   — CENTRE BROCA/MUTUELLE GENERALE
   — CHU BREST
   — Hôpital des 15-20
   — CHU de NANTES
   — FONDATION ROTHSCHILD
   — MARSEILLE CLINIQUE MONTECELLI

— CHU de NICE/Hôpital Pasteur 2

Patient recruitment progress for the 14 Ophthalmology departments at the end of March 2023 is as Fig. 3.4:

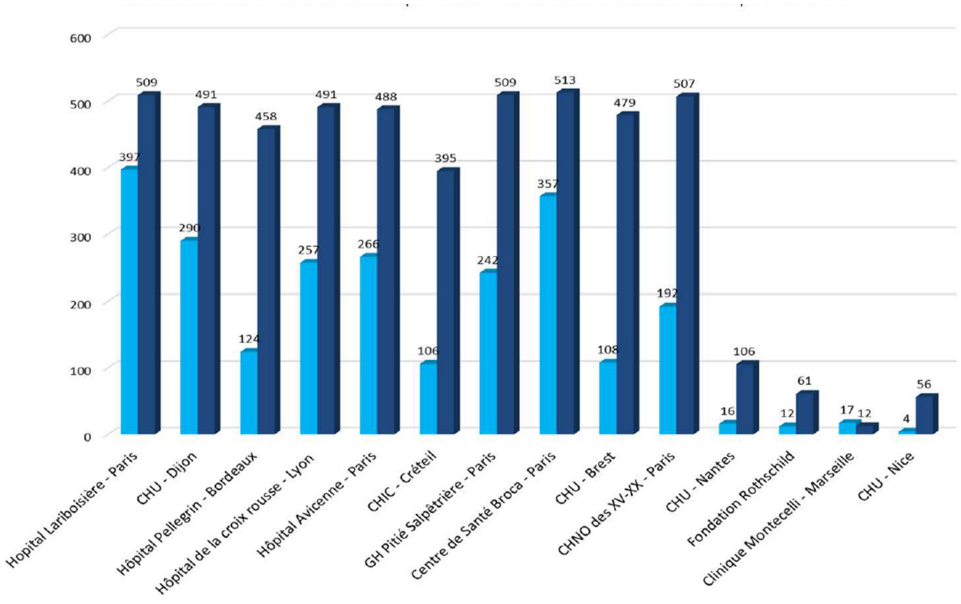


Figure 3.4 – Recruitment of patients (end of March 2023). Dark blue is the planned recruitment numbers, and light blue is the current recruitment numbers.

5. Management and monitoring of the clinical trials

   The Sponsor will coordinate the overall clinical studies management. A quality control system will be established under the responsibility of the Sponsor. Monitoring will be performed by the PARTNERS F-CRIN platform (AP-HP).

6. Data management and Statistical Analysis

   Data management will begin after the inclusion of the first patients according to a specific data-monitoring plan. Standard operating procedures for data management issued from the FCRIN-PARTNERS-APHP platform (currently in the process of data-center certification by ECRIN (European clinical research infrastructures network)).

7. Statistical analysis will be made using SAS version 9

   A detailed statistical analysis plan will be written before freezing the database. All steps of data transfer between eCRF and the Statistical Analysis System (SAS) database will be traced according to the procedures proposed by FCRIN-PARTNERS-APHP for its data-center ECRIN certification.

3.3.3 Data stored and annotated

All the images and data were collected thanks to a common platform and centralized on a server. The images were annotated by graders in the virtual reading center run by the Ophthalmology Department in Lariboisiere, using the Annotation Software product developed during the project's first year.

Storage of data

Evolucare/ADCIS developed client-side integration software to interact with each OCT, OCTA, and color-fundus system: "Annotate". Specific software products had to be installed on computers connected to the acquisition systems to retrieve the data from each acquisition device and upload them to a server. Such a tool also requires specific development to interface with each device and collect data. Second, Evolucare developed secured cloud-based software and infrastructure ("the platform") to host the data acquired by the different sites. A server ensures compatibility with client computers and current standards. The platform, in particular, presents 2D and 3D image visualization capabilities, with the availability of host rendering for live-shared visualizations, allowing the sharing of expertise and reaching consensus between professionals as well as live decisions for annotations. Various users' typologies' annotations are also managed,

statistically controlled for multicomparative annotation and quality control, and used for feedback and further reinforcement and learning by the algorithms. The platform was designed to handle at least 500 terabytes (Tb) of data in a centralized secured host with one local buffer server per acquisition center. It will integrate algorithms after their industrialization.

Annotation of data

Based on "Annotate", AP-HP set up a virtual Reading Center in the Ophthalmology department of Lariboisière, where the retinal images produced by the clinical studies are being read, visually interpreted, and labeled. Images provided by recruiting centers are graded by several independent observers after the creation of a reading center. Two non-ophthalmologist "graders" (trained orthoptists) read each image independently. Then, one ophthalmologist "super-grader" analyzes the two sets of annotations to produce a final consolidated version. All were trained for the use of the digital platform and tools to use them with high efficiency and quality. They were also involved in cloud-based grading software specifications and development. Annotations generated by the graders were used to develop the algorithms. At the end of the EviRed project, EviRed will study the feasibility of maintaining the reading center at its size or developing it to accept external image reading works for other clinical trials.

Data storage servers

The hosting structure of the dataset for EviRed is primarily based on the robust infrastructure provided by OVH Cloud services [1], a certified health data host. This assures the security and accessibility of the data

1. `https://www.ovhcloud.com/`

throughout the process of this research.
The dataset infrastructure utilizes a total of 10 servers, each with a specific role. These include:

— Two storage servers are responsible for the storage of incoming data. These servers maintain the main body of the dataset and ensure efficient data management and retrieval.
— An annotation server dedicated to the process of data annotation. This server carries anonymized data, making it accessible to the annotation team.
— A master server employed by the AI team. This server contains the learning database, which is essential for the team's training tasks.
— Six computing servers, each equipped with 4 Tesla V100s GPUs and 192 GB of RAM. These servers are primarily used for deep learning model training and preprocessing tasks.

The data for this research is primarily sourced from recruitment centers. It is directly sent to the storage servers, ensuring a streamlined data input flow. For data security and redundancy, a copy of all the incoming data is also saved on a backup server. This safeguards the data against any unforeseen loss or server failure.
To ensure the privacy and security of the data subjects, only anonymized data is made available on the annotation server. The annotation teams are given access to this server for the purpose of data annotation. In this way, we maintain a balance between accessibility for data processing and privacy for the data subjects.
The AI team, on the other hand, is provided access to the training dataset on the master server. This server serves as the team's main access point for machine learning operations and tasks.

3.3.4 Data description

There will be two groups of patients in the EviRed cohort: one group of 4000, which will be used to develop algorithms, and one group of 1000, which will be used to validate these algorithms (evaluation cohort). The algorithms will be trained using the data of the remaining 4000 patients (training cohort). Both general data and ophthalmological data will be collected. In addition to the usual clinical care, patients will undergo retinal imaging with two devices of different brands (Zeiss and/or another brand) instead of one device for color fundus, OCT, and OCT angiography.

Images modalities

The EviRed project will use Ultra-WideField Color Fundus Photography (UWF-CFP) and OCTA images taken by different brands for the development of the algorithm. Here is the information regarding the two modalities:

— UWF-CFP

As discussed in Chapter 2, Ultra-widefield retinography examination is a technological evolution of conventional retinography examination by allowing extensive retina views and providing more information about signs of pathology in the periphery. Different UWF devices exist that produce visually different images. Three different brands of UWF images are available in the EviRed project: Clarus, Optos, and Eidon. As shown in Fig.3.5, both sets demonstrate the variations in imaging quality and detail between the respective systems.

— OCTA

As discussed in Chapter 2, OCTA provides a detailed view of the retinal and choroid vascularization at the microvascular level. Five

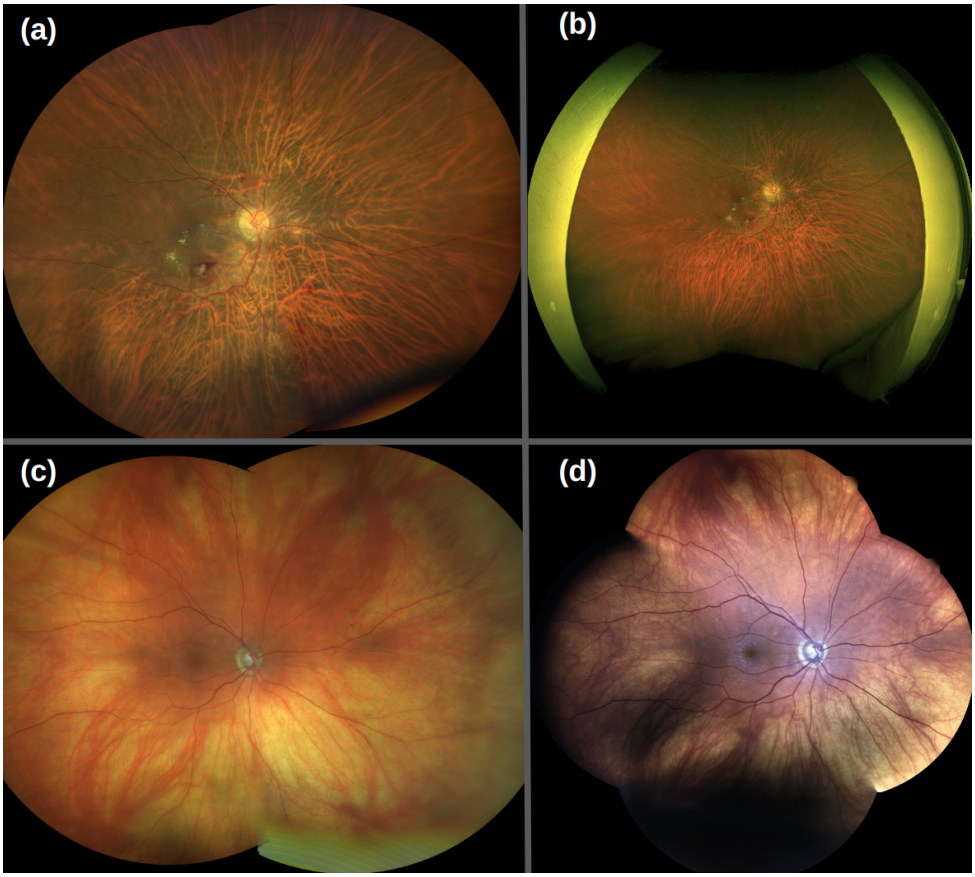


Figure 3.5 – Comparison of imaging modalities for two patients. (a) Clarus image of Patient 1, (b) Optos image of Patient 1, (c) Clarus image of Patient 2, and (d) Eidon image of Patient 2.

different brands of OCTA images are available in the EviRed project: PlexElite, Cirrus, Spectralis, Triton, and Optovue. For the time being, OCTA data is dominated by PlexElite, and data for other brands is limited. Furthermore, we have not yet finished developing a parser and viewer for other brands besides PlexElite. Therefore, in my thesis research, we only used data from the PLEX®Elite 9000 model specifically.

The PLEX®Elite 9000 has a scanning frequency of 200 kHz and is capable of acquiring both 15×15 mm^2 and 6×6 mm^2 SS-OCTA images. Following the EviRed study protocol, each patient's ocular data often contains two specifications of acquisitions: 6×6 mm^2 high-resolution SS-OCTA and 15×15 mm^2 UWF-SS-OCTA. Each OCTA image encompasses both structural (Structure) and flow (Flow) information. Fig. 3.6 shows en-face images and their corresponding B-scan images (pre-processed in Chapitre 6) of the Structure and Flow from the same patient acquired for different specifications.

The EviRed raw data size is $500 \times 1536 \times 500 \times 2$ voxels for the 6×6 mm^2 SS-OCTA and $834 \times 3072 \times 834 \times 2$ voxels for the 15×15 mm^2 SS-OCTA. The last dimension (channels) presents the information of Structure and Flow, respectively.

Fig.3.7 presents the lesions observed in a patient with severe non-proliferative diabetic retinopathy. The LSO image (a) provides an initial overview of the center of the retina using OCTA, while the high-resolution Clarus image (c) delivers a more detailed panorama of both the center of the retina and its periphery. Structural images (d-f) underscore the anatomical intricacies and potential changes, with macular edema being evident in images (d) and (f). Meanwhile, the flow images (b, e, g) illuminate the dynamic vascular activities

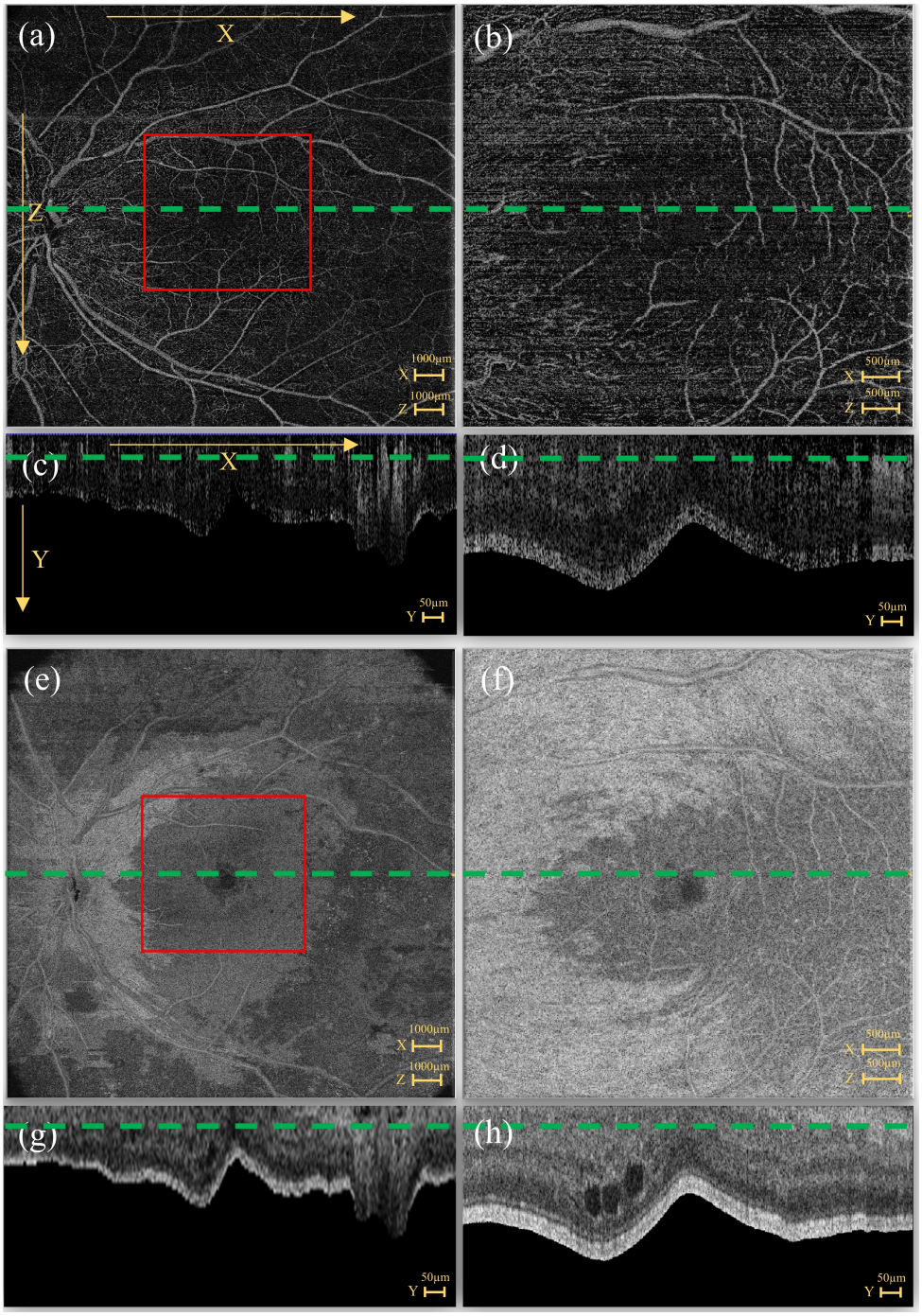


Figure 3.6 – Structure and Flow en-face slices and B-scan images from 6×6 mm^2 SS-OCTA and 15×15 mm^2 SS-OCTA. (a,c) Flow of 15×15 mm^2 SS-OCTA. (b,d) Flow of 6×6 mm^2 SS-OCTA. (e,g) Structure of 15×15 mm^2 SS-OCTA. (f,h) Structure of 6×6 mm^2 SS-OCTA. The area on the 6×6 mm^2 SS-OCTA is in the center of the 15×15 mm^2 SS-OCTA image (red bounding box). The green line in the en-face slice shows the source of the B-scan, and the green line in the B-scan image shows the intercept direction of the en-face slice.

and potential discrepancies. Together, these imaging modalities complement one another, not only showcasing the severity of the pathology but also offering a comprehensive perspective, emphasizing the nuanced variations and valuable insights each imaging method contributes.

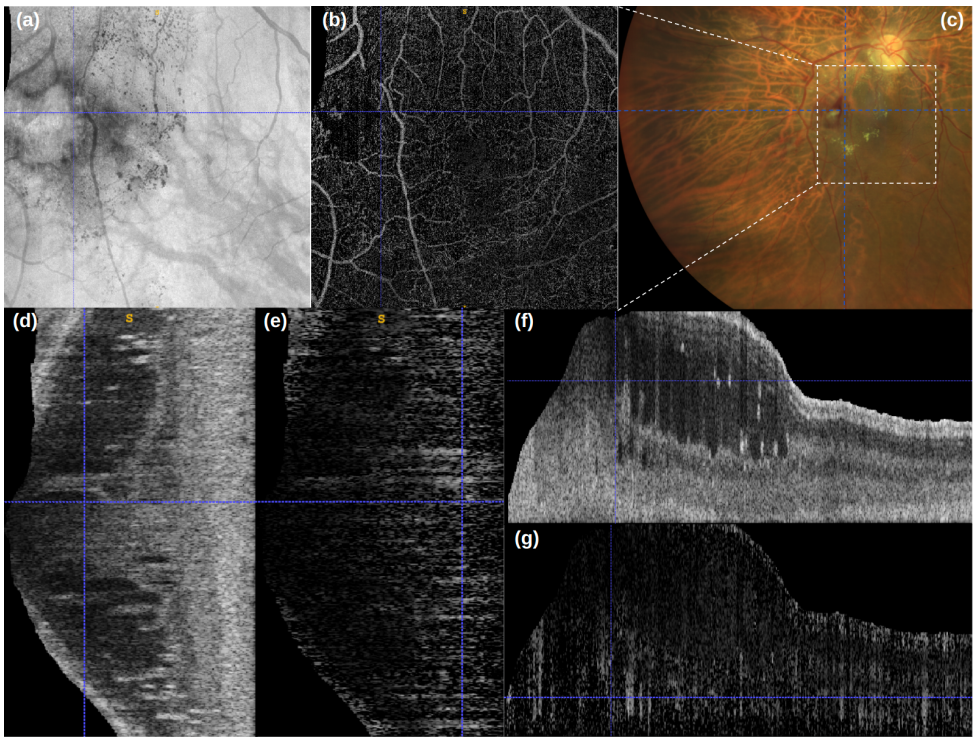


Figure 3.7 – Multimodal imaging of a patient with severe non-proliferant diabetic retinopathy. (a) LSO image, (c) Corresponding Clarus image, (b, e, g) Flow images, and (d, f) Structure images.

Contextual data

In addition to imaging modalities, at inclusion, the following data will be recorded: Age, gender, type and duration of diabetes, race/ethnicity, occupational group, history of poor glycemic control, current medication (oral or injectable diabetes medication including insulin, aspirin, statin, oral blood pressure lowering drugs), tobacco smoking, BMI, history of car-

dio vascular disease (Stroke, Coronary event or revascularization, Peripheral arterial disease/lower extremities arterial disease or revascularization), nephropathy (Proteinuria, Renal failure, Dialysis, Graft), Plantar foot ulcer and/or amputation, History of nephropathy, neuropathy, pregnancy, sleep apnea, bariatric surgery, environmental data (address, occupational group), ophthalmological history (cataract surgery, ophthalmological history, vitrectomy, IOP lowering drugs, laser photocoagulation (PRP/focal), intravitreal injections (anti VEGFs/steroids).

Training cohort

The Training cohort of EviRed's program consists of data collected from up to 4,000 patients. However, the recruitment and annotation of EviRed data is still in progress, and I received a total of three phases of data in my thesis study. The first stage of data containing 103 patients was received in January 2022. However, these data were not annotated since the annotation process had not been completed. As of November 2022, we received the second stage data, comprising 532 patients with detailed annotated e-CRF files. The 532 patients included the 103 patients from the first stage. In July 2023, we received data and annotations for an additional 500 patients, which, together with 532 patients from the second stage, comprise the third stage data for 1032 patients. It is important to note that the EviRed project is ongoing and that more data will be added to the training cohort and the data collected in the three stages described above. These are the imaging information of the three stages of data:

1. First stage: 103 patients without annotations

   The data information of the different devices in 103 patients is shown in Fig. 3.8. For the first stage, we only used data from PlexElite for

102 patients with 243 eyes.

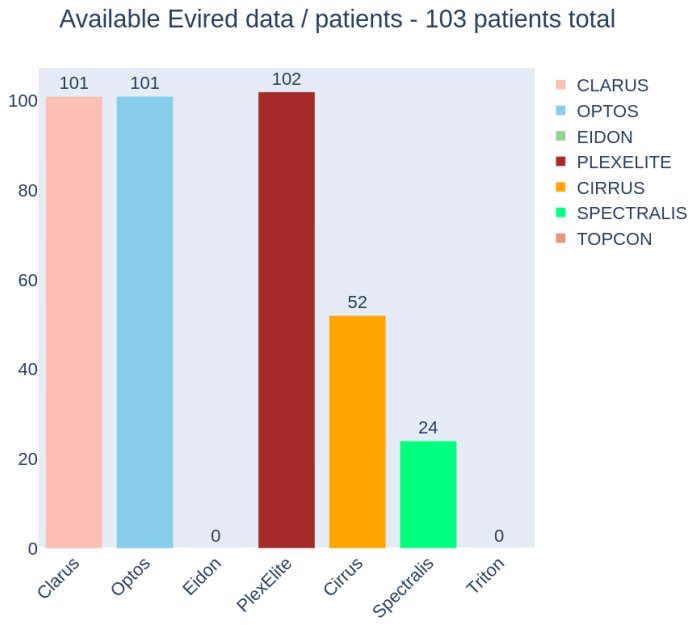


Figure 3.8 – Patient information for different devices in the first stage data.

2. Second stage: 532 patients with annotations

   The data information of the different devices in 532 patients is shown in Fig. 3.9. For the second stage, based on the number of photographs taken with different brands of devices, we used the data from Clarus and PlexElite to perform as the source of UWF and OCTA in multimodal fusion.

3. Third stage: 1032 patients with annotations

   The data information of the different devices in 1032 patients is shown in Fig. 3.10. The third stage of the data contains longitudinal time-series data for the patient, where V1 represents the patient's first visit and V2 represents the patient's second follow-up visit 12 months later. Due to the fact that the third stage data were received at the end of the third year of the thesis, the work on the thesis did not

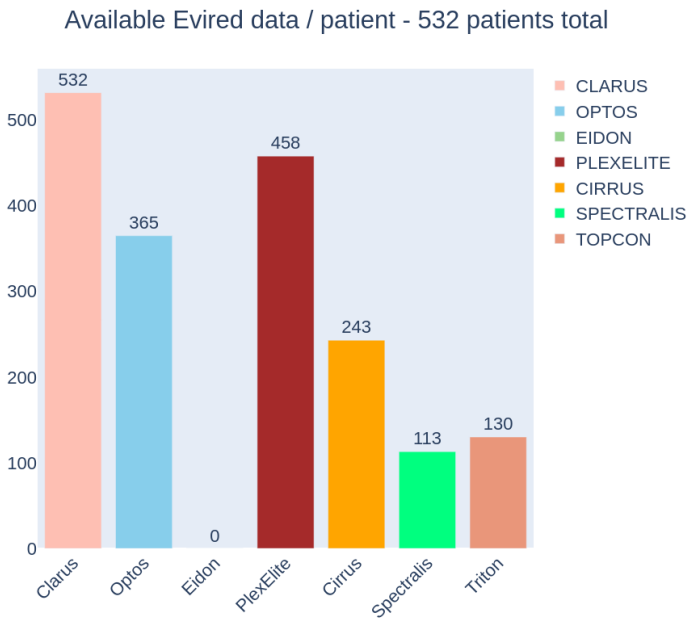


Figure 3.9 – Patient information for different devices in the second stage data.

involve the testing of the third stage data.

Evaluation cohort

EviRed project will use the 1000 diabetic patient data (evaluation cohort) to validate the algorithms that are developed. Algorithms will be evaluated by comparing the automatic progression prediction provided by the algorithm to the effective progression observed after one year. The main outcome measures will be sensitivity, specificity, and the Area Under the Curve (AUC) of the algorithm to predict DR progression toward a severe form of DR (defined by the presence of proliferative DR and/or severe macular edema involving the center of the macula, or the need for laser photocoagulation, vitrectomy, or intravitreal injection) in the following year.

The evaluation cohort will be used to evaluate our fusion algorithm in order to ensure its accuracy and robustness at the end of the project. At

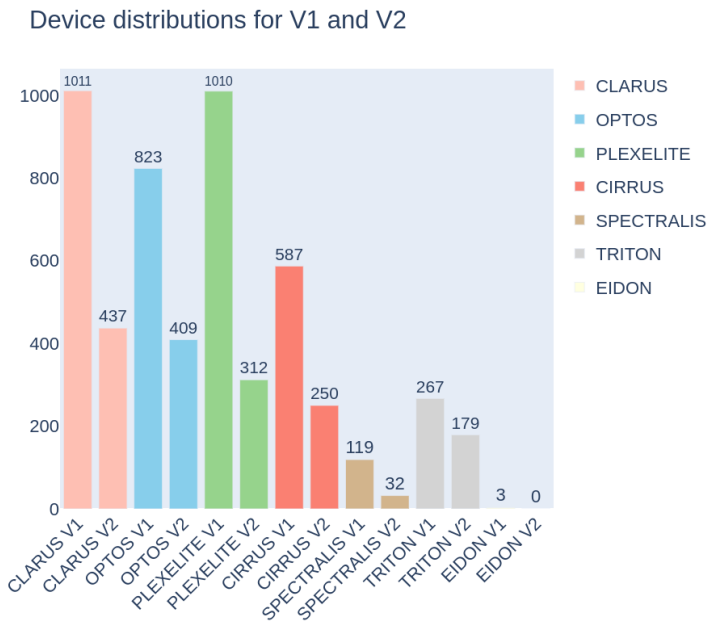


Figure 3.10 – Patient information for different devices in the third stage data.

this time, the EviRed project is still in a developmental stage, we have not received all of the training cohort data yet, and the algorithm is still awaiting further testing. Besides, the LaTIM is not expected to test the algorithms on the evaluation cohort: ADCIS is in charge of testing the final algorithms independently.

3.4 Supplemental dataset

In order to assess the effectiveness and robustness of the multimodal algorithm developed based on EviRed retrospective data before the arrival of the prospective datasets, we selected an ophthalmic multimodal classification dataset, Glaucoma grAding from Multi-Modality imAges (GAMMA) [1], in order to test the algorithm's effectiveness and robustness.

GAMMA dataset is provided by Sun Yat-sen Ophthalmic Center, Sun

1. `https://aistudio.baidu.com/aistudio/competition/detail/119/0/introduction`

Yat-sen University, Guangzhou, China. There are 200 pairs of clinical modality images in the dataset, 100 pairs in the training set, and 100 pairs in the test set. Each pair contains a 45° fundus image and an OCT volume. The OCT volumes were acquired using a Topcon DRI OCT Triton machine. The OCT was centered on the macula and had a 3 × 3 mm en-face field of view. The Kowa 2000 × 2992 and Topcon TRC-NW400 cameras were used to acquire fundus images [335].

It aims to analyze clinical data of two modalities, 2D fundus images and 3D OCT scanning volumes, and classify the samples into three categories according to visual features: no glaucoma, early glaucoma, and moderate or advanced glaucoma as shown in Fig. 3.11. Please note that the samples without glaucoma are not normal samples without disease but patients with other eye diseases.

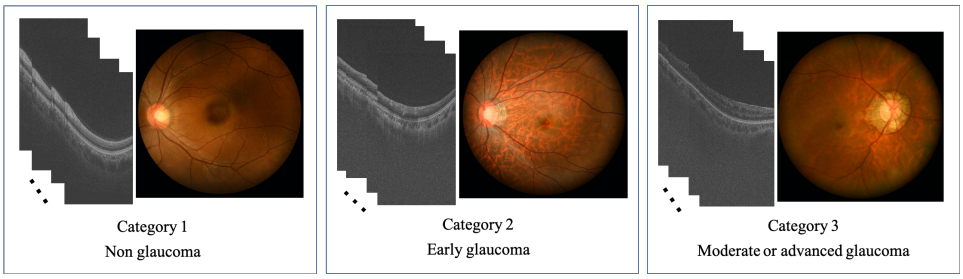


Figure 3.11 – GAMMA dataset and classification.

3.5 Conclusion

The process of our thesis progressed with the EviRed project, but the data's delayed arrival has also caused many difficulties for our thesis. For this reason, combining the goals of our thesis with the arrival times of different datasets, we made the following plans as we progressed through the thesis:

1. In the first year and a half of the thesis, the fusion algorithm for 2D LSO, 3D structural OCT, and 3D OCT angiography was developed using the EviRed retrospective dataset in Chapter 4, and the effectiveness of the fusion algorithm was validated on the additional dataset GAMMA.
2. At the beginning of the second year of the thesis, the first stage data of the EviRed prospective dataset (103 patients without annotations) arrived, for which we tried to explore the unlabeled data, hoping to combine the unlabeled data to improve the performance of the classification. In Chapter 5, we added the arrival of the EviRed prospective dataset (103 patients without annotations) to the current dataset for further multimodal fusion testing.
3. At the beginning of the third year of the thesis, the second stage data of the EviRed prospective dataset (532 patients with annotations) arrived. We proposed a new hybrid fusion algorithm for the joint analysis of different specifications of OCTA acquisitions in Chapter 6.
4. Meanwhile, in the second half of the third year, we tried to test the joint analysis of OCTA and UWF-CFP using the second-stage data of the EviRed prospective dataset (532 patients with annotations) in Chapter 7.

Therefore, we were able to successfully manage our time and schedule even as the data were delayed, and thus achieve our thesis objective.

Chapter 4

Multimodal information fusion in OCTA

"I imagine a world in which AI is going to make us work less."

— *Gwenolé Quellec*

As part of the early stages of the thesis, we explored the fusion meth-

ods of different modalities in OCTA. In OCTA, we tested fusion for three types of information: 2D line scanning ophthalmoscope (LSO), 3D structural OCT (Structure), and 3D OCT angiography (Flow). To solve retinal analysis tasks, we investigated three multimodal information fusion strategies based on deep learning using the EviRed retrospective and public GAMMA datasets: input fusion, single-level fusion, and hierarchical fusion. In addition, we experimented with the performance of the attention mechanism on hierarchical fusion.

4.1 Introduction

In recent years, algorithms for diagnosing glaucoma and DR have emerged with the development of deep learning and improved computer equipment. Fundus photography and optical coherence tomography (OCT) are the two most cost-effective screening tools for glaucoma and DR [335]. For two-dimensional fundus photographs, powerful convolutional neural networks (CNN), such as ResNet or GoogleNet Inception models, were used to achieve pathology detection [336–338]. It should be noted that 2D fundus data are more accessible than other modalities, so datasets are generally larger, and thus, models can be trained more efficiently. OCT data are more sensitive to structural pathological features. Both 3D-CNN networks and 2D-CNN networks operating on 2D slices were used to achieve feature extraction from OCT volumes [339–341]. In addition, optical coherence tomography angiography (OCTA) is a new, non-invasive imaging technique that generates volumetric angiography images in seconds. It can display both structural and blood flow information [342]. The effectiveness of CNN networks in classifying DR using OCTA data was also demonstrated [343].

All the previous algorithms are usually based on information from only

one modality. However, multi-modality screening is often recommended to reach a more accurate and reliable diagnosis [17]. This is why multimodal algorithms are needed in ophthalmic pathology diagnosis.

Humans are capable of recognizing salient regions in complex scenes naturally and effectively. This observation led to the introduction of attention mechanisms into computer vision in order to mimic this aspect of the human visual system [344]. The attention mechanism described above can be regarded as a process of dynamic weight adjustment based on features contained in the input image [345]. In recent years, attention mechanisms have become a common element of neural architectures and have been applied to a variety of tasks [346], such as image classification [128, 347], text classification [348, 349], machine translation [350], action recognition [351, 352] and image caption generation [353]. It is worthwhile exploring the possibility of incorporating attention mechanisms into multimodal fusion architectures.

4.2 Material and methods

This chapter presents three fusion algorithms for multimodal data in ophthalmology: input fusion, single-level fusion, and hierarchical fusion. They enable the fusion of 2D and 3D modal data. Specifically, the innovative hierarchical fusion algorithm we developed (Fig. 4.2) achieves excellent glaucoma and DR classification results. We will examine the challenges of applying different fusion methods to ophthalmic data and the structural aspects of our network. Furthermore, we performed some exploratory tests of the attention mechanism in the fusion network architecture. We evaluated the proposed method using the EviRed retrospective dataset for diabetic retinopathy classification and the public GAMMA challenge dataset for

glaucoma classification.

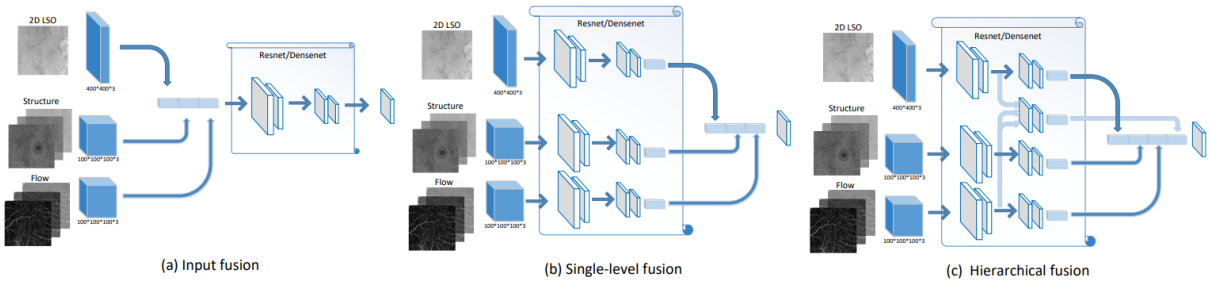


Figure 4.1 – Different fusion methods for fusing LSO, Structure, and Flow modality information in the EviRed retrospective dataset.

4.2.1 Input fusion

As described in section , for input fusion, data from different modalities are fed into a classification network as different channels [195]. Specifically, multi-modality images are fused channel by channel to form multi-channel inputs. Then, a classification network is trained to learn a fused feature representation from these inputs. Many of today's medical fusion strategies use input fusion [354, 355].

Let $X \times Y \times Z$ denote the size of the 3D volumes in voxels. In the input fusion solution, the 2D images are resized to $X \times Y$ pixels and duplicated Z times to form a $X \times Y \times Z$ voxel channel. Feature extraction from the multimodal input of size $C \times X \times Y \times Z$, where C denotes the number of channels, is then performed using a 3D-CNN network as shown in Fig. 4.2 (a). In addition, the alignment of different modalities is crucial to input fusion.

Input fusion is a simple method, but it is not very effective due to the semantic gap between the modalities of ophthalmic data. For example, fundus photographs give an overall en-face view of the retina in 2D, and OCT volumes provide structural information about the retina in 3D.

However, there is a significant gap between these two modalities regarding the equipment used to capture them, imaging methods, and data information. In particular, when we convert 2D data into 3D volumes, we cannot guarantee that the modalities are accurately aligned.

4.2.2 Single-level fusion

As described in section , in contrast with input fusion, single fusion does not assume spatially aligned modalities. Instead, each modality data is used as an input to a single classification branch, and the outputs from each branch are integrated to produce a final result [356, 357]. The single-level fusion strategy fuses features before the final decision layer, as shown in Fig. 4.2 (b). In contrast, late fusion fuses the decision results, ignoring any correlation between the different modalities [358].

As we use different independent branches to extract feature information from each modality, we do not need to consider the consistency of the input data. Using different 2D and 3D CNN branches to extract different features for 2D and 3D data is possible.

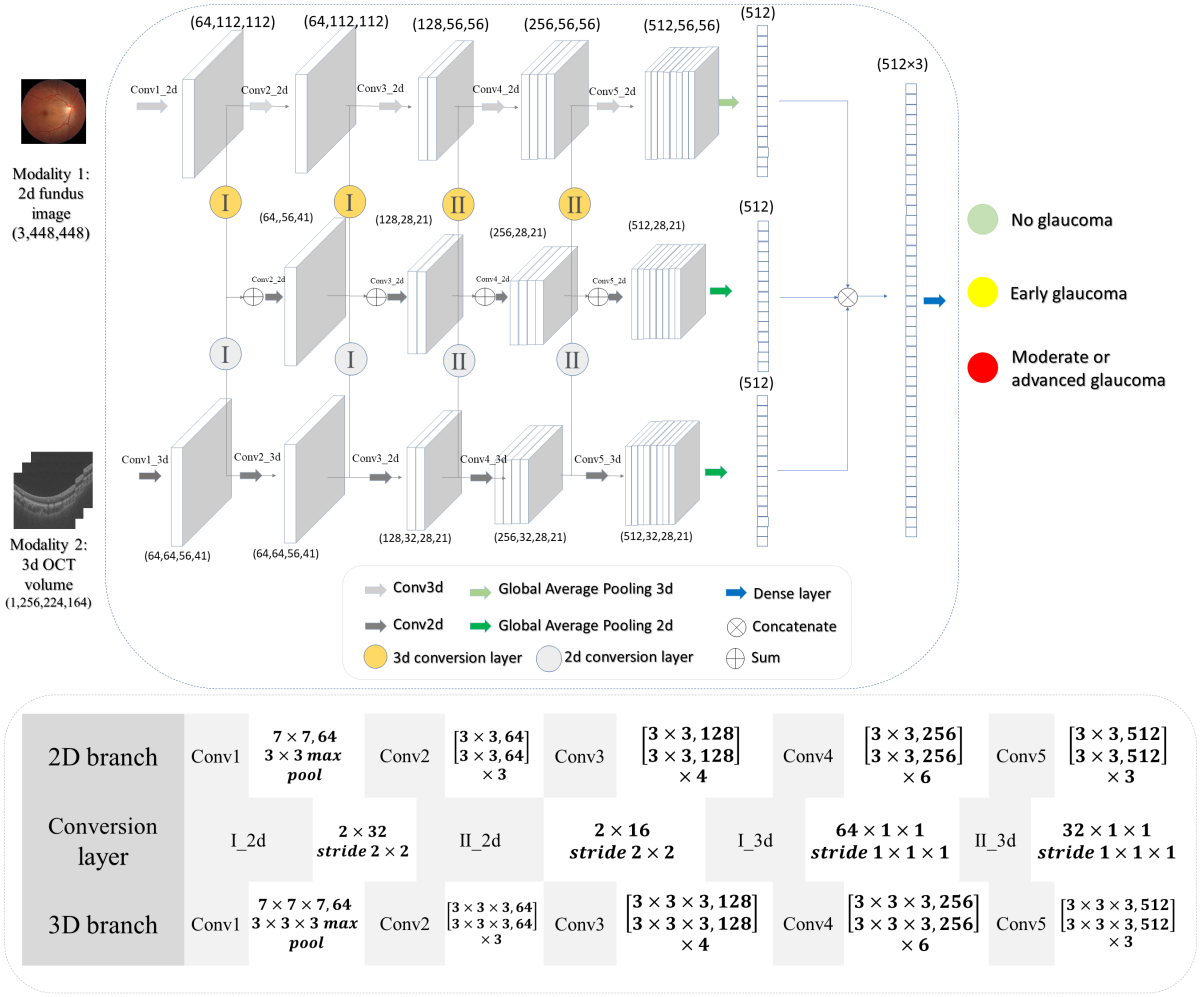
Single-level fusion is a simple yet effective method for feature fusion. The method effectively bridges the significant gaps between different modalities in ophthalmology (2D fundus images and 3D OCT or OCTA volumes). In particular, most participants in Task 1 of the GAMMA Challenge employed this method to classify glaucoma and achieved good results [17]. Nevertheless, as single-level fusion is a mere concatenation of high-dimensional features, the correlation information inevitably gets lost, adversely impacting classification performance.

4.2.3 Hierarchical fusion

In this work, we have extended the network structure of single-level fusion to address its shortcomings. As described in section , hierarchical fusion works by using each modality image as an input of a single classification branch and then fusing these learned individual feature representations in the deeper layers of the network. However, unlike single-level fusion, an additional branch performs feature fusion at different scales. A decision layer is then applied to the fused result to obtain the final label [262].

Fusion between modality-specific features of different dimensions in a network structure is challenging. Prior studies have generally focused on simpler problems. For example, the fused modalities are all 3D data of the same size in [359]. In that case, multimodal features always have the same shape at each scale so that feature fusion can be easily achieved through concatenation. For ophthalmic data, the size and dimensionality of the features are modality-dependent: 3D tensors for 2D images and 4D tensors for 3D images.

A solution is proposed hereafter and illustrated in Fig. 4.2. Two CNN branches are used to extract features from a multichannel 2D image and a multichannel 3D volume, respectively. Furthermore, we use a third fusion branch to achieve feature fusion at different scales. Since the dimensional features are of different dimensions and sizes, to allow alignment between 3-D and 2-D features, we introduce the concept of "conversion layers", which can be used to harmonize their shape before concatenating them. In these conversion convolution layers, the parameters are calculated according to the size of the modality-specific features.



2D branch	Conv1	$7\times7, 64$ 3×3 ***max pool***	Conv2	$\begin{bmatrix}3\times3, 64\\3\times3, 64\end{bmatrix}\times3$	Conv3	$\begin{bmatrix}3\times3, 128\\3\times3, 128\end{bmatrix}\times4$	Conv4	$\begin{bmatrix}3\times3, 256\\3\times3, 256\end{bmatrix}\times6$	Conv5	$\begin{bmatrix}3\times3, 512\\3\times3, 512\end{bmatrix}\times3$	
Conversion layer	I_2d	2×32 ***stride*** 2×2	II_2d	2×16 ***stride*** 2×2	I_3d	$64\times1\times1$ ***stride*** $1\times1\times1$	II_3d	$32\times1\times1$ ***stride*** $1\times1\times1$			
3D branch	Conv1	$7\times7\times7, 64$ $3\times3\times3$ ***max pool***	Conv2	$\begin{bmatrix}3\times3\times3, 64\\3\times3\times3, 64\end{bmatrix}\times3$	Conv3	$\begin{bmatrix}3\times3\times3, 128\\3\times3\times3, 128\end{bmatrix}\times4$	Conv4	$\begin{bmatrix}3\times3\times3, 256\\3\times3\times3, 256\end{bmatrix}\times6$	Conv5	$\begin{bmatrix}3\times3\times3, 512\\3\times3\times3, 512\end{bmatrix}\times3$	

Figure 4.2 – Proposed hierarchical fusion configuration, illustrated using 2D and 3D ResNet34, for glaucoma classification from 2D fundus photography and 3D OCT. I and II are different types of conversion layers, and their configurations are shown in the list.

$$F_{3D}\left(C \times Z_{3D} \times X_{3D} \times Y_{3D}\right) \Longrightarrow F'_{3D}\left(C \times 1 \times X_{3D} \times Y_{3D}\right)$$
$$F_{2D}\left(C \times X_{2D} \times Y_{2D}\right) \longrightarrow F'_{2D}\left(C \times X_{3D} \times Y_{3D}\right)$$

Where F is the feature of modality, X, Y, Z, C represent the length, width, depth, and number of channels of the features. $\Longrightarrow$ and $\longrightarrow$ represent the 3D and 2D conversion convolutional layers respectively. The convolution kernel size and stride of 3D conversion convolutional layers are $(Z_{3D} \times 1 \times 1)$ and $(1 \times 1 \times 1)$. For the 2D conversion layer, the stride is set to (2, 2) and the filter size is set to $(X_{2D} - 2[X_{3D} - 1], Y_{2D} - 2[Y_{3D} - 1])$, without padding, to ensure that F'_{2D} matches the size of F'_{3D}. The parameters of each convolutional layer are shown in Fig. 4.2 for ResNet34.

We also extract the features from the 2D CNN block to reduce the number of parameters of the fusion branch. In the end, the high-dimensional features of the three branches are concatenated, and the classification layer is used to make the final classification.

In addition to the advantages of single-level fusion, hierarchical fusion also considers features from different scales, enhancing the correlation between different modalities and increasing the accuracy of diagnosis.

4.2.4 Attention mechanism

Channel attention block

In multi-modal fusion tasks, attention complementary strategies can be applied to extract synergies between multi-modal images. Based on the architecture of hierarchical fusion, [262] proposed a channel attention block as in Fig. 4.3. The channel attention block can selectively extract features from different branches and suppress irrelevant information. The

fusion ratio of each modality is automatically determined according to the importance of the data in the attention block.

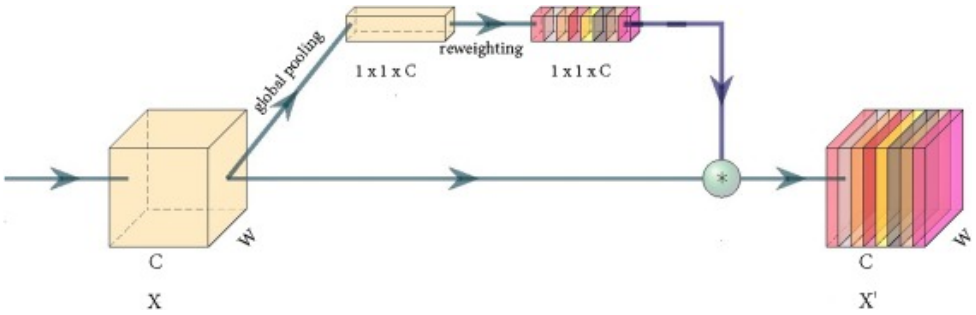


Figure 4.3 – The architecture of the channel attention block. X represents the input feature map. X' represents the output feature map after channel reweighting.

The deployment of the attention model is completed in three steps for an input feature map X with C channels.

(1) The two-dimensional features of each channel are compressed along the spatial dimension, transforming them into scalars. The real number has a global receptive field to some extent, and the output dimensions are determined by the number of input feature channels. Global pooling can be used to implement this process, and the following equation depicts the squeeze operation.

$$Z_C = F_{se} = \frac{1}{H * W} \Sigma_i^H \Sigma_j^W X_C(i, j)$$

Where H, W, and C represent each feature map's height, width, and number of channels, respectively.

(2) To explicitly model the correlation between feature channels, new weights are generated for each channel to map its importance with a compressed set of scalars. A 1 x 1 convolution can be used to examine the correlation among different channels in order to determine their weight

distribution. The corresponding calculation is shown as follows.

$$S_C = \delta(conv(Z_C))$$

Where δ represents the Sigmoid activation function.

(3) As a result of the second step, weights are regenerated to reflect the importance of each channel. The original features are then multiplied gradually in order to complete the redistribution of the original features in the channel dimension. The transition of the input feature map X_C to X'_C can be expressed as follows.

$$X'_C = S_C \otimes X_C$$

As a result, feature maps X_C are transformed into feature maps X'_C, which contain reweighted channel information. It can be considered that the attention model essentially introduces additional dynamic characteristics on the input, which can be viewed as self-attention functions. Therefore, we combined hierarchical fusion and channel attention to develop the fusion model structure shown in Fig. 4.4.

Dual attention fusion block

Considering that not all the features extracted from the encoders are useful for diagnosis, [360] proposed to recalibrate the features along the modality and spatial paths using dual attention-based fusion blocks as in Fig. 4.5, which suppresses less informative features and emphasizes more informative ones.

(1) In the modality attention module, a global average pooling is first performed to produce a tensor $g \in R^{1\times1\times4}$, which represents the global

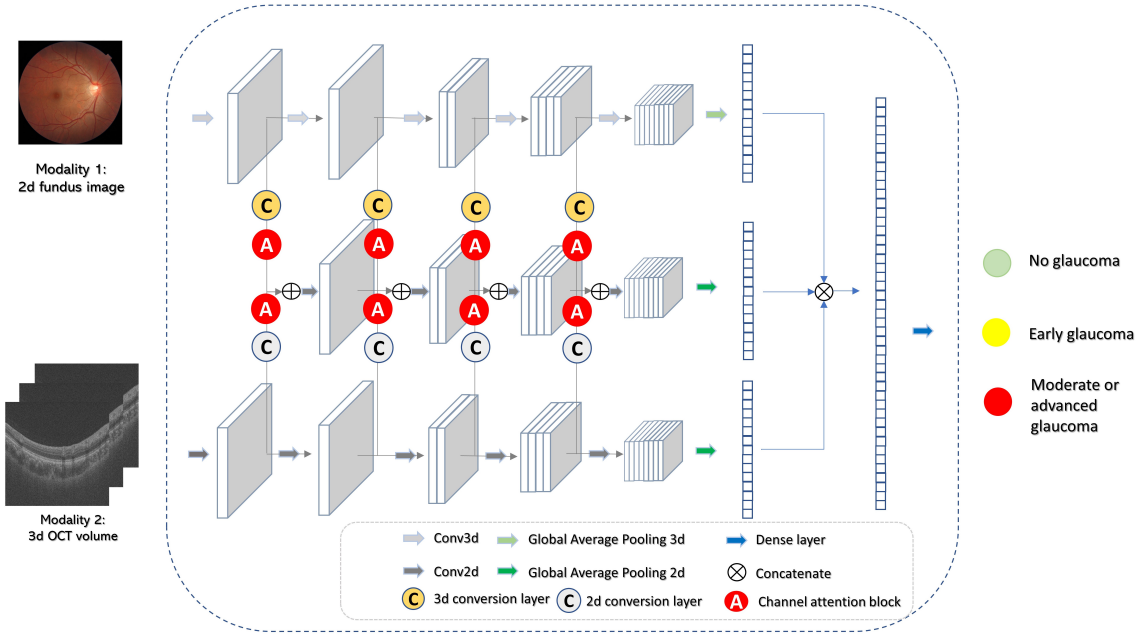


Figure 4.4 – Proposed hierarchical fusion with channel attention blocks, for glaucoma classification from 2D fundus photography and 3D OCT.

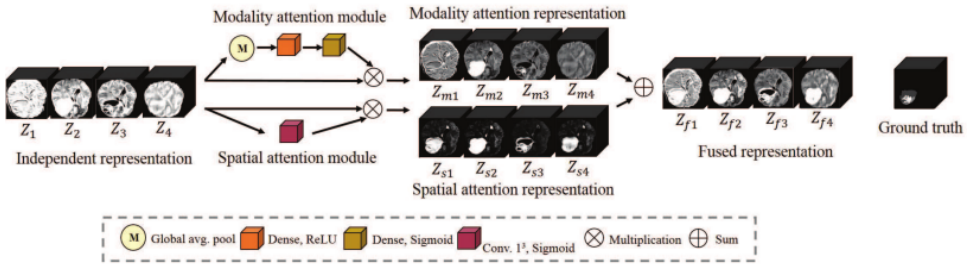


Figure 4.5 – The architecture of the dual attention fusion block. The individual feature representations (Z_1, Z_2, Z_3, Z_4 are first concatenated, then they are recalibrated along modality attention module and spatial attention module to achieve the modality attention representation Z_m and spatial attention representation Z_s, final they are added to obtain the fused feature representation Z_f.

spatial information of the feature representation, with its k^{th} element.

$$g_k = \frac{1}{H \times W} \sum_{i}^{H} \sum_{j}^{W} Z_k(i,j)$$

Then two fully-connected layers are applied to encode the modality-wise dependencies, $\hat{g} = W_1(\delta(W_2 g))$, with $W_1 \in R^{4\times 2}$, $W_2 \in R^{2\times 4}$, being weights of two fully-connected layers and the ReLU operator $\delta(\cdot)$, $\hat{g}$ is then passed through the sigmoid layer to obtain the modality-wise weights, which will be applied to the input representation Z through multiplication to achieve the modality-wise features Z_m, and the $\sigma(\hat{g_k})$ indicates the importance of the i modality of the feature representation.

$$Z_m = [\sigma(\hat{g_1})Z_1, \sigma(\hat{g_2})Z_2, \sigma(\hat{g_3})Z_3, \sigma(\hat{g_4})Z_4]$$

(2) In the spatial attention module, the feature representation can be considered as $Z = [Z^{1,1}, Z^{1,2}, ..., Z^{i,j}, ..., Z^{H,W}]$, $Z^{i,j} \in R^{1\times 1\times 4}$, $i \in 1, 2, ..., H$, $j \in 1, 2, ..., W$ and then a convolution operation $q = W_s \star Z$, $q \in R^{H\times W}$with weight $W_s \in R^{1\times 1\times 4\times 1}$, is used to squeeze the spatial domain, and to produce a projection tensor, which represents the linearly combined representation for all modalities for a spatial location. The tensor is finally passed through a sigmoid layer to obtain the space-wise weights, $\sigma(q_{i,j})$ indicates the importance of the spatial information (i, j) of the feature representation.

$$Z_s = [\sigma(q_{1,1})Z^{1,1}, ..., \sigma(q_{i,j})Z^{i,j}, ..., \sigma(q_{H,W})Z^{H,W}]$$

(3) Finally, the learned fused feature representation is obtained by adding the modality- and space-wise feature representation.

$$Z_f = Z_m + Z_s$$

Due to the fact that the dual attention fusion block can be directly applied to any multimodal fusion problem, we applied it to our hierarchical fusion, as shown in Fig. 4.6.

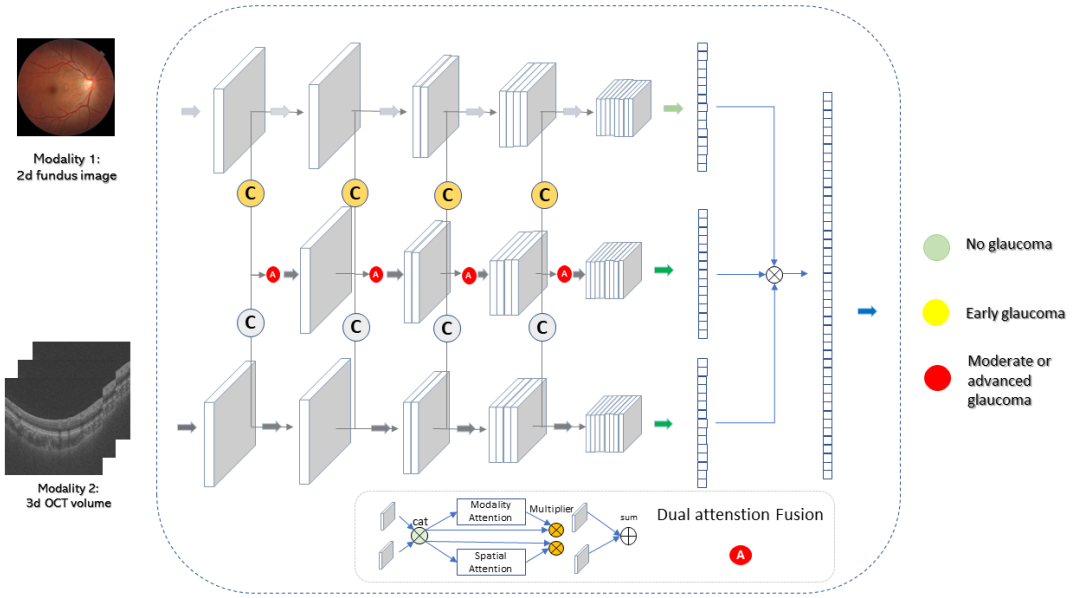


Figure 4.6 – Proposed hierarchical fusion with dual attention fusion blocks, for glaucoma classification from 2D fundus photography and 3D OCT.

4.2.5 Data and classification tasks and metrics

EviRed retrospective dataset

For the EviRed retrospective dataset, we investigated the fusion of 3 modalities: 3D Structure, 3D Flow, and 2D LSO for the classification of

Table 4.1 – Distribution of eyes with different levels of severity in different datasets.

Severity	Train set	validation set	Test set
Absence of diabetic retinopathy	15	4	5
Mild NPDR	19	7	7
Moderate NPDR	13	4	4
Severe NPDR	25	8	10
PDR or PRP	16	5	9
Total	88	28	35

diabetic retinopathy. DR severity was assessed by a retina specialist using fundus photographs, according to the International Clinical Diabetic Retinopathy Disease Severity Scale (ICDR): absence of diabetic retinopathy, mild nonproliferative diabetic retinopathy (NPDR), moderate NPDR, severe NPDR, proliferative diabetic retinopathy (PDR) and panretinal photocoagulation (PRP). We performed four binary classification tasks: task0 (detecting mild NPDR or more), task1 (detecting moderate NPDR or more), task2 (detecting severe NPDR or more), and task3 (detecting PDR or PRP). To assess the performance of four binary classifications, we used Area Under the ROC Curve (AUC) described in Section 2.2.3: AUC0 ($\geq$ mild NPDR), AUC1 ($\geq$ moderate NPDR), AUC2 ($\geq$ severe NPDR) and AUC3 ($\geq$ PDR).

After removing patient data with incomplete modalities and absent annotations from the EviRed retrospective dataset described in Section 3.2, 151 acquisitions from 64 diabetic patients were collected for the binary classification. This collection was divided as follows: 88 acquisitions (from 31 patients) for training, 28 acquisitions (from 14 patients) for validation and 35 acquisitions (from 19 patients) for testing. The severity distribution is displayed in Tab. 4.1.

GAMMA dataset

For the GAMMA dataset, we analyzed clinical data from 2D fundus images and 3D OCT scans to classify glaucoma into three groups based on visual features: no glaucoma, early glaucoma, and moderate or advanced glaucoma. Cohen's kappa described in Section 2.2.3 was also used to evaluate the EviRed dataset's six-category results as a standard evaluation metric for the multi-category classification task. Based on the confusion matrix, the Kappa coefficient is calculated with a value between -1 (worse than chance agreement) and 1 (perfect agreement).

There are 50 pairs of no-glaucoma patients in the training set, 26 pairs of early glaucoma patients, and 24 pairs of moderate or advanced glaucoma patients in the training set. These pairs were divided as follows: 80 pairs for training (41 pairs no glaucoma, 21 pairs early glaucoma, 18 pairs moderate or advanced glaucoma) and 20 pairs for validation (9 pairs no glaucoma, 6 pairs early glaucoma, 5 pairs moderate or advanced glaucoma).

4.2.6 Data pre-processing

The original 2D and 3D images were too large to train a fusion network. To reduce the volume under consideration, we intercepted the OCTA image located between the internal limiting membrane (ILM) and retinal pigment epithelium (RPE) layers along the depth (y-axis) and flattened the ILM layer. The following dimensions were used: $X = Y = Z = 100$ for the EviRed retrospective dataset, $X = 224$, $Y = 164$, and $Z = 256$ for the GAMMA dataset. For single-level and hierarchical fusion, 2D images could be larger than 3D images: they were resized to 400 $\times$ 400 pixels for the EviRed retrospective dataset and 448 $\times$ 448 pixels for the GAMMA dataset. Note that 2D and 3D data are not spatially registered in GAMMA;

they are only approximately centered on the same anatomical structure (the optic nerve head). All modalities are natively registered in the EviRed retrospective dataset.

4.2.7 Implementation details

Experiments were performed using 2D and 3D versions of ResNet [361], and DenseNet [362]. These networks were used as is or adapted for each fusion strategy. RandomGamma, GaussianNoise, and flipping were applied for all tests to augment the data. Gradient descent was performed with the Adam optimizer, which has an initial learning rate of 1e-4 and a weight decay rate of 1e-4. The network training and testing were carried out using one NVIDIA Titan GPU unit with 32 GB memory.

4.3 Results

4.3.1 EviRed retrospective dataset

For the EviRed retrospective dataset, the following backbones were investigated for each method: ResNet50, ResNet101, DenseNet121, and DenseNet169. Tab. 4.2 shows the fusion results for different backbones.

According to the results in Tab. 4.2, task0, task1, and task2 perform averagely and are very unstable. The model is not robust on diagnostic tasks that are pathologically similar due to the small amount of data we have. In comparison, the model performs well and is stable on task 3. As a result, we focused on the classification of PDR in the current dataset. Moreover, it plays a crucial role in clinical diagnosis. The treatment of diabetic retinopathy is of prime importance for patients at the PDR stage. For the PDR classification, we summarized the above results in Tab. 4.3,

Table 4.2 – Backbones tests results with different fusion methods on the test set.

Backbone	Fusion methods	AUC0	AUC1	AUC2	AUC3
Resnet50	Structure (Unimodal)	0.434	0.500	0.628	0.815
	Flow (Unimodal)	0.505	0.573	0.638	0.705
	LSO (Unimodal)	0.592	0.493	0.451	0.650
	Input fusion	0.493	0.526	0.614	0.753
	Single-level fusion	0.485	0.503	0.632	0.739
	Hierarchical fusion	0.587	0.547	0.549	0.812
Resnet101	Structure (Unimodal)	0.546	0.612	0.625	0.859
	Flow (Unimodal)	0.665	0.530	0.579	0.650
	LSO (Unimodal)	0.582	0.370	0.408	0.568
	Input fusion	0.534	0.507	0.605	0.709
	Single-level fusion	0.740	0.497	0.493	0.726
	Hierarchical fusion	0.546	0.583	0.691	0.846
Densenet121	Structure (Unimodal)	0.434	0.513	0.648	0.774
	Flow (Unimodal)	0.469	0.747	0.753	0.585
	LSO (Unimodal)	0.684	0.480	0.546	0.662
	Input fusion	0.584	0.527	0.642	0.865
	Single-level fusion	0.536	0.430	0.671	0.744
	Hierarchical fusion	0.434	0.530	0.668	0.911
Densenet169	Structure (Unimodal)	0.505	0.467	0.523	0.620
	Flow (Unimodal)	0.607	0.643	0.618	0.816
	LSO (Unimodal)	0.689	0.500	0.395	0.628
	Input fusion	0.645	0.542	0.575	0.693
	Single-level fusion	0.709	0.523	0.566	0.726
	Hierarchical fusion	0.531	0.477	0.493	0.679

Table 4.3 – Results of different fusion methods on the test set.

Method	Backbone	AUC	Sensitivity	Specificity	Improvement
Single modality (Structure)	ResNet101	0.859	0.78	0.77	Baseline
Single modality (Flow)	DenseNet169	0.816	0.78	0.85	-0.043
Single modality (LSO)	DenseNet121	0.662	0.67	0.74	-0.197
Hierarchical fusion	DenseNet121	**0.911**	**0.86**	**0.88**	**+0.052**
Input Fusion	DenseNet121	0.865	0.78	0.85	+0.006
Single-level fusion	DenseNet121	0.744	0.67	0.85	-0.115

and its corresponding ROC curve is shown in Fig. 4.7.

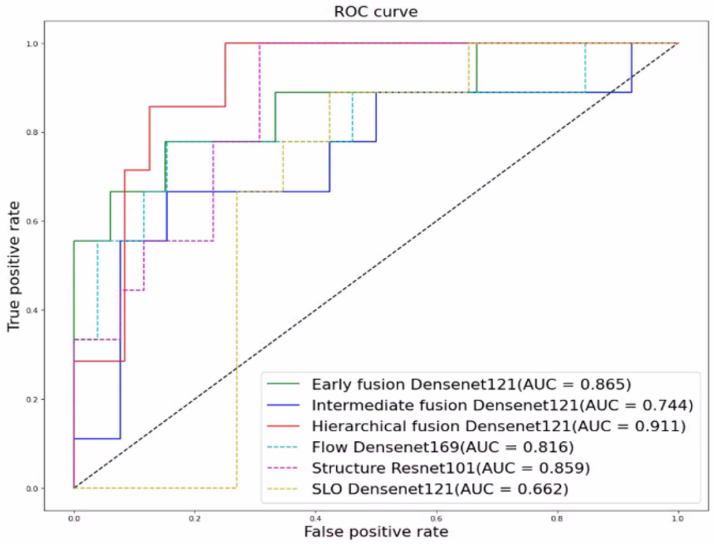


Figure 4.7 – ROC curves of different fusion methods on the test set.

The structure data achieved the best performance using a single modality: AUC reaches 0.859 using ResNet101 (our baseline). Single-level fusion performed worse than baseline. The three modalities are spatially aligned in the EviRed retrospective dataset, so the input fusion approach achieves good results. Hierarchical fusion achieves the best results: AUC reaches 0.911 using DenseNet121. The LSO images do not provide very distinct pathological details, compared to fundus images, hence a more limited impact of information fusion.

Table 4.4 – Kappa results of different fusion methods on the GAMMA dataset

Backbone	Single modality (fundus image)	Single modality (OCT)	Input fusion	Single-level fusion	Hierarchical fusion
ResNet34	0.6997	0.6841	0.6718	0.7547	0.7684
ResNet50	0.6555	0.5952	0.6896	0.7690	**0.8404**
ResNet101	0.6767	0.5794	0.7113	0.7551	0.8255
ResNet152	0.5207	0.4646	0.4642	0.6570	0.7816
Average	0.6382	0.5808	0.6342	0.7340	0.8040

4.3.2 GAMMA dataset

We tested the performance of four ResNet networks on the same dataset: ResNet34, ResNet50, ResNet101, and ResNet152. The best-performing models were selected from the validation set and tested on the 100 pairs test set. The final Kappa results on the test set were computed independently by the PaddlePaddle deep learning platform [1], which is the host platform for the GAMMA challenge.

We tested each modality separately, as well as the three fusion methods, and the results are shown in Table 4.4.

The Kappa results above show that color fundus images outperform OCT volumes when using data from a single modality. In addition, ResNet34 has better performance, possibly because simple features of a single modality are easy to learn. Although, according to the average of different backbones, 0.6382 is still far from a result that can be useful for diagnosis. Thus, single-modality glaucoma classification is very ineffective.

Results for the input fusion were not significantly improved. This probably is because fundus and OCT images are not spatially registered in this dataset.

Single-level fusion is a more suitable fusion algorithm in this case be-

1. `https://aistudio.baidu.com/aistudio/competition/detail/119/0/introduction`

cause of the disparity between fundus images and OCT volumes, and the dual feature extraction branch can effectively handle the large differences between modalities. As a result, the performance of single-level fusion is greatly improved compared to the single-modality scenario. In addition, for ResNet152, we had to reduce the batch size during training to prevent the device from exceeding the memory limit, which is one reason for the poor performance of ResNet152.

The GAMMA challenge also uses single-level fusion as its baseline [335]. In the official baseline, two convolutional branches are used for single-level fusion. Based on 3D OCT, retinal thickness is used as a channel for the input of the 2D convolutional branch in the algorithm. By contrast, we utilize 3D convolutional branches to extract 3D OCT features, which allows us to fully utilize the spatial features of 3D data. This is why our Kappa value of 0.734 for single-level fusion is higher than the official single-level fusion result of 0.702.

Comparatively to single-level fusion, hierarchical fusion is able to exploit better correlations between features of different dimensions: the Kappa value increased by 0.0700. These results support the efficiency of our hierarchical fusion.

Specifically, our hierarchical fusion performs very well on ResNet50 and ResNet101. To achieve a higher score in the GAMMA challenge, we selected the models ResNet50 and ResNet101 for further training. The training and validation sets were re-divided, and the checkpoint obtained from the previous test was fine-tuned. Finally, we achieved a Kappa value of 0.8662 for ResNet50 and 0.8745 for ResNet101. For our hierarchical fusion, we improved the final Kappa to 0.8996 by ensembling the predicted values of ResNet50 and ResNet101 models.

Table 4.5 – Results hierarchical fusion with different attention blocks on the test set.

Dataset	Metric	Without block	Channel attention block	Dual attention fusion block
EviRed retrospective	AUC	0.91	0.87	0.90
	Sensitivity	0.86	0.79	0.84
	Specificity	0.88	0.83	0.85
GAMMA	Kappa	0.8404	0.7942	0.8159

4.3.3 Attention mechanism

We incorporated two attention mechanism blocks into the hierarchical fusion architecture and compared their performance, and the results are shown in Tab. 4.5. Dual attention fusion blocks outperformed channel attention blocks, but neither was as effective as without attention blocks. Our attention modules do not work primarily because we fused 2D and 3D data, making it difficult to calibrate the weights based on the importance of two-dimensional features. Furthermore, before using the attention modules, we used the additional conversion convolutional layers to downscale the features from the 3D data in the Z-axis direction, which negatively impacts the performance of the channel attention blocks and the modality attention blocks.

4.4 Discussion and conclusions

This chapter presented three fusion strategies based on deep learning: input fusion, single-level fusion, and hierarchical fusion. The commonly used input and single-level fusion are simple but do not fully exploit the complementary information between modalities. We developed the hierarchical fusion approach that focuses on combining features across multiple dimensions of the network, as well as exploring the correlation between

modalities. Our hierarchical fusion method performed the best in different tasks and paved the way for better clinical diagnosis. The results of these tests not only validate the effectiveness of our fusion method discussed in Sect. 2, but also provide significant theoretical and experimental evidence for the subsequent fusion of more modalities in the EviRed dataset. On glaucoma and diabetic retinopathy classification tasks, they clearly outperform classification using a single modality. The novel hierarchical fusion approach is particularly promising, both for glaucoma grading and proliferative DR detection.

However, these experiments should be replicated in larger datasets to demonstrate clinically useful detection performance. It is of great interest for clinical diagnosis to be able to diagnose pathology at different stages. However, the performance of the existing dataset is poor. The further dataset of EviRed has a sufficient number of patient records to assist us in the task of multi-classification. In Chapter 5, we added the arrived EviRed prospective dataset (103 patients without annotations) to the current dataset for further multimodal fusion testing.

One of the factors that may limit the performance of the model is the current input size. Compressing the raw data to fit the inputs into the network is not particularly suitable. The compression of images results in a loss of much pathology information, which in turn affects diagnostic results. The next step in our research should involve the development of better image-processing methods. Further, the use of GPU technology to increase the input size is also a promising direction to pursue in addition to hardware improvements.

Furthermore, we should continue to focus on attention mechanisms. Despite the fact that our proposed attention module performs poorly in fusion networks using 2D and 3D data, there are some transformer-based models

[121] currently emerging that have shown good performance for multimodal fusion tasks using data of different dimensions [363–365]. Our multimodal fusion task may benefit from these models.

Chapter 5

Unlabeled data exploration

"In a world of diminishing mystery, the unknown persists."

— *Jhumpa Lahiri*

During the early stages of the second year of my thesis, we received the EviRed prospective dataset (First stage: 103 patients without annotations) from hospitals that ophthalmologists had not yet annotated. We explored these unlabeled data while we awaited the annotation. Two methods were tested in order to improve diagnostic performance using these unlabeled data:

1. Self-Supervised Learning: Pretext task
2. Semi-Supervised Learning: FixMatch

5.1 Introduction

As two fundamental pillars of machine learning, supervised and unsupervised learning offer distinct insights and capacities for extracting information from data [366].

Supervised learning involves training a model on a labeled dataset in which each input is paired with its corresponding output [367]. This approach requires a large amount of labeled data for training and is well-suited to tasks such as classification, segmentation, and regression. The unsupervised learning process involves training a model on unlabeled data without explicitly labeling the outputs. Through this method, meaningful relationships are discovered between data without the need for specific guidance [368].

In spite of the fact that both supervised and unsupervised learning are powerful in their own right, they are not mutually exclusive. By incorporating both labeled and unlabeled data, semi-supervised learning bridges the gap between these paradigms [369]. In addition to combining insights from both supervised and unsupervised approaches, semi-supervised learning uses the advantages of labeled data and the larger, more readily available pool of unlabeled data [370]. Models can thus generalize more effectively, particularly in situations where labeled data are not available or are impractical [371]. For example, the Consistency Regularization algorithm is a technique used in semi-supervised learning to improve model performance by encouraging consistency between predictions made on augmented versions of unlabeled data [372].

Self-supervised learning also is a middle ground between supervised learning and unsupervised learning. By designing tasks where the model generates its own supervisory signals from the input data, this approach creates an effective "self-created supervision" [373]. This approach takes advantage of the intrinsic relationships within the data, presenting a middle ground between explicit supervision and autonomous discovery [374]. As an example, when predicting the missing areas of an image (contextual prediction), the model learns meaningful representations based on the relationship between different parts of the input [375]. It is possible to predict missing frames in a video using context prediction.

There were two datasets available: the EviRed retrospective dataset and the EviRed prospective dataset (First stage: 103 patients without annotations). Self-supervised learning and semi-supervised learning were employed to enable the exploration of the unlabeled data. Self-supervised learning was achieved by using a pretext task to generate a pre-trained model based on the EviRed prospective dataset and then evaluating the downstream classification task based on the EviRed retrospective dataset. As part of the semi-supervised learning process, we utilized the FixMatch [376] approach to train with both datasets simultaneously. We expected that these two approaches could fully utilize the unlabeled data, which would improve the diagnostic performance of the models.

5.2 Material and methods

5.2.1 Pretext task for self-supervised learning

In general, self-supervised learning pipelines involve the execution of two tasks: a pretext task and a downstream task. Downstream tasks utilize the

knowledge acquired during the pretext task to perform application-specific tasks. Fig. 5.1 illustrates how knowledge is transferred from the pretext task to the downstream task. By fine-tuning the parameters, the learned parameters serve as a pretrained model for transferring to other downstream computer vision tasks [377]. The performance of transfer learning on these high-level vision tasks demonstrates the generalization ability of the learned features [373].

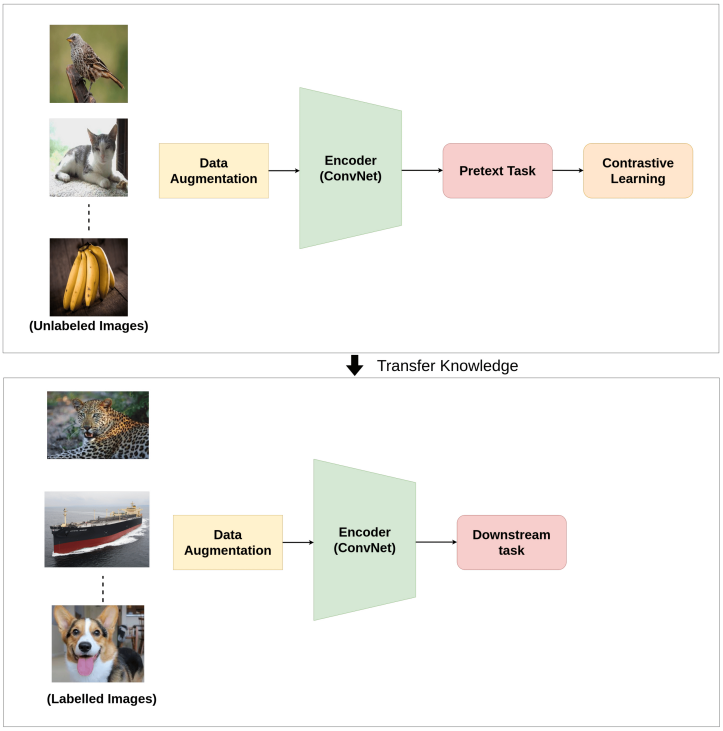


Figure 5.1 – An overview of downstream task for images [373].

In order to verify the effectiveness of the pretext task in multimodal fusion networks, we first designed the pretext and the downstream tasks, as shown in Fig. 5.2. Our pretext task for analyzing the unlabeled data in the EviRed prospective dataset consists of determining if the input data for the Structure and Flow modalities are from the same patient. Contrastive loss was applied to features derived from different branches in the single-

level fusion network. It takes as input a pair of features that are either similar or dissimilar, and it brings similar features closer and dissimilar features far apart. The contrastive loss is defined as:

$$L = (1 - Y) * \| x_i - x_j \|^2 + Y * max(0, m - \| x_i - x_j \|^2)$$

Where m is a hyperparameter, defining the lower bound distance between dissimilar features. x_i is the feature of Structure and x_j is the feature of Flow. Label Y that is equal to 0 if the Structure and Flow modalities are from the same patient and 1 otherwise.

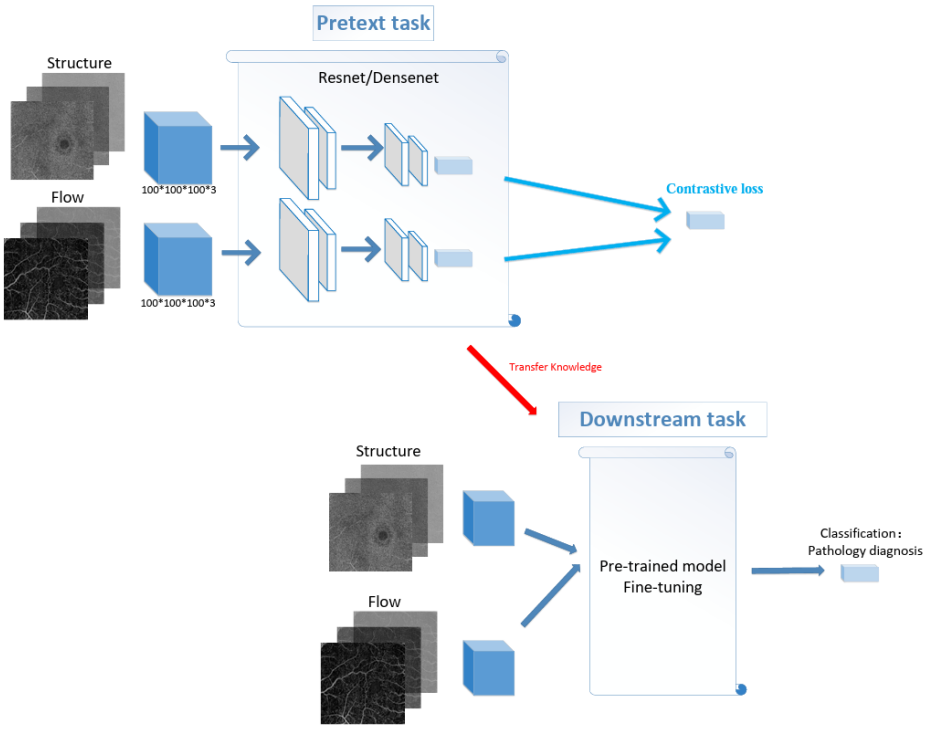


Figure 5.2 – Proposed pretext and downstream tasks.

The downstream task is the same as the diagnosis task in the last chapter for the classification of PDR. To assess the validity of the pretext task, the encoder part of the single-level fusion model that has been trained with

the pretext task is used as a pre-trained model for the downstream task, while the single-level fusion model without the pretext task is used for baseline performance. If the pretext task for single-level fusion improves the downstream task's performance, the pretext task for more complex hierarchical fusion will be designed.

5.2.2 FixMatch: a semi-supervised learning algorithm

FixMatch is a semi-supervised method that uses consistency regularization [378] and pseudo-labeling [379] to enhance model performance by training on both labeled and unlabeled data, which is particularly beneficial for tasks that do not have sufficient labeled examples [376].

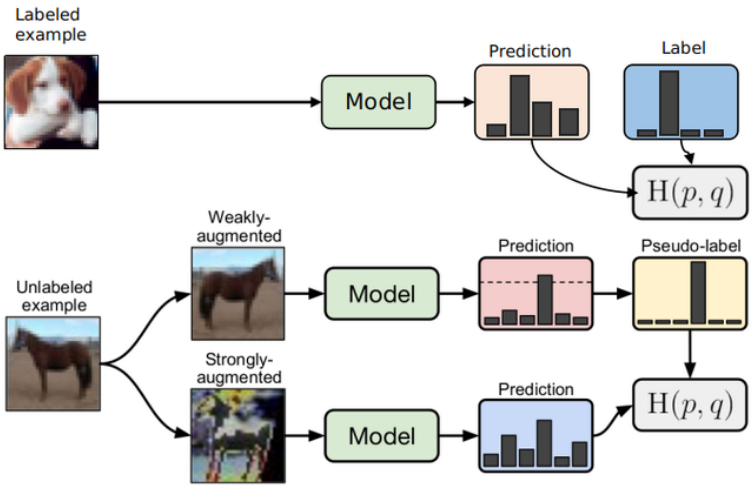


Figure 5.3 – Diagram of FixMatch. For the unsupervised part, it is first necessary to feed a weakly augmented version of an unlabeled image (top) into the model in order to obtain its predictions (red box). It is converted to a one-hot pseudo-label when the model assigns a probability to a class above the threshold (dotted line). In the next step, we compute the model's prediction for a version of the image that has been strongly augmented (bottom). Through a standard cross-entropy loss, the model is trained to make predictions for the strongly augmented version that matches the pseudo-label [376].

The training process of FixMatch is divided into supervised and un-

supervised parts, as shown in Fig. 5.3. The labeled example involves supervised training. The unlabeled example generates a pseudo-label if the weakly enhanced output exceeds a certain threshold. This pseudo-label is then used to supervise the strongly enhanced output values. The unsupervised training process involves consistency regularization and pseudo-labeling [376]:

— Consistency Regularization: Unlabeled data is augmented multiple times, generating different versions of each data point. The model's predictions on these augmented versions are encouraged to be consistent with each other, promoting stable and reliable predictions across variations of the same instance [380, 381].

— Pseudo-Labeling: The most confident predictions made by the model on the augmented unlabeled data are treated as pseudo-labels. These pseudo-labeled examples are then included in the training process as if they were labeled data, allowing the model to learn from them and improve its predictions [382, 383].

FixMatch enhances model generalization by combining these two principles. Through a combination of supervised and unsupervised training, it gradually improves the classification performance of the model.

There are only two cross-entropy loss terms in FixMatch: a supervised loss ℓ_s applied to labeled data and an unsupervised loss ℓ_u. For a sample $b \in (1, ..., B)$ in a batch of B labeled examples, where x_b are the training examples and p_b are one-hot labels. Let $p_m(y \mid x)$ be the predicted class distribution produced by the model for input x. The cross-entropy between two probability distributions p and q is $H(p, q)$. Two types of augmentations are part of FixMatch: strong and weak, denoted by $\mathcal{A}(\cdot)$ and $\alpha(\cdot)$, respectively. In particular, ℓ_s is simply the cross-entropy loss on weakly augmented labeled examples:

$$\ell_s = \frac{1}{B} \sum_{b=1}^{B} H(p_b, p_m(y \mid \alpha(x_b)))$$

For a sample $b \in (1, ..., \mu B)$ in a batch of μB unlabeled examples, FixMatch computes an artificial label for each example of unlabeled data u_b, which is then used in a standard cross-entropy loss. In order to create an artificial label, the model's predicted class distribution is first computed given a weakly augmented version of an unlabeled image: $q_b = p_m(y \mid \alpha(u_b))$. Using $\hat{q}_b = argmax(q_b)$ as a pseudo-label, the cross-entropy loss is applied to the output of a strongly augmented version of u_b as follow:

$$l_u = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathbb{1}(max(q_b \geq \tau)) H(\hat{q}_b, p_m(y \mid \mathcal{A}(u_b)))$$

Where τ is a scalar hyperparameter denoting the threshold above which we retain a pseudo-label. In sum, the loss minimized by FixMatch is simply $\ell_s + \lambda_u \ell_u$ where λ_u is a fixed scalar hyperparameter denoting the relative weight of the unlabeled loss.

5.2.3 Dataset

After removing patient data with incomplete modalities and absent annotations from the EviRed prospective dataset (First stage: 103 patients without annotations) described in Section 3.3.4, 170 acquisitions from 71 diabetic patients were collected for unlabeled exploration. In addition, 151 acquisitions (64 patients) with annotations from the EviRed retrospective dataset were added to the dataset. Tab. 5.1 is information about the data used for self-supervised and semi-supervised learning:

Table 5.1 – Datasets used for self-supervised and semi-supervised learning.

Method	Datasets	Patient	Eyes
Pretext task	EviRed prospective dataset (First stage)	71	170
Downstream task	EviRed retrospective dataset	64	151
FixMatch	EviRed retrospective dataset + EviRed prospective dataset (First stage)	64+71=135	151+170= 321

For the pretext task, we randomly selected 136 acquisitions (80% of the unlabeled prospective dataset) as the training set and the remaining 34 acquisitions (20%) as the validation set. For the downstream task, we used the same training, validation, and test sets as in Section 4.2.5 (from the labeled retrospective dataset). For FixMatch, we treat all data from the EviRed prospective dataset (First stage) as unlabeled data in the training set, and the validation and test sets remain the same as before.

5.2.4 Implementation details

The data was processed in the same way described in Section 4.3.2. Self-supervised learning is performed using a single-level fusion model of 3D DenseNet121. The hyperparameter $m = 1.0$ is used for the pretext task. To augment the data, RandomGamma, GaussianNoise, and flipping were applied for the pretext task and downstream task. Gradient descent was performed with the Adam optimizer, which has an initial learning rate of 1e-4 and a weight decay rate of 1e-4.

In the paper of FixMatch, weak augmentation is a standard flip-and-shift augmentation strategy. Specifically, they randomly flip images horizontally with a probability of 50% and translate images vertically and hor-

Table 5.2 – 3D data strong augmentation pool.

Operator	Parameters	Probability
Flip	horizontal, vertical	0.5
Rotate	x_limit = (-15, 15)	0.5
ElasticTransform	deformation_limits = (0, 0.25), interpolation = 2	0.5
RandomRotate90	axes = (1, 2)	0.5
GaussianNoise	var_limit = (0, 5)	0.5
RandomGamma	gamma_limit = (0.5, 1.5)	0.5
GridDropout	ratio = 0.5,unit_size_min = 50, unit_size_max = 60, holes_number_x = 3, holes_number_y = 2 ,holes_number_z = 2	0.5
CutoutAbs	ratio=0.5	1.0

izontally by up to 12.5%. They tested two approaches based on AutoAugment [384] for strong augmentation: RandAugment [385] and CTAugment [386]. However, the 3D structural and flow modalities in our fusion model require the reconstruction of an augmentation strategy. With the help of the Volumentations 3D library [387], we constructed a strong enhancement pool as in Tab. 5.2. Three data augmentation operations are randomly selected from the pool for data augmentation when a strong augmentation operation is performed on unlabeled data.

FixMatch is performed using a hierarchical fusion model of 3D DenseNet121. The hyperparameters τ, μ, and λ_u are 0.8, 7.0, and 1.0, respectively. The other training configurations are the same as for self-supervised learning. The network training and testing were carried out on the OVH cluster using one NVIDIA Tesla V100S units with 32 GB memory.

5.3 Results

5.3.1 Pretext task

In order to visually verify our pretext task, we concatenated output features from different branches of the single-level fusion model and visualized them using t-SNE [388], as in Fig. 5.4. A certain distance exists between the

distributions of purple dots (Structure and Flow from the same patients) and yellow dots (Structure and Flow from different patients). The model can achieve an accuracy of 0.92 on the validation set of the pretext task. The model fulfills the task goal excellently and can explicitly distinguish whether the input Structure and Flow are from the same patient.

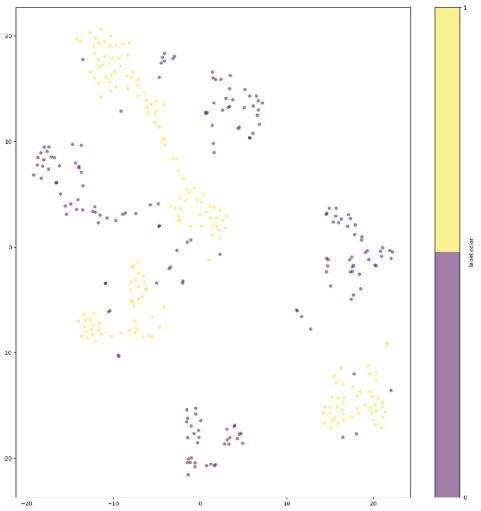


Figure 5.4 – The image of the pretext task using t-SNE visualization features. Purple dots (Class 0) represent features from the same patient for Structure and Flow, and yellow dots (Class 1) represent features from a different patient.

After that, we tested the performance of single-level fusion by using the model of the pretext task as a pre-trained model. Despite the good performance of the pretext task, Tab. 5.3 indicates that there is no significant improvement in the downstream task. The pretrained model for the pretext task improves AUC and Specificity slightly, but Sensitivity does not perform as well as the baseline model. In general, the pretrained model of the pretext task performs similarly to the baseline model.

Table 5.3 – Performance of different versions of single-level fusion on the test set for PDR classification.

Method	Pretext task	AUC	Sensitivity	Specificity
Single-level fusion	✗	0.735	0.67	0.83
Single-level fusion	✓	0.737	0.66	0.85

Table 5.4 – Performance of different versions of hierarchical fusion on the test set for PDR classification.

Method	FixMatch	AUC	Sensitivity	Specificity
Hierarchical fusion	✗	0.911	0.86	0.88
Hierarchical fusion	✓	0.907	0.85	0.88

5.3.2 FixMatch

The model performance of FixMatch was compared to baseline supervised learning as shown in Tab. 5.4. Unfortunately, FixMatch did not enhance the diagnostic performance of the model. This is primarily due to the fact that our unlabeled dataset is not large enough. The amount of data for unlabeled images in the FixMatch paper is 20 times that for labeled images. For us, however, the amount of unlabeled images is comparable to the amount of labeled images, which undoubtedly affects the performance of FixMatch.

5.4 Discussion and conclusions

In this chapter, we explored the unlabeled EviRed data through self-supervised and semi-supervised learning. Due to the limited number of data patients, these two methods did not significantly improve the performance of the model. Unfortunately, the pretext task and FixMatch used

did not improve the diagnostic performance of the fusion model on the PDR classification task.

Insufficient unlabeled data was the primary cause of the poor performance. The EviRed prospective dataset (First stage) we used is comparable to the EviRed retrospective dataset, and exploring the unlabeled data is not very beneficial for the fusion model. Furthermore, many of the hyperparameters in FixMatch may require further testing.

Fortunately, as the project progressed, we received more data from the EviRed prospective dataset. And, unlike the previous release of data, all subsequent releases associate data with annotations (Second stage: 532 patients with annotations). Considering the poor performance of semi-supervised and self-supervised learning and the fact that all images are now annotated, we have decided to continue using supervised learning for DR classification. In Chapters 6 and 7, we used the Second stage dataset for the exploration of multimodal fusion methods with supervised learning. Even so, the exploratory process for unlabeled data is informative for the future development of the EviRed project.

Chapter 6

Hybrid fusion of high-resolution and ultra-widefield OCTA acquisitions

---

> "The coming era of Artificial Intelligence will not be the era of war, but be the era of deep compassion, non-violence, and love."
>
> — *Amit Ray*

The second stage of the EviRed prospective dataset containing 532 patients and annotations was received in the third year of my thesis. Diabetic patients were examined with 6×6 mm^2 high-resolution OCTA and 15×15 mm^2 UWF-OCTA using PLEX®Elite 9000. The OCTA acquisition specifications used for testing in Chapter 4 were mixed. However, due to the small number of patients, the model performs poorly on classification tasks of severity levels. We were not able to study the effect of different types of OCTA acquisitions on the classification of severity levels in Chapter 4, but the newly arrived dataset for phase *"EviRed prospective dataset"* (Second stage) offers this possibility. This chapter evaluated a deep learning (DL) algorithm for automatic DR severity assessment using high-resolution and ultra-widefield (UWF) OCTA. A novel DL algorithm was trained for automatic DR severity inference using both OCTA acquisitions.

6.1 Introduction

In this chapter, we used high-resolution 6×6 mm^2 SS-OCTA and 15×15 mm^2 UWF-SS-OCTA images obtained from a PLEX® Elite 9000 (Carl Zeiss Meditec Inc., Dublin, CA, USA) for the diagnosis of DR. Each OCTA image encompassed both structural (Structure) and flow (Flow) information.

The 6×6mm^2 high-resolution SS-OCTA provides superior visualization of the capillary network and the central avascular zone [389]. Consequently, it enables the calculation of metrics such as vascular density (the ratio of vessel area with respect to the total area) [390–392], fractal dimensions [393], and intercapillary spaces [394]. However, its limitation lies in its focus on the macular region, potentially neglecting global retinal damage.

On the other hand, the 15×15 mm^2 UWF-SS-OCTA provides a more extensive view of the retina, allowing the detection of relevant abnormalities, such as the presence of an intraretinal microvascular abnormality (IRMA) or a preretinal vascular anomaly (neovessel) [46, 49, 395]. Furthermore, the absence of capillary networks on important surfaces in the 15×15 mm^2 image can be easily observed [396], which are considered an important biomarker of proliferating diabetic retinopathy [397, 398].

Overall, 6×6 mm^2 SS-OCTA allows an accurate calculation of certain vascular metrics and an analysis of the central avascular zone. Still, it only explores a small part of the retina, while 15×15 mm^2 SS-OCTA allows a broader investigation of vascular anomalies and areas of non-perfusion. The two specifications complement each other quite well in clinical practice.

This chapter presents an innovative approach to improve the accuracy of DR diagnosis by leveraging the complementary information provided by 6×6 mm^2 and 15×15 mm^2 SS-OCTA images. We meticulously investigated the use of the information from each acquisition and tested the performance of the fusion on structural and flow information for OCTA. Our proposed hybrid fusion network utilizes the structural and flow information of each acquisition and fusing images from both acquisitions to enhance DR diagnostic performance significantly. As the first study exploring the fusion of different OCTA acquisitions using deep learning methods, this work paves the way for future diagnosis applications from OCTA images.

6.2 Material and methods

6.2.1 Hybrid fusion workflow

This study aimed to find the best hybrid fusion network structure for the fusion of 6×6 mm^2 SS-OCTA data with 15×15 mm^2 SS-OCTA data. To achieve this, we organized the workflow into the following four stages:

(1) Data processing. The first step involved exploring a variety of approaches to process the raw data from different acquisitions and adapt it to the input specifications of the CNN network.

(2) Backbones. Subsequently, we investigated the most effective backbone for the Structure and Flow separately for both acquisitions of OCTA data.

(3) Fusion of Structure and Flow. After selecting the most effective backbone from three deep learning architectures for each modality, we evaluated four different fusion strategies—input fusion, single-level fusion, output fusion (leveraging averaging strategies), and hierarchical fusion using Structure and Flow.

(4) Fusion of 6×6 mm^2 and 15×15 mm^2 acquisitions. Based on the best optimal fusion structure for each acquisition, we assessed two strategies, namely single-level fusion and output fusion, on information derived from both 6×6 mm^2 and 15×15 mm^2 SS-OCTA acquisitions:

— For the single-level fusion strategy, we utilized the model parameters obtained in the previous fusion step and conducted two types of fine-tuning—(a) fine-tuning the entire network (network fine-tuning), and (b) freezing all convolutional layers and fine-tuning the classification layer (layer fine-tuning).

— For the output fusion strategy, we implemented and tested both averaging (Avg) and maximization (Max) strategies.

This comprehensive process led us to a hybrid fusion network structure that facilitates the fusion of single-acquisition multimodal information with multiple-acquisition information. This hybrid fusion structure maximized the diagnosis performance of DR by integrating the complementary information from both 6×6 mm^2 and 15×15 mm^2 SS-OCTA acquisitions. This method fully leveraged the structural and flow information derived from each acquisition, thus optimizing our diagnosis process. Figure 6.1 illustrates the workflow of this study.

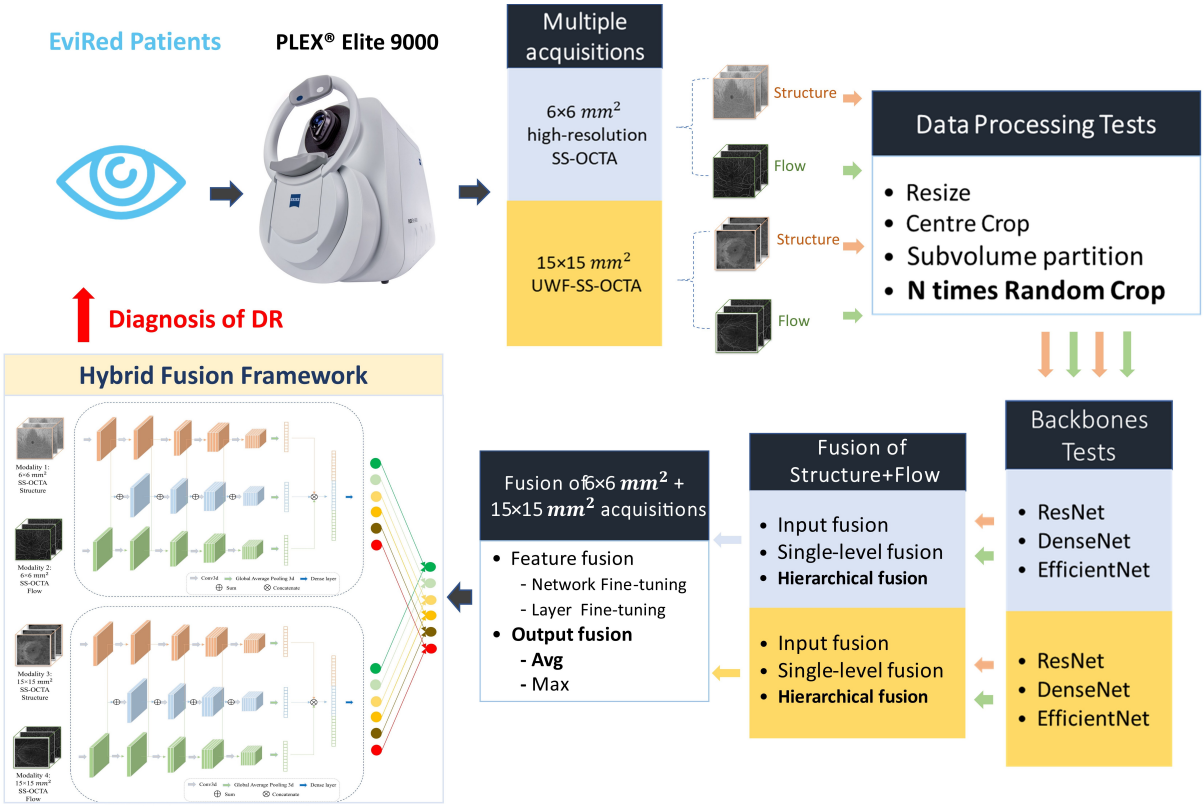


Figure 6.1 – Proposed workflow.

6.2.2 Data processing

EviRed prospective dataset

In the second stage of the EviRed prospective dataset described in Section 3.3.4, OCTA data were gathered for 875 eyes from a total of 444 patients after removing patient data without OCTA modalities and absent annotations. This substantial dataset was used to train and test our deep learning models.

Following the EviRed study protocol, each patient's ocular data often contained two specifications of acquisitions: 6×6 mm^2 high-resolution SS-OCTA and 15×15 mm^2 UWF-SS-OCTA with a wavelength of 1060 nanometers. Figure 3.6 shows en-face images and their corresponding B-scan images (pre-processed) of the Structure and Flow from the same patient acquired for different specifications.

The EviRed raw data size was $500 \times 1536 \times 500 \times 2$ voxels for the 6×6 mm^2 SS-OCTA and $834 \times 3072 \times 834 \times 2$ voxels for the 15×15 mm^2 SS-OCTA. The last channel presented the information of Structure and Flow, respectively. To reduce the volume under consideration, we used the same preprocessing as in Section 4.2.6. The EviRed raw data were resized to dimensions of $500 \times 224 \times 500 \times 2$ voxels for the 6×6 mm^2 SS-OCTA and $834 \times 224 \times 834 \times 2$ voxels for the 15×15 mm^2 SS-OCTA. Figure 3.6(a,c) illustrates the orientation of each dimension. For the 6×6 mm^2 SS-OCTA, the en-face images had a size of 500×500 pixels, and the B-scan images were 500×224 pixels. The 15×15 mm^2 SS-OCTA had en-face images and B-scan images of sizes 834×834 pixels and 834×224 pixels, respectively.

OCTA Cropping

Due to graphics processing unit (GPU) hardware limitations (NVIDIA Tesla V100S with 32 GB memory), our 3D deep learning backbones could only accommodate inputs up to $224 \times 224 \times 224 \times 2$ voxels. The patch extraction method is commonly used to address hardware limitations in 3D medical images [142, 243]. Nevertheless, it is difficult to ensure that each patch contains pathology information. Based on the idea of test time augmentation [399], the model synthesized and analyzed multiple predictions in order to avoid making inaccurate predictions. As a result, a global prediction of multiple patches was an effective method under the limitations of our hardware. In this context, we proposed a strategy, named N times Random Crop method, for processing images as shown in Figure 6.2. We compared our proposed method with other commonly used methods of data processing. For this comparison, we used the input fusion of ResNet [361] with the 15×15 mm^2 OCTA in order to verify its effectiveness. The following methods were tested for prediction:

(1) N times Random Crop (proposed). During the training of the deep learning network, Random Crop processing was employed, while in the prediction process, we utilized multiple volumes extracted from the OCTA image (N times Random Crop) simultaneously to make predictions. Considering that the patch size was $224 \times 224 \times 224 \times 2$ voxels, it would take at least 9 batches ($\lceil \frac{500}{224} \rceil \times \lceil \frac{224}{224} \rceil \times \lceil \frac{500}{224} \rceil \times \lceil \frac{2}{2} \rceil$) to traverse the $500 \times 224 \times 500 \times 2$ voxel 6×6 mm^2 SS-OCTA images, while 16 batches ($\lceil \frac{834}{224} \rceil \times \lceil \frac{224}{224} \rceil \times \lceil \frac{834}{224} \rceil \times \lceil \frac{2}{2} \rceil$) would be required to traverse the $834 \times 224 \times 834 \times 2$ voxel 15×15 mm^2 SS-OCTA images. By comparing the performance of the ResNet input fusion model on the validation set with different N times Random Crop methods, we

determined the N values for the two SS-OCTA acquisitions. The final prediction for an OCTA image was based on the severest prediction among these N predictions.

(2) Resize. This method compressed the original volume of $834 \times 224 \times 834 \times 2$ voxels into $224 \times 224 \times 224 \times 2$ voxels for both training and prediction.

(3) Center Crop. This approach selected a random patch of $224 \times 224 \times 224 \times 2$ voxels from the original $834 \times 834 \times 834 \times 2$ voxel OCTA for training. For prediction, a central patch was selected.

(4) Subvolume Crop. This technique traversed the OCTA using a window, predicting all subvolumes of $224 \times 224 \times 224 \times 2$ voxels and determining the maximum value.

It is worth noting that for single-acquisition fusion, we ensured the registration of data across different modalities. However, when fusing data from different acquisitions, Random Crop generated data from varying regions. Having processed the data, our next step was to use these images to extract meaningful features and combine them for our classification task.

6.2.3 Multimodal information fusion

In this section, we describe three fusion network structures commonly used in multimodal research as shown in Figure 6.3: input fusion, single-level fusion, and output fusion [176]. Furthermore, we introduce hierarchical fusion, which is our extension of traditional feature fusion. Input, single-level, and hierarchical fusions are the same as described in Section 4.2.

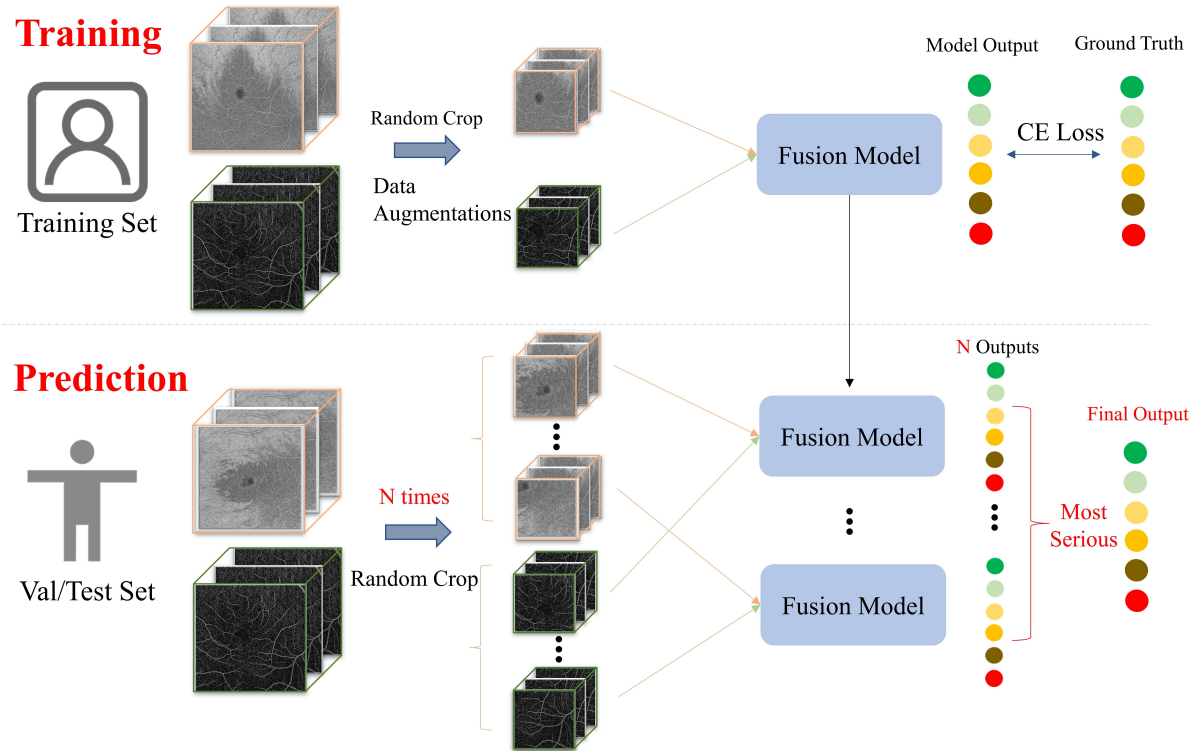


Figure 6.2 – Our proposed data processing approach, where N is 10 for 6×6 mm^2 SS-OCTA and 20 for 15×15 mm^2 SS-OCTA. Predictions were based on the same fusion model as for training. Colored discs indicate the DR severity categories.

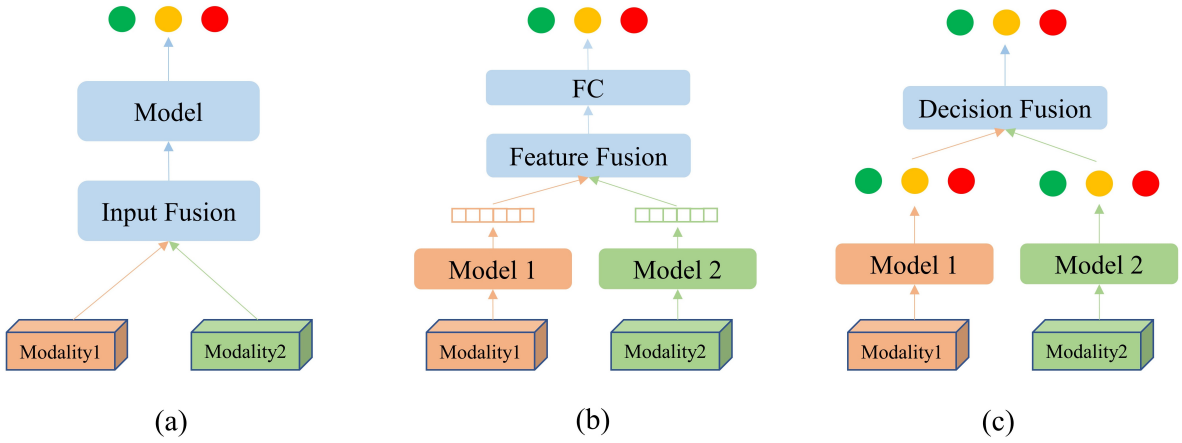


Figure 6.3 – An illustration of the three types of multimodal fusion networks: (**a**) input fusion, (**b**) single-level fusion, (**c**) output fusion.

Output Fusion

Output fusion involves extracting features and making decisions through separate deep learning backbones, and the results are combined into one final decision, as shown in Figure 6.3c. Many fusion strategies have been proposed for output fusion [400]. Most of them are based on averaging and majority voting [270, 271]. Due to the absence of single-level fusion, exploiting the complementary information between different modalities is difficult.

6.2.4 Classification tasks

DR severity was assessed by a retina specialist using fundus photographs, according to the International Clinical Diabetic Retinopathy Disease Severity Scale (ICDR): the absence of diabetic retinopathy, mild nonproliferative diabetic retinopathy (NPDR), moderate NPDR, severe NPDR, proliferative diabetic retinopathy (PDR), and panretinal photocoagulation (PRP). Compared to previous chapters, we now have access to more training data, which allows us to tackle a more ambitious challenge: classifying all DR severity levels, and not simply PDR. In addition to the six-category multiclass classification, we also performed four binary classification tasks: task0 (detecting mild NPDR or more), task1 (detecting moderate NPDR or more), task2 (detecting severe NPDR or more), and task3 (detecting PDR or PRP). To assess the performance of the four binary classifications, we used the area under the ROC curve (AUC) described in Section 2.2.3: AUC0 ($\geq$mild NPDR), AUC1 ($\geq$moderate NPDR), AUC2 ($\geq$severe NPDR) and AUC3 ($\geq$PDR). As a standard evaluation metric for the multicategory classification task, Cohen's kappa was also used to evaluate the EviRed dataset's six-category results. Based on the confusion matrix,

the Kappa coefficient described in Section 2.2.3 was calculated with a value between -1 (worse than chance agreement) and 1 (perfect agreement).

6.2.5 Dataset splitting

During the data acquisition process, there were instances when the collection of OCTA data for each patient's eyes could not be guaranteed due to factors such as operator errors or the patient's physical condition. Similarly, not all patients were able to provide both 6×6 mm^2 and 15×15 mm^2 SS-OCTA data. Despite these constraints, in order to make full use of the dataset to train different model frameworks for different acquisitions and to test the performance of the fusion model, we split the data as follows: Initially, we selected 53 patients out of the 444 in the EviRed dataset who had both 6×6 mm^2 and 15×15 mm^2 SS-OCTA data in each eye to form a test set. The remaining patients were shared out randomly between a training set and a validation set. Depending on the fusion task, subsets of the training and validation sets were used in each experiment—all 6×6 mm^2 acquisitions or all 15×15 mm^2 acquisitions for single-acquisition tasks, all matched pairs of 6×6 mm^2 and 15×15 mm^2 acquisitions for multiple-acquisition tasks. All fusion tests were trained and validated using five-fold cross-validation (four-fold training and one-fold validation), and performance scores were derived from the same test sets. In the training, validation, and test sets, the distribution of data was identical to the original distribution. The patient and eye data statistics for different fusion datasets are shown in Table 6.1, and the severity distribution is displayed in Table 6.2.

Table 6.1 – Statistics on the number of patients and eyes in the dataset. For the fusion of Structure+Flow for 6×6 mm^2 SS-OCTA, the fusion of Structure+Flow for 15×15 mm^2 SS-OCTA and the fusion of 6×6 mm^2 + 15×15 mm^2 SS-OCTA, the test sets are identical and fixed. Dataset6×6, Dataset15×15, and Dateset6×6+15×15 represent the corresponding training and validation sets.

Dataset Type	Patients	Eyes
Total (EviRed dataset)	444	875
Test set (for all fusion tests)	53	97
Dataset6×6 (for fusion of 6×6 mm^2 OCTA: Structure + Flow)	386	753
Dataset15×15 (for fusion of 15×15 mm^2 OCTA: Structure + Flow)	372	701
Dataset6×6+15×15 (for fusion of 6×6 mm^2 + 15×15 mm^2 OCTA)	364	676

Table 6.2 – Distribution of eyes with different levels of severity in different datasets.

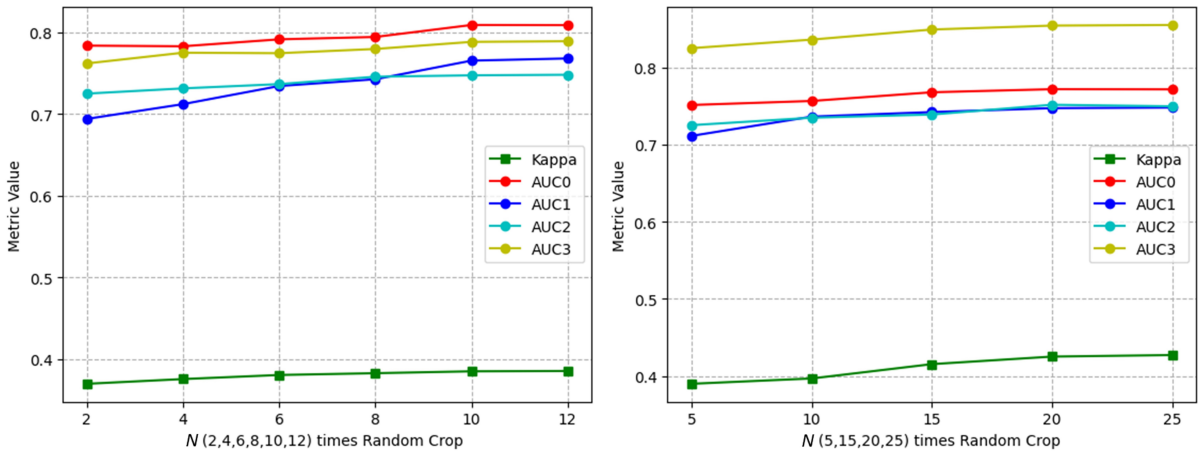
Severity	Dataset6×6	Dataset15×15	Dataset6×6+15×15	Test set
Absence of diabetic retinopathy	151	128	127	17
Mild NPDR	76	69	68	12
Moderate NPDR	348	334	321	39
Severe NPDR	111	107	97	18
PDR	20	20	20	3
PRP	47	43	43	8

6.2.6 Implementation details

The experiments were carried out with 3D versions of ResNet50 [361], DenseNet121 [362], and EfficientNetB0 [401] trained from scratch. Data augmentation techniques such as random Gamma transformations, Gaussian noise injection, and image flipping were employed to enhance the robustness of these models. For model training, we utilized the Adam optimizer for gradient descent with an initial learning rate of 1×10^{-4}. ExponentialLR with a gamma equal to 0.99 was the learning rate decay strategy. The number of training epochs was set to 500, and the batch size was set to 2. The network training and testing were carried out on the OVH cluster using four NVIDIA Tesla V100S units with 32 GB memory. For training large models such as the hierarchical fusion used in this experiment, model parallelism was used. The validation set was used to select the best backbones and the best checkpoint of each backbone. It was also used to select the best data cropping and information fusion strategies. However, for simplicity, performance is illustrated solely on the test set hereafter.

6.3 Results

Figure 6.4 shows the test results for different N times Random Crop methods. The performance of the fusion model on different metrics improved with an increase in N. For the 6×6 mm^2 SS-OCTA, the performance at $N = 10$ and $N = 12$ was comparable. For the 15×15 mm^2 SS-OCTA, the performance at $N = 20$ and $N = 25$ was essentially unchanged. As a result, we chose $N = 10$ for 6×6 mm^2 SS-OCTA and 20 for 15×15 mm^2 SS-OCTA as reasonable tradeoffs between computation times and classification scores.



(a) 6×6 mm^2 SS-OCTA **(b) 15×15 mm^2 SS-OCTA**

Figure 6.4 – The results of the different N times Random Crop methods on the validation set for the input fusion of ResNet with the two SS-OCTA acquisitions.

Table 6.3 – The results of the different data cropping methods on the test set for the input fusion of ResNet with the 15×15 mm^2 SS-OCTA.

Data Cropping Method	**Kappa**	**AUC0**	**AUC1**	**AUC2**	**AUC3**
Resize	0.2913	0.6485	0.6557	0.6836	0.7074
Center Crop	0.3270	0.7257	0.7059	0.6850	0.6903
Subvolume Crop	0.4048	0.7596	0.7429	0.7449	0.8340
N times Random Crop (proposed)	**0.4252**	**0.7721**	**0.7474**	**0.7519**	**0.8546**

Table 6.3 compares cropping methods: it shows that the Resize and Center Crop methods performed poorly due to a significant loss of information. As a result of the data compression of Resize, many pathological details were rendered invisible, while Center Crop focused only on the information obtained from the center patch. Although Subvolume Crop performed relatively well, manually extracting subvolumes may have omitted key pathological features, affecting the model's judgment. In the validation and test sets, our proposed data cropping method, namely Random Crop, outperformed the others in each classification task, demonstrating its effectiveness in handling the large original volume of OCTA images.

6.3.1 Backbones

A total of four modalities of information from Structure and Flow with different acquisitions were tested, along with three deep learning backbones: ResNet, DenseNet, and EfficientNet. Table 6.4 presents the results of the backbone tests. In the validation and test sets, ResNet demonstrated superior performance across all classification tasks for both the structure modality from the 6×6 mm^2 SS-OCTA images and the flow modality from the 15×15 mm^2 SS-OCTA images. The performance of the other backbones varied across the remaining modalities, and it was difficult to determine which backbone was the most effective. EfficientNet was effective for the multiclass classification as well as early pathology detection in the Flow from 6×6 mm^2 SS-OCTA images, while ResNet excelled in the more severe pathology detection tasks. Interestingly, DenseNet surpassed ResNet on task 0 when using Structure from 15×15 mm^2 SS-OCTA images. Based on these results, we selected the best-performing backbones (in bold in Table 6.4) for different tasks as baselines for the subsequent fusion schemes of Structure and Flow.

Table 6.4 – Backbone test results with different modalities on the test set.

Modality	Backbone	Kappa	AUC0	AUC1	AUC2	AUC3
6×6 mm^2 SS-OCTA—Structure	**ResNet**	**0.4150**	**0.8375**	**0.7659**	**0.7889**	**0.8104**
	DenseNet	0.3597	0.8285	0.7462	0.7368	0.7040
	EfficientNet	0.4149	0.8246	0.7521	0.7438	0.7788
6×6 mm^2 SS-OCTA—Flow	**ResNet**	0.3768	0.7931	0.7653	**0.7566**	**0.7863**
	DenseNet	0.3399	0.7972	0.7700	0.7525	0.7653
	EfficientNet	**0.4085**	**0.8306**	**0.7775**	0.7446	0.7150
15×15 mm^2 SS-OCTA—Structure	**ResNet**	**0.3900**	0.8118	**0.7604**	0.7462	0.8700
	DenseNet	0.3589	**0.8251**	0.7527	**0.7923**	**0.8732**
	EfficientNet	0.3230	0.8046	0.7407	0.7757	0.8671
15×15 mm^2 SS-OCTA—Flow	**ResNet**	**0.4189**	**0.7927**	**0.7627**	**0.7911**	**0.8774**
	DenseNet	0.3261	0.7770	0.7517	0.7788	0.8125
	EfficientNet	0.3259	0.7848	0.7557	0.7545	0.8397

6.3.2 Fusion of Structure and Flow

We combined Structure and Flow from different acquisitions using the top-performing backbones from the previous section. We tested input fusion, single-level fusion, and hierarchical fusion. Tables 6.5 and 6.6 show the fusion results for 6×6 mm^2 and 15×15 mm^2 SS-OCTA acquisitions, respectively.

Table 6.5 – Results of Structure + Flow fusion for 6×6 mm^2 SS-OCTA acquisitions on the test set. The unimodal results are baselines derived from the previous step.

Fusion Method	**Backbone**	**Kappa**	**AUC0**	**AUC1**	**AUC2**	**AUC3**
Structure (unimodal)	ResNet	0.4150	0.8375	0.7659	**0.7889**	**0.8104**
Flow (unimodal)	ResNet	0.3768	0.7931	0.7653	0.7566	0.7863
Flow (unimodal)	EfficientNet	0.4085	0.8306	0.7775	0.7446	0.7150
Input Fusion	ResNet	0.3849	0.8093	0.7656	0.7476	0.7886
Input Fusion	EfficientNet	0.3885	0.8192	0.7755	0.7496	0.7321
Single-level Fusion	ResNet + ResNet	0.4329	0.8246	0.7763	0.7577	0.7900
Single-level Fusion	ResNet + EfficientNet	0.3959	0.8132	0.7637	0.7023	0.7622
Output Fusion	ResNet + ResNet	0.3814	0.8074	0.7757	0.7530	0.7868
Output Fusion	ResNet + EfficientNet	0.4227	0.8446	0.7770	0.7500	0.7478
Hierarchical Fusion	ResNet + ResNet	**0.4752**	**0.8462**	**0.7793**	0.7607	0.8013
Hierarchical Fusion	ResNet + EfficientNet	0.4205	0.8206	0.7662	0.7186	0.7743

In the validation and test sets, the hierarchical fusion outperformed other methods for 6×6 mm^2 OCTA. Based on two ResNet branches, the hierarchical fusion method achieved a Kappa value of 0.4752 for the six-category multiclass classification, a significant improvement over the unimodal baseline. Furthermore, hierarchical fusion improved diagnostic performance for both task 0 and task 1. In contrast, hierarchical fusion did not perform as well as unimodal fusion in tasks 2 and 3. There was a significant difference in performance between Structure and Flow in tasks 2 and 3. As a result, fusion was not effective since Flow did not provide additional complementary information to Structure. Also, the hierarchical fusion of ResNet and EfficientNet was not effective, likely due to the

structural differences between these backbones.

Table 6.6 – Results of Structure + Flow fusion for 15×15 mm^2 SS-OCTA images on the test set. The multimodal results are baselines derived from the previous step.

Fusion Method	**Backbone**	**Kappa**	**AUC0**	**AUC1**	**AUC2**	**AUC3**
Structure (unimodal)	ResNet	0.3900	0.8118	0.7604	0.7462	0.8700
Structure (unimodal)	DenseNet	0.3589	**0.8251**	0.7527	0.7923	0.8732
Flow (unimodal)	ResNet	0.4189	0.7927	0.7627	0.7911	0.8774
Input Fusion	ResNet	0.4252	0.7721	0.7475	0.7519	0.8546
Input Fusion	DenseNet	0.3286	0.7108	0.7072	0.7235	0.8175
Single-level Fusion	ResNet + ResNet	0.3982	0.8029	0.7627	0.7876	0.8630
Single-level Fusion	DenseNet + ResNet	0.3227	0.7437	0.7366	0.7546	0.8429
Output Fusion	ResNet + ResNet	0.4124	0.7949	0.7688	0.7688	0.8728
Output Fusion	DenseNet + ResNet	0.4376	0.8205	0.7583	0.7726	0.8754
Hierarchical Fusion	ResNet + ResNet	**0.4430**	0.8187	**0.7745**	**0.7967**	**0.8786**
Hierarchical Fusion	DenseNet + ResNet	0.4137	0.8088	0.7662	0.7794	0.8719

Similarly, for the 15×15 mm^2 SS-OCTA acquisitions, hierarchical fusion was the most effective in the validation and test sets. The hierarchical fusion of two ResNet branches significantly improved performance for six-category multiclass classification and tasks 1, 2, and 3 compared to the unimodal baseline results. Specifically, hierarchical fusion achieved an AUC of 0.8786 for task 3. Due to the similar performance of Structure and Flow, the hierarchical fusion was able to take advantage of the complementary information provided by the different modalities and performed well.

From the above results, the 6×6 mm^2 SS-OCTA was very effective for diagnosing early diabetic retinal lesions, while the 15×15 mm^2 SS-OCTA was more effective in diagnosing more advanced pathology, which is consistent with our clinical prior knowledge. As shown in [16], hierarchical fusion proves effective since the Structure and Flow-based hierarchical fusion can utilize complementary information to enhance the strengths of each acquisition individually, thereby facilitating the subsequent fusion of different acquisitions.

6.3.3 Fusion of 6×6 mm^2 SS-OCTA and 15×15 mm^2 SS-OCTA

To maximize the complementary strengths of the 6×6 mm^2 SS-OCTA and 15×15 mm^2 SS-OCTA acquisitions for different tasks, we further tested single-level fusion and output fusion on the hierarchical fusion architectures. The unimodal results of the 6×6 mm^2 SS-OCTA and 15×15 mm^2 SS-OCTA images were used as baselines. The results of this fusion are shown in Table 6.7.

Table 6.7 – Results of the 6×6 mm^2 SS-OCTA + 15×15 mm^2 SS-OCTA fusion on the test set. The 6×6 mm^2 SS-OCTA and 15×15 mm^2 SS-OCTA rows show the best performance of single acquisitions on different tasks.

Modality	**Fusion Method**	**Kappa**	**AUC0**	**AUC1**	**AUC2**	**AUC3**	**Inference Time (seconds/eye)**
6×6 mm^2 SS-OCTA	Structure (Unimodal)	0.4150	0.8375	0.7659	**0.7889**	**0.8104**	0.9729
	Hierarchical Fusion	**0.4752**	**0.8462**	**0.7793**	0.7607	0.8013	1.8041
15×15 mm^2 SS-OCTA	Structure (Unimodal)	0.3589	**0.8251**	0.7527	0.7923	0.8732	1.6394
	Hierarchical Fusion	**0.4430**	0.8187	**0.7745**	**0.7967**	**0.8786**	3.84655
6×6 mm^2 SS-OCTA + 15×15 mm^2 SS-OCTA	Single-level Fusion - Fine-tuning	0.4637	0.8469	0.8004	0.7989	0.8670	4.9233
	Single-level Fusion - Freezing layers	0.5132	0.8741	0.7853	0.7555	0.8207	4.8410
	Output fusion - Max	0.5218	0.8801	0.8027	0.8083	0.8911	4.0686
	Output fusion - Avg (Hybrid fusion proposed)	**0.5593**	**0.8868**	**0.8276**	**0.8367**	**0.9070**	3.9679

Output fusion with average aggregation was the best strategy for the validation and test sets. The process of single-level fusion showed improvements over single acquisitions on certain tasks but did not achieve its goal for tasks 2 and 3. This discrepancy could have been due to the differing volumes of the acquisitions. Without using registered image information for random cropping, the pathological features between acquisition branches could vary significantly, potentially impairing the fusion model's judgment. Conversely, output fusion had the capacity to address this issue effectively. The output fusion process operated only on the final output probability after each branch had independently made its assessments. This allowed

for the integration of information without being affected by image registration at the same time. As shown in Table 6.7, the proposed output fusion method based on averaging performed well.

The inference time of the different fusion methods was also compared. Inference took longer for 15×15 mm^2 SS-OCTA ($N = 20$), since the N times Random Crop was twice as large as for 6×6 mm^2 SS-OCTA ($N = 10$). Due to the complexity of its structure, hierarchical fusion requires more time for inference. Despite this, because of the parallel nature of the model, Our hybrid fusion method did not take longer than hierarchical fusion. The resulting four-second inference time per eye is acceptable and can provide reliable results for ophthalmologists within a short period of time.

6.4 Discussion and conclusions

This chapter investigated a deep learning algorithm to classify diabetic retinopathy severity using 6×6 mm^2 high-resolution SS-OCTA and 15×15 mm^2 UWF-SS-OCTA acquisitions. It relied on a hybrid fusion architecture that utilized complementary structure and flow information from both acquisitions. In detail, this architecture combined hierarchical fusion to jointly analyze Flow and Structure from the same acquisition and output fusion to merge predictions from both acquisitions. This algorithm was evaluated on preliminary data from the EviRed project.

The algorithm employed a unique hybrid fusion framework, integrating structural and flow information from both acquisitions. It was trained on data from 875 eyes of 444 patients. Tested on 53 patients (97 eyes), the algorithm achieved a good area under the receiver operating characteristic curve (AUC) for detecting DR (0.8868), moderate non-proliferative DR

(0.8276), severe non-proliferative DR (0.8376), and proliferative/treated DR (0.9070).

Our experiments showed that the 6×6 mm^2 SS-OCTA acquisitions were highly effective for the detection of early-stage pathology, while 15×15 mm^2 SS-OCTA acquisitions performed better in terms of advanced pathology detection (see Table 6.7). This was consistent with the perceived usefulness of these acquisitions by ophthalmologists: in the early stages, anomalies are generally small and are therefore better seen in high-resolution SS-OCTA images, while in the advanced stages, anomalies are larger, and an ultra-widefield image becomes more beneficial than a high-resolution image. The suggested hybrid fusion system demonstrated significant improvements over single acquisitions (see Table 6.7). The hybrid fusion approach integrated the strengths of both acquisitions: it delivered excellent performance in both early and late pathological diagnosis while significantly improving the accuracy of the six-category multiclass classification. Therefore, this study clearly validated the relevance of jointly analyzing multiple acquisitions. To a lesser extent, this study also validated the relevance of analyzing multiple modalities: combining Flow and Structure always outperformed analyzing a single modality, although the performance gain was limited (see Tables 6.5 and 6.6).

In recent times, transformer-based models [121] have shown good performance on classification tasks, such as the Vision Transformer (ViT) [128], which we also tested. The performance of the structure and flow modalities of 6×6 mm^2 SS-OCTA images was tested using 3D ViT models (patch size = (32, 32, 32)) from the Monai library [1]. Table 6.8 illustrates the test results for ViT.

In all tasks, ViT performed very poorly. A large dataset and pre-trained

1. `https://monai.io/`

Table 6.8 – Results for 3D ViT with different modalities on the test set.

Modality	Backbone	Kappa	AUC0	AUC1	AUC2	AUC3
6×6 mm^2 SS-OCTA—Structure	ViT	0.1122	0.6774	0.6490	0.4900	0.5912
6×6 mm^2 SS-OCTA—Flow	ViT	0.0854	0.6696	0.6474	0.5487	0.5843

models contribute significantly to the excellent performance of ViT [128]. In addition to the limited number of patients in our dataset, there was no publicly available pre-training model for 3D ViT, which was likely the major reason for its poor performance. Nevertheless, extensive testing is still required for the hyperparameter configuration of 3D transformer models.

It should be noted, however, that some transformer-based models are increasingly used to perform multimodal tasks in the medical field [264, 402, 403]. It has been observed that these models often combine a CNN structure with a transformer structure, resulting in excellent classification performance with limited medical datasets; this is one of the directions that may be pursued by the team, during the end of the Evired project.

One limitation of this study was that the current dataset is insufficiently large, resulting in suboptimal performance on the six-category multiclass classification task. Furthermore, too small a dataset may adversely affect the robustness of a model. The EviRed project is expected to collect clinical data from thousands of patients, and so more datasets will be tested in the near future. Further studies will be conducted to test the model's stability and fine-tune it to improve its performance on the six-category multiclass classification task.

The current EviRed prospective dataset also contains ultra-widefield color fundus photography (UWF-CFP) data alongside OCTA data from different acquisitions, which may aid in further improving the accuracy of DR diagnosis. In [37], the use of UWF-OCTA in conjunction with UWF-

CFP was recommended for the screening and follow-up of DR. Conversely, UWF-OCTA alone had some limitations. Identifying microaneurysm and intraretinal hemorrhage from OCTA en-face images is difficult and sometimes ambiguous. To facilitate diagnosis, searching for corresponding lesions on B-scan images is often necessary, a time-consuming process. The use of UWF-CFP images would make this task much easier. The joint analysis of OCTA and UWF-CFP images will direct our next investigation in Chapter 7.

Chapter 7

Multimodal Fusion of UWF-CFP and OCTA Images

"Every great advance in science has issued from a new audacity of imagination."

— *John Dewey*

Recent advancements in imaging technologies, such as Ultra-WideField Color Fundus Photography (UWF-CFP) imaging and Optical Coherence Tomography Angiography (OCTA), provide opportunities for the early detection of Diabetic Retinopathy (DR) but also pose significant challenges given the disparate nature of the data they produce. We have included both OCTA and UWF-CFP data for each patient in the second

stage of the EviRed prospective dataset. It has been demonstrated clinically that these two modalities are complementary [37, 404]. This chapter introduces a novel multimodal approach that leverages these imaging modalities to notably enhance DR classification, as proposed in Section 6.4. Our approach integrates 2D UWF-CFP images and 3D high-resolution 6x6 mm^2 OCTA (both structure and flow) images using a fusion of ResNet50 and 3D-ResNet50 models, with Squeeze-and-Excitation (SE) blocks to amplify relevant features. Additionally, to increase the model's generalization capabilities, a multimodal extension of Manifold Mixup, applied to concatenated multimodal features, is implemented.

7.1 Introduction

Recent advances in imaging techniques have significantly enhanced the ability to detect and classify DR. UWF-CFP imaging and OCTA are two such techniques that have shown great promise. UWF-CFP imaging offers a panoramic view of the retina, allowing for a more comprehensive assessment [405], while OCTA provides depth-resolved images of retinal blood flow, revealing detailed microvascular changes indicative of DR [406].
Despite the individual merits of these imaging modalities, each offers a unique perspective on retinal pathology. Leveraging the information from both could potentially enhance the diagnosis and classification of DR based on artificial intelligence (AI) [407, 408]. However, the integration of these modalities poses a significant challenge due to the disparate nature of the data they produce, especially in terms of dimensionality (2D versus 3D) and field of view.

In the quest to enhance deep learning (DL) models, the field has bene-

fited significantly from incorporating innovative techniques like the Manifold Mixup [409]. Through its unique method of generating virtual training examples via the convex combinations of hidden state representations, this technique has made a profound impact by significantly reducing a model's sensitivity to the data distribution and encouraging smoother decision boundaries.

Building upon these advanced techniques, several proposed methods in the state of the art have employed multimodal imaging [410, 411]. These methods aim to utilize the complementary information available in different types of images. Recent works have effectively used mixing strategies to enhance multimodal DL models. For example, the M^{3}ixup approach [412] leverages a mixup strategy to enhance multimodal representation learning and increase robustness against missing modalities by mixing different modalities and aligning mixed views with original multimodal representations. The LeMDA (Learning Multimodal Data Augmentation) [413] method automatically learns to jointly augment multimodal data in feature space, enhancing the performance of multimodal deep learning architectures and achieving good results across various applications. MixGen [414] introduces a joint data augmentation for vision-language representation learning to boost data efficiency, generating new image-text pairs while preserving semantic relationships. This method has shown remarkable performance improvements across various vision-language tasks. Furthermore, TMMDA (Token Mixup Multimodal Data Augmentation) [415] for Multimodal Sentiment Analysis (MSA) generates virtual modalities from the mixed token-level representation of raw modalities, enhancing representation learning on limited labeled datasets.

Despite the significant results obtained, these methods are proposed

for vision-language and vision-audio fusion but are not suitable for 2D image/3D volume fusion. This study proposes a new multimodal DL approach for DR classification, integrating 2D UWF-CFP and 3D OCTA images and incorporating a custom mixing strategy. Regarding the modalities used in this work, recent research has used UWF-CFP and OCTA imaging to diagnose diseases such as Alzheimer [416]. However, to the best of our knowledge, our study is the first to develop a DL model for the classification of DR using both UWF-CFP and OCTA imaging modalities, which contributes significantly to the existing body of knowledge.

7.2 Material and methods

7.2.1 Model architecture

We utilize two separate CNN architectures, ResNet50 and 3D-ResNet50, designed to process 2D UWF-CFP and 3D OCTA images to extract features from each imaging modality. ResNet50 was chosen as a backbone for feature extraction due to its remarkable performance in computer vision tasks. Its structure provides a balance between depth and complexity, allowing the network to learn complex patterns without suffering from overfitting. To further improve such models' performance, Squeeze-and-Excitation (SE) blocks have gained attention in the DL community [417]. As shown in Fig.6.3(d), the SE blocks dynamically recalibrate channel-wise feature responses by explicitly modeling the interdependencies between channels, thus helping the model focus on more informative features. They have been demonstrated to significantly improve the representational power of deep networks without a significant additional computational cost.

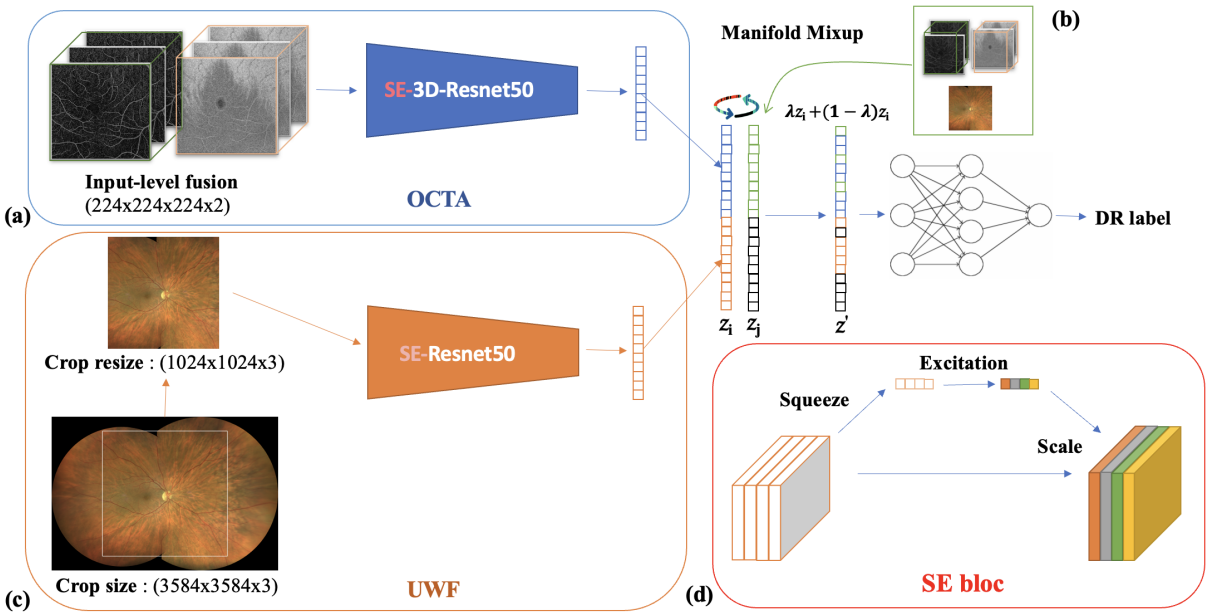


Figure 7.1 – Proposed pipeline.

The 3D-ResNet50, a 3D extension of the ResNet50 architecture, integrated with SE blocks, is applied to process OCTA images (Fig.7.1(a)). This model expands traditional 2D convolution operations into the 3D space, making it particularly appropriate for volumetric image data. This enables the network to decipher spatial hierarchies inherent in volumetric data, thus facilitating a comprehensive feature extraction from OCTA volumes. SE blocks in the 3D-ResNet50 model perform a similar role as in the 2D ResNet50 model, thus enhancing the performance of the 3D backbone. For the rest of the paper, we will refer to these models as SE-ResNet50 and SE-3D-ResNet50.

7.2.2 Fusion strategy

The fusion of multiple modalities has been an area of active research due to the enhanced performances it offers [143, 198, 248]. Such fusion can be executed at input, feature, and decision levels, each offering distinct

advantages and disadvantages.
In this work, we employ an input-level fusion for merging the structure and flow information embedded in OCTA images. Numerous studies affirm that merging these distinct types of information can significantly enhance the accuracy of DR diagnosis [410, 418]. Input-level fusion involves integrating multiple modalities into a single data tensor subsequently processed by a DL model Fig. 6.3(a). This method is effective without the need for registration, as the structure and flow data align with each other by design.

On the other hand, the fusion of UWF-CFP and OCTA images is performed through a different approach, primarily due to the absence of inherent alignment between these imaging modalities. Here, a feature-level fusion strategy is adopted, which allows us to use different backbones for each modality (SE-ResNet50 and 3D-SE-ResNet50), thus effectively addressing the alignment challenge. We have chosen feature-level fusion over decision-level fusion to capitalize on the rich interplay between the modalities at the feature level. This strategy facilitates the extraction of features and the fusion of high-dimensional feature-level information, making it especially suited for unregistered or dimensionally diverse data [17, 228, 252, 419].

7.2.3 Manifold Mixup

To enhance the model's robustness and generalization capabilities, we implemented a multimodal extension of Manifold Mixup into our training process. The original Manifold Mixup method [409] is a recently introduced regularization technique. It generates virtual training examples by forming convex combinations of the hidden state representations of two randomly

chosen training examples and their associated labels.
Extending the concept of Input Mixup [420] to the hidden layers, Manifold Mixup serves as a robust regularization method that provokes neural networks to predict interpolated hidden representations with lesser confidence. It leverages semantic interpolations as an auxiliary training signal, leading to the cultivation of neural networks with smoother decision boundaries across multiple representation levels. Consequently, neural networks trained with Manifold Mixup can learn class representations with reduced directions of variance, thus yielding a model that exhibits enhanced performance on unseen data[409]. The operational process of the Manifold Mixup approach is as follows:

1. The original Manifold Mixup performs the mixing of the hidden representation randomly on a set of predefined eligible layers. Instead, in our proposed implementation, we have purposefully selected the layer containing the concatenated feature maps from UWF-CFP and OCTA images to process the Manifold Mixup. This strategic choice is not only the simplest way to introduce Manifold Mixup but also ensures we are capitalizing on a layer that encapsulates a high-dimensional, multimodal feature space. Creating numerous virtual training samples from the fusion layer significantly improves the model's ability to generalize to new data.
2. Feed two images into the neural network until the selected layer is reached.
3. Extract the feature representations (z_i for multimodal data x_i and z_j for multimodal data x_j).
4. Mix the extracted feature representations according to the following equation in order to derive the new representation (new features z'

associated with new label y').

$$(z', y') = (\lambda z_i + (1 - \lambda) z_j, \lambda y_i + (1 - \lambda) y_j)$$

where z_i and z_j are the features of two random training examples, and y_i and y_j are their corresponding labels. $\lambda \in [0, 1]$ is a Mixup coefficient sampled from a Beta distribution $Beta(\alpha, \alpha)$, where α is a hyperparameter that determines the shape of the Beta distribution.

5. Carry out the forward pass in the network for the remaining layers with the mixed data.
6. Use the output of the mixed data to compute the loss and gradients. Given $\mathcal{L}$ the original loss function, the new loss $\mathcal{L}'$ is computed as:

$$\mathcal{L}' = \lambda \mathcal{L}(y_i, y') + (1 - \lambda) \mathcal{L}(y_j, y')$$

Through this process, Manifold Mixup enhances our fusion strategy by operating on the joint feature representation (Fig.6.3(b)), thereby ensuring that the model can generalize from the learned features of UWF-CFP and OCTA images.

7.2.4 Dataset

After removing patient data with incomplete modalities and absent annotations, 875 eyes belonging to 444 patients from the second stage of the EviRed prospective dataset described in Section 3.3.4 were divided into training sets, validation sets, and test sets in the same manner as in Section 6.2.5. Each patient's eye was labeled by an ophthalmologist into one of the

6 DR classes: Normal, mild nonproliferative diabetic retinopathy (NPDR), moderate NPDR, severe NPDR, proliferative DR (PDR), or Pan-Retinal Photocoagulation (PRP). The UWF-CFP images in the dataset, captured using the Clarus 500 (Carl Zeiss Meditec Inc., Dublin, CA, USA), varied in size, ranging from 3900×3900 to 7900×4900 pixels. This size variation arises from the image stitching process for montage creation, not from changes in the device's resolution. Considering the clinicians' focus on the seven Early Treatment Diabetic Retinopathy Study (ETDRS) fields [421], we carried out center cropping on each image to 3584×3584. This process ensured that all seven fields were included in the image. Subsequently, we resized these cropped images to 1024×1024, a size that guarantees no loss of details.

The high-resolution 6x6 mm^2 OCTA images, offering 500×224×500 voxels and centered on the macula, were captured using the Zeiss PLEX Elite 9000. Each OCTA volume includes 2-D en-face localizer, structural, and flow 3D volumes. Due to the restrictions posed by the graphics processing unit (32Gb GPU) hardware, our 3D-SE-ResNet50 could only accommodate inputs up to $224 \times 224 \times 224 \times 2$ input tensors. This limitation guided our data pre-processing. The OCTA images were preprocessed in the same method as in Section 6.2.2. During the prediction process, we extracted multiple volumes from the OCTA image using N=10 times random crop, which were simultaneously processed with the full UWF-CFP image to make predictions. The final prediction for an examination was determined based on the severest prediction among these N predictions (test-time augmentation).

7.2.5 Implementation details

Our models were implemented using the PyTorch [1] deep learning library. This experiment was performed on the OVH cluster (one NVIDIA Tesla V100s GPU was used). For UWF-CFP images, we used the SE-ResNet50 architecture with weights pre-trained on ImageNet, while for OCTA images, we trained from scratch our implementation of the 3D-SE-ResNet50 backbone with input-level fusion for structure and flow volumes. The key to our model enhancement process included incorporating SE blocks in both ResNet models and using Manifold Mixup on multimodal features for model regularization. In our implementation, we set the reduction ratio, a crucial SE hyperparameter, to 16, following the practice from the original SE network paper [417]. For Mixup, we carried out a grid search focusing on the α parameter, which is essential for deriving the adequate Beta distribution $Beta(\alpha, \alpha)$ for sampling the right λ interpolation parameter during Manifold Mixup training. This comprehensive exploration determined 0.2 as the optimal value for α, which yielded the best model performance. The two models were trained jointly on the UWF-CFP and OCTA datasets, using a cross-entropy loss function and an AdamW optimizer. During training, we used a learning rate of 0.001 with the OneCycle scheduler, a decay factor of 0.0001, and a batch size of 4 over 200 epochs.

7.3 Results

To compare the performance of our proposed method with the individual modalities, we trained standalone models using either UWF-CFP or OCTA images with the same training settings as described above. This provided

1. https://pytorch.org/

a baseline performance for each modality, against which the performance of the multimodal approach was compared. In addition, an ablation study was conducted to further understand each component's impact and contribution to our pipeline. We compared the performance of our model without the Manifold Mixup and the SE blocks.

The performance of the proposed models was evaluated in terms of the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) described in Section 2.2.3. This metric was chosen due to its ability to provide an aggregate measure of performance across the four DR severity cutoffs ($\geq$ mild NPDR, $\geq$ moderate NPDR, $\geq$ severe NPDR, $\geq$ PDR).

Tab.7.1 presents the performance of the different models: the ResNet50 model trained on UWF-CFP images, the 3D-ResNet50 model trained on OCTA images, the proposed multimodal pipeline, the multimodal models without SE, the pipeline without Manifold Mixup (MM in the table), and the pipeline without SE and Manifold Mixup.

Data	SE	MM	$\geq$ mild NPDR	$\geq$ moderate NPDR	$\geq$ severe NPDR	$\geq$ PDR
UWF-CFP	✗	✗	0.7983	0.7925	0.7906	**0.9159**
OCTA	✗	✗	0.8316	0.7627	0.7338	0.7576
Multimodal	✓	✓	**0.8566**	**0.8037**	**0.7922**	0.8820
Multimodal	✗	✓	0.8241	0.7969	0.7682	0.8522
Multimodal	✓	✗	0.8431	0.7782	0.7566	0.8420
Multimodal	✗	✗	0.8140	0.7775	0.7525	0.8164

Table 7.1 – Performance of Models in DR Classification

Our approach that combines both UWF-CFP and OCTA images using a multimodal pipeline notably outperformed models based on individual modalities. Specifically, when evaluating DR severity cutoffs, the multimodal model achieved an AUC score of 0.8566 for $\geq$ mild NPDR, notably higher than 0.7983 for UWF-CFP alone and 0.8316 for OCTA alone. This trend continued with $\geq$ moderate NPDR and $\geq$ severe NPDR, where our

multimodal model attained AUC scores of 0.8037 and 0.7922, respectively, compared to 0.7925 and 0.7906 for UWF-CFP and 0.7627 and 0.7338 for OCTA.

7.4 Discussion and conclusions

Experimental results demonstrate a remarkable enhancement in DR classification performance with the proposed multimodal approach compared to methods relying on a single modality only. The methodology laid out in this work holds substantial promise for facilitating more accurate, early detection of DR, potentially improving clinical outcomes for patients.

These outcomes underscore the importance of capitalizing on diverse image modalities to provide a more comprehensive, holistic analysis, thereby enhancing the robustness and accuracy of DR classification. Our study suggests that each imaging modality captures distinct aspects of DR, and the concurrent utilization of both modalities in our models appears to improve the diagnosis, which is aligned with clinical studies [407, 408].
The greater success of UWF-CFP in identifying the cutoff $\geq$ PDR can be attributed to its wide-field view of the retina, which allows for the detection of peripheral lesions and signs of PRP laser impacts. Conversely, OCTA images proved to be particularly useful for $\geq$ mild NPDR detection due to their central focus on the macula and the high-resolution imaging of the microvasculature.
Regarding the added components in our pipeline, the Manifold Mixup and the SE blocks were proven to enhance the model's performance. For example, omitting the SE blocks caused a decrease in AUC scores across all DR severities. This indicates the critical role of SE blocks in bolstering feature

representations and overall model robustness. Similarly, when the Manifold Mixup was excluded, there was a noticeable drop in performance, corroborating the effectiveness of such a regularization technique in improving model generalization.

Our findings demonstrate the efficacy of the proposed multimodal model in improving DR classification. This model, which integrates UWF-CFP and OCTA images using a feature-level fusion strategy and employing both our proposed adaption of the Manifold Mixup technique and SE blocks, delivers a compelling performance. The ablation study further attests to the significance of each component within our pipeline. These findings reiterate the necessity and potency of multimodal approaches coupled with advanced regularization techniques, such as Manifold Mixup and SE blocks, for medical image classification tasks.

To the best of our knowledge, our study is the first to propose a pipeline for the classification of DR using both UWF-CFP and OCTA images. However, we believe several improvements and extensions could further enhance the classification performance. The application of cross-modal attention mechanisms may provide a more effective way of fusing features from different modalities by focusing on the most relevant information from each. Similarly, implementing Manifold Mixup at different levels of the model, rather than solely at the concatenation layer, could provide further regularization and performance improvements. Moreover, introducing novel components, such as Transformer blocks, might prove beneficial in capturing complex relationships within and across modalities.

Additionally, due to the late arrival of data, we were only able to explore

the multimodal fusion of UWF-CFP and OCTA during the final stages of the thesis. Currently, the exploration of the fusion of these two modalities is in its infancy, and a number of tests will be implemented in the future. In Chapter 6, results indicate that the fusion of the 6×6 mm^2 OCTA and 15×15 mm^2 OCTA can enhance diagnostic performance, and how to incorporate 15×15 mm^2 OCTA into our current fusion network needs to be explored. Further, we have tested only the single-level fusion architecture at this time; hierarchical fusion or the more complex hybrid fusion will be tested in the future.

Conclusions and future works

> "The future is an unknown, but a somewhat predictable unknown. To look to the future, we must first look back upon the past. That is where the seeds of the future were planted."
>
> — *Albert Einstein*

Conclusions

Diabetic retinopathy (DR), which affects 422 million people worldwide, including 3.3 million in France, is a leading cause of blindness in the country. The current diagnostic challenge lies in the outdated DR classification system based on Color Fundus Photography (CFP), which fails to predict disease progression effectively. Although modern imaging techniques like Ultra-Wide-Field CFP (UWF-CFP), Optical Coherence Tomography (OCT), and Optical Coherence Tomography Angiography (OCTA) offer richer, more detailed data, they generate complex datasets requiring specialized analysis. The Évaluation Intelligente de la Rétinopathie diabétique (EviRed) project emerges within this landscape, aspiring to revolutionize DR diagnosis. EviRed's expert system leverages contemporary imaging and patient data to replace the current classification, aiming to improve

predictions of disease evolution and ensure timely treatments. This innovation addresses the growing need for enhanced DR diagnostics amidst the increasing volume of data from advanced imaging, a challenge for many ophthalmologists.

As a component of the EviRed project, this thesis delved into the application of artificial intelligence for the purpose of seamlessly integrating extensive datasets. Its overarching goal was to streamline the diagnostic process, improve prediction accuracy, and enhance the decision-making capabilities of ophthalmologists in their DR case follow-ups. More precisely, the focus lay in crafting deep learning network structures that could effectively harness the unique advantages offered by diverse imaging modalities, ultimately elevating the quality of diagnostic outcomes.

During the thesis, we explored multiple fusion techniques between different modalities. We summarized and proposed multiple multimodal fusion deep learning frameworks and conducted extensive evaluations. Experimental results indicate that our multimodal fusion network can effectively utilize the complementary information between the different modalities, thus improving the accuracy of DR diagnosis. Overall, the three main objectives of the thesis have been met:

— For the **joint analysis of multi-modal information in OCTA**, we examined fusion techniques across three types of data: 2D line scanning ophthalmoscope (LSO), 3D structural OCT, and 3D OCT angiography. In addressing retinal analysis challenges, we explored three multimodal information fusion strategies grounded in deep learning: input, single-level, and hierarchical fusion. While input and single-level fusion methods are straightforward, they fail to fully capitalize on the complementary information inherent in these modalities. In response, we developed a hierarchical fusion approach emphasizing

feature combination across various network dimensions and the exploration of modality correlations. Our hierarchical fusion approach consistently outperformed others across different tasks, promising significant advancements in clinical diagnosis.

— For the **joint analysis of different specifications of OCTA acquisitions**, we investigated a deep learning algorithm using both high resolution and ultra-widefield (UWF) OCTA for assessing DR severity automatically. It relied on a hybrid fusion architecture that utilized complementary structure and flow information from both acquisitions. In detail, this architecture combined hierarchical fusion to jointly analyze Flow and Structure from the same acquisition and output fusion to merge predictions from both acquisitions. The hybrid fusion approach integrated the strengths of both acquisitions: it delivered excellent performance in both early and late pathological diagnosis while significantly improving the accuracy of the six-category multiclass classification. Therefore, this study clearly validated the relevance of jointly analyzing multiple acquisitions. The suggested hybrid fusion system demonstrated significant improvements over single acquisitions.

— For the **joint analysis of OCTA and UWF-CFP**, we proposed a novel single-level fusion network that leverages these imaging modalities to notably enhance DR classification. Our approach integrates 2D UWF-CFP images and 3D high-resolution OCTA images using a fusion of ResNet50 and 3D-ResNet50 models with Squeeze-and-Excitation (SE) blocks to amplify relevant features. Additionally, to increase the model's generalization capabilities, a multimodal extension of Manifold Mixup, applied to concatenated multimodal features, is implemented. Experimental results demonstrate a remark-

able enhancement in DR classification performance with the proposed multimodal approach compared to methods relying on a single modality only. The methodology laid out in this work holds substantial promise for facilitating more accurate, early detection of DR, potentially improving clinical outcomes for patients.

We have demonstrated that combining multimodal data, from the same acquisition or from different acquisitions, improves performance in the context of DR severity assessment, provided that a suitable fusion framework is used. Based on our extensive literature review, we believe this finding would likely generalize to several other clinical tasks. Therefore, it is advisable to explore multiple fusion frameworks when addressing a clinical decision support problem with multimodal data. Our work meets the basic requirements of the EviRed project for multimodal fusion tasks and provides a reference and direction for the project's future development. At the same time, the algorithms we developed can be easily integrated into the EviRed system, thus providing ophthalmologists with timely and reliable assistance in their diagnostic procedures. Further, some exploration of unlabeled data was also undertaken during the PhD. Even though their performance on diagnosis was not significantly improved, they still provide experience and references for future work in the EviRed project. Amidst the growing enthusiasm for computer-aided diagnosis, our work has made a meaningful contribution to this evolving landscape.

Future works

The EviRed project is currently in an intermediate stage, with much exploratory work still to be completed. Specifically, the multimodal fusion work requires further testing as more patients are recruited, and more data

is collected. Further, the rapid development of deep learning algorithms and the upgrading of hardware devices also provide additional directions for future multimodal fusion methods. Some of the perspectives we wish to address in the future:

1. Replication of experiments

   One limitation of our study was that the current dataset is insufficiently large, resulting in suboptimal performance on the DR severity multi-classification task. Furthermore, too small a dataset may adversely affect the robustness of a model. The third phase of the EviRed prospective dataset is currently being collected, containing data on 1032 patients in total. The proposed fusion methods need to be retrained and tested using this dataset. These experiments should be replicated in larger datasets to demonstrate clinically useful detection performance.

2. Further exploration of the multimodal fusion of UWF-CFP and OCTA Images

   Our research on the fusion of UWF-CFP and OCTA Images began at the end of the paper. In comparison to other modalities, these two contain more complementary information, but the structure of the fusion network between 2D and 3D modalities is also more complex. Although preliminary experimental results demonstrate the effectiveness of our proposed fusion method, we have only tested the single-level fusion framework between the two modalities. Fusion methods with more complexity, such as hierarchical and hybrid fusion, require further testing. Further, our experiments until now have only used high-resolution OCTA images and UWF-CFP for fusion. Considering that ultra-widefield OCTA images contain additional information, we

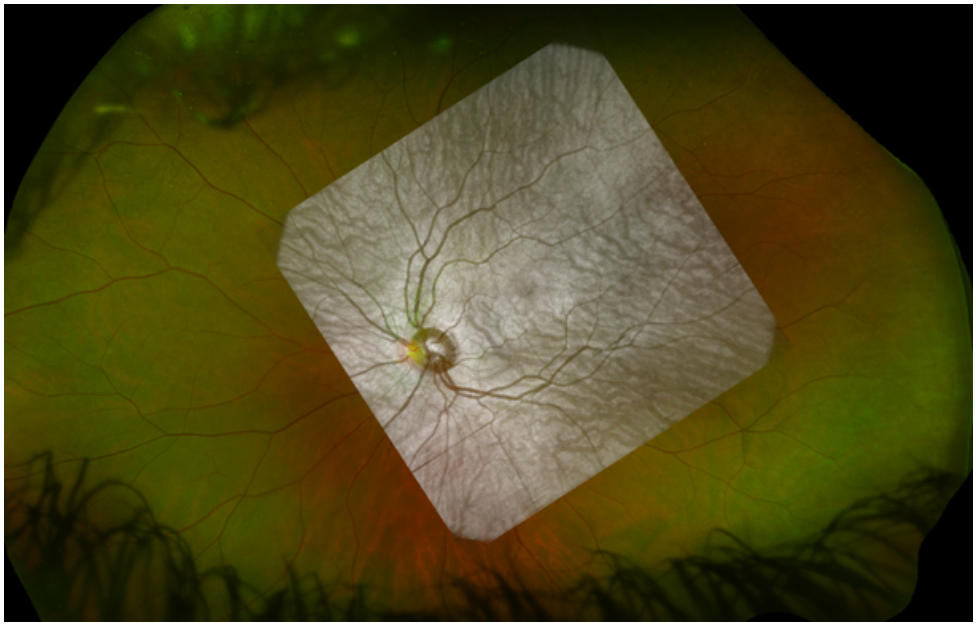


Figure 7.2 – Registration of UWF-CFP and LSO (from OCTA image) in the en-face image direction.

need to investigate how to incorporate ultra-widefield OCTA images into our fusion framework. We will first validate the approach of hierarchical fusion by using different branches to extract features from each of the three modalities and then adding an additional fusion branch to fuse the features at different levels.

The results of our experiments have also demonstrated that the registration between different modalities may affect the performance of hierarchical fusion. In order to improve information fusion between different modalities, the EviRed project is working on developing a registration algorithm for UWF-CFP and OCTA Images. The registration of UWF-CFP and OCTA images in the en-face image direction can be roughly realized using LSO images from OCTA Images, as shown in Fig. 7.2. Meanwhile, we are developing a mechanism for feature registration within the hierarchical fusion architecture on the fusion branch.

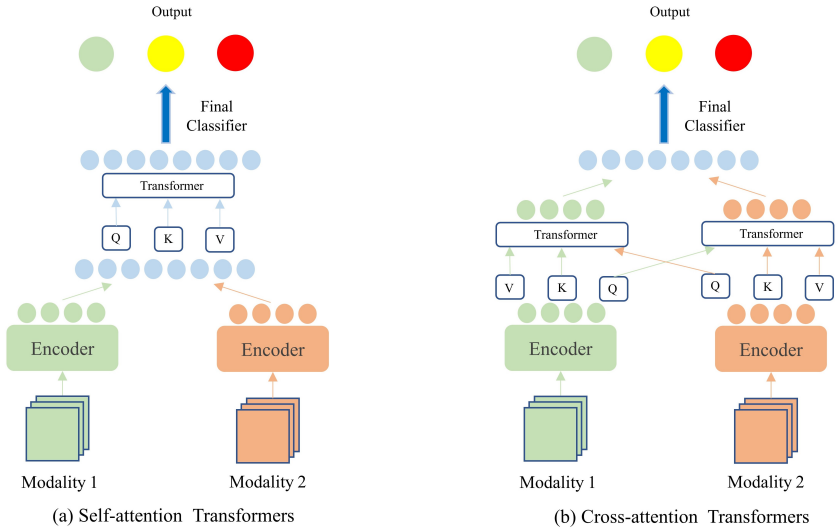


Figure 7.3 – Two architectures of Transformer-based fusion.

3. Investigation of Transformer-based fusion networks

   Transformer is one of the most popular network architectures, and multimodal fusion based on Transformer has developed rapidly in the past two years. In particular, for visual-language tasks, Transformer can handle the fusion of images, languages, and text very effectively. Based on the research conducted in different fields, we classify Transformer-based multimodal fusion networks into self-attention Transformers [363, 364, 422–427] and cross-attention Transformers [428–435], as shown in Fig. 7.3. Following the extraction of features using encoders, self-attention Transformers concatenate features from different modalities and compute the attention relationship between the fused features using Transformer blocks. Alternatively, cross-attentional Transformers compute the attentional relationships among different modalities in order to achieve information fusion. Nowadays, these two architectures are the most popular multimodal fusion networks.

Compared to CNNs, Transformers have the advantage of efficiently identifying long-range relationships between sequences. In medical images, most visual representations are ordered due to the similarity of human organs. Medical images contain more information regarding sequence relationships than natural images [264]. This indicates that Transformer-based multimodal medical image fusion is a promising approach, and the above two network architectures are worth exploring.

In Section 6.4, we did not obtain satisfactory results using ViT. It should be noted, however, that some transformer-based models are increasingly used to perform multimodal tasks in the medical field [264, 402, 403]. It has been observed that these models often combine a CNN structure with a transformer structure, resulting in excellent classification performance with limited medical datasets; this is one of the directions that we plan to pursue in the future.

4. Problem of incomplete or heterogeneous modality

   The problem of modality incompleteness is one of the most pressing challenges in multimodality medical research. In the EviRed prospective dataset, there are dozens of patients with modal incompleteness problems. The most common approach to solving the modality incompleteness problem is to discard the modality incomplete subjects [186, 201, 225, 243], just like we did, but this approach reduces the number of trainable subjects for the deep learning model, resulting in reduced classification performance.

   Generative Adversarial Networks (GAN) [436] is a type of generative model used to produce data of a modality from another modality [437]. With the development of GAN, more and more fields are us-

ing this technology to generate images. The modal incompleteness problem has recently been solved through the use of GAN in many studies [141, 247, 259, 438–440]. The GAN is used to generate the missing data, and then the generated data is used for multimodal classification. It significantly increases the number of subjects in the dataset, improves the model's classification performance, and is an effective solution when dealing with multimodal incompleteness.

Volume-to-volume GAN has been implemented in the Challenge APIS (Section 7.4 Challenges), which can be used directly on OCTA images to generate incomplete OCTA acquisitions. Although the large volume of OCTA images makes model training difficult, more possibilities can be explored by upgrading GPU hardware.

In addition, we have OCTA and UWF-CFP from different devices with different specifications. The use of GAN to generate incomplete or heterogeneous modalities is also worth exploring.

5. Time series analysis of multimodal examinations

   As part of the EviRed project, another PhD student is examining time-series analysis with unimodal data. Through the use of longitudinal fundus images, we have achieved excellent results in the diagnosis of DR [441, 442]. Our next objective is to conduct a joint analysis of multimodal longitudinal data.

As a final note, we believe the work presented in this thesis will result in a more complete and reliable automated diagnosis for the EviRed project. This will facilitate a much wider adoption of artificial intelligence, thereby allowing the shortage of ophthalmologists and the continuous increase in the number of people at risk of developing DR to be addressed.

Diffusion of research results and extension of thesis

"If you have knowledge, let others light their candles in it."

— *Margaret Fuller*

Below are the list of publications and challenges completed during this PhD:

Journal papers

1. Li, Y., El Habib Daho, M., Conze, P.-H., Zeghlache, R., Le Boité, H., Bonnin, S., Cosette, D., Magazzeni, S., Lay, B., Le Guilcher, A., Tadayoni, R., Cochener, B., Lamard, M. & Quellec, G., Hybrid Fusion of High-Resolution and Ultra-Widefield OCTA Acquisitions for the Automatic Diagnosis of Diabetic Retinopathy, Diagnostics 2023, 13, 2770. https://doi.org/10.3390/diagnostics13172770
2. Li, Y., El Habib Daho, M., Conze, P-H., Zeghlache, R., Le Boité, H., Tadayoni, R., Cochener, B., Lamard, M. & Quellec, G., A review

of deep learning-based information fusion techniques for multimodal medical image classification. Submitted to Computers in Biology and Medicine.

3. El Habib Daho, M., Li, Y., Zeghlache, R., Le Boité H., Deman P., Borderie, L., Ren, H., Mannivanan, N., Lepicard, C., Cochener, B., Couturier, A., Tadayoni, R., Conze, P.-H., Lamard, M. & Quellec, G., DISCOVER: 2-D Multiview Summarization of Optical Coherence Tomography Angiography for Automatic Diabetic Retinopathy Diagnosis. Submitted to Artificial Intelligence In Medicine.

4. (Challenge paper) Qian, B., Chen, H., Wang, X., Che, H., Kwon, G., Kim, J., Choi, S., Shin, S., Krause, F., Unterdechler M., Hou, J., Feng, R., Li, Y., El Habib Daho, M., et al., DRAC: Diabetic Retinopathy Analysis Challenge with Ultra-Wide Optical Coherence Tomography Angiography Images. Submitted to Patterns, arXiv:2304.02389 (2023).

Conference papers

1. Li, Y., El Habib Daho, M., Conze, P.-H., Al Hajj, H., Bonnin, S., Ren, H., Manivannan, N., Magazzeni, S., Tadayoni, R., Cochener, B., Lamard, M. & Quellec, G., Multimodal information fusion for glaucoma and diabetic retinopathy classification, International Workshop on Ophthalmic Medical Image Analysis, Cham: Springer International Publishing, 2022: 53-62.

2. Li, Y., Zeghlache, R., Brahim, I., Xu, H., Tan, Y., Conze, P.-H., Lamard, M., Quellec, G. & El Habib Daho, M., Segmentation, Classification, and Quality Assessment of UW-OCTA Images for the Diag-

nosis of Diabetic Retinopathy, MICCAI Challenge on Mitosis Domain Generalization, Cham: Springer Nature Switzerland, 2022: 146-160.

3. Li, Y., Zhang, P., Tan, Y., Zhang, J., Wang, Z., Jiang, W. Conze, P.-H., Lamard, M., Quellec, G. & El Habib Daho, M., Automated Detection of Myopic Maculopathy in MMAC 2023: Achievements in Classification, Segmentation, and Spherical Equivalent Prediction. Submitted to MICCAI MMAC 2023 Myopic Maculopathy Analysis Challenge.

4. El Habib Daho, M., Li, Y., Zeghlache, R., Atse, Y.C., Le Boité H., Bonnin, S., Cosette, D., Deman, P., Borderie, L., Lepicard, C., Tadayoni, R., Cochener, B., Conze, P.-H., Lamard, M. & Quellec, G., Improved Automatic Diabetic Retinopathy Severity Classification Using Deep Multimodal Fusion of UWF-CFP and OCTA Images. Accepted by International Workshop on Ophthalmic Medical Image Analysis 2023.

5. Xu H., Li, Y., Zhao, W., Quellec, G., Lu, L. & Hatt, M. Joint nnU-Net and radiomics approaches for segmentation and prognosis of head and neck cancers with PET/CT images, 3D Head and Neck Tumor Segmentation in PET/CT Challenge, Cham: Springer Nature Switzerland, 2022: 154-165.

6. Zeghlache, R., Conze, P.-H., El Habib Daho, M., Li, Y., Le Boité H., Massin, P., Tadayoni, R., Cochener, B., Brahim, I., Quellec, G. & Lamard, M., LMT: Longitudinal Mixing Training a framework for the prediction of disease progression using a single image. Accepted by International Workshop on Machine Learning in Medical Imaging 2023.

7. Zeghlache, R., Conze, P.-H., El Habib Daho, M., Li, Y., Le Boité

H., Massin, P., Tadayoni, R., Cochener, B., Brahim, I., Quellec, G. & Lamard, M., Longitudinal self-supervised learning using neural ordinary differential equation. Accepted by International Workshop on Predictive Intelligence in Medicine 2023.

Conference abstract

1. Li, Y., El Habib Daho, M., Conze, P.-H., Zeghlache, R., Ren, H., Lepicard, C., Deman, P., Le Guilcher. A., Cochener, B., Tadayoni, R., Lamard, M. & Quellec, G., 3-D analysis of multiple OCTA acquisitions for the automatic diagnosis of diabetic retinopathy. ARVO 2023.
2. Li, Y., Al Hajj, H., Conze, P.-H., Bonnin, S., Ren, H., Manivannan, N., Magazzeni, S., Tadayoni, R., Lamard, M. & Quellec, G., Multimodal information fusion for the diagnosis of diabetic retinopathy. ARVO 2022.
3. Li, Y., Al Hajj, H., Conze, P.-H., El Habib Daho, M., Bonnin, S., Ren, H., Manivannan, N., Magazzeni, S., Tadayoni, R., Cochener, B., Lamard, M. & Quellec, G., Multimodal information fusion for the diagnosis of diabetic retinopathy. RITS 2022.
4. Quellec, G., Li, Y., Al Hajj, H., Bonnin, S., Ren, H., Manivannan, N., Magazzeni, S., Tadayoni, R., Conze, P.-H. & Lamard, M., 3-D style transfer between structure and flow channels in OCT angiography. ARVO 2022.
5. El Habib Daho, M., Zeghlache, R., Li, Y., Le Boité, H., Bonnin, S., Magazzeni, S., Borderie, L., Lay, B., Tadayoni, R., Cochener, B., Conze, P.-H., Lamard, M. & Quellec, G., Performance of two ultra-

widefield retinal imaging systems for the automatic diagnosis of diabetic retinopathy. ARVO 2023.

6. Zeghlache, R., Conze, P.-H., El Habib Daho, M., Li, Y., Brahim, I., Le Boité, H., Massin, P., Tadayoni, R., Cochener, B., Quellec, G. & Lamard, M., Time-aware deep models for predicting diabetic retinopathy progression. ARVO 2023.

Challenges

As an extension of my thesis work, I participated in several challenges related to ophthalmic pathology or multimodal fusion during my PhD. I have been able to improve my work on the project by using the methods used in the challenges.

1. MICCAI DRAC2022: Diabetic Retinopathy Analysis Challenge 2022 [1]
   With rising popularity, OCT angiography (OCTA) has the capability of visualizing the retinal and choroidal vasculature at a microvascular level in great detail [443]. Specially, swept-source (SS)-OCTA allows additionally the individual assessment of the choroidal vasculature. Further, ultra-wide optical coherence tomography angiography imaging (UW-OCTA) modality showed higher burden of pathology in the retinal periphery that was not captured by typical OCTA [444]. However, there are currently no works capable of automatic DR analysis using UW-OCTA. Thus, it is crucial to build a flexible and robust model to realize automatic image quality assessment, lesion segmentation and PDR detection. In order to promote the application of machine learning and deep learning algorithms in (1) automatic le-

1. https://drac22.grand-challenge.org/

sion segmentation, (2) image quality assessment and (3) PDR detection using UW-OCTA images, and promote the application of corresponding technologies in clinical diagnosis of DR, DRAC2022 provide a standardized ultra-wide (swept-source) optical coherence tomography angiography (UW-OCTA) data set for testing the effectiveness of various algorithms.

In the challenge, we achieved **fifth place** in the segmentation task using nnU-Net and Vnet. Inspire by the semi-supervised learning, we developed a pseudo-labeling method based on FixMatch method (Section 5.2.2). In the classification task, the pseudo labeling learning method we proposed significantly improved the performance of the model. Our team achieved **fourth place** out of 45 teams in task two (image quality detection), and **third place** out of 45 teams in task three (diabetic retinopathy classification).

2. MICCAI MMAC2023: Myopic Maculopathy Analysis Challenge 2023 [2]

Myopia is a common eye disease that affects large populations in the world [445]. More seriously, myopia may further develop into high myopia in which the visual impairment mainly results from the development of different types of myopic maculopathy [443, 446]. According to the severity, myopic maculopathy can be classified into five categories: no macular lesions, tessellated fundus, diffuse chorioretinal atrophy, patchy chorioretinal atrophy and macular atrophy [447]. In addition, three additional "Plus" lesions are also defined and added to these categories: lacquer cracks (Lc), choroidal neovascularization (CNV), and Fuchs spot (Fs). Aiming to advance the state-of-the-art in automatic myopic maculopathy analysis, MMAC2023

2. https://zenodo.org/record/7866585

organize the myopic maculopathy analysis challenge. The challenge encourages researchers to develop algorithms for different tasks in myopic maculopathy analysis using fundus photography, including (1) classification of myopic maculopathy, (2) segmentation of myopic maculopathy plus lesions and (3) spherical equivalent prediction.

Through the use of the pretext task (Section 5.2.1), we were able to improve the classification performance of our model through self-supervised learning. Among the 17 teams that participated in the challenge, our pretrained model based on Pretext task helped us to achieve the **eighth place** in task one (classification of myopic maculopathy). In addition, we proposed a MAnet-based Test Time Augmentation (TTA) method (Section 6.2.2), which achieved the **second place** in task two (segmentation of myopic maculopathy plus lesions). Finally, in task three (spherical equivalent prediction), we were able to achieve **first place** with our multi-model ensemble method.

3. MICCAI2023 STAGE Challenge: Structural-Functional Transition in Glaucoma Assessment [3]

   STAGE challenge uses OCT images centered on the fovea to predict the results of the Visual Field (VF) test. OCT is the most widely used imaging method in ophthalmic examination. VF test is a reference standard examination to assess visual function. This is a subjective test that requires the subject to remain calm and focused and cooperate with the doctor. Monocular perimetry takes about 15 minutes. In contrast, a monocular OCT scan takes only about three seconds. Therefore, STAGE challenge focuses on how to use objective and easily accessible OCT images of structures to predict functional VF

3. https://aistudio.baidu.com/competition/detail/968/0/introduction

information. Based on this research, three VF information prediction tasks are proposed: (1) Prediction of Mean Deviation (MD), (2) Sensitivity map prediction and (3) Pattern deviation probability map prediction.

Our proposed single-level fusion network (Section 2.3.4) utilizes the complementary information of the 3D OCT data and the 1D glaucoma classification annotations to perform well on different tasks. Our team achieved **first place** in the preliminary round among 22 teams.

4. ISBI2023 APIS: A Paired CT-MRI Dataset for Ischemic Stroke Segmentation Challenge [4]

   Stroke represents the second leading cause of mortality worldwide. The key component for immediate diagnosis is the localization (over CT scans) and delineation of lesions (over MRI studies). The lesions are nonetheless poorly delineated, only visible at advanced stages, and analysis uses manual delineation. This challenge introduces a paired dataset of CT and ADC studies. The researchers are invited to propose computational strategies that approach paired data, during training, and deal with lesion segmentation over CT onset sequences.

   To resolve the incomplete multimodal problem, we proposed a Transformer-based Volume-to-Volume GAN network to generate the corresponding MRI volume based on 3D CT images. We then performed ischemic stroke segmentation using input fusion method (Section 2.3.4) and achieved **fifth place** among 41 teams.

5. MICCAI2022 GOALS Challenge: Glaucoma Oct Analysis and Layer Segmentation [5]

4. https://bivl2ab.uis.edu.co/challenges/apis
5. https://aistudio.baidu.com/competition/detail/230/0/introduction

Optical Coherence Tomography (OCT) is a powerful tool for the diagnosis of ocular diseases, since the image acquisition consists in a contactless, non-invasive method which gives a set of images of the main retinal structures in real time. Segmentation and quantification of layer thickness is useful in the diagnosis of many retinal and optic nerve disorders, for example, glaucoma, macular degeneration or diabetic retinopathy. In the diagnosis of glaucoma, it is easier to detect early cases using OCT than using fundus color images. GOALS design two tasks around OCT images: (1) A segmentation task to determine the retinal nerve fiber layer, ganglion cellinner plexiform layer, and choroidal layer, which are helpful for diagnosis and differentiation of glaucoma and (2) An automatic diagnosis task of glaucoma.

As a result of using a multi-model ensemble and an adaptive post-processing method, our proposed approach performs well on different tasks, resulting in a final ranking of **10th place** out of 100 participating teams.

6. MICCAI HECKTOR2022: HEad and neCK TumOR segmentation and outcome prediction 2022 [6]

   Head and Neck (H&N) cancers are among the most common cancers worldwide [448]. Recently, several radiomics studies based on Positron Emission Tomography (PET) and Computed Tomography (CT) imaging were proposed to better identify patients with a worse prognosis in a non-invasive fashion and by exploiting already available images such as these acquired for diagnosis and treatment planning [449–451]. Two tasks are proposed by HECKTOR: (1) Primary tumor (GTVp) and lymph nodes (GTVn) segmentation in PET/CT

6. https://hecktor.grand-challenge.org/

images and (2) Recurrence-Free Survival (RFS) prediction relying on PET/CT images and/or available clinical information.

As a multimodal segmentation task, we used nn-UNet-based [452] input fusion method to exploit the complementary information of PET volume and CT volume for tumor segmentation with excellent results.

APPENDIX

Table 7.2: List of publications for different fusion networks.

Research work	Year	Multimodal combination	Fusion Methods	Information Fusion Technique	DL Backbone	Final Classifier	Body Organ	Dataset
[227]	2015	MRI, PET, CSF	Input Fusion	Concatenation (Inputs)	RBM	SVM	Brain	ADNI
[186]	2015	MRI, PET	Input Fusion	Concatenation (Inputs)	Manual	Softmax	Brain	ADNI
[198]	2018	sMRI, fMRI	Input Fusion	Concatenation (Inputs)	DBN	FC Layer	Brain	ABIDE
[246]	2020	MRI (ADC, DWI, T2)	Input Fusion	Concatenation (Inputs)	CNN	FC Layer	Prostate	TCIA
[143]	2020	US (US B-mode, US color Doppler)	Input Fusion	Concatenation (Inputs)	CNN	FC Layer	Breast	Private Data
[248]	2020	MRI (ADC, DWI, T2)	Input Fusion	Concatenation (Inputs)	CNN	FC Layer	Prostate	TCIA
[453]	2020	MRI (T2, ADC, High-b)	Input Fusion	Merge (Inputs)	Resnet	FC Layer	Prostate	TCIA
[247]	2021	MRI, PET	Input Fusion	Concatenation (Inputs)	CNN	FC Layer	Brain	ADNI
[250]	2021	MRI, PET	Input Fusion	Merge (Inputs)	CNN	FC Layer	Brain	ADNI
[184]	2021	MRI (T1, T1C, T2M FLAIR)	Input Fusion	Merge (Inputs)	CNN	FC Layer	Brain	TCIA, BraTS
[140]	2022	MRI, PET	Input Fusion	Merge (Inputs)	CNN	FC Layer	Brain	ADNI
[141]	2022	MRI, PET	Input Fusion	Concatenation (Inputs)	Resnet	FC Layer	Brain	ADNI
[249]	2023	MRI (PEI, DWI)	Input Fusion	Concatenation (Inputs)	CNN	FC Layer	Breast	Private Data
[234]	2013	MRI, PET, CSF	Single-level Fusion	Concatenation (Classic)	SAE	SVM	Brain	ADNI
[243]	2014	MRI, PET	Single-level Fusion	Concatenation (Classic)	DBM	SVM	Brain	ADNI
[235]	2015	MRI, PET	Single-level Fusion	Concatenation (Classic)	SAE	SVM	Brain	ADNI

[253]	2016	Photograph of the cervix, Pap tests, HPV tests	Single-level Fusion	Concatenation (Classic)	CNN	FC Layer	Cervix	TCIA
[148]	2017	MRI (ADC, DWI, DCE)	Single-level Fusion	Concatenation (Classic)	CNN	FC Layer	Prostate	TCIA
[454]	2017	MRI, PET	Single-level Fusion	Concatenation (Classic)	CNN	FC Layer	Brain	ADNI
[224]	2017	sMRI, fMRI	Single-level Fusion	Concatenation (Classic)	CNN	FC Layer	Brain	ADHD-200
[147]	2017	MRI (ADC, T2)	Single-level Fusion	Merge (Classic)	CNN	FC Layer	Prostate	Private Data
[225]	2017	MRI, PET	Single-level Fusion	Concatenation (Network)	MM-SDPN	FC Layer	Brain	ADNI
[256]	2017	MRI, PET	Single-level Fusion	Concatenation (Network)	DNN	Score Merge	Brain	ADNI
[185]	2017	MRI (TI, T2, T1C, FLAIR)	Single-level Fusion	Merge (Classic)	CNN	FC Layer	Brain	BraTS
[257]	2017	MRI, PET	Single-level Fusion	Concatenation (Network)	CNN	FC Layer	Brain	ADNI
[146]	2017	MRI (ADC, T2WI)	Single-level Fusion	Merge (Classic)	CNN	SVM	Prostate	TCIA, Private Data
[438]	2018	MRI, PET	Single-level Fusion	Concatenation (Classic)	CNN	FC Layer	Brain	ADNI
[455]	2018	MRI (TI, T2, FLAIR)	Single-level Fusion	Merge (Classic)	CNN	FC Layer	Brain	BraTS, TCIA
[142]	2018	MRI, PET	Single-level Fusion	Concatenation (Classic)	CNN	FC Layer	Brain	ADNI
[223]	2018	MRI (ADC, T2)	Single-level Fusion	Concatenation (Classic)	CNN	Softmax	Prostate	TCIA
[226]	2018	MRI, PET, CSF	Single-level Fusion	Concatenation (Network)	sELM-AE	KELM	Brain	ADNI
[190]	2018	Dsc, Clinical Image, Metadata	Single-level Fusion	Concatenation (Classic)	CNN	FC Layer	Skin	SPC
[456]	2018	MRI, PET	Single-level Fusion	Concatenation (Network)	DNN	Score Merge	Brain	ADNI
[260]	2018	MRI, PET, SNP	Single-level Fusion	Concatenation (Network)	DNN	Score Merge	Brain	ADNI
[457]	2019	MRI, PET	Single-level Fusion	Concatenation (Classic)	CNN, LSTM	Softmax	Brain	ADNI

[144]	2019	MRI (ADC, T2, TWIST)	Single-level Fusion	Concatenation (Classic)	CNN	Random Forest	Breast	Private Data
[203]	2019	MRI, PET	Single-level Fusion	Concatenation (Classic)	VGG	FC Layer	Brain	ADNI
[458]	2019	MRI, CSF, Demographic Information, Cognitive Performance	Single-level Fusion	Concatenation (Classic)	GRU	LR	Brain	ADNI
[255]	2019	MRI, PET	Single-level Fusion	Merge (Classic)	CNN	FC Layer	Brain	ADNI
[275]	2019	MRI, PET	Single-level Fusion	Concatenation (Classic)	CNN	Softmax	Brain	ADNI
[188]	2020	PET, CT	Single-level Fusion	Merge (Classic)	CNN	FC Layer	Lung	Private Data
[439]	2020	MRI, PET	Single-level Fusion	Concatenation (Classic)	CNN	FC Layer	Brain	ADNI
[228]	2020	MRI, PET, Clinical Datas	Single-level Fusion	Concatenation (Classic)	CNN, LSTM	FC Layer	Brain	ADNI
[214]	2021	MRI, EHR, SNP	Single-level Fusion	Concatenation (Classic)	CNN, SAE	FC Layer	Brain	ADNI
[145]	2021	US (US B-mode, US color Doppler, US elastography images)	Single-level Fusion	Concatenation (Classic)	CNN	FC Layer	Breast	Private Data
[459]	2021	MRI (T1, FA, MD)	Single-level Fusion	Concatenation (Classic)	ResNet	FC Layer	Brain	OASIS
[460]	2021	MRI, PET, SNP	Single-level Fusion	Concatenation (Classic)	CNN	FC Layer	Brain	ADNI
[461]	2021	CT, EMR	Single-level Fusion	Concatenation (Classic)	CNN, HoFN	FC Layer	Lung	Private Data
[440]	2021	MRI, PET	Single-level Fusion	Concatenation (Classic)	mDSNet	FC Layer	Brain	ADNI
[236]	2021	EMR, Pathological images	Single-level Fusion	Concatenation (Classic)	VGG, AE	FC Layer	Breast	Private Data
[462]	2021	MRI, PET	Single-level Fusion	Concatenation (Network)	CNN	Score Merge	Brain	ADNI

[254]	2021	MRI (T1, T2), Clinical Information	Single-level Fusion	Concatenation (Classic)	ResNet	FC Layer	Breast	Private Data
[251]	2021	sMRI, fMRI, SNP	Single-level Fusion	Concatenation (Classic)	DNN	FC Layer	Brain	COBRE
[276]	2022	MRI, PET	Single-level Fusion	Concatenation (Network)	CNN	FC Layer	Brain	ADNI
[463]	2022	Echocardiography, CMR	Single-level Fusion	Merge (Network)	CNN	SVM	Heart	Private Data
[419]	2022	DXA, CFP	Single-level Fusion	Concatenation (Network)	CNN	FC Layer	Heart	Private Data
[258]	2022	sMRI, fMRI, Genomic Sequence	Single-level Fusion	Merge (Network)	AE, MLP, LSTM	Softmax	Brain	COBRE
[17]	2022	CFP, OCT	Single-level Fusion	Concatenation (Classic)	CNN	FC Layer	Eye	GAMMA
[259]	2022	MRI, PET	Single-level Fusion	Concatenation (Network)	ResNet	FC Layer	Brain	ADNI
[252]	2022	VF, OCT	Single-level Fusion	Merge (Classic)	CNN	FC Layer	Eye	Private Data
[150]	2022	VF, CFP	Single-level Fusion	Concatenation (Classic)	CNN	FC Layer	Eye	Private Data
[237]	2022	PET-AV45, PET-FDG	Single-level Fusion	Concatenation (Classic)	ViT	FC Layer	Brain	ADNI
[464]	2022	sMRI, fMRI, SNP	Single-level Fusion	Concatenation (Classic)	CNN	FC Layer	Brain	ADNI
[261]	2018	White Light RGB, NBI	Hierarchical Fusion	Merge (Network), Concatenation (Classic)	CNN	FC Layer	Digestive tract	ISIT-UMR
[262]	2020	MRI, PET	Hierarchical Fusion	Merge (Network), Merge (Classic)	CNN	Softmax	Brain	ADNI
[229]	2021	MRI, PET	Hierarchical Fusion	Merge (Network),Merge (Classic)	RBM	Softmax	Brain	ADNI
[242]	2021	MRI (T1C, FLAIR)	Hierarchical Fusion	Merge (Network), Merge (Classic)	CNN	FC Layer	Brain	TCIA, BraTS

[263]	2021	MRI, PET	Hierarchical Fusion	Concatenation (Network), Concatenation (Classic)	CNN	FC Layer	Brain	ADNI
[16]	2022	CFP, OCT	Hierarchical Fusion	Merge (Network), Concatenation (Classic)	CNN	FC Layer	Eye	GAMMA, Private Data
[199]	2022	sMRI, fMRI	Hierarchical Fusion	Merge (Network), Concatenation (Classic)	MGF	FC Layer	Brain	ABIDE, ADHD-200, COBRE
[264]	2021	MRI (T1,T2)	Attention-based Fusion	Attention Fusion	TransMed	FC Layer	Parotid, Knee	MRNet
[465]	2022	Visual Data, Clinical Data	Attention-based Fusion	Attention Fusion	CLIMAT	FFN	Brain, Knee	ADNI
[267]	2022	Neuroimaging Data, Clinical Data	Attention-based Fusion	Attention Fusion	3MT	FC Layer	Brain	ADNI
[266]	2022	Image, Text	Attention-based Fusion	Attention Fusion	MMIF	FC Layer	Uterus	CTU-UHB
[268]	2022	Multi-parametric MRI	Attention-based Fusion	Attention Fusion	AGDAF	FC Layer	Liver	Private Data
[269]	2022	MRI (T1,T2)	Attention-based Fusion	Attention Fusion	MMNet	FC Layer	Parotid, Prostate	MRNet, TCIA
[265]	2022	Genomic Data, Pathology Data	Attention-based Fusion	Attention Fusion	AHM-Fusion	FC Layer	Brain, Lung	TCGA
[272]	2020	sMRI, fMRI	Output Fusion	Merge (Outputs)	CNN	SVM, KNN, LDA	Brain	ADHD-200
[466]	2020	MRI (T1C, T2), Clinical Data	Output Fusion	Merge (Outputs)	ResNet	Bagging	Kidney	Private Data
[270]	2020	US (ROI, Tumor image, TSI, Fused image)	Output Fusion	Merge (Outputs)	CNN	Score Merge	Breast	BUSI [467]

[187]	2020	MRI, PET	Output Fusion	Merge (Outputs)	CNN	AdaBoost	Brain	ADNI
[468]	2020	CT, EMR	Output Fusion	Merge (Outputs)	DNN, CNN	Score Merge	Lung	Private Data
[469]	2021	MRI, SNP	Output Fusion	Merge (Outputs)	CNN, MLP	Ensemble Gate	Brain	ADNI
[149]	2022	CFP, Clinical Data	Output Fusion	Merge (Outputs)	ResNet	XGBoost	Eye	Private Data
[213]	2022	MRI, EHR	Output Fusion	Merge (Outputs)	CNN, AE	Score Merge	Brain	ADNI
[271]	2022	MRI (TI, T2, T1C, FLAIR)	Output Fusion	Merge (Outputs)	CNN	Score Merge	Brain	CPM-RadPath
[273]	2022	Laryngeal image, Voice	Output Fusion	Merge (Outputs)	CNN	Decision Tree	Larynx	Private Data
[245]	2020	MRI (DCE, T2)	Input Fusion, Single-level Fusion, Output Fusion	Merge (Input), Merge (Classic), Merge (Outputs)	VGG	Score Merge	Breast	Private Data
[244]	2022	WSI, MRI (T1, T1-Gd, T2, FLAIR)	Input Fusion, Output Fusion	Concatenation (Input), Merge (Outputs)	CNN	Score Merge	Brain	CPM-RadPath
[189]	2022	Dsc, Clinical Image, Metadata	Single-level Fusion, Output Fusion	Concatenation (Classic), Merge (Outputs)	CNN	Score Merge	Skin	SPC

BIBLIOGRAPHY

1. Roglic, G., *Global report on diabetes* (World Health Organization, 2016).
2. Alam, U., Asghar, O., Azmi, S. & Malik, R. A., General aspects of diabetes mellitus, *Handbook of clinical neurology* **126,** 211–222 (2014).
3. Forouhi, N. G. & Wareham, N. J., Epidemiology of diabetes, *Medicine* **38,** 602–606 (2010).
4. Bastaki, S., Diabetes mellitus and its treatment, *Dubai Diabetes And Endocrinology Journal* **13,** 111–134 (2005).
5. Association, A. D. *et al.,* Gestational diabetes mellitus, *Diabetes care* **27,** S88 (2004).
6. Kanguru, L., Bezawada, N., Hussein, J. & Bell, J., The burden of diabetes mellitus during pregnancy in low-and middle-income countries: a systematic review, *Global Health Action* **7,** 23987 (2014).
7. Stitt, A. W. *et al.,* The progress in understanding and treatment of diabetic retinopathy, *Progress in retinal and eye research* **51,** 156–186 (2016).
8. Derakhshan, R. *et al.,* Increased circulating levels of SDF-1 (CXCL12) in type 2 diabetic patients are correlated to disease state but are unrelated to polymorphism of the SDF-1β gene in the Iranian population, *Inflammation* **35,** 900–904 (2012).
9. Guariguata, L., Whiting, D., Weil, C. & Unwin, N., The International Diabetes Federation diabetes atlas methodology for estimating global and national prevalence of diabetes in adults, *Diabetes research and clinical practice* **94,** 322–332 (2011).
10. Rathmann, W. & Giani, G., Global Prevalence of Diabetes: Estimates for the Year 2000 and Projections for 2030: Response to Wild et al. *Diabetes care* **27,** 2568–2569 (2004).
11. Sivaprasad, S., Gupta, B., Crosby-Nwaobi, R. & Evans, J., Prevalence of diabetic retinopathy in various ethnic groups: a worldwide perspective, *Survey of ophthalmology* **57,** 347–370 (2012).

12. Yau, J. W. *et al.,* Global prevalence and major risk factors of diabetic retinopathy, *Diabetes care* **35,** 556–564 (2012).

13. Zheng, Y., He, M. & Congdon, N., The worldwide epidemic of diabetic retinopathy, *Indian journal of ophthalmology* **60,** 428 (2012).

14. Heintz, E., Wiréhn, A.-B., Peebo, B. B., Rosenqvist, U. & Levin, L.-Å., Prevalence and healthcare costs of diabetic retinopathy: a population-based register study in Sweden, *Diabetologia* **53,** 2147–2154 (2010).

15. Chakrabarti, R., Harper, C. A. & Keeffe, J. E., Diabetic retinopathy management guidelines, *Expert review of ophthalmology* **7,** 417–439 (2012).

16. Li, Y. *et al., Multimodal information fusion for glaucoma and diabetic retinopathy classification in Ophthalmic Medical Image Analysis: 9th International Workshop, OMIA 2022, Held in Conjunction with MICCAI 2022, Singapore, Singapore, September 22, 2022, Proceedings* (2022), 53–62.

17. Wu, J. *et al.,* Gamma challenge: glaucoma grading from multi-modality images, *arXiv preprint arXiv:2202.06511* (2022).

18. Group, E. T. D. R. S. R. *et al.,* Fundus photographic risk factors for progression of diabetic retinopathy: ETDRS report number 12, *Ophthalmology* **98,** 823–833 (1991).

19. Group, E. T. D. R. S. R. *et al.,* Grading diabetic retinopathy from stereoscopic color fundus photographs—an extension of the modified Airlie House classification: ETDRS report number 10, *Ophthalmology* **98,** 786–806 (1991).

20. Williams, G. A. *et al.,* Single-field fundus photography for diabetic retinopathy screening: a report by the American Academy of Ophthalmology, *Ophthalmology* **111,** 1055–1062 (2004).

21. Tanguy, T., *Détection automatisée par apprentissage profond de pathologies rétiniennes en tomographie par cohérence optique et en rétinophotographies* fr, PhD thesis (Université de Bretagne Occidentale, Brest, 2018).

22. Qummar, S. *et al.,* A Deep Learning Ensemble Approach for Diabetic Retinopathy Detection, *IEEE Access* **7,** 150530–150539 (2019).

23. Wang, W. & Lo, A. C., Diabetic retinopathy: pathophysiology and treatments, *International journal of molecular sciences* **19,** 1816 (2018).

24. Mohamed, Q., Gillies, M. C. & Wong, T. Y., Management of diabetic retinopathy: a systematic review, *Jama* **298,** 902–916 (2007).

25. Chew, E. Y. *et al.,* Association of elevated serum lipid levels with retinal hard exudate in diabetic retinopathy: Early Treatment Diabetic Retinopathy Study (ETDRS) Report 22, *Archives of ophthalmology* **114,** 1079–1084 (1996).

26. Wang, M., Garg, I. & Miller, J. B., *Wide field swept source optical coherence tomography angiography for the evaluation of proliferative diabetic retinopathy and associated lesions: a review in Seminars in Ophthalmology* **36** (2021), 162–167.

27. Taylor, R. & Batey, D., *Handbook of retinal screening in diabetes: diagnosis and management* (John Wiley & Sons, 2012).

28. Alyoubi, W. L., Shalash, W. M. & Abulkhair, M. F., Diabetic retinopathy detection through deep learning techniques: A review, *Informatics in Medicine Unlocked* **20,** 100377 (2020).

29. Scanlon, P. H., Sallam, A. & Van Wijngaarden, P., *A practical manual of diabetic retinopathy management* (John Wiley & Sons, 2017).

30. Bandello, F., Zarbin, M. A., Lattanzio, R. & Zucchiatti, I., *Clinical strategies in the management of diabetic retinopathy* (Springer, 2016).

31. Massin, P. *et al.,* Recommendations of the ALFEDIAM (French Association for the Study of Diabetes and Metabolic Diseases) for the screening and surveillance of diabetic retinopathy, *Journal francais d'ophtalmologie* **20,** 302–310 (1997).

32. Wilkinson, C. P. *et al.,* Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales, *Ophthalmology* **110,** 1677–1682 (2003).

33. Islam, M. M., Yang, H.-C., Poly, T. N., Jian, W.-S. & Li, Y.-C. J., Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis, *Computer Methods and Programs in Biomedicine* **191,** 105320 (2020).

34. Falavarjani, K. G., Wang, K., Khadamy, J. & Sadda, S. R., Ultra-wide-field imaging in diabetic retinopathy; an overview, *Journal of Current Ophthalmology* **28,** 57–60 (2016).

35. Liu, T. A. & Arevalo, J. F., Wide-field imaging in proliferative diabetic retinopathy, *International Journal of Retina and Vitreous* **5,** 1–4 (2019).

36. Silva, P. S. *et al.,* Diabetic retinopathy severity and peripheral lesions are associated with nonperfusion on ultrawide field angiography, *Ophthalmology* **122,** 2465–2472 (2015).

37. Li, J. *et al.,* Ultra-widefield color fundus photography combined with high-speed ultra-widefield swept-source optical coherence tomography angiography for non-invasive detection of lesions in diabetic retinopathy, *Frontiers in Public Health* **10** (2022).

38. Huang, D. *et al.,* Optical coherence tomography, *science* **254,** 1178–1181 (1991).

39. Virgili, G. *et al.,* Optical coherence tomography (OCT) for detection of macular oedema in patients with diabetic retinopathy, *Cochrane Database of Systematic Reviews* (2015).

40. Liu, G., Xu, D. & Wang, F., New insights into diabetic retinopathy by OCT angiography, *Diabetes Research and Clinical Practice* **142,** 243–253 (2018).

41. Lains, I. *et al.,* Retinal applications of swept source optical coherence tomography (OCT) and optical coherence tomography angiography (OCTA), *Progress in Retinal and Eye Research* **84,** 100951 (2021).

42. Cui, Y. *et al.,* Comparison of widefield swept-source optical coherence tomography angiography with ultra-widefield colour fundus photography and fluorescein angiography for detection of lesions in diabetic retinopathy, *British Journal of Ophthalmology* **105,** 577–581 (2021).

43. Russell, J. F. *et al.,* Longitudinal wide-field swept-source OCT angiography of neovascularization in proliferative diabetic retinopathy after panretinal photocoagulation, *Ophthalmology Retina* **3,** 350–361 (2019).

44. Pichi, F. *et al.,* Wide-field optical coherence tomography angiography for the detection of proliferative diabetic retinopathy, *Graefe's Archive for Clinical and Experimental Ophthalmology* **258,** 1901–1909 (2020).

45. Khalid, H. *et al.,* Widefield optical coherence tomography angiography for early detection and objective evaluation of proliferative diabetic retinopathy, *British Journal of Ophthalmology* **105,** 118–123 (2021).

46. Sawada, O. *et al.,* Comparison between wide-angle OCT angiography and ultra-wide field fluorescein angiography for detecting non-perfusion areas and retinal neovascularization in eyes with diabetic retinopathy, en, *Graefe's Archive for Clinical and Experimental Ophthalmology* **256,** 1275–1280, ISSN: 0721-832X, 1435-702X, `http://link.springer.com/10.1007/s00417-018-3992-y` (2022) (July 2018).

47. Shiraki, A. *et al.,* Evaluation of retinal nonperfusion in branch retinal vein occlusion using wide-field optical coherence tomography angiography, *Acta ophthalmologica* **97,** e913–e918 (2019).

48. Li, M. *et al.,* Different scan areas affect the detection rates of diabetic retinopathy lesions by high-speed ultra-widefield swept-source optical coherence tomography angiography, *Frontiers in Endocrinology* **14,** 350 (2023).

49. Hirano, T. *et al.,* Wide-field en face swept-source optical coherence tomography angiography using extended field imaging in diabetic retinopathy, *British Journal of Ophthalmology* **102,** 1199–1203 (2018).

50. Xuan, Y. *et al.,* Clinical observation of choroidal osteoma using swept-source optical coherence tomography and optical coherence tomography angiography, *Applied Sciences* **12,** 4472 (2022).

51. Zhang, W. *et al.,* Advanced ultrawide-field optical coherence tomography angiography identifies previously undetectable changes in biomechanics-related parameters in nonpathological myopic fundus, *Frontiers in Bioengineering and Biotechnology* **10** (2022).

52. Lombardo, M., Serrao, S., Devaney, N., Parravano, M. & Lombardo, G., Adaptive optics technology for high-resolution retinal imaging, *Sensors* **13,** 334–366 (2012).

53. Arichika, S. *et al.,* Retinal hemorheologic characterization of early-stage diabetic retinopathy using adaptive optics scanning laser ophthalmoscopy, *Investigative ophthalmology & visual science* **55,** 8513–8522 (2014).

54. Roorda, A. & Williams, D. R., The arrangement of the three cone classes in the living human eye, *Nature* **397,** 520–522 (1999).

55. Tam, J., Tiruveedhula, P. & Roorda, A., Characterization of single-file flow through human retinal parafoveal capillaries using an adaptive optics scanning laser ophthalmoscope, *Biomedical optics express* **2,** 781–793 (2011).

56. Uji, A. *et al.,* The source of moving particles in parafoveal capillaries detected by adaptive optics scanning laser ophthalmoscopy, *Investigative ophthalmology & visual science* **53,** 171–178 (2012).

57. Tam, J., Martin, J. A. & Roorda, A., Noninvasive visualization and analysis of parafoveal capillaries in humans, *Investigative ophthalmology & visual science* **51,** 1691–1698 (2010).

58. Chui, T. Y., Gast, T. J. & Burns, S. A., Imaging of vascular wall fine structure in the human retina using adaptive optics scanning laser ophthalmoscopy, *Investigative ophthalmology & visual science* **54,** 7115–7124 (2013).

59. Takayama, K. *et al.,* High-resolution imaging of the retinal nerve fiber layer in normal eyes using adaptive optics scanning laser ophthalmoscopy, *PloS one* **7,** e33158 (2012).

60. Ogura, Y., In vivo evaluation of leukocyte dynamics in the retinal and choroidal circulation, *Japanese journal of ophthalmology* **44,** 322–323 (2000).

61. Puyo, L., Paques, M., Fink, M., Sahel, J.-A. & Atlan, M., In vivo laser Doppler holography of the human retina, *Biomedical optics express* **9,** 4113–4129 (2018).

62. Puyo, L., Paques, M., Fink, M., Sahel, J.-A. & Atlan, M., Waveform analysis of human retinal and choroidal blood flow with laser Doppler holography, *Biomedical optics express* **10,** 4942–4963 (2019).

63. Screening, g., American College Of Physicians, American Diabetes Association And American Academy Of Ophthalmology, *Ann Int Med.* **116,** 683–5 (1992).

64. Group, E. T. D. R. S. R. *et al.,* Early photocoagulation for diabetic retinopathy: ETDRS report number 9, *Ophthalmology* **98,** 766–785 (1991).

65. Gardner, T. W. *et al.,* Diabetic retinopathy: more than meets the eye, *Survey of ophthalmology* **47,** S253–S262 (2002).

66. Klein, R., Klein, B. E. & Moss, S. E., Visual impairment in diabetes, *Ophthalmology* **91,** 1–9 (1984).

67. Sjolie, A. *et al.,* EURODIAB Complications Study Group, Retinopathy and vision loss in insulin-dependent diabetes in Europe, *Ophthalmology* **104,** 252 (1997).

68. Delcourt, C. *et al.,* Visual impairment in type 2 diabetic patients: a multicentre study in France, *Acta Ophthalmologica Scandinavica* **73,** 293–298 (1995).

69. Massin, P. *et al.,* OPHDIAT©: A telemedical network screening system for diabetic retinopathy in the Île-de-France, *Diabetes & metabolism* **34,** 227–234 (2008).

70. Chabouis, A. *et al.,* Benefits of Ophdiat®, a telemedical network to screen for diabetic retinopathy: A retrospective study in five reference hospital centres, *Diabetes & metabolism* **35,** 228–232 (2009).

71. Massin, P. *et al.,* Recommandations de l'ALFEDIAM pour le dépistage et la surveillance de la rétinopathie diabétique, *Journal français d'ophtalmologie* **20,** 302–310 (1997).

72. LAVERSIN, S. & DUROCHER, A., Suivi du patient diabétique de type 2 à l'exclusion du suivi des complications: recommandations de l'ANAES, *Diabetes & metabolism* **25** (1999).

73. Delcourt, C., Massin, P. & Rosilio, M., Epidemiology of diabetic retinopathy: expected vs reported prevalence of cases in the French population, *Diabetes & metabolism* **35,** 431–438 (2009).

74. FAGOT CAMPAGNA, A. *et al.,* Caractéristiques des personnes diabétiques traitées et adéquation du suivi médical du diabète aux recommandations officielles. Entred 2001, *Bulletin épidémiologique hebdomadaire,* 238–239 (2003).

75. Wild, S., Roglic, G., Green, A., Sicree, R. & King, H., Global prevalence of diabetes: estimates for the year 2000 and projections for 2030, *Diabetes care* **27,** 1047–1053 (2004).

76. Des professions de santé (France), M. D. & Berland, Y., *" Démographie des professions de santé"....* (Mission Démographie des professions de santé, 2003).

77. Williams, R., Nussey, S., Humphry, R. & Thompson, G., Assessment of non-mydriatic fundus photography in detection of diabetic retinopathy. *Br Med J (Clin Res Ed)* **293,** 1140–1142 (1986).

78. Pugh, J. A. *et al.,* Screening for diabetic retinopathy: the wide-angle retinal camera, *Diabetes care* **16,** 889–895 (1993).

79. Scanlon, P. H. *et al.,* Comparison of two reference standards in validating two field mydriatic digital photography as a method of screening for diabetic retinopathy, *British journal of ophthalmology* **87,** 1258–1263 (2003).

80. Lin, D. Y., Blumenkranz, M. S., Brothers, R. J., Grosvenor, D. M. & Group, T. D. D. S., The sensitivity and specificity of single-field nonmydriatic monochromatic digital fundus photography with remote image interpretation for diabetic retinopathy screening: a comparison with ophthalmoscopy and standardized mydriatic color photography, *American journal of ophthalmology* **134,** 204–213 (2002).

81. Cavallerano, J. & Aiello, L. M., Emerging trends in ocular telemedicine: the diabetic retinopathy model, *Journal of telemedicine and telecare* **11,** 163–166 (2005).

82. Gómez-Ulla, F. *et al.,* Digital retinal images and teleophthalmology for detecting and grading diabetic retinopathy, *Diabetes Care* **25,** 1384–1389 (2002).

83. Beynat, J. *et al.,* Screening for diabetic retinopathy in a rural French population with a mobile non-mydriatic camera, *Diabetes & metabolism* **35,** 49–56 (2009).

84. Cuadros, J. & Bresnick, G., EyePACS: an adaptable telemedicine system for diabetic retinopathy screening, *Journal of diabetes science and technology* **3,** 509–516 (2009).

85. Lemmetty, R. & Mäkelä, K., Mobile digital fundus screening of type 2 diabetes patients in the Finnish county of South-Ostrobothnia, *Journal of Telemedicine and Telecare* **15,** 68–72 (2009).

86. Nathoo, N., Ng, M., Rudnisky, C. J. & Tennant, M. T., The prevalence of diabetic retinopathy as identified by teleophthalmology in rural Alberta, *Canadian Journal of Ophthalmology* **45,** 28–32 (2010).

87. Gibelalde, A., Ruiz-Miguel, M., Mendicute, J., Ayerdi, S. & Martinez-Zabalegi, D., *Prevalence of diabetic retinopathy using non-mydriatic retinography in Anales del Sistema Sanitario de Navarra* **33** (2010), 271–276.

88. Coquin, Y., Arrêté du 13 juillet 2004 relatif à la pratique de la vaccination par le vaccin antituberculeux BCG et aux tests tuberculiniques, *Médecine thérapeutique/Pédiatrie* **7,** 222–223 (2004).

89. Schulze-Döbold, C., Erginay, A., Robert, N., Chabouis, A. & Massin, P., Ophdiat®: five-year experience of a telemedical screening programme for diabetic retinopathy in Paris and the surrounding area, *Diabetes & metabolism* **38,** 450–457 (2012).

90. Bandello, F., Lattanzio, R., Zucchiatti, I. & Del Turco, C., Pathophysiology and treatment of diabetic retinopathy, *Acta diabetologica* **50,** 1–20 (2013).

91. Saaddine, J. B. *et al.,* Projection of diabetic retinopathy and other major eye diseases among people with diabetes mellitus: United States, 2005-2050, *Archives of ophthalmology* **126,** 1740–1747 (2008).

92. Kozak, I. *et al.,* Clinical evaluation and treatment accuracy in diabetic macular edema using navigated laser photocoagulator NAVILAS, *Ophthalmology* **118,** 1119–1124 (2011).

93. Group, D. R. S. R. *et al.,* Preliminary report on effects of photocoagulation therapy, *American journal of ophthalmology* **81,** 383–396 (1976).

94. Zhang, X., Bao, S., Hambly, B. D. & Gillies, M. C., Vascular endothelial growth factor-A: a multifunctional molecular player in diabetic retinopathy, *The international journal of biochemistry & cell biology* **41,** 2368–2371 (2009).

95. Bandello, F. *et al.,* New approaches for the treatment of diabetic macular oedema: recommendations by an expert panel, *Eye* **26,** 485–493 (2012).

96. Lee, S. S., Hughes, P. M. & Robinson, M. R., Recent advances in drug delivery systems for treating ocular complications of systemic diseases, *Current opinion in ophthalmology* **20,** 511–519 (2009).

97. Nauck, M. *et al.,* Induction of vascular endothelial growth factor by platelet-activating factor and platelet-derived growth factor is downregulated by corticosteroids. *American journal of respiratory cell and molecular biology* **16,** 398–406 (1997).

98. Blankenship, G. W., Evaluation of a single intravitreal injection of dexamethasone phosphate in vitrectomy surgery for diabetic retinopathy complications, *Graefe's archive for clinical and experimental ophthalmology* **229,** 62–65 (1991).

99. Gillies, M. C. *et al.,* Five-year results of a randomized trial with open-label extension of triamcinolone acetonide for refractory diabetic macular edema, *Ophthalmology* **116,** 2182–2187 (2009).

100. Massin, P. *et al.,* Safety and Efficacy of Ranibizumab in Diabetic Macular Edema (RESOLVE Study) A 12-month, randomized, controlled, double-masked, multicenter phase II study, *Diabetes care* **33,** 2399–2405 (2010).

101. Nguyen, Q. D. *et al.,* Ranibizumab for diabetic macular edema: results from 2 phase III randomized trials: RISE and RIDE, *Ophthalmology* **119,** 789–801 (2012).

102. Nguyen, Q. D. *et al.,* Two-year outcomes of the ranibizumab for edema of the mAcula in diabetes (READ-2) study, *Ophthalmology* **117,** 2146–2151 (2010).

103. Paccola, L. *et al.,* Intravitreal triamcinolone versus bevacizumab for treatment of refractory diabetic macular oedema (IBEME study), *British Journal of Ophthalmology* **92,** 76–80 (2008).

104. Kumagai, K. *et al.,* Long-term follow-up of vitrectomy for diffuse nontractional diabetic macular edema, *Retina* **29,** 464–472 (2009).

105. Gandorfer, A., Messmer, E. M., Ulbig, M. W. & Kampik, A., Resolution of diabetic macular edema after surgical removal of the posterior hyaloid and the inner limiting membrane, *Retina* **20,** 126–133 (2000).

106. Yanase, J. & Triantaphyllou, E., A systematic survey of computer-aided diagnosis in medicine: Past and present developments, *Expert Systems with Applications* **138,** 112821 (2019).

107. Doi, K., Computer-aided diagnosis in medical imaging: historical review, current status and future potential, *Computerized medical imaging and graphics* **31,** 198–211 (2007).

108. Fujita, H., AI-based computer-aided diagnosis (AI-CAD): the latest review to read first, *Radiological physics and technology* **13,** 6–19 (2020).

109. Klang, E., Deep learning and medical imaging, *Journal of thoracic disease* **10,** 1325 (2018).

110. Castiglioni, I. *et al.,* AI applications to medical images: From machine learning to deep learning, *Physica Medica* **83,** 9–24 (2021).

111. Litjens, G. *et al.,* A survey on deep learning in medical image analysis, *Medical image analysis* **42,** 60–88 (2017).

112. Chartrand, G. *et al.,* Deep learning: a primer for radiologists, *Radiographics* **37,** 2113–2131 (2017).

113. BRAHIM, I., *Automatic quantification of ocular dryness by artificial intelligence in the context of Sjögren's syndrome* en, PhD thesis (Université de Bretagne Occidentale, Brest, 2022).

114. Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F. & Campbell, J. P., Introduction to machine learning, neural networks, and deep learning, *Translational Vision Science & Technology* **9,** 14–14 (2020).

115. Klang, E., Deep learning and medical imaging, *Journal of Thoracic Disease* **10,** ISSN: 2077-6624, https://jtd.amegroups.com/article/view/19648 (2018).

116. Celard, P. *et al.,* A survey on deep learning applied to medical images: from simple artificial neural networks to generative models, *Neural Computing and Applications* **35,** 2291–2323 (2023).

117. Cai, L., Gao, J. & Zhao, D., A review of the application of deep learning in medical image classification and segmentation, *Annals of translational medicine* **8** (2020).

118. LeCun, Y. *et al.,* Handwritten digit recognition with a back-propagation network, *Advances in neural information processing systems* **2** (1989).

119. Krizhevsky, A., Sutskever, I. & Hinton, G. E., Imagenet classification with deep convolutional neural networks, *Communications of the ACM* **60,** 84–90 (2017).

120. Litjens, G. *et al.,* A survey on deep learning in medical image analysis, *Medical Image Analysis* **42,** 60–88, ISSN: 1361-8415, https://www.sciencedirect.com/science/article/pii/S1361841517301135 (2017).

121. Vaswani, A. *et al.,* Attention is all you need, *Advances in neural information processing systems* **30** (2017).

122. Dong, L., Xu, S. & Xu, B., *Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition in 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (2018), 5884–5888.

123. Li, N., Liu, S., Liu, Y., Zhao, S. & Liu, M., *Neural speech synthesis with transformer network in Proceedings of the AAAI conference on artificial intelligence* **33** (2019), 6706–6713.

124. Vila, L. C., Escolano, C., Fonollosa, J. A. & Costa-Jussa, M. R., *End-to-end speech translation with the transformer. in IberSPEECH* (2018), 60–63.

125. Topal, M. O., Bas, A. & van Heerden, I., Exploring transformers in natural language generation: Gpt, bert, and xlnet, *arXiv preprint arXiv:2102.08036* (2021).

126. Shamshad, F. *et al.,* Transformers in medical imaging: A survey, *Medical Image Analysis,* 102802 (2023).

127. Parvaiz, A. *et al.,* Vision transformers in medical computer vision—A contemplative retrospection, *Engineering Applications of Artificial Intelligence* **122,** 106126 (2023).

128. Dosovitskiy, A. *et al.,* An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint arXiv:2010.11929* (2020).

129. Li, J. *et al.,* Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives, *Medical image analysis,* 102762 (2023).

130. Wang, R. *et al.,* Medical image segmentation using deep learning: A survey, *IET Image Processing* **16,** 1243–1267 (2022).

131. Liu, Z. *et al., A convnet for the 2020s in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), 11976–11986.

132. Naseer, M. M. *et al.,* Intriguing properties of vision transformers, *Advances in Neural Information Processing Systems* **34,** 23296–23308 (2021).

133. Zhai, X., Kolesnikov, A., Houlsby, N. & Beyer, L., *Scaling vision transformers in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), 12104–12113.

134. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C. & Dosovitskiy, A., Do vision transformers see like convolutional neural networks?, *Advances in Neural Information Processing Systems* **34,** 12116–12128 (2021).

135. Chen, J. *et al.,* Transunet: Transformers make strong encoders for medical image segmentation, *arXiv preprint arXiv:2102.04306* (2021).

136. Hatamizadeh, A. *et al., Unetr: Transformers for 3d medical image segmentation in Proceedings of the IEEE/CVF winter conference on applications of computer vision* (2022), 574–584.

137. Chen, J. *et al.,* Transmorph: Transformer for unsupervised medical image registration, *Medical image analysis* **82,** 102615 (2022).

138. Zhang, Z., Yu, L., Liang, X., Zhao, W. & Xing, L., *TransCT: dual-path transformer for low dose computed tomography in Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24* (2021), 55–64.

139. He, K. *et al.,* Transformers in medical image analysis: A review. arXiv 2022, *arXiv preprint arXiv:2202.12165.*

140. Kong, Z. *et al.,* Multi-modal data Alzheimer's disease detection based on 3D convolution, *Biomedical Signal Processing and Control* **75,** 103565 (2022).

141. Zhang, J., He, X., Qing, L., Gao, F. & Wang, B., BPGAN: brain PET synthesis from MRI using generative adversarial network for multi-modal Alzheimer's disease diagnosis, *Computer Methods and Programs in Biomedicine* **217,** 106676 (2022).

142. Liu, M., Cheng, D., Wang, K., Wang, Y. & ADNI, Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis, *Neuroinformatics* **16,** 295–308 (2018).

143. Qian, X. *et al.,* A combined ultrasonic B-mode and color Doppler system for the classification of breast masses using neural network, *European Radiology* **30,** 3023–3033 (2020).

144. Dalmis, M. U. *et al.,* Artificial intelligence–based classification of breast lesions imaged with a multiparametric breast MRI protocol with ultrafast DCE-MRI, T2, and DWI, *Investigative radiology* **54,** 325–332 (2019).

145. Qian, X. *et al.,* Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning, *Nature biomedical engineering* **5,** 522–532 (2021).

146. Le, M. H. *et al.,* Automated diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks, *Physics in Medicine & Biology* **62,** 6497 (2017).

147. Yang, X. *et al.,* Co-trained convolutional neural networks for automated detection of prostate cancer in multi-parametric MRI, *Medical image analysis* **42,** 212–227 (2017).

148. Mehrtash, A. *et al., Classification of clinical significance of MRI prostate findings using 3D convolutional neural networks in Medical imaging 2017: computer-aided diagnosis* **10134** (2017), 589–592.

149. Yoo, T. K. *et al.,* DeepPDT-Net: predicting the outcome of photodynamic therapy for chronic central serous chorioretinopathy using two-stage multimodal transfer learning, *Scientific Reports* **12,** 18689 (2022).

150. Huang, X. *et al.,* Detecting glaucoma from multi-modal data using probabilistic deep learning, *Frontiers in Medicine* **9** (2022).

151. Muhammad, G. *et al.,* A comprehensive survey on multimodal medical signals fusion for smart healthcare systems, *Information Fusion* **76,** 355–375 (2021).

152. Azam, M. A. *et al.,* A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics, *Computers in biology and medicine* **144,** 105253 (2022).

153. Hermessi, H., Mourali, O. & Zagrouba, E., Multimodal medical image fusion review: Theoretical background and recent advances, *Signal Processing* **183,** 108036 (2021).

154. Kline, A. *et al.,* Multimodal machine learning in precision health: A scoping review, *npj Digital Medicine* **5,** 171 (2022).

155. El-Gamal, F. E.-Z. A., Elmogy, M. & Atwan, A., Current trends in medical image registration and fusion, *Egyptian Informatics Journal* **17,** 99–124 (2016).

156. Stokking, R., Zuiderveld, K. J. & Viergever, M. A., Integrated volume visualization of functional image data and anatomical surfaces using normal fusion, *Human Brain Mapping* **12,** 203–218 (2001).

157. Bhatnagar, G., Wu, Q. J. & Liu, Z., Directive contrast based multimodal medical image fusion in NSCT domain, *IEEE transactions on multimedia* **15,** 1014–1024 (2013).

158. He, C., Liu, Q., Li, H. & Wang, H., Multimodal medical image fusion based on IHS and PCA, *Procedia Engineering* **7,** 280–285 (2010).

159. Bashir, R., Junejo, R., Qadri, N. N., Fleury, M. & Qadri, M. Y., SWT and PCA image fusion methods for multi-modal imagery, *Multimedia Tools and Applications* **78,** 1235–1263 (2019).

160. Princess, M. R., Kumar, V. S. & Begum, M. R., Comprehensive and comparative study of different image fusion techniques, *Int. J. Adv. Res. Electr. Electron. Instrum. Eng,* 11800–11806 (2014).

161. Parmar, K. & Kher, R., *A comparative analysis of multimodality medical image fusion methods in 2012 Sixth Asia Modelling Symposium* (2012), 93–97.

162. Sadjadi, F., *Comparative image fusion analysais in 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)-workshops* (2005), 8–8.

163. Das, S. & Kundu, M. K., A neuro-fuzzy approach for medical image fusion, *IEEE transactions on biomedical engineering* **60,** 3347–3353 (2013).

164. Liu, Y., Yang, J. & Sun, J., *PET/CT medical image fusion algorithm based on multiwavelet transform in 2010 2nd International Conference on Advanced Computer Control* **2** (2010), 264–268.

165. Xi, X.-X., Luo, X.-Q., Zhang, Z.-C., You, Q.-J. & Wu, X., *Multimodal medical volumetric image fusion based on multi-feature in 3-D shearlet transform in 2017 International Smart Cities Conference (ISC2)* (2017), 1–6.

166. Zhang, Q., Liu, Y., Blum, R. S., Han, J. & Tao, D., Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review, *Information Fusion* **40,** 57–75 (2018).

167. Zhu, Z., Zheng, M., Qi, G., Wang, D. & Xiang, Y., A phase congruency and local Laplacian energy based multi-modality medical image fusion method in NSCT domain, *IEEE Access* **7,** 20811–20824 (2019).

168. Liu, Y., Liu, S. & Wang, Z., A general framework for image fusion based on multi-scale transform and sparse representation, *Information fusion* **24,** 147–164 (2015).

169. Mishra, D. & Palkar, B., Image fusion techniques: a review, *International Journal of Computer Applications* **130,** 7–13 (2015).

170. Bhat, S. & Koundal, D., Multi-focus image fusion techniques: a survey, *Artificial Intelligence Review* **54,** 5735–5787 (2021).

171. Sharma, A. M., Dogra, A., Goyal, B., Vig, R. & Agrawal, S., From pyramids to state-of-the-art: a study and comprehensive comparison of visible–infrared image fusion techniques, *IET Image Processing* **14,** 1671–1689 (2020).

172. Lee, J. *et al.,* Machine learning-based prediction of clinical pain using multimodal neuroimaging and autonomic metrics, *pain* **160,** 550 (2019).

173. Tang, X. *et al.,* Elaboration of a multimodal MRI-based radiomics signature for the preoperative prediction of the histological subtype in patients with non-small-cell lung cancer, *Biomedical engineering online* **19,** 1–17 (2020).

174. Quellec, G., Lamard, M., Cazuguel, G., Roux, C. & Cochener, B., Case retrieval in medical databases by fusing heterogeneous information, *IEEE Transactions on Medical Imaging* **30,** 108–118 (2010).

175. Lalousis, P. A. *et al.,* Heterogeneity and classification of recent onset psychosis and depression: a multimodal machine learning approach, *Schizophrenia bulletin* **47,** 1130–1140 (2021).

176. Ramachandram, D. & Taylor, G. W., Deep multimodal learning: A survey on recent advances and trends, *IEEE signal processing magazine* **34,** 96–108 (2017).

177. Boulahia, S. Y., Amamra, A., Madi, M. R. & Daikh, S., Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition, *Machine Vision and Applications* **32,** 121 (2021).

178. Singh, B., Gautam, R., Kumar, S. & Umapathy, S., Application of vibrational microspectroscopy to biology and medicine (2012).

179. Plewes, D. B. & Kucharczyk, W., Physics of MRI: a primer, *Journal of magnetic resonance imaging* **35,** 1038–1054 (2012).

180. Bailey, D. L., Maisey, M. N., Townsend, D. W. & Valk, P. E., *Positron emission tomography* (Springer, 2005).

181. Buzug, T. M., *Computed tomography* (Springer, 2011).

182. Leighton, T. G., What is ultrasound?, *Progress in biophysics and molecular biology* **93,** 3–83 (2007).

183. MacKie, R., Fleming, C., McMahon, A. & Jarrett, P., The use of the dermatoscope to identify early melanoma using the three-colour test, *British Journal of Dermatology* **146,** 481–484 (2002).

184. Decuyper, M., Bonte, S., Deblaere, K. & Van Holen, R., Automated MRI based pipeline for segmentation and prediction of grade, IDH mutation and 1p19q co-deletion in glioma, *Computerized Medical Imaging and Graphics* **88,** 101831 (2021).

185. Ye, F., Pu, J., Wang, J., Li, Y. & Zha, H., *Glioma grading based on 3D multimodal convolutional neural network and privileged learning in 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (2017), 759–763.

186. Liu, S. *et al.,* Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease, *IEEE transactions on biomedical engineering* **62,** 1132–1140 (2014).

187. Fang, X., Liu, Z. & Xu, M., Ensemble of deep convolutional neural networks based multi-modality images for Alzheimer's disease diagnosis, *IET Image Processing* **14,** 318–326 (2020).

188. Qin, R. *et al.,* Fine-grained lung cancer classification from PET and CT images based on multidimensional attention mechanism, *Complexity* **2020,** 1–12 (2020).

189. Tang, P. *et al.,* FusionM4Net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification, *Medical Image Analysis* **76,** 102307 (2022).

190. Kawahara, J., Daneshvar, S., Argenziano, G. & Hamarneh, G., Seven-point checklist and skin lesion classification using multitask multimodal neural nets, *IEEE journal of biomedical and health informatics* **23,** 538–546 (2018).

191. Pai, P. P. *et al.,* BRAHMA: Population specific t1, t2, and FLAIR weighted brain templates and their impact in structural and functional imaging studies, *Magnetic resonance imaging* **70,** 5–21 (2020).

192. Lindig, T. *et al.,* Evaluation of multimodal segmentation based on 3D T1-, T2- and FLAIR-weighted images–the difficulty of choosing, *Neuroimage* **170,** 210–221 (2018).

193. Hecht, M. *et al.,* MRI-FLAIR images of the head show corticospinal tract alterations in ALS patients more frequently than T2-, T1-and proton-density-weighted images, *Journal of the neurological sciences* **186,** 37–44 (2001).

194. Kuban, D. A. *et al.,* Long-term multi-institutional analysis of stage T1–T2 prostate cancer treated with radiotherapy in the PSA era, *International Journal of Radiation Oncology* Biology* Physics* **57,** 915–928 (2003).

195. Zhou, T., Ruan, S. & Canu, S., A review: Deep learning for medical image segmentation using multi-modality fusion, *Array* **3,** 100004 (2019).

196. Preston, D. C., Magnetic resonance imaging (mri) of the brain and spine: Basics, *MRI Basics, Case Med* **30** (2006).

197. Shen, J.-M., Xia, X.-W., Kang, W.-G., Yuan, J.-J. & Sheng, L., The use of MRI apparent diffusion coefficient (ADC) in monitoring the development of brain infarction, *BMC Medical Imaging* **11,** 1–4 (2011).

198. Akhavan Aghdam, M., Sharifi, A. & Pedram, M. M., Combination of rs-fMRI and sMRI data to discriminate autism spectrum disorders in young children using deep belief network, *Journal of digital imaging* **31,** 895–903 (2018).

199. Liu, R. *et al.,* Attention-Like Multimodality Fusion With Data Augmentation for Diagnosis of Mental Disorders Using MRI, *IEEE Transactions on Neural Networks and Learning Systems* (2022).

200. Bandettini, P. A., Twenty years of functional MRI: the science and the stories, *Neuroimage* **62,** 575–588 (2012).

201. Calhoun, V. D. & Sui, J., Multimodal fusion of brain imaging data: a key to finding the missing link (s) in complex mental illness, *Biological psychiatry: cognitive neuroscience and neuroimaging* **1,** 230–244 (2016).

202. Liu, M., Gao, Y., Yap, P.-T. & Shen, D., Multi-hypergraph learning for incomplete multimodality data, *IEEE journal of biomedical and health informatics* **22,** 1197–1208 (2017).

203. Huang, Y. *et al.,* Diagnosis of Alzheimer's disease via multi-modality 3D convolutional neural network, *Frontiers in neuroscience* **13,** 509 (2019).

204. Xu, H. *et al.,* Joint nnU-Net and Radiomics Approaches for Segmentation and Prognosis of Head and Neck Cancers with PET/CT images, *arXiv preprint arXiv:2211.10138* (2022).

205. Andrearczyk, V. *et al., in Head and Neck Tumor Segmentation and Outcome Prediction: Second Challenge, HECKTOR 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings* 1–37 (Springer, 2022).

206. Muehllehner, G. & Karp, J. S., Positron emission tomography, *Physics in Medicine & Biology* **51,** R117 (2006).

207. Zwiebel, W. J. & Pellerito, J. S., *Introduction to vascular ultrasonography* (Elsevier Saunders Philadelphia, 2005).

208. Abdelgawad, E. A., Abu-samra, M. F., Abdelhay, N. M. & Abdel-Azeem, H. M., B-mode ultrasound, color Doppler, and sonoelastography in differentiation between benign and malignant cervical lymph nodes with special emphasis on sonoelastography, *Egyptian Journal of Radiology and Nuclear Medicine* **51,** 1–10 (2020).

209. Lu, M.-H. *et al.,* A comparative study of clinical value of single B-mode ultrasound guidance and B-mode combined with color doppler ultrasound guidance in mini-invasive percutaneous nephrolithotomy to decrease hemorrhagic complications, *Urology* **76,** 815–820 (2010).

210. Schelling, M. *et al.,* Combined transvaginal B-mode and color Doppler sonography for differential diagnosis of ovarian tumors: results of a multivariate logistic regression analysis, *Gynecologic oncology* **77,** 78–86 (2000).

211. Schelling, M. *et al.,* Optimized differential diagnosis of breast lesions by combined B-mode and color Doppler sonography, *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology* **10,** 48–53 (1997).

212. Ge, Z., Demyanov, S., Chakravorty, R., Bowling, A. & Garnavi, R., *Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images in Medical Image Computing and Computer Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20* (2017), 250–258.

213. Prabhu, S. S. *et al., Multi-Modal Deep Learning Models for Alzheimer's Disease Prediction Using MRI and EHR in 2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE)* (2022), 168–173.

214. Venugopalan, J., Tong, L., Hassanzadeh, H. R. & Wang, M. D., Multimodal deep learning models for early detection of Alzheimer's disease stage, *Scientific reports* **11,** 3254 (2021).

215. Yap, J., Yolland, W. & Tschandl, P., Multimodal skin lesion classification using deep learning, *Experimental dermatology* **27,** 1261–1267 (2018).

216. Tomczak, K., Czerwińska, P. & Wiznerowicz, M., Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge, *Contemporary Oncology/Współczesna Onkologia* **2015,** 68–77 (2015).

217. Sleeman IV, W. C., Kapoor, R. & Ghosh, P., Multimodal classification: Current landscape, taxonomy and future directions, *ACM Computing Surveys* **55,** 1–31 (2022).

218. Cuingnet, R. *et al.,* Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database, *neuroimage* **56,** 766–781 (2011).

219. Davatzikos, C., Bhatt, P., Shaw, L. M., Batmanghelich, K. N. & Trojanowski, J. Q., Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification, *Neurobiology of aging* **32,** 2322–e19 (2011).

220. Kohannim, O. *et al.,* Boosting power for clinical trials using classifiers based on multiple biomarkers, *Neurobiology of aging* **31,** 1429–1442 (2010).

221. Liu, M., Zhang, D., Shen, D., ADNI, *et al.,* Ensemble sparse classification of Alzheimer's disease, *NeuroImage* **60,** 1106–1116 (2012).

222. Liu, M., Zhang, D., Shen, D. & ADNI, Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis, *Human brain mapping* **35,** 1305–1319 (2014).

223. Wang, Z. *et al.,* Automated detection of clinically significant prostate cancer in mp-MRI images based on an end-to-end deep neural network, *IEEE transactions on medical imaging* **37,** 1127–1139 (2018).

224. Zou, L., Zheng, J., Miao, C., Mckeown, M. J. & Wang, Z. J., 3D CNN based automatic diagnosis of attention deficit hyperactivity disorder using functional and structural MRI, *Ieee Access* **5,** 23626–23636 (2017).

225. Shi, J., Zheng, X., Li, Y., Zhang, Q. & Ying, S., Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease, *IEEE journal of biomedical and health informatics* **22,** 173–183 (2017).

226. Kim, J. & Lee, B., Identification of Alzheimer's disease and mild cognitive impairment using multimodal sparse hierarchical extreme learning machine, *Human brain mapping* **39,** 3728–3741 (2018).

227. Li, F. *et al.,* A robust deep model for improved classification of AD/MCI patients, *IEEE journal of biomedical and health informatics* **19,** 1610–1616 (2015).

228. El-Sappagh, S., Abuhmed, T., Islam, S. R. & Kwak, K. S., Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data, *Neurocomputing* **412,** 197–215 (2020).

229. Zhou, P. *et al.,* Use of a sparse-response deep belief network and extreme learning machine to discriminate Alzheimer's disease, mild cognitive impairment, and normal controls based on amyloid PET/MRI images, *Frontiers in Medicine* **7,** 621204 (2021).

230. Simonyan, K. & Zisserman, A., Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556* (2014).

231. Szegedy, C. *et al., Going deeper with convolutions in Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), 1–9.

232. He, K., Zhang, X., Ren, S. & Sun, J., *Deep Residual Learning for Image Recognition* 2015, `https://arxiv.org/abs/1512.03385`.

233. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q., *Densely Connected Convolutional Networks* 2016, `https://arxiv.org/abs/1608.06993`.

234. Suk, H.-I. & Shen, D., *Deep learning-based feature representation for AD/MCI classification in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16* (2013), 583–590.

235. Suk, H.-I., Lee, S.-W., Shen, D. & ADNI, Latent feature representation with stacked auto-encoder for AD/MCI diagnosis, *Brain Structure and Function* **220,** 841–859 (2015).

236. Yan, R. *et al.,* Richer fusion network for breast cancer classification based on multimodal data, *BMC Medical Informatics and Decision Making* **21,** 1–15 (2021).

237. Xing, X. *et al., Advit: Vision transformer on multi-modality pet images for alzheimer disease diagnosis in 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)* (2022), 1–4.

238. Deng, J. *et al., Imagenet: A large-scale hierarchical image database in 2009 IEEE conference on computer vision and pattern recognition* (2009), 248–255.

239. Goodfellow, I., Bengio, Y. & Courville, A., *Deep learning* (MIT press, 2016).

240. Ballard, D. H., *Modular learning in neural networks. in Aaai* **647** (1987), 279–284.

241. Vaswani, A. *et al., Attention Is All You Need* 2017, `https://arxiv.org/abs/1706.03762`.

242. He, M., Han, K., Zhang, Y. & Chen, W., Hierarchical-order multimodal interaction fusion network for grading gliomas, *Physics in Medicine & Biology* **66,** 215016 (2021).

243. Suk, H.-I., Lee, S.-W., Shen, D., ADNI, *et al.,* Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis, *NeuroImage* **101,** 569–582 (2014).

244. Wang, X. *et al.,* Combining Radiology and Pathology for Automatic Glioma Classification, *Frontiers in Bioengineering and Biotechnology* **10** (2022).

245. Hu, Q., Whitney, H. M. & Giger, M. L., A deep learning methodology for improved breast cancer diagnosis using multiparametric MRI, *Scientific reports* **10,** 10536 (2020).

246. Aldoj, N., Lukas, S., Dewey, M. & Penzkofer, T., Semi-automatic classification of prostate cancer on multi-parametric MR imaging using a multi-channel 3D convolutional neural network, *European radiology* **30,** 1243–1253 (2020).

247. Lin, W. *et al.,* Bidirectional mapping of brain MRI and PET with 3D reversible GAN for the diagnosis of Alzheimer's disease, *Frontiers in Neuroscience* **15,** 646013 (2021).

248. Zong, W. *et al.,* A deep dive into understanding tumor foci classification using multiparametric MRI based on convolutional neural network, *Medical physics* **47,** 4077–4086 (2020).

249. Zhou, Z. *et al.,* Prediction of pathologic complete response to neoadjuvant systemic therapy in triple negative breast cancer using deep learning on multiparametric MRI, *Scientific Reports* **13,** 1171 (2023).

250. Song, J. *et al.,* An effective multimodal image fusion method using MRI and PET for Alzheimer's disease diagnosis, *Frontiers in digital health* **3,** 637386 (2021).

251. Rahaman, M. A. *et al., Multi-modal deep learning of functional and structural neuroimaging and genomic data to predict mental illness in 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (2021), 3267–3272.

252. Xiong, J. *et al.,* Multimodal machine learning using visual fields and peripapillary circular OCT scans in detection of glaucomatous optic neuropathy, *Ophthalmology* **129,** 171–180 (2022).

253. Xu, T., Zhang, H., Huang, X., Zhang, S. & Metaxas, D. N., *Multimodal deep learning for cervical dysplasia diagnosis in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19* (2016), 115–123.

254. Joo, S. *et al.,* Multimodal deep learning models for the prediction of pathologic response to neoadjuvant chemotherapy in breast cancer, *Scientific reports* **11,** 18800 (2021).

255. Punjabi, A. *et al.,* Neuroimaging modality fusion in Alzheimer's classification using convolutional neural networks, *PloS one* **14,** e0225759 (2019).

256. Zhou, T., Thung, K.-H., Zhu, X. & Shen, D., *Feature learning and fusion of multimodality neuroimaging and genetic data for multi-status dementia diagnosis in Machine Learning in Medical Imaging: 8th International Workshop, MLMI 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 10, 2017, Proceedings 8* (2017), 132–140.

257. Cheng, D. & Liu, M., *CNNs based multi-modality classification for AD diagnosis in 2017 10th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI)* (2017), 1–5.

258. Rahaman, M. A. *et al., Two-Dimensional Attentive Fusion for Multi-Modal Learning of Neuroimaging and Genomics Data in 2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)* (2022), 1–6.

259. Jin, L., Zhao, K., Zhao, Y., Che, T. & Li, S., A Hybrid Deep Learning Method for Early and Late Mild Cognitive Impairment Diagnosis With Incomplete Multimodal Data, *Frontiers in Neuroinformatics* **16** (2022).

260. Zhou, T., Thung, K.-H., Zhu, X. & Shen, D., Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis, *Human brain mapping* **40,** 1001–1016 (2019).

261. Mahmood, F., Yang, Z., Ashley, T. & Durr, N. J., Multimodal densenet, *arXiv preprint arXiv:1811.07407* (2018).

262. Zhang, T. & Shi, M., Multi-modal neuroimaging feature fusion for diagnosis of Alzheimer's disease, *Journal of Neuroscience Methods* **341,** 108795 (2020).

263. Gao, X., Shi, F., Shen, D. & Liu, M., Task-induced pyramid and attention GAN for multimodal brain image imputation and classification in alzheimer's disease, *IEEE journal of biomedical and health informatics* **26,** 36–43 (2021).

264. Dai, Y., Gao, Y. & Liu, F., Transmed: Transformers advance multi-modal medical image classification, *Diagnostics* **11,** 1384 (2021).

265. Qiu, L. *et al.,* Hierarchical multimodal fusion framework based on noisy label learning and attention mechanism for cancer classification with pathology and genomic features, *Computerized Medical Imaging and Graphics,* 102176 (2023).

266. Zhang, Y. *et al.,* Multimodal learning for fetal distress diagnosis using a multimodal medical information fusion framework, *Frontiers in Physiology,* 2362 (2022).

267. Liu, L. *et al.,* Cascaded Multi-Modal Mixing Transformers for Alzheimer's Disease Classification with Incomplete Data, *arXiv preprint arXiv:2210.00255* (2022).

268. Li, S., Xie, Y., Wang, G., Zhang, L. & Zhou, W., Attention guided discriminative feature learning and adaptive fusion for grading hepatocellular carcinoma with Contrast-enhanced MR, *Computerized Medical Imaging and Graphics* **97,** 102050 (2022).

269. Dai, Y., Gao, Y., Liu, F. & Fu, J., Mutual Attention-based Hybrid Dimensional Network for Multimodal Imaging Computer-aided Diagnosis, *arXiv preprint arXiv:2201.09421* (2022).

270. Moon, W. K. *et al.,* Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks, *Computer methods and programs in biomedicine* **190,** 105361 (2020).

271. Guo, S. *et al.,* Multimodal MRI image decision fusion-based network for glioma classification, *Frontiers in Oncology* **12** (2022).

272. Abdolmaleki, S. & Abadeh, M. S., *Brain MR image classification for ADHD diagnosis using deep neural networks in 2020 international conference on machine vision and image processing (MVIP)* (2020), 1–5.

273. Kwon, I. *et al.,* Diagnosis of Early Glottic Cancer Using Laryngeal Image and Voice Based on Ensemble Learning of Convolutional Neural Network Classifiers, *Journal of Voice* (2022).

274. Dubois, B. *et al.,* Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria, *The Lancet Neurology* **6,** 734–746 (2007).

275. Zhang, F. *et al.,* Multi-modal deep learning model for auxiliary diagnosis of Alzheimer's disease, *Neurocomputing* **361,** 185–195 (2019).

276. Abdelaziz, M., Wang, T. & Elazab, A., Fusing Multimodal and Anatomical Volumes of Interest Features Using Convolutional Auto-Encoder and Convolutional Neural Networks for Alzheimer's Disease Diagnosis, *Frontiers in Aging Neuroscience* **14** (2022).

277. Tymchenko, B., Marchenko, P. & Spodarets, D., Deep learning approach to diabetic retinopathy detection, *arXiv preprint arXiv:2003.02261* (2020).

278. Decencière, E. *et al.,* Feedback on a publicly distributed image database: the Messidor database, *Image Analysis & Stereology* **33,** 231–234 (2014).

279. Li, T. *et al.,* Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening, *Information Sciences* **501,** 511–522 (2019).

280. Decenciere, E. *et al.,* TeleOphta: Machine learning and image processing methods for teleophthalmology, *Irbm* **34,** 196–203 (2013).

281. Seth, S. & Agarwal, B., A hybrid deep learning model for detecting diabetic retinopathy, *Journal of Statistics and Management Systems* **21,** 569–574 (2018).

282. Bellemo, V. *et al.,* Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study, *The Lancet Digital Health* **1,** e35–e44 (2019).

283. Abràmoff, M. D. *et al.,* Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning, *Investigative ophthalmology & visual science* **57,** 5200–5206 (2016).

284. Ghasemi Falavarjani, K., Wang, K., Khadamy, J. & Sadda, S. R., Ultra-wide-field imaging in diabetic retinopathy; an overview, *Journal of Current Ophthalmology* **28,** 57–60, ISSN: 2452-2325, `https://www.sciencedirect.com/science/article/pii/S2452232516300312` (2016).

285. Price, L. D., Au, S. & Chong, N. V., Optomap ultrawide field imaging identifies additional retinal abnormalities in patients with diabetic retinopathy, *Clinical Ophthalmology,* 527–531 (2015).

286. Talks, S. J., Manjunath, V., Steel, D. H., Peto, T. & Taylor, R., New vessels detected on wide-field imaging compared to two-field and seven-field imaging: implications for diabetic retinopathy screening image analysis, *British Journal of Ophthalmology* **99,** 1606–1609 (2015).

287. Liu, H., Teng, L., Fan, L., Sun, Y. & Li, H., A new ultra-wide-field fundus dataset to diabetic retinopathy grading using hybrid preprocessing methods, *Computers in Biology and Medicine* **157,** 106750 (2023).

288. Chandak, S., *Classification of Severity Levels in Diabetic Retinopathy in Ultra-wide Field Colour Fundus Images using Hybrid Deep Learning Models* PhD thesis (Dublin, National College of Ireland, 2022).

289. Viedma, I. A., Alonso-Caneiro, D., Read, S. A. & Collins, M. J., Deep learning in retinal optical coherence tomography (OCT): A comprehensive survey, *Neurocomputing* (2022).

290. Zhang, L., Van Dijk, E. H., Borrelli, E., Fragiotta, S. & Breazzano, M. P., OCT and OCT Angiography Update: Clinical Application to Age-Related Macular Degeneration, Central Serous Chorioretinopathy, Macular Telangiectasia, and Diabetic Retinopathy, *Diagnostics* **13,** 232 (2023).

291. Kwan, C. C. & Fawzi, A. A., Imaging and biomarkers in diabetic macular edema and diabetic retinopathy, *Current diabetes reports* **19,** 1–10 (2019).

292. Lains, I. *et al.,* Choroidal thickness in diabetic retinopathy assessed with swept-source optical coherence tomography, *Retina* **38,** 173–182 (2018).

293. Shen, Z.-J. *et al.,* Association of choroidal thickness with early stages of diabetic retinopathy in type 2 diabetes, *International journal of ophthalmology* **10,** 613 (2017).

294. Karn, P. K. & Abdulla, W. H., On Machine Learning in Clinical Interpretation of Retinal Diseases Using OCT Images, *Bioengineering* **10,** 407 (2023).

295. Elgafi, M. *et al.,* Detection of Diabetic Retinopathy Using Extracted 3D Features from OCT Images, *Sensors* **22,** 7833 (2022).

296. Elsharkawy, M. *et al.,* A Novel Computer-Aided Diagnostic System for Early Detection of Diabetic Retinopathy Using 3D-OCT Higher-Order Spatial Appearance Model, *Diagnostics* **12,** ISSN: 2075-4418, `https://www.mdpi.com/2075-4418/12/2/461` (2022).

297. Tan, B. *et al.,* Approaches to quantify optical coherence tomography angiography metrics, *Annals of Translational Medicine* **8** (2020).

298. Baumann, B. *et al.,* Total retinal blood flow measurement with ultrahigh speed swept source/Fourier domain OCT, *Biomed Opt Express* **2,** 1539–1552, ISSN: 2156-7085, (2023) (June 2011).

299. Yang, Z., Tan, T.-E., Shao, Y., Wong, T. & Li, X., Classification of diabetic retinopathy: Past, present and future, *Front Endocrinol* **13,** 1079217, ISSN: 1664-2392 (Dec. 2022).

300. Vujosevic, S. *et al.,* Standardization of Optical Coherence Tomography Angiography Imaging Biomarkers in Diabetic Retinal Disease, *Ophthalmic Res* **64,** 871–887, ISSN: 0030-3747 (Dec. 2021).

301. Sun, Z., Yang, D., Tang, Z., Ng, D. & Cheung, C., Optical coherence tomography angiography in diabetic retinopathy: An updated review, *Eye* **35,** 149–161, ISSN: 0950-222X (Jan. 2021).

302. Guo, Y. *et al.,* Automated segmentation of retinal fluid volumes from structural and angiographic optical coherence tomography using deep learning, *Transl Vis Sci Technol* **9,** 54, ISSN: 2164-2591 (Oct. 2020).

303. Alam, M., Le, D., Lim, J., Chan, R. & Yao, X., Supervised machine learning based multi-task artificial intelligence classification of retinopathies, *J Clin Med* **8,** 872, ISSN: 2077-0383 (June 2019).

304. Lo, J. *et al.,* Federated Learning for Microvasculature Segmentation and Diabetic Retinopathy Classification of OCT Data, *Ophthal Sci* **1,** 100069, ISSN: 2666-9145 (Oct. 2021).

305. Khalili Pour, E. *et al.,* Automated machine learning–based classification of proliferative and non-proliferative diabetic retinopathy using optical coherence tomography angiography vascular density maps, *Graefes Arch Clin Exp Ophthalmol* **261,** 391–399, ISSN: 0721-832X (Feb. 2023).

306. Guo, M. *et al.,* Automatic quantification of superficial foveal avascular zone in optical coherence tomography angiography implemented with deep learning, *Vis Comput Ind Biomed Art* **2,** 21, ISSN: 2096-496X (Dec. 2019).

307. Li, Q. *et al.,* Diagnosing Diabetic Retinopathy in OCTA Images Based on Multilevel Information Fusion using a Deep Learning Framework, *Comput Math Methods Med* **2022,** 4316507, ISSN: 1748-670X (Aug. 2022).

308. Vaz-Pereira, S., Morais-Sarmento, T. & Engelbert, M., Update on optical coherence tomography and optical coherence tomography angiography imaging in proliferative diabetic retinopathy, *Diagnostics* **11,** 1869, ISSN: 2075-4418 (Oct. 2021).

309. Carrera-Escalé, L. *et al.,* Radiomics-Based Assessment of OCT Angiography Images for Diabetic Retinopathy Diagnosis, *Ophthalmol Sci* **3,** 100259, ISSN: 2666-9145 (June 2023).

310. Ryu, G. *et al.,* A Deep Learning Algorithm for Classifying Diabetic Retinopathy using Optical Coherence Tomography Angiography, *Transl Vis Sci Technol* **11,** 39, ISSN: 2164-2591 (Feb. 2022).

311. Le, D. *et al.,* Transfer learning for automated OCTA detection of diabetic retinopathy, *Transl Vis Sci Technol* **9,** 35, ISSN: 2164-2591 (July 2020).

312. Andreeva, R., Fontanella, A., Giarratano, Y. & Bernabeu, M., *DR Detection using Optical Coherence Tomography Angiography (OCTA): A Transfer Learning Approach with Robustness Analysis in Proc MICCAI OMIA Works* **12069 LNCS** (Lima, Peru, Oct. 2020), 11–20.

313. Ryu, G., Lee, K., Park, D., Park, S. & Sagong, M., A deep learning model for identifying diabetic retinopathy using optical coherence tomography angiography, *Sci Rep* **11,** 23024, ISSN: 2045-2322 (Nov. 2021).

314. Heisler, M. *et al.,* Ensemble deep learning for diabetic retinopathy detection using optical coherence tomography angiography, *Transl Vis Sci Technol* **9,** 20, ISSN: 2164-2591 (Apr. 2020).

315. Yasser, I. *et al.,* Automated Diagnosis of Optical Coherence Tomography Angiography (OCTA) based on Machine Learning Techniques, *Sensors* **22,** 2342, ISSN: 1424-8220 (Mar. 2022).

316. Zang, P. *et al.,* A Diabetic Retinopathy Classification Framework Based on Deep-Learning Analysis of OCT Angiography, *Transl Vis Sci Technol* **11,** 10, ISSN: 2164-2591 (July 2022).

317. Liu, R. *et al.,* Application of artificial intelligence-based dual-modality analysis combining fundus photography and optical coherence tomography in diabetic retinopathy screening in a community hospital, *BioMedical Engineering OnLine* **21,** 1–11 (2022).

318. Lim, W. S. *et al.,* Use of multimodal dataset in AI for detecting glaucoma based on fundus photographs assessed with OCT: focus group study on high prevalence of myopia, *BMC Medical Imaging* **22,** 1–14 (2022).

319. He, K., Zhang, X., Ren, S. & Sun, J., Deep residual learningfor image recognition, *CoRR, abs/1512* **3385,** 2 (2015).

320. Ren, S., He, K., Girshick, R. & Sun, J., Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv 2015, *arXiv preprint arXiv:1506.01497* (2015).

321. Gorczynska, I., Migacz, J. V., Zawadzki, R. J., Capps, A. G. & Werner, J. S., Comparison of amplitude-decorrelation, speckle-variance and phase-variance OCT angiography methods for imaging the human retina and choroid, *Biomed Opt Express* **7,** 911–942, ISSN: 2156-7085, (2023) (Feb. 2016).

322. Makita, S., Hong, Y., Yamanari, M., Yatagai, T. & Yasuno, Y., Optical coherence angiography, *Optics express* **14,** 7821–7840 (2006).

323. An, L. & Wang, R. K., In vivo volumetric imaging of vascular perfusion within human retina and choroids with optical micro-angiography, *Optics express* **16,** 11438–11452 (2008).

324. Jia, Y. *et al.,* Split-spectrum amplitude-decorrelation angiography with optical coherence tomography, *Optics express* **20,** 4710–4725 (2012).

325. DaCosta, J., Bhatia, D., Crothers, O. & Talks, J., Utilisation of optical coherence tomography and optical coherence tomography angiography to assess retinal neovascularisation in diabetic retinopathy, *Eye* **36,** 827–834 (2022).

326. Hwang, T. S. *et al.,* Optical coherence tomography angiography features of diabetic retinopathy, *Retina (Philadelphia, Pa.)* **35,** 2371 (2015).

327. Adhi, M. & Duker, J. S., Optical coherence tomography–current and future applications, *Current opinion in ophthalmology* **24,** 213 (2013).

328. Zang, P. *et al.,* Deep-Learning–Aided Diagnosis of Diabetic Retinopathy, Age-Related Macular Degeneration, and Glaucoma Based on Structural and Angiographic OCT, *Ophthalmology Science* **3,** 100245 (2023).

329. Hazin, R., Barazi, M. K. & Summerfield, M., Challenges to establishing nationwide diabetic retinopathy screening programs, *Current opinion in ophthalmology* **22,** 174–179 (2011).

330. Zang, P. *et al.,* A Diabetic Retinopathy Classification Framework Based on Deep-Learning Analysis of OCT Angiography, *Translational Vision Science & Technology* **11,** 10–10, ISSN: 2164-2591, eprint: https://arvojournals.org/arvo/content_public/journal/tvst/938598/i2164-2591-11-7-10_1657608897.74855.pdf, https://doi.org/10.1167/tvst.11.7.10 (July 2022).

331. Heisler, M. *et al.,* Ensemble Deep Learning for Diabetic Retinopathy Detection Using Optical Coherence Tomography Angiography, *Translational Vision Science & Technology* **9,** 20–20, ISSN: 2164-2591, eprint: https://arvojournals.org/arvo/content_public/journal/tvst/938366/i2164-2591-344-1-1985.pdf, https://doi.org/10.1167/tvst.9.2.20 (Apr. 2020).

332. Zhu, Y. *et al.,* Different Scan Protocols Affect the Detection Rates of Diabetic Retinopathy Lesions by Wide-Field Swept-Source Optical Coherence Tomography Angiography, *American Journal of Ophthalmology* **215,** 72–80, ISSN: 0002-9394, https://www.sciencedirect.com/science/article/pii/S0002939420301021 (2020).

333. Wang, X.-n. *et al.,* Optical coherence tomography angiography for the detection and evaluation of ptic disc neovascularization: a retrospective, observational study, *BMC ophthalmology* **22,** 1–8 (2022).

334. Menean, M. *et al.,* Combined Wide-Field Imaging in Grading Diabetic Retinopathy (2022).

335. Wu, J. *et al., GAMMA Challenge:Glaucoma grAding from Multi-Modality imAges* 2022.

336. Shibata, N. *et al.,* Development of a deep residual learning algorithm to screen for glaucoma from fundus photography, *Scientific Reports* **8,** 14665, ISSN: 2045-2322 (Oct. 2018).

337. Ahn, J. M. *et al.,* A deep learning model for the detection of both advanced and early glaucoma using fundus photography, *PloS one* **13,** e0207982 (2018).

338. Li, F. *et al.,* Automatic detection of diabetic retinopathy in retinal fundus photographs based on deep learning algorithm, *Translational vision science & technology* **8,** 4–4 (2019).

339. Asaoka, R. *et al.,* Using deep learning and transfer learning to accurately diagnose early-onset glaucoma from macular optical coherence tomography images, *American journal of ophthalmology* **198,** 136–145 (2019).

340. Muhammad, H. *et al.,* Hybrid Deep Learning on Single Wide-field Optical Coherence tomography Scans Accurately Classifies Glaucoma Suspects, eng, *Journal of glaucoma* **26,** PMC5716847[pmcid], 1086–1094, ISSN: 1536-481X (Dec. 2017).

341. Perdomo, O. *et al.,* Classification of diabetes-related retinal diseases using a deep learning approach in optical coherence tomography, *Computer methods and programs in biomedicine* **178,** 181–189 (2019).

342. De Carlo, T. E., Romano, A., Waheed, N. K. & Duker, J. S., A review of optical coherence tomography angiography (OCTA), *International journal of retina and vitreous* **1,** 1–15 (2015).

343. Ryu, G., Lee, K., Park, D., Park, S. H. & Sagong, M., A deep learning model for identifying diabetic retinopathy using optical coherence tomography angiography, *Scientific reports* **11,** 23024 (2021).

344. Guo, M.-H. *et al.,* Attention mechanisms in computer vision: A survey, *Computational visual media* **8,** 331–368 (2022).

345. Brauwers, G. & Frasincar, F., A general survey on attention mechanisms in deep learning, *IEEE Transactions on Knowledge and Data Engineering* (2021).

346. Niu, Z., Zhong, G. & Yu, H., A review on the attention mechanism of deep learning, *Neurocomputing* **452,** 48–62 (2021).

347. Wang, F. *et al., Residual attention network for image classification in Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), 3156–3164.

348. Liu, G. & Guo, J., Bidirectional LSTM with attention mechanism and convolutional layer for text classification, *Neurocomputing* **337,** 325–338 (2019).

349. Li, Y., Yang, L., Xu, B., Wang, J. & Lin, H., Improving user attribute classification with text and social network attention, *Cognitive Computation* **11,** 459–468 (2019).

350. Sutskever, I., Vinyals, O. & Le, Q. V., Sequence to sequence learning with neural networks, *Advances in neural information processing systems* **27** (2014).

351. Song, S., Lan, C., Xing, J., Zeng, W. & Liu, J., *An end-to-end spatio-temporal attention model for human action recognition from skeleton data in Proceedings of the AAAI conference on artificial intelligence* **31** (2017).

352. Tian, Y., Hu, W., Jiang, H. & Wu, J., Densely connected attentional pyramid residual network for human pose estimation, *Neurocomputing* **347,** 13–23 (2019).

353. Xu, K. *et al., Show, attend and tell: Neural image caption generation with visual attention in International conference on machine learning* (2015), 2048–2057.

354. Clèrigues, A. *et al.,* Acute and sub-acute stroke lesion segmentation from multimodal MRI, *Computer methods and programs in biomedicine* **194,** 105521 (2020).

355. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M. & Maier-Hein, K. H., *Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge in Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3* (2018), 287–297.

356. Gao, X. W., Hui, R. & Tian, Z., Classification of CT brain images based on deep learning networks, *Computer methods and programs in biomedicine* **138,** 49–56 (2017).

357. Zhang, C., Zhao, J., Niu, J. & Li, D., New convolutional neural network model for screening and diagnosis of mammograms, *PLoS One* **15,** e0237674 (2020).

358. Benzebouchi, N. E., Azizi, N., Ashour, A. S., Dey, N. & Sherratt, R. S., Multimodal classifier fusion with feature cooperation for glaucoma diagnosis, *Journal of Experimental & Theoretical Artificial Intelligence* **31,** 841–874 (2019).

359. Dolz, J. *et al.,* HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation, *IEEE transactions on medical imaging* **38,** 1116–1126 (2018).

360. Zhou, T., Canu, S., Vera, P. & Ruan, S., *3d medical multi-modal segmentation network guided by multi-source correlation constraint in 2020 25th International Conference on Pattern Recognition (ICPR)* (2021), 10243–10250.

361. He, K., Zhang, X., Ren, S. & Sun, J., *Deep residual learning for image recognition in Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 770–778.

362. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q., *Densely connected convolutional networks in Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), 4700–4708.

363. Akbari, H. *et al.,* Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text, *Advances in Neural Information Processing Systems* **34,** 24206–24221 (2021).

364. Appalaraju, S., Jasani, B., Kota, B. U., Xie, Y. & Manmatha, R., *Docformer: End-to-end transformer for document understanding in Proceedings of the IEEE/CVF international conference on computer vision* (2021), 993–1003.

365. Likhosherstov, V. *et al.,* Polyvit: Co-training vision transformers on images, videos and audio, *arXiv preprint arXiv:2111.12993* (2021).

366. Sathya, R., Abraham, A., *et al.,* Comparison of supervised and unsupervised learning algorithms for pattern classification, *International Journal of Advanced Research in Artificial Intelligence* **2,** 34–38 (2013).

367. Berry, M. W., Mohamed, A. & Yap, B. W., *Supervised and unsupervised learning for data science* (Springer, 2019).

368. Hastie, T. *et al.,* Unsupervised learning, *The elements of statistical learning: Data mining, inference, and prediction,* 485–585 (2009).

369. Zhu, X. J., Semi-supervised learning literature survey (2005).

370. Van Engelen, J. E. & Hoos, H. H., A survey on semi-supervised learning, *Machine learning* **109,** 373–440 (2020).

371. Ouali, Y., Hudelot, C. & Tami, M., An overview of deep semi-supervised learning, *arXiv preprint arXiv:2006.05278* (2020).

372. Luo, X., Chen, J., Song, T. & Wang, G., *Semi-supervised medical image segmentation through dual-task consistency in Proceedings of the AAAI conference on artificial intelligence* **35** (2021), 8801–8809.

373. Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D. & Makedon, F., A survey on contrastive self-supervised learning, *Technologies* **9,** 2 (2020).

374. Krishnan, R., Rajpurkar, P. & Topol, E. J., Self-supervised learning in medicine and healthcare, *Nature Biomedical Engineering* **6,** 1346–1352 (2022).

375. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T. & Efros, A. A., *Context encoders: Feature learning by inpainting in Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 2536–2544.

376. Sohn, K. *et al.,* Fixmatch: Simplifying semi-supervised learning with consistency and confidence, *Advances in neural information processing systems* **33,** 596–608 (2020).

377. Albelwi, S., Survey on self-supervised learning: auxiliary pretext tasks and contrastive learning methods in imaging, *Entropy* **24,** 551 (2022).

378. Zhang, H., Zhang, Z., Odena, A. & Lee, H., Consistency regularization for generative adversarial networks, *arXiv preprint arXiv:1910.12027* (2019).

379. Zhang, B. *et al.,* Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling, *Advances in Neural Information Processing Systems* **34,** 18408–18419 (2021).

380. Sajjadi, M., Javanmardi, M. & Tasdizen, T., Regularization with stochastic transformations and perturbations for deep semi-supervised learning, *Advances in neural information processing systems* **29** (2016).

381. Laine, S. & Aila, T., Temporal ensembling for semi-supervised learning, *arXiv preprint arXiv:1610.02242* (2016).

382. McLachlan, G. J., Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis, *Journal of the American Statistical Association* **70,** 365–369 (1975).

383. Scudder, H., Probability of error of some adaptive pattern-recognition machines, *IEEE Transactions on Information Theory* **11,** 363–371 (1965).

384. Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V. & Le, Q. V., *Autoaugment: Learning augmentation strategies from data in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), 113–123.

385. Cubuk, E. D., Zoph, B., Shlens, J. & Le, Q. V., *Randaugment: Practical automated data augmentation with a reduced search space in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (2020), 702–703.

386. Berthelot, D. *et al.,* Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring, *arXiv preprint arXiv:1911.09785* (2019).

387. Solovyev, R., Kalinin, A. A. & Gabruseva, T., 3D convolutional neural networks for stalled brain capillary detection, *Computers in biology and medicine* **141,** 105089 (2022).

388. Van der Maaten, L. & Hinton, G., Visualizing data using t-SNE. *Journal of machine learning research* **9** (2008).

389. Salz, D. A. *et al.,* Select Features of Diabetic Retinopathy on Swept-Source Optical Coherence Tomographic Angiography Compared With Fluorescein Angiography and Normal Eyes, en, *JAMA Ophthalmology* **134,** 644, ISSN: 2168-6165, http://archopht.jamanetwork.com/article.aspx?doi=10.1001/jamaophthalmol.2016.0600 (2022) (June 2016).

390. Agemy, S. A. *et al.,* RETINAL VASCULAR PERFUSION DENSITY MAPPING USING OPTICAL COHERENCE TOMOGRAPHY ANGIOGRAPHY IN NORMALS AND DIABETIC RETINOPATHY PATIENTS, en, *Retina* **35,** 2353–2363, ISSN: 0275-004X, https://journals.lww.com/00006982-201511000-00024 (2021) (Nov. 2015).

391. Al-Sheikh, M., Akil, H., Pfau, M. & Sadda, S. R., Swept-Source OCT Angiography Imaging of the Foveal Avascular Zone and Macular Capillary Network Density in Diabetic Retinopathy, en, *Investigative Opthalmology & Visual Science* **57,** 3907, ISSN: 1552-5783, http://iovs.arvojournals.org/article.aspx?doi=10.1167/iovs.16-19570 (2022) (July 2016).

392. Hwang, T. S. *et al.,* Automated Quantification of Capillary Nonperfusion Using Optical Coherence Tomography Angiography in Diabetic Retinopathy, en, *JAMA Ophthalmology* **134,** 367, ISSN: 2168-6165, http://archopht.jamanetwork.com/article.aspx?doi=10.1001/jamaophthalmol.2015.5658 (2021) (Apr. 2016).

393. Fayed, A. E., Abdelbaki, A. M., El Zawahry, O. M. & Fawzi, A. A., Optical coherence tomography angiography reveals progressive worsening of retinal vascular geometry in diabetic retinopathy and improved geometry after panretinal photocoagulation, en, *PLOS ONE* **14** (ed Vavvas, D. G.) e0226629, ISSN: 1932-6203, https://dx.plos.org/10.1371/journal.pone.0226629 (2022) (Dec. 2019).

394. Schottenhamml, J. *et al.,* AN AUTOMATIC, INTERCAPILLARY AREA-BASED ALGORITHM FOR QUANTIFYING DIABETES-RELATED CAPILLARY DROPOUT USING OPTICAL COHERENCE TOMOGRAPHY ANGIOGRAPHY, en, *Retina*

36, S93–S101, ISSN: 0275-004X, https://journals.lww.com/00006982-201612001-00010 (2021) (Dec. 2016).

395. Ishibazawa, A. *et al.,* Optical Coherence Tomography Angiography in Diabetic Retinopathy: A Prospective Pilot Study, en, *American Journal of Ophthalmology* **160,** 35–44.e1, ISSN: 00029394, https://linkinghub.elsevier.com/retrieve/pii/S000293941500224X (2021) (July 2015).

396. Couturier, A. *et al.,* Widefield OCT-Angiography and Fluorescein Angiography Assessments of Nonperfusion in Diabetic Retinopathy and Edema Treated with Anti–Vascular Endothelial Growth Factor, en, *Ophthalmology* **126,** 1685–1694, ISSN: 01616420, https://linkinghub.elsevier.com/retrieve/pii/S0161642019303045 (2021) (Dec. 2019).

397. Alibhai, A. Y. *et al.,* QUANTIFICATION OF RETINAL CAPILLARY NONPERFUSION IN DIABETICS USING WIDE-FIELD OPTICAL COHERENCE TOMOGRAPHY ANGIOGRAPHY, en, *Retina* **40,** 412–420, ISSN: 0275-004X, https://journals.lww.com/10.1097/IAE.0000000000002403 (2021) (Mar. 2020).

398. Jia, Y. *et al.,* Quantitative optical coherence tomography angiography of vascular abnormalities in the living human eye, en, *Proceedings of the National Academy of Sciences* **112,** E2395–E2402, ISSN: 0027-8424, 1091-6490, http://www.pnas.org/lookup/doi/10.1073/pnas.1500185112 (2021) (May 2015).

399. Shanmugam, D., Blalock, D., Balakrishnan, G. & Guttag, J., *Better aggregation in test-time augmentation in Proceedings of the IEEE/CVF international conference on computer vision* (2021), 1214–1223.

400. Rokach, L., Ensemble-based classifiers, *Artificial intelligence review* **33,** 1–39 (2010).

401. Tan, M. & Le, Q., *Efficientnet: Rethinking model scaling for convolutional neural networks in International conference on machine learning* (2019), 6105–6114.

402. Liu, L. *et al.,* Cascaded multi-modal mixing transformers for alzheimer's disease classification with incomplete data, *NeuroImage* **277,** 120267 (2023).

403. Nguyen, H. H., Blaschko, M. B., Saarakkala, S. & Tiulpin, A., Clinically-Inspired Multi-Agent Transformers for Disease Trajectory Forecasting from Multimodal Data, *arXiv preprint arXiv:2210.13889* (2022).

404. Yang, J., Zhang, B., Wang, E., Xia, S. & Chen, Y., Ultra-wide field swept-source optical coherence tomography angiography in patients with diabetes without clinically detectable retinopathy, *BMC ophthalmology* **21,** 1–8 (2021).

405. Silva, P. S. *et al.,* Diabetic Retinopathy Severity and Peripheral Lesions Are Associated with Nonperfusion on Ultrawide Field Angiography, *Ophthalmology* **122,** 2465–2472, ISSN: 0161-6420 (2015).

406. Sun, Z., Yang, D., Tang, Z. & et al., Optical coherence tomography angiography in diabetic retinopathy: an updated review, *Eye* **35,** 149–161 (2021).

407. Yang, J., Zhang, B., Wang, E. & et al., Ultra-wide field swept-source optical coherence tomography angiography in patients with diabetes without clinically detectable retinopathy, *BMC Ophthalmology* **21,** 192 (2021).

408. Li, J. *et al.,* Ultra-widefield color fundus photography combined with high-speed ultra-widefield swept-source optical coherence tomography angiography for noninvasive detection of lesions in diabetic retinopathy, *Frontiers in Public Health* **10,** ISSN: 2296-2565 (2022).

409. Verma, V. *et al., Manifold Mixup: Better Representations by Interpolating Hidden States* 2019, arXiv: 1806.05236 [stat.ML].

410. Li, Y. *et al., Multimodal Information Fusion for Glaucoma and Diabetic Retinopathy Classification in Ophthalmic Medical Image Analysis* (Cham, 2022), 53–62.

411. Sleeman, W. C., Kapoor, R. & Ghosh, P., Multimodal Classification: Current Landscape, Taxonomy and Future Directions, *ACM Comput. Surv.* **55,** ISSN: 0360-0300 (Dec. 2022).

412. Lin, R. & Hu, H., Adapt and Explore: Multimodal Mixup for Representation Learning, *Available at SSRN* (2023).

413. Liu, Z. *et al., Learning Multimodal Data Augmentation in Feature Space* 2023, arXiv: 2212.14453 [cs.LG].

414. Hao, X. *et al., MixGen: A New Multi-Modal Data Augmentation* 2023, arXiv: 2206.08358 [cs.CV].

415. Zhao, X. *et al., TMMDA: A New Token Mixup Multimodal Data Augmentation for Multimodal Sentiment Analysis in Proceedings of the ACM Web Conference 2023* (Association for Computing Machinery, Austin, TX, USA, 2023), 1714–1722, ISBN: 9781450394161.

416. Wisely, C. E. *et al.,* Convolutional neural network to identify symptomatic Alzheimer's disease using multimodal retinal imaging, *British Journal of Ophthalmology* **106,** 388–395, ISSN: 0007-1161 (2022).

417. Hu, J., Shen, L. & Sun, G., *Squeeze-and-Excitation Networks in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 7132–7141.

418. Zang, P. *et al.,* A Diabetic Retinopathy Classification Framework Based on Deep-Learning Analysis of OCT Angiography, *Translational Vision Science & Technology* **11,** 10–10 (2022).

419. Al-Absi, H. R., Islam, M. T., Refaee, M. A., Chowdhury, M. E. & Alam, T., Cardiovascular disease diagnosis from DXA scan and retinal images using deep learning, *Sensors* **22,** 4310 (2022).

420. Zhang, H., Cissé, M., Dauphin, Y. N. & Lopez-Paz, D., mixup: Beyond Empirical Risk Minimization, *CoRR* **abs/1710.09412,** arXiv: 1710.09412, http://arxiv.org/abs/1710.09412 (2017).

421. Early Treatment Diabetic Retinopathy Study Design and Baseline Patient Characteristics: ETDRS Report Number 7, *Ophthalmology* **98,** 741–756, ISSN: 0161-6420 (1991).

422. Nagrani, A. *et al.,* Attention bottlenecks for multimodal fusion, *Advances in Neural Information Processing Systems* **34,** 14200–14213 (2021).

423. Shi, B., Hsu, W.-N., Lakhotia, K. & Mohamed, A., Learning audio-visual speech representation by masked multimodal cluster prediction, *arXiv preprint arXiv:2201.02184* (2022).

424. Li, R., Yang, S., Ross, D. A. & Kanazawa, A., *Ai choreographer: Music conditioned 3d dance generation with aist++ in Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), 13401–13412.

425. Pashevich, A., Schmid, C. & Sun, C., *Episodic transformer for vision-and-language navigation in Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), 15942–15952.

426. Steitz, J.-M. O., Pfeiffer, J., Gurevych, I. & Roth, S., *TxT: Crossmodal end-to-end learning with transformers in Pattern Recognition: 43rd DAGM German Conference, DAGM GCPR 2021, Bonn, Germany, September 28–October 1, 2021, Proceedings* (2022), 405–420.

427. Wu, P. Y. & Mebane Jr, W. R., MARMOT: A deep learning framework for constructing multimodal representations for vision-and-language tasks, *Computational Communication Research* **4** (2022).

428. Lu, J., Batra, D., Parikh, D. & Lee, S., Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, *Advances in neural information processing systems* **32** (2019).

429. Chen, R. J. *et al., Multimodal co-attention transformer for survival prediction in gigapixel whole slide images in Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), 4015–4025.

430. Tan, H. & Bansal, M., Lxmert: Learning cross-modality encoder representations from transformers, *arXiv preprint arXiv:1908.07490* (2019).

431. Zhu, L. & Yang, Y., *Actbert: Learning global-local video-text representations in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), 8746–8755.

432. Ramesh, K., Xing, C., Wang, W., Wang, D. & Chen, X., *Vset: A multimodal transformer for visual speech enhancement in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2021), 6658–6662.

433. Rahman, T., Yang, M. & Sigal, L., Tribert: Full-body human-centric audio-visual representation learning for visual sound separation, *arXiv preprint arXiv:2110.13412* (2021).

434. Chen, S., Guhur, P.-L., Schmid, C. & Laptev, I., History aware multimodal transformer for vision-and-language navigation, *Advances in neural information processing systems* **34,** 5834–5847 (2021).

435. Li, Y. *et al., Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), 17182–17191.

436. Goodfellow, I. *et al.,* Generative adversarial networks, *Communications of the ACM* **63,** 139–144 (2020).

437. Goodfellow, I. J. *et al., Generative Adversarial Networks* 2014, `https://arxiv.org/abs/1406.2661`.

438. Pan, Y. *et al., Synthesizing missing PET from MRI with cycle-consistent generative adversarial networks for Alzheimer's disease diagnosis in Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11* (2018), 455–463.

439. Pan, Y., Liu, M., Lian, C., Xia, Y. & Shen, D., Spatially-constrained fisher representation for brain disease identification with incomplete multi-modal neuroimages, *IEEE transactions on medical imaging* **39,** 2965–2975 (2020).

440. Pan, Y., Liu, M., Xia, Y. & Shen, D., Disease-image-specific learning for diagnosis-oriented neuroimage synthesis with incomplete multi-modality data, *IEEE transactions on pattern analysis and machine intelligence* **44,** 6839–6853 (2021).

441. Zeghlache, R. *et al., Longitudinal self-supervised learning using neural ordinary differential equation in International Workshop on Predictive Intelligence in Medicine* (Vancouver (Canada), Canada, Oct. 2023), `https://imt-atlantique.hal.science/hal-04171357`.

442. Zeghlache, R. *et al., Detection of Diabetic Retinopathy Using Longitudinal Self-supervised Learning in Ophthalmic Medical Image Analysis* (eds Antony, B. *et al.*) (Springer International Publishing, Cham, 2022), 43–52.

443. Spaide, R. F., Fujimoto, J. G., Waheed, N. K., Sadda, S. R. & Staurenghi, G., Optical coherence tomography angiography, *Progress in retinal and eye research* **64,** 1–55 (2018).

444. Zhang, Q. *et al.,* Ultra-wide optical coherence tomography angiography in diabetic retinopathy, *Quantitative imaging in medicine and surgery* **8,** 743 (2018).

445. Holden, B. A. *et al.,* Global prevalence of myopia and high myopia and temporal trends from 2000 through 2050, *Ophthalmology* **123,** 1036–1042 (2016).

446. Dai, L. *et al.,* A deep learning system for detecting diabetic retinopathy across the disease spectrum, *Nature communications* **12,** 3242 (2021).

447. Schaal, K. B. *et al.,* Vascular abnormalities in diabetic retinopathy assessed with swept-source optical coherence tomography angiography widefield imaging, *Retina* **39,** 79–87 (2019).

448. Parkin, D. M., Bray, F., Ferlay, J. & Pisani, P., Global cancer statistics, 2002, *CA: a cancer journal for clinicians* **55,** 74–108 (2005).

449. Vallieres, M. *et al.,* Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer, *Scientific reports* **7,** 10117 (2017).

450. Bogowicz, M. *et al.,* Comparison of PET and CT radiomics for prediction of local tumor control in head and neck squamous cell carcinoma, *Acta oncologica* **56,** 1531–1536 (2017).

451. Castelli, J. *et al.,* A PET-based nomogram for oropharyngeal cancers, *European journal of cancer* **75,** 222–230 (2017).

452. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H., nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, *Nature methods* **18,** 203–211 (2021).

453. Sanford, T. *et al.,* Deep-learning-based artificial intelligence for PI-RADS classification to assist multiparametric prostate MRI interpretation: A development study, *Journal of Magnetic Resonance Imaging* **52,** 1499–1507 (2020).

454. Vu, T. D., Yang, H.-J., Nguyen, V. Q., Oh, A.-R. & Kim, M.-S., *Multimodal learning using convolution neural network and sparse autoencoder in 2017 IEEE international conference on big data and smart computing (BigComp)* (2017), 309–312.

455. Ge, C., Gu, I. Y.-H., Jakola, A. S. & Yang, J., *Deep learning and multi-sensor fusion for glioma classification using multistream 2D convolutional networks in 2018 40th Annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (2018), 5894–5897.

456. Lu, D., Popuri, K., Ding, G. W., Balachandar, R. & Beg, M. F., Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images, *Scientific reports* **8,** 5697 (2018).

457. Feng, C. *et al.,* Deep learning framework for Alzheimer's disease diagnosis via 3D-CNN and FSBi-LSTM, *IEEE Access* **7,** 63605–63618 (2019).

458. Lee, G., Nho, K., Kang, B., Sohn, K.-A. & Kim, D., Predicting Alzheimer's disease progression using multi-modal deep learning approach, *Scientific reports* **9,** 1952 (2019).

459. Massalimova, A. & Varol, H. A., *Input agnostic deep learning for Alzheimer's disease classification using multimodal MRI images in 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (2021), 2875–2878.

460. Abdelaziz, M., Wang, T. & Elazab, A., Alzheimer's disease diagnosis framework from incomplete multimodal data using convolutional neural networks, *Journal of Biomedical Informatics* **121,** 103863 (2021).

461. Zhou, J. *et al.,* Cohesive multi-modality feature learning and fusion for COVID-19 patient severity prediction, *IEEE Transactions on Circuits and Systems for Video Technology* **32,** 2535–2549 (2021).

462. Zhang, X., Lin, W., Xiao, M. & Ji, H., Multimodal 2.5 D Convolutional Neural Network for Diagnosis of Alzheimer's Disease with Magnetic Resonance Imaging and Positron Emission Tomography. *Progress In Electromagnetics Research* **171** (2021).

463. Puyol-Antón, E. *et al.,* A multimodal deep learning model for cardiac resynchronisation therapy response prediction, *Medical Image Analysis* **79,** 102465 (2022).

464. Dolci, G. *et al., A deep generative multimodal imaging genomics framework for Alzheimer's disease prediction in 2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE)* (2022), 41–44.

465. Hoang Nguyen, H., Blaschko, M. B., Saarakkala, S. & Tiulpin, A., Clinically-Inspired Multi-Agent Transformers for Disease Trajectory Forecasting from Multimodal Data, *arXiv e-prints,* arXiv–2210 (2022).

466. Xi, I. L. *et al.,* Deep learning to distinguish benign from malignant renal lesions based on routine MR ImagingDeep learning for characterization of renal lesions, *Clinical Cancer Research* **26,** 1944–1952 (2020).

467. Al-Dhabyani, W., Gomaa, M., Khaled, H. & Fahmy, A., *Dataset of breast ultrasound images. Data Brief 28, 104863 (2020)* 2019.

468. Huang, S.-C., Pareek, A., Zamanian, R., Banerjee, I. & Lungren, M. P., Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection, *Scientific reports* **10,** 1–9 (2020).

469. Ying, Q. *et al., Multi-modal data analysis for alzheimer's disease diagnosis: An ensemble model using imagery and genetic features in 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (2021), 3586–3591.

470. Selvachandran, G., Quek, S. G., Paramesran, R., Ding, W. & Son, L. H., Developments in the detection of diabetic retinopathy: a state-of-the-art review of computer-aided diagnosis and machine learning methods, *Artificial Intelligence Review* **56,** 915–964 (2023).

471. Barth, T. & Helbig, H., Diabetische Retinopathie, *Augenheilkunde up2date* **11,** 231–247 (2021).

472. Saeedi, P. *et al.,* Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition, *Diabetes Research and Clinical Practice* **157,** 107843, ISSN: 0168-8227 (2019).

473. Grzybowski, A., Singhanetr, P., Nanegrungsunk, O. & Ruamviboonsuk, P., Artificial intelligence for diabetic retinopathy screening using color retinal photographs: from development to deployment, *Ophthalmology and Therapy,* 1–19 (2023).

474. Le, D. *et al.,* Transfer learning for automated OCTA detection of diabetic retinopathy, *Translational Vision Science & Technology* **9,** 35–35 (2020).

475. Pereira, S., Pinto, A., Alves, V. & Silva, C. A., Brain tumor segmentation using convolutional neural networks in MRI images, *IEEE transactions on medical imaging* **35,** 1240–1251 (2016).

476. Cui, S., Mao, L., Jiang, J., Liu, C. & Xiong, S., Automatic semantic segmentation of brain gliomas from MRI images using a deep cascaded neural network, *Journal of healthcare engineering* **2018** (2018).

477. Wang, G., Li, W., Ourselin, S. & Vercauteren, T., *Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks in Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3* (2018), 178–190.

478. Xu, H. *et al., in Head and Neck Tumor Segmentation and Outcome Prediction: Third Challenge, HECKTOR 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings* 154–165 (Springer, 2023).

479. Heisler, M. *et al.,* Ensemble deep learning for diabetic retinopathy detection using optical coherence tomography angiography, *Translational Vision Science & Technology* **9,** 20–20 (2020).

480. Dubow, M. *et al.,* Classification of human retinal microaneurysms using adaptive optics scanning light ophthalmoscope fluorescein angiography, *Investigative ophthalmology & visual science* **55,** 1299–1309 (2014).

481. Mookiah, M. R. K. *et al.,* Computer-aided diagnosis of diabetic retinopathy: A review, *Computers in biology and medicine* **43,** 2136–2155 (2013).

482. Group, D. R. V. S. R. *et al.,* Early vitrectomy for severe proliferative diabetic retinopathy in eyes with useful vision: results of a randomized trial—Diabetic Retinopathy Vitrectomy Study report 3, *Ophthalmology* **95,** 1307–1320 (1988).

483. Lee, J.-G. *et al.,* Deep learning in medical imaging: general overview, *Korean journal of radiology* **18,** 570–584 (2017).

484. Liu, S. *et al.,* Deep learning in medical ultrasound analysis: a review, *Engineering* **5,** 261–275 (2019).

485. Masood, S., Sharif, M., Yasmin, M., Shahid, M. A. & Rehman, A., Image Fusion Methods: A Survey. *Journal of Engineering Science & Technology Review* **10** (2017).

486. Ronneberger, O., Fischer, P. & Brox, T., *U-Net: Convolutional Networks for Biomedical Image Segmentation* 2015, `https://arxiv.org/abs/1505.04597`.

487. Liu, Y. *et al.,* Assessing clinical progression from subjective cognitive decline to mild cognitive impairment with incomplete multi-modal neuroimages, *Medical image analysis* **75,** 102266 (2022).

488. Shoeibi, A. *et al.,* Diagnosis of brain diseases in fusion of neuroimaging modalities using deep learning: A review, *Information Fusion* (2022).

489. Liu, S. *et al.,* Multimodal neuroimaging computing: a review of the applications in neuropsychiatric disorders, *Brain informatics* **2,** 167–180 (2015).

490. Xie, G. *et al., Cross-Modality Neuroimage Synthesis: A Survey* 2022, `https://arxiv.org/abs/2202.06997`.

491. Kamnitsas, K. *et al.,* Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation, *Medical image analysis* **36,** 61–78 (2017).

492. Kamnitsas, K. *et al., Ensembles of multiple models and architectures for robust brain tumour segmentation in Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3* (2018), 450–462.

493. Jack Jr, C. R. *et al.,* The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods, *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* **27,** 685–691 (2008).

494. Kharazmi, P., Kalia, S., Lui, H., Wang, Z. & Lee, T., A feature fusion system for basal cell carcinoma detection through data-driven feature learning and patient profile, *Skin research and technology* **24,** 256–264 (2018).

495. Bien, N. *et al.,* Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet, *PLoS medicine* **15,** e1002699 (2018).

496. Nevitt, M., Felson, D. & Lester, G., The osteoarthritis initiative, *Protocol for the cohort study* **1** (2006).

497. Goldberger, A. L. *et al.,* PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, *circulation* **101,** e215–e220 (2000).

498. Chang, Y.-J., Huang, T.-Y., Liu, Y.-J., Chung, H.-W. & Juan, C.-J., Classification of parotid gland tumors by using multimodal MRI and deep learning, *NMR in Biomedicine* **34,** e4408 (2021).

499. Yin, Y., Huang, S. & Zhang, X., *Bm-nas: Bilevel multimodal neural architecture search in Proceedings of the AAAI Conference on Artificial Intelligence* **36** (2022), 8901–8909.

500. Singh, A. & Nair, H., A Neural Architecture Search for Automated Multimodal Learning, *Expert Systems with Applications* **207,** 118051 (2022).

501. Pérez-Rúa, J.-M., Vielzeuf, V., Pateux, S., Baccouche, M. & Jurie, F., *Mfas: Multimodal fusion architecture search in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), 6966–6975.

502. Yu, Z. *et al., Deep multimodal neural architecture search in Proceedings of the 28th ACM International Conference on Multimedia* (2020), 3743–3752.

503. Wang, F., Neural architecture search for gliomas segmentation on multimodal magnetic resonance imaging, *arXiv preprint arXiv:2005.06338* (2020).

504. Chatzianastasis, M., Ilias, L., Askounis, D. & Vazirgiannis, M., Neural Architecture Search with Multimodal Fusion Methods for Diagnosing Dementia, *arXiv preprint arXiv:2302.05894* (2023).

505. Sterne, J. A. *et al.,* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls, *Bmj* **338** (2009).

506. Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T. & Moons, K. G., A gentle introduction to imputation of missing values, *Journal of clinical epidemiology* **59,** 1087–1091 (2006).

507. Narazani, M. *et al., Is a PET All You Need? A Multi-modal Study for Alzheimer's Disease Using 3D CNNs in Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I* (2022), 66–76.

508. Khagi, B. & Kwon, G.-R., 3D CNN design for the classification of Alzheimer's disease using brain MRI and PET, *IEEE Access* **8,** 217830–217847 (2020).

509. Salvador, R. *et al.,* Multimodal integration of brain images for MRI-based diagnosis in schizophrenia, *Frontiers in neuroscience* **13,** 1203 (2019).

510. Xu, B., Kocyigit, D., Grimm, R., Griffin, B. P. & Cheng, F., Applications of artificial intelligence in multimodality cardiovascular imaging: a state-of-the-art review, *Progress in cardiovascular diseases* **63,** 367–376 (2020).

511. Lipkova, J. *et al.,* Artificial intelligence for multimodal data integration in oncology, *Cancer Cell* **40,** 1095–1110 (2022).

512. Baltrušaitis, T., Ahuja, C. & Morency, L.-P., Multimodal machine learning: A survey and taxonomy, *IEEE transactions on pattern analysis and machine intelligence* **41,** 423–443 (2018).

513. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M. & Maier-Hein, K. H., *No new-net in Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4* (2019), 234–244.

514. Dolz, J., Desrosiers, C. & Ben Ayed, I., *IVD-Net: Intervertebral disc localization and segmentation in MRI with a multi-modal UNet in Computational Methods and Clinical Applications for Spine Imaging: 5th International Workshop and Challenge, CSI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers* (2019), 130–143.

515. Chen, L. *et al., MRI tumor segmentation with densely connected 3D CNN in Medical Imaging 2018: Image Processing* **10574** (2018), 357–364.

516. Andrade-Miranda, G. *et al., Pure versus hybrid Transformers for multi-modal brain tumor segmentation: a comparative study in 2022 IEEE International Conference on Image Processing (ICIP)* (2022), 1336–1340.

517. Aygün, M., Şahin, Y. H. & Ünal, G., Multi modal convolutional neural networks for brain tumor segmentation, *arXiv preprint arXiv:1809.06191* (2018).

518. Li, J., Bu, C. & Qian, C., *A cross-attention based image fusion Network for prediction of mild cognitive impairment in Journal of Physics: Conference Series* **2284** (2022), 012002.

519. Cheng, J., Liu, J., Kuang, H. & Wang, J., A fully automated multimodal MRI-based multi-task learning for glioma segmentation and IDH genotyping, *IEEE Transactions on Medical Imaging* **41,** 1520–1532 (2022).

520. Mohit Prabhushankar *et al., OLIVES Dataset: Ophthalmic Labels for Investigating Visual Eye Semantics* 2022, `https://zenodo.org/record/7105232`.

521. Cai, Z., Lin, L., He, H. & Tang, X., *Corolla: An Efficient Multi-Modality Fusion Framework with Supervised Contrastive Learning for Glaucoma Grading in 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)* (2022), 1–4.

522. Gutiérrez, Y., Arevalo, J. & Martánez, F., *Multimodal Contrastive Supervised Learning to Classify Clinical Significance MRI Regions on Prostate Cancer in 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (2022), 1682–1685.

523. Xing, X. *et al., Discrepancy and Gradient-Guided Multi-modal Knowledge Distillation for Pathological Glioma Grading in Medical Image Computing and Computer Assisted Intervention – MICCAI 2022* (eds Wang, L., Dou, Q., Fletcher, P. T., Speidel, S. & Li, S.) (Springer Nature Switzerland, Cham, 2022), 636–646, ISBN: 978-3-031-16443-9.

524. Wei, Y. *et al., Multi-modal learning for predicting the genotype of glioma* 2022, arXiv: 2203.10852 [cs.CV].

525. Taleb, A., Kirchler, M., Monti, R. & Lippert, C., *ContIG: Self-Supervised Multimodal Contrastive Learning for Medical Imaging With Genetics in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2022), 20908–20921.

526. Hager, P., Menten, M. J. & Rueckert, D., *Best of Both Worlds: Multimodal Contrastive Learning with Tabular and Imaging Data* 2023, arXiv: 2303.14080 [cs.CV].

527. AbdelMaksoud, E., Barakat, S. & Elmogy, M., A computer-aided diagnosis system for detecting various diabetic retinopathy grades based on a hybrid deep learning technique, *Medical & Biological Engineering & Computing* **60,** 2015–2038 (2022).

528. Teo, Z. L. *et al.,* Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis, *Ophthalmology* **128,** 1580–1591 (2021).

529. Patel, S. N., Shi, A., Wibbelsman, T. D. & Klufas, M. A., *Ultra-widefield retinal imaging: An update on recent advances* Jan. 2020.

530. Baumann, B. *et al.,* Total retinal blood flow measurement with ultrahigh speed swept source/Fourier domain OCT, *Biomed Opt Express* **2,** 1539–1552, ISSN: 2156-7085, (2023) (June 2011).

531. Or, C., Sabrosa, A., Sorour, O. & et al., Use of OCTA, FA, and Ultra-Widefield Imaging in Quantifying Retinal Ischemia: A Review, *Asia-Pacific Journal of Ophthalmology* **7,** 46–51 (2021).

532. Mohite, A. A., Perais, J. A., McCullough, P. & Lois, N., Retinal Ischaemia in Diabetic Retinopathy: Understanding and Overcoming a Therapeutic Challenge, *Journal of Clinical Medicine* **12,** ISSN: 2077-0383 (2023).

533. Lahsaini, I., El Habib Daho, M. & Chikh, M. A., Deep transfer learning based classification model for covid-19 using chest CT-scans, *Pattern Recognition Letters* **152,** 122–128, ISSN: 0167-8655 (2021).

534. Quellec, G. *et al.,* ExplAIn: Explanatory artificial intelligence for diabetic retinopathy diagnosis, *Medical Image Analysis* **72,** 102118, ISSN: 1361-8415 (2021).

535. Mirikharaji, Z. *et al.,* A survey on deep learning for skin lesion segmentation, *Medical Image Analysis* **88,** 102863, ISSN: 1361-8415 (2023).

536. Yun, S. *et al.,* CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features, *CoRR* **abs/1905.04899,** eprint: 1905.04899, http://arxiv.org/abs/1905.04899 (2019).

537. Li, X., Wei, Y., Wang, L., Fu, S. & Wang, C., MSGSE-Net: Multi-scale guided squeeze-and-excitation network for subcortical brain structure segmentation, *Neurocomputing* **461,** 228–243, ISSN: 0925-2312 (2021).

538. Bodapati, J. D., Shareef, S. N., Naralasetti, V. & Mundukur, N. B., MSENet: Multi-Modal Squeeze-and-Excitation Network for Brain Tumor Severity Prediction, *International Journal of Pattern Recognition and Artificial Intelligence* **35,** 2157005 (2021).

539. Li, T. *et al., Applications of Deep Learning in Fundus Images: A Review* 2021, arXiv: 2101.09864 [eess.IV], https://arxiv.org/abs/2101.09864.

540. Shamshad, F. *et al.,* Transformers in medical imaging: A survey, *Medical Image Analysis* **88,** 102802, ISSN: 1361-8415 (2023).

COLLEGE DOCTORAL BRETAGNE SCIENCES DE LA VIE ET DE LA SANTE

Université de Bretagne Occidentale

Titre : Fusion d'informations multimodales pour le diagnostic de la rétinopathie diabétique

Mot clés : classification de la RD, fusion d'informations multimodales, apprentissage profond

Résumé : Le diabète touche 422 millions de personnes dans le monde et 3,3 millions en France, provoquant des complications comme la rétinopathie diabétique (RD) et la cécité. La classification actuelle de la RD, basée sur la rétinophotographie couleur (CFP), peine à prédire l'évolution de la maladie. Les techniques d'imagerie modernes comme la rétinophotographie couleur ultra-grand champ (UWF-CFP), la tomographie en cohérence optique et angiographique (OCTA) fournissent des données complètes mais complexes, nécessitant de l'expertise pour l'analyse. Le projet EviRed vise le développement d'un système expert utilisant des images modernes et des données patients pour prédire la progression de la RD et assurer des traitements opportuns. Cette thèse, qui s'inscrit dans EviRed, explore l'utilisation de l'intelligence artificielle (IA) pour combiner ces différentes données, afin d'améliorer le diagnostic et la prédiction. Différents scénarios sont étudiés : l'analyse conjointe d'informations multimodales issues de l'OCTA, l'analyse de plusieurs spécifications d'acquisition OCTA ou encore l'analyse de l'OCTA avec l'UWF-CFP. De nouvelles architectures neuronales sont proposées pour cela. La validation clinique confirme l'efficacité de la fusion, qui améliore nettement la précision diagnostique par rapport aux images unimodales. L'algorithme proposé va renforcer le projet EviRed, contribuant à la révolution imminente du dépistage, du diagnostic et de la gestion de la RD.

Title: Multimodal information fusion for the diagnosis of diabetic retinopathy

Keywords: diabetic retinopathy classification, multimodal information fusion, deep learning

Abstract: Diabetes, affecting 422 million globally and 3.3 million in France, leads to complications like diabetic retinopathy (DR), causing blindness. The existing DR classification, based on outdated Color Fundus Photography (CFP), cannot predict disease progression accurately. Modern imaging techniques such as Ultra-Wide-Field CFP (UWF-CFP) and Optical Coherence Tomography Angiography (OCTA) generate comprehensive but complex fundus data needing expert analysis. The EviRed project aims to develop an expert system using updated imaging and patient data to predict DR progression and ensure timely treatments. As part of EviRed, this thesis investigates artificial intelligence (AI) to integrate the data, enhancing diagnosis and prediction. Deep learning network architectures combining imaging modalities are designed for better diagnostic performance. Scenarios involving joint analysis of multimodal OCTA information, different OCTA acquisition specifications, and OCTA with UWF-CFP analysis are examined, proposing new architectures for DR diagnosis. Clinical validation confirms the fusion's effectiveness, significantly improving diagnostic accuracy compared to unimodal images. The algorithm will strengthen EviRed, contributing to the imminent revolution in DR screening, diagnosis, and management.