THÈSE DE DOCTORAT

DE L'ÉTABLISSEMENT UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

PRÉPARÉE À L'UNIVERSITÉ DE BOURGOGNE

École doctorale n°37

Sciences Pour l'Ingénieur et Microtechniques

Doctorat d'Instrumentation et Informatique de l'Image

par

MME SISWADI ANNEKE ANNASSIA PUTRI

Computer-Aided-Diagnosis for Ocular Abnormalities from A Single Color Fundus Photography with Deep Learning

Microaneurysms and Multi-Label Detections

Thèse présentée et soutenue à Le Creusot, le 21 Mars 2023

Composition du Jury :

M SIDIBE DÉSIRÉ	Professeur à l'Université d'Evry	Président
M LAMARD MATHIEU	Ingénieur de recherche HDR à l'Université de Bretagne Occidentale	Rapporteur
M URRUTY THIERRY	Maître de Conférences HDR à l'Université de Poitiers	Rapporteur
M MADENDA SARIFUDDIN	Professeur à Universitas Gunadarma	Examinateur
M MERIAUDEAU FABRICE	Professeur à l'Université de Bourgogne	Directeur de thèse
MME BRICQ STÉPHANIE	Maitresse de Conférences HDR à l'Université de Bourgogne	Codirectrice de thèse

N° | X | X | X |

ACKNOWLEDGEMENTS

All praise and gratefulness are raised to Allah (SWT), who gave me the wisdom to finish my research. This thesis is intended to complete the requirement to finish my research at the ImViA laboratory, University of Burgundy. The completion of this thesis would not have been possible without the support and encouragement of many people. Therefore, I would like to thank:

- Gunadarma University, who gave me the opportunity and financial support to do this PhD research.
- Fabrice MERIAUDEAU and Stéphanie BRICQ, my PhD supervisors, for their guidance and support in this research.
- My beloved mother (Farida), father (Didi SISWADI), and the family for their prayers and support.
- Everyone who helped directly or indirectly to complete this thesis.

Contents

List of Figures

List of Tables

List of Abbreviations

AB	Asteroid Bodies
AH	Asteroid Hyalosis
AI	Artificial Intelligence
AION	Anterior Ischemic Optic Neuropathy
AP	Average Precision
ARMD	Age-Related Macular Degeneration
ASL	ASymmetric Loss
AUC	Area Under Curve
AUC (PR)	Area Under the PR Curve
AV	Arteriovenous
BERT	Bidirectional Encoder Representations from Transformers
BRAO	Branch Retinal Artery Occlusion
BRVO	Branch retinal vein occlusion
CAD	Computer-Aided Diagnosis
CALM	Confident Adaptive Language Modelling
CB	Coloboma
CBOW	Continuous Bag-of-Words
CF	Choroidal Folds
CFP	Color Fundus Photography
CL	Collateral
CLAHE	Contrast Limited Adaptive Histogram Equalization
CME	Cystoid Macular Edema
CNN	Convolutional Neural Network
CRAO	Central Retinal Artery Occlusion
CRS	Chorioretinitis
CRVO	Central Retinal Vein Occlusion
CSR	Central Serous Retinopathy
CWS	Cotton-Wool Spots
DCNN	Deep Convolutional Neural Network
DD	Disc Diameter
DETR	DEtection TRansformer
DLC	Directional Local Contrast

DME	Diabetic Macular Edema
DN	Drusens
DR	Diabetic Retinopathy
DRIVE	Digital Retinal Images for Vessel Extraction
ED	Euclidean Distance
EDN	Exudation
ERM	Epiretinal Membrane
EX	Hard Exudates
FA	Fluorescein Angiography
FFN	Feed-Forward Network
FN	False Negative
FOV	Field of View
FP	False Positive
FPI	False Positive per Image
FPN	Feature Pyramid Networks
FPR	False Positive Rate
FROC	Free-Receiver Operating Characteristic
GAN	Generative Adversarial Network
GCN	Graph Convolutional Network
GPT	Generative Pre-trained Transformer
GT	Ground-Truth
HE	Hemorrhages
HPED	Hemorrhagic Pigment Epithelial Detachment
HR	Hemorrhagic Retinopathy
IDRiD	Indian Diabetic Retinopathy Image Dataset
ISNS	Inverse of the Square root of a Number of Samples
LGE	Linguistic Guided Enhancement
LP	Large Positive
LS	Laser Scars
MA	Microaneurysm
MAE	Masked Autoencoder
mAP	Mean Average Precision
MAs	Microaneurysms
MCA	Macroaneurysm
MESSIDOR	Methods to Evaluate Segmentation and Indexing Techniques in the field of Retinal Ophthalmology
MH	Media Haze
MHL	Macular Hole

MLP	Multi-Layer Perceptron
MNF	Myelinated Nerve Fibers
MS	Macular Scar
MSE	Optimized-Mean Square Error
MTL	Multi-Task-Learning
MViT	Multi-scale Vision Transformers
MYA	Myopia
NLP	Natural Language Processing
NNLM	Neural Network Language Model
OCT	Optical Coherence Tomography
OCTA	Optical Coherence Tomography Angiography
OD	Optic Disc
ODC	Optic Disc Cupping
ODE	Optic Disc Edema
ODP	Optic Disc Pallor
ODPM	Optic Disc Pit Maculopathy
OOV	Out-Of-Vocabulary
PDR	Proliferative Diabetic Retinopathy
PE	Positional Encoding
PLQ	Plaque
PR	Precision-Recall
PRH	Preretinal Hemorrhage
PT	Parafoveal Telangiectasia
PTCR	Post-Traumatic Choroidal Rupture
ResNet	Residual Network
RFMiD	Retinal Fundus Multi-Disease Image Dataset
ROC	Receiver Operating Characteristic
ROI	Region Of Interest
RP	Retinitis Pigmentosa
RPE	Retinal Pigment Epithelium
RPEC	Retinal Pigment Epithelium Changes
RS	Retinitis
RT	Retinal Traction Detachment
S-TSDL	Symptoms-based TSDL
SCM	Spatial Correlation Module
SE	Soft Exudates
SGD	Stochastic Gradient Descent
SP	Small Positive

ST	Optociliary Shunt
STARE	STructured Analysis of the REtina
SWIN	Shifted WINdow transformers
TD	Tilted Disc
TN	True Negative
TP	True Positive
TPR	True Positive Rate
TSDL	Transformers Semantic Dictionary Learning
TSLN	Tessellation
TV	Tortuous Vessels
VH	Vitreous Hemorrhage
ViT	Vision Transformer
VS	Vasculitis

I

Context and Background

1

Introduction

The eyes are one of the most sensitive organs. They are responsible for four-fifths of all the information that is received by the human brain [Acharya et al., 2008]. Persistent ocular diseases give a high risk of retinal damage which can lead to blindness [Congdon et al., 2004]. Early detection and timely diagnosis of ocular pathologies are effective ways to prevent visual impairment. The detection of ocular diseases with the help of computer-aided diagnosis (CAD) is facilitated by the availability of retinal imaging modalities such as optical coherence tomography (OCT), optical coherence tomography angiography (OCTA), fluorescein angiography (FA), and color fundus photography (CFP). Among these modalities, CFP is the most cost-effective and simple approach for eye screening.

Diabetic Retinopathy (DR), one of the critical ocular abnormalities, is the primary focus of this research. DR is the most frequent complication of diabetes that can lead to blindness [Stitt et al., 2016] [WHO, 2021]. Diabetes Mellitus, commonly known as diabetes, is a condition when the body can not utilize glucose properly to provide energy. It occurs due to the lack of insulin produced by the pancreas. Uncontrolled diabetes can cause hyperglycemia, too much glucose in the bloodstream, which can cause vital damage to many of the body systems, like nerves and blood vessels. The number of adult people (aged 20-79 years) having diabetes has grown more than 60% over the past ten years, and it is estimated to reach 578 million by 2030, and 700 million by 2045 [Saeedi et al., 2019]. DR is a direct result of chronic hyperglycemia that causes damage to retinal capillaries which leads to capillary leakage and blockage [Hollow, 2015]. Approximately one in three people with diabetes have diabetic retinopathy, and one in ten will develop a vision-threatening form of the disease. DR affects 80% of patients who have had diabetes for more than ten years [Kertes et al., 2007]. DR is clinically diagnosed by analyzing the presence of microaneurysm (MA), hemorrhages (HE), soft exudates (SE), and hard exudates (EX) in color fundus images obtained from eye screening test [Yau et al., 2012]. The risk of vision loss caused by DR is preventable by early treatment, and the first visible symptom of DR in the CFP test is microaneurysms (MAs).

On the other hand, the possibility of a patient getting affected by more than one ocular disease increases the necessity to detect multiple ocular abnormalities from a single CFP. Considering efficiency, a single high-performance multi-label detection model is essential. However, one of the main challenges in this multi-label detection is to detect both frequent and rare ocular abnormalities from only a single CFP since rare ocular abnormalities are usually ignored in the detection because of the limited availability of training data [Pachade et al., 2021].

Machine learning and computer vision algorithms have been widely used to build CAD systems for various applications. The information from the input imaging and/or non-imaging data are extracted and interpreted to predict the outcome for a given task [Chan et al., 2020]. However, CAD systems built with machine learning generate more false positives than physicians and thus led to an increment in assessment time and unnecessary biopsies. Interestingly, deep learning technology overcomes these problems with great accuracy [Kim et al., 2019]. Deep learning is a neural network with multiple layers. It focuses to learn the relevant features needed to predict the output based on the input. Deep learning architecture consists of the input layer, hidden layer(s), and output layer. Deep learning frameworks are classified depending on the task and the type of data and they are used in various medical applications such as image classification, object detection, and image segmentation.

1.1/ Objectives of The Thesis

This research follows two main objectives:

1. To build a model for microaneurysms detection from a single color fundus photography.
2. To build a multi-label detection model from a single color fundus photography for both frequent and rare ocular abnormalities.

1.2/ Outline of the PhD Thesis Dissertation

The dissertation is written with five Chapters. The introduction and the objectives are described in Chapter 1. The clinical and theoretical backgrounds are presented in Chapter 2. Chapter 3 presents the first objective of the thesis, microaneurysms (MAs) detection, while the second objective, multi-label detection, is presented in Chapter 4. Finally, the conclusion and perspectives are explained in the last chapter, Chapter 5.

1.3/ PUBLICATIONS

Part of the materials contained in this thesis has been published in the following works:

- Siswadi, Anneke Annassia Putri, Stéphanie Bricq, and Fabrice Meriaudeau. "A Survey on Microaneurysms Detection in Color Fundus Images." *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS).* IEEE, 2020.
- Siswadi, Anneke Annassia Putri, Stéphanie Bricq, and Fabrice Meriaudeau. "Microaneurysms Detection in Color Fundus Image with Feature-based Background Suppression." *2022 26th International Conference on Pattern Recognition (ICPR).* IEEE, 2022.
- Siswadi, Anneke Annassia Putri, Stéphanie Bricq, and Fabrice Meriaudeau. "Multi-Label Ocular Abnormalities Detection with Semantic Dictionary Learning." *SPIE Medical Imaging: Computer-Aided Diagnosis.* SPIE, 2023.

2

BACKGROUND

This chapter is sub-divided into three sections: **Clinical Background** - consists of information about the retina and related diseases, imaging modalities, and the datasets; **Theoretical Background** - consists of the underlying methods of this research; **Evaluation Metrics** - consists of various methods that are used to evaluate the performance of the proposed methods.

2.1/ CLINICAL BACKGROUND

2.1.1/ HUMAN EYE

Eyes are the most sensitive and delicate organs. Figure 2.1 shows the structure of the human eye. The human eye consists of six main regions: cornea, aqueous humor, iris, lens, vitreous humor, and sclera. Other ocular domains include the retina and the choroid [Acharya et al., 2008].

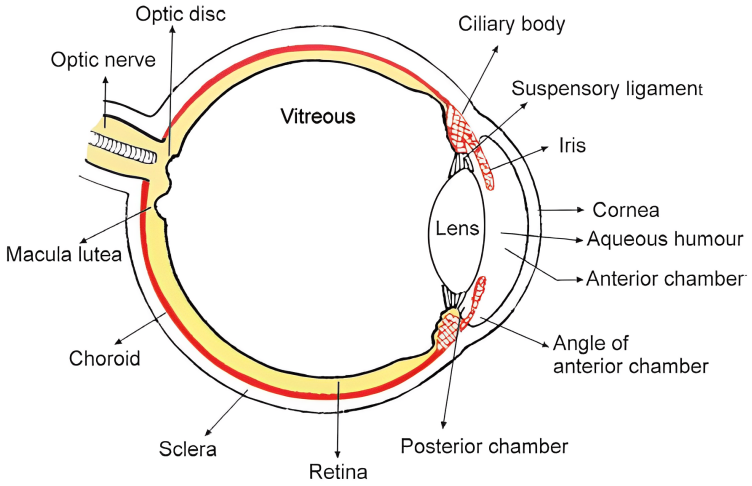


Figure 2.1: Structure of the human eye [Jogi, 2003].

The details of each main vision region are listed as follows:

- *Cornea* — the transparent, ellipsoid, anterior part of the eyeball. The cornea is the main refracting surface of the eye [Jogi, 2003].
- *Aqueous Humor* — an optically clear solution of water and electrolytes similar to tissue fluids except that aqueous humor usually has a low protein content [Galloway et al., 2016].
- *Iris* — divides the anterior part of the eye into anterior and posterior chambers. It contains aqueous humor secreted by the ciliary body [Jogi, 2003].
- *Lens* — has a thick elastic capsule, which prevents molecules (e.g., proteins) from moving in or out [Galloway et al., 2016].
- *Vitreous Humor* — a transparent gel consisting of a three-dimensional network of collagen fibers with inter-spaces filled with polymerized hyaluronic acid molecules and water [Galloway et al., 2016].
- *Sclera* — maintains the shape of the eye and gives attachment to the extraocular muscles [Jogi, 2003].
- *Retina* — the layer that converts light rays into electrical signals (transduction) for transmission to the brain [Galloway et al., 2016].
- *Choroid* — a highly vascularized structure in the human eye that accounts for 85% of the total ocular blood flow [Acharya et al., 2008].

The visual image which is produced by the optical system of the eye is received by the retina. It converts the light ray into an electrical signal, which undergoes initial processing, and then is transmitted through the optic nerve to the visual cortex, where the structural (form, color, and contrast) and spatial (position, depth, and motion) attributes are perceived. The fovea (the center of the macula) is responsible for good spatial resolution (visual acuity) and color vision [Vaughan et al., 1977]. On the other hand, the optic disc is a blind spot in human vision since there are no light-sensitive cells present [Galloway et al., 2016].

2.1.2/ Fundus Examination

Fundus examination is necessary to diagnose ocular abnormalities [Vaughan et al., 1977]. It is a routine examination performed by doctors and ophthalmologists to have an exclusive inspection of the patient's eye [Walker et al., 1990]. The fundus is the inside, back surface of the eye which is made up of the retina,

blood vessels, optic disc, macula, and fovea [Turbert, 2020]. Direct ophthalmoscopy and indirect ophthalmoscopy are the general types of fundus examination. Direct ophthalmoscopy corresponds to the hand-held ophthalmoscope that provides a monocular image, including a 15× magnified view of the fundus. The ophthalmoscope is held by the ophthalmologist to be close to the patient's pupil in 2-5 cm approximately [Vaughan et al., 1977]. In contrast, indirect ophthalmoscopy provides a much clear and wider field inverted view of the mid-air fundus image. The dilated pupil of an eye is examined through a mirror with a hole in it. The patient stays at arm's length from the observer and the mirror is held close to the observer's eye [Galloway et al., 2016]. However, clinical information and retinal imaging modalities are highly required to assist the ophthalmologist in diagnosing and monitoring some ocular abnormalities.

2.1.3/ Retinal Imaging Modalities

Retinal imaging holds an important role in the diagnosis of eye diseases. There are multiple retinal imaging modalities that are used by ophthalmologists to diagnose the eye diseases such as optical coherence tomography (OCT), fluorescein angiography (FA), color fundus photography (CFP), etc. Each imaging modality provides different features of the retina.

2.1.3.1/ Optical Coherence Tomography (OCT)

Optical Coherence Tomography (OCT) is a non-invasive imaging modality. The OCT image provides a cross-sectional view of the retina with unprecedented high resolution and allows detailed structures to be differentiated [Fujimoto et al., 2000]. OCT is used to examine and measure intraocular structures in three dimensions. It can be performed through an undilated pupil. Posterior segment OCT enables detailed analysis of the optic disc, retinal nerve fiber layer, and macula whereas the anterior segment provides high-resolution images and measurements of the cornea, iris, and intraocular devices and lenses [Riordan-Eva et al., 2011]. Figure 2.2 shows the layers of normal retina.

OCT serves as vital guidance for surgical interventional procedures. The OCT image can be used to detect and diagnose some retinal diseases such as glaucoma, macular edema, macular hole, central serous chorioretinopathy, age-related macular degeneration, epiretinal membranes, optic disc pits, and choroidal tumors [Fujimoto et al., 2000]. However, clinicians should be aware that scans and analyses are not without fault. Poor ocular media, patient compliance, and even saccadic movement can introduce image artifacts that can masquerade as pathology.

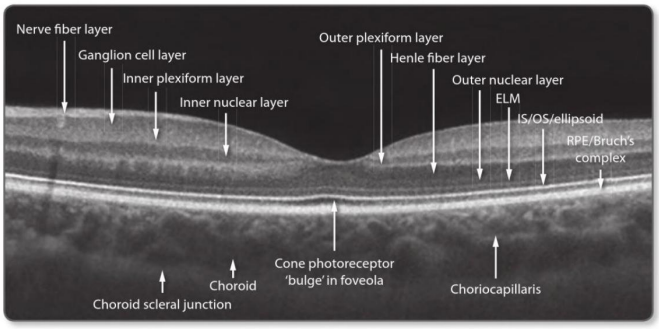


Figure 2.2: OCT image of normal retina [Reichel et al., 2015].

2.1.3.2/ Optical Coherence Tomography Angiography (OCTA)

Optical Coherence Tomography Angiography (OCTA) is a non-invasive imaging modality for the microvasculature of the choroid and retina with depth resolution. The OCTA image is generated using laser light reflectance of the surface of moving red blood cells to accurately depict vessels through different segmented areas of the eye, thus eliminating the need for intravascular dyes [Koustenis et al., 2017]. Figure 2.3 shows an OCTA image of a normal eye with details of microvasculature in the macula.

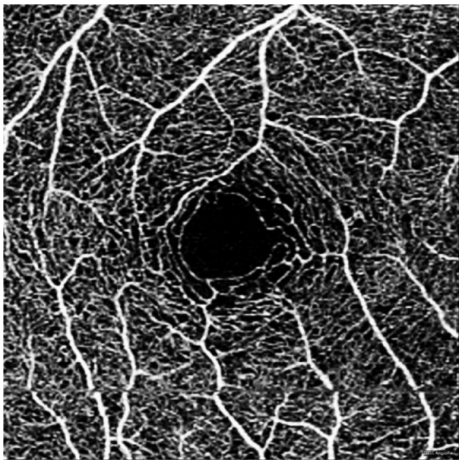
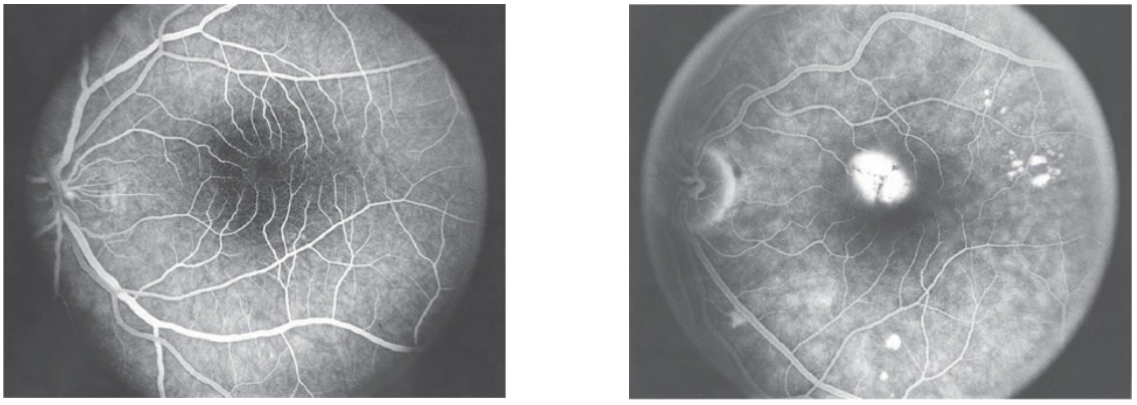


Figure 2.3: OCTA image of a normal eye [O'Keefe, 2022].

OCTA image provides the blood flow of all the vascular layers of the retina which also offers the possibility of imaging the radial peripapillary capillary network and the intermediate and deep capillary plexuses. OCTA images are used to have disease description and quantification, research into pathogenesis of disease, and development and evaluation of new treatments. OCTA can be performed much more rapidly than Fluorescein Angiography (FA). However, OCTA imaging protocols require re-scanning the same retinal position multiple times so it requires higher imaging speeds (A-scan rates) or longer imaging times than structural OCT [Spaide et al., 2018].

2.1.3.3/ Fluorescein Angiography (FA)

Fluorescein Angiography is an eye examination that uses fluorescein dye and a special camera to examine the circulation of the retina and choroid [Reichel et al., 2015]. Fluorescein dye emits green light when stimulated by blue light. It highlights vascular and anatomic details of the fundus photograph [Riordan-Eva et al., 2011].



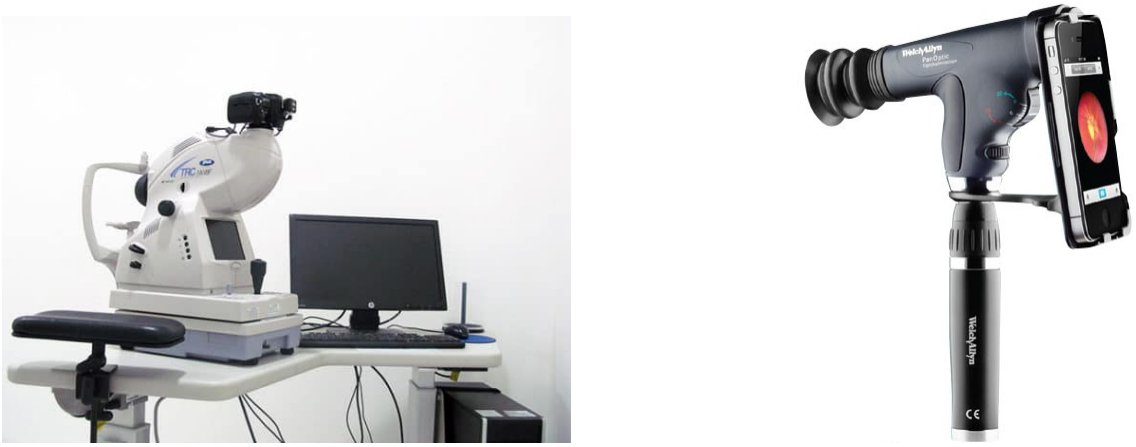
(a) Normal fundus **(b)** Abnormal fundus

Figure 2.4: Fluorescein angiogram images [Vaughan et al., 1977].

In a normal fundus FA image (see Figure 2.4a), the choroidal and retinal circulations are anatomically separated by a thin, homogeneous monolayer of the pigmented cell (pigment epithelium), whereas an abnormal fundus image (see Figure 2.4b) shows a darker central zone because of the macula denser pigmentation and an abnormal increase in visibility of the background fluorescence [Vaughan et al., 1977]. FA examination is necessary to plan the laser treatment of retinal vascular disease [Riordan-Eva et al., 2011]. FA is invaluable in the assessment of the choroidal vasculature [Ruia et al., 2021] and the anatomy, physiology, and pathology of the retinal such as retinal neovascularization, intraretinal microvascular abnormalities, and capillary nonperfusion [Baumal, 2018].

2.1.3.4/ Color Fundus Photography (CFP)

Compared to other optical imaging modalities, color fundus photography which is taken by the fundus camera is cost-effective and simple [Yao et al., 2022][Besenczi et al., 2016]. Despite its limitation in the field of view (FOV), CFP is frequently required for the screening purpose of eye diseases [Xiao et al., 2020]. CFP is a non-invasive technique to record the fundus image so that it can be referred to in another location or time in a wide variety of ophthalmic conditions [Besenczi et al., 2016].



(a) (b)

Figure 2.5: Digital fundus camera. (a) Topcon TRC-NW8Fplus [Topcon,]. (b) iExaminer [Allyn,].

A fundus camera is an indirect ophthalmoscope-based camera. It is built with an attached camera and a specialized low-power microscope. Fundus cameras are described by the field of view (FOV) - the optical angle of acceptance of the lens. The normal FOV, 30°, visualizes a retina image that is 2.5 times larger than the actual retina. Fundus camera provides capturing the images between 45° and 140° [Saine et al., 2002].

The fundus image is captured when a patient sitting upright at the fundus camera while an ophthalmic photographer focuses and aligns the camera [Panwar et al., 2016]. The conventional traditional fundus cameras have a limited FOV and frequently require pupillary dilation for reliable examination of eye conditions [Yao et al., 2022]. Despite the advantage of digital retinal imaging that provides rapidly acquired, high-resolution, reproducible images [Baumal, 2018], the fundus camera is still actively innovated to provide higher quality and lower cost [Panwar et al., 2016]. Figure 2.5 shows two different types of digital fundus cameras.

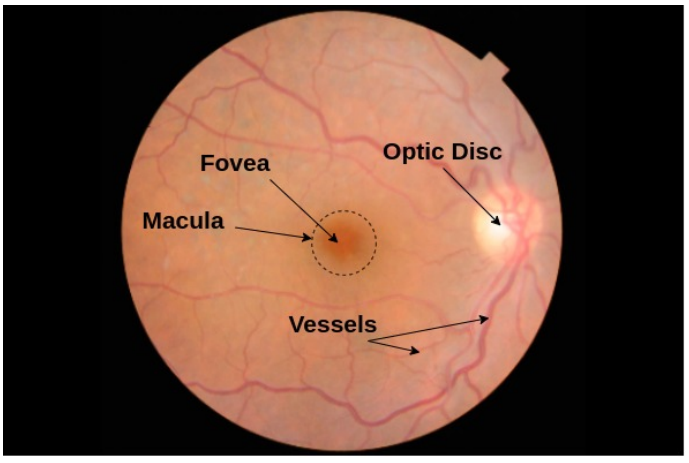


Figure 2.6: Color fundus image.

The color fundus image consists of three channels (Red-Green-Blue). As seen in Figure 2.6, the optic disc, macula, fovea, and vessels are clearly visible in a color fundus image. The color fundus image is also helpful in interpreting the FA image since some retinal landmarks are visible in the color fundus image but not in the FA image [Saine et al., 2002]. CFP is used for clinical studies, disease documentation, telemedicine, and patient education [Baumal, 2018].

2.1.4/ Ocular Abnormalities

There are several ocular abnormalities out of which 45 abnormalities are discussed in this section. All the following ocular abnormalities can be diagnosed by analyzing the color fundus image. RFMiD dataset provides color fundus images annotated with these abnormalities for the multi-label classification (see details in Section 2.1.5.5).

2.1.4.1/ Diabetic Retinopathy (DR)

Diabetic retinopathy is the most common microvascular complication of diabetes mellitus [Vaughan et al., 1977]. The worst condition caused by DR is blindness and having a painful eye which makes removal of the eye as the only treatment option [Galloway et al., 2016]. DR can be diagnosed by analyzing the presence of one or more lesions that are visible in the color fundus image such as microaneurysms, hemorrhage, soft exudates, and hard exudates (see Figure 2.7).

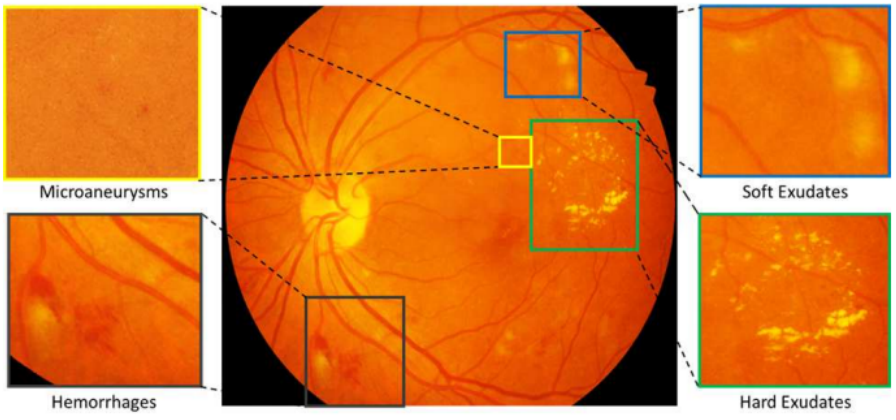


Figure 2.7: The symptoms of diabetic retinopathy [Porwal et al., 2018].

2.1.4.2/ Age-Related Macular Degeneration (ARMD)

The development and progression of age-related macular degeneration (ARMD) are caused by several factors such as advanced age, white race, heredity, and a history of

smoking. The symptoms include blurred vision, decreased contrast sensitivity, abnormal dark adaptation, and the need for brighter light or additional magnification to read smaller prints [Jager et al., 2008]. A color fundus image is utilized to grade the ARMD with the drusen characteristics (size, type, area), pigmentary abnormalities, and the presence of the characteristic of neovascular abnormalities [Davis et al., 2005].

2.1.4.3/ MEDIA HAZE (MH)

Media haze is a condition of opacity in the media called lenticular regions, which mostly results in bad visual interpretation of the viewed object or any entity [Sengar et al., 2021]. The opacity of media can work as a hallmark for the occurrence of cataracts, vitreous opacities, corneal edema, or small pupils [Chen et al., 2011]. In order to decrease the risk of blindness through cataracts, earlier diagnosis is required to facilitate early state treatment.

2.1.4.4/ DRUSENS (DN)

Drusens (DN) is the earliest abnormal sign of ARMD that is visible in the fundus image. DN appears as white-yellow deposits on the retina [Sathiyamurthy et al., 2007]. It is an extracellular material lying between the basement membrane of the retinal pigment epithelium (RPE) and the inner collagenous zone of Bruch's membrane [Gella et al., 2016].

2.1.4.5/ MYOPIA (MYA)

Myopia (MYA) is the thinning and tractional changes to various layers and structures within the eye. The complication of MYA may result in retinal detachment, posterior staphyloma, retinal pigment epithelium and choroidal atrophy, and subretinal hemorrhages [Ho et al., 2017].

2.1.4.6/ BRANCH RETINAL VEIN OCCLUSION (BRVO)

Branch retinal vein occlusion (BRVO) is a group of abnormal conditions such as arteriovenous (AV) crossing with vein compression, degenerative changes of the vessel wall, and hematological factors that constitute the primary mechanism of vessel occlusion [Rehak et al., 2008]. BRVO is associated with hypertension, diabetes mellitus, hyperlipidemia, systemic and inflammatory diseases, thrombophilia and hypercoagulation, medications, and ocular conditions. The visible clinical characteristic of BRVO in fundus images are flame-shaped, dot and blot hemorrhage, soft exudates, hard exudates, retinal edema, and dilated tortuous veins [Jaulim et al., 2013].

2.1.4.7/ Tessellation (TSLN)

Tessellation (TSLN) is a characteristic of myopic eyes. It is an important sign of the development of retinochoroidal changes. TSLN is a retinal condition in which the choroidal vessels are visible through the retina owing to reduced pigmentation or hypoplasia of the retinal pigment epithelium (RPE) [Yoshihara et al., 2014].

2.1.4.8/ Epiretinal Membrane (ERM)

Epiretinal membrane (ERM) is a delicate cellular membrane occurring at the vitreoretinal juncture [Foos, 1974]. ERM is a thin sheet of glial cells. It appears as an irregular light reflex and occurs in the macular region [Pachade et al., 2021].

2.1.4.9/ Laser Scars (LS)

Laser scars (LS) are the damages that occurred in the retinal neurons as a result of laser therapy. It affects central vision, central acuity, and night vision [Zhang et al., 2011]. LS appears as circular or irregular-shaped scars on the retinal surface [Pachade et al., 2021].

2.1.4.10/ Macular Scar (MS)

Macular scars (MS) are caused by inflammatory intraocular reaction or infection that causes slight injury to RPE [Pachade et al., 2021]. MS appears as an egg yolk-shaped, dense homogeneous yellow deposit [Frangieh et al., 1982].

2.1.4.11/ Central Serous Retinopathy (CSR)

Central serous retinopathy (CSR) is an eye disease caused by the watery fluids behind the retina. It affects the vision causing abnormalities such as obscured vision, metamorphopsia, and a decrease in shading vision. CSR is characterized as a dark spot of a blister of fluid appearing in the macula [Hassan et al., 2021].

2.1.4.12/ Optic Disc Cupping (ODC)

Optic disc cupping (ODC) is a sign of glaucoma. ODC is characterized by the increase in the size of the optic cup and in the cup-to-disc ratio. It is also characterized by enlarged chorioretinal atrophy, backward bowing or excavation of the lamina cribrosa, and vertical elongation of the optic cup. Cupping also occurs as a result of inconsistent and infrequent optic neuropathies other than glaucoma [Piette et al., 2006].

2.1.4.13/ Central Retinal Vein Occlusion (CRVO)

Central Retinal Vein Occlusion (CRVO) is characterized by the appearance of dilatation, retinal hemorrhage, optic disc edema, and tortuosity of retinal veins [Fong et al., 1993]. Cotton wool spots and cystoid macular edema are other signs of CRVO. The risk factors causing CRVO are diabetes, glaucoma, increasing age, and hypertension [McAllister, 2012].

2.1.4.14/ Tortuous Vessels (TV)

Tortuous vessels (TV) appear as a marked tortuosity of the retinal vessels. It is associated with diabetes, hypertension, and other genetic disorders [Pachade et al., 2021].

2.1.4.15/ Asteroid Hyalosis (AH)

Asteroid hyalosis (AH) is also called Benson disease. It is a condition in which asteroid bodies (AB) are present in the vitreous body. AB appears as small yellow-white and spherical particles. AH is associated with vitreous opacities, hereditary pigmentary retinopathies, and diabetes mellitus. AH is diagnosed by detecting the presence of AB in the anterior and central vitreous body in the fundus image [Khoshnevis et al., 2019].

2.1.4.16/ Optic Disc Pallor (ODP)

ODC and optic disc pallor (ODP) are considered as the visible signs of glaucoma, especially if associated with loss of visual field [Schwartz et al., 1973]. The changes in tissue transflucency and reflectance following the axonal loss and glial reorganization is the cause of ODP. ODP can be diagnosed by focusing the observation on the characteristics of the optic disc [Yang et al., 2019].

2.1.4.17/ Optic Disc Edema (ODE)

Optic disc edema (ODE) is known as optic disc swelling. ODE is a condition in which the axonal is swelling and the fluid surrounding the axons is increasing [Van et al., 2007]. ODE may be associated with optic disc ischemia, cotton wool spots, exudates, Paton's lines, choroidal folds, and venous pulsation [Yan et al., 2021].

2.1.4.18/ Optociliary Shunt (ST)

Optociliary shunt (ST) vessels are the vessels on the optic nerve head which connect the retinal and choroidal circulations [Haskes et al., 1995]. ST is associated with optic nerve sheath meningioma, spheno-orbital meningiomas, triad symptoms of optociliary veins, disc pallor, CRVO, papilloedema, optic nerve glioma, optic disc drusen, arachnoid cyst of the optic nerve, phakomatosis, and chronic glaucoma [Lee et al., 2004].

2.1.4.19/ Anterior Ischemic Optic Neuropathy (AION)

Anterior ischaemic optic neuropathy (AION) is the change in color of the optic disc (into pale), followed by peripapillary hemorrhages, and invariable optic atrophy in later stages. AION is characterized by observing the changes in the optic disc and retina. It shows as chalky-white swelling with a rare hemorrhage or pale pink edema with flame-shaped hemorrhages depending on the severity of the AION [Hayreh et al., 1981].

2.1.4.20/ Parafoveal Telangiectasia (PT)

Parafoveal telangiectasia (PT) is a microvascular abnormality of the macula [Millay et al., 1986]. It is detected by changes in macula such as yellow, lipid-rich exudation, parafoveal graying, abnormalities from distention, and tortuous blood vessels [Pachade et al., 2021].

2.1.4.21/ Retinal Traction Detachment (RT)

Retinal traction detachment is the separation of the neurosensory retina from the retinal pigment epithelium (RPE) due to the traction caused by proliferative membranes present over the retinal surface. The proliferative membranes can occur due to proliferative retinopathies, the most common being proliferative diabetic retinopathy (PDR). Retinal traction detachments are caused by tractional forces and the detached retina takes a concave shape [Mishra et al., 2022].

2.1.4.22/ Retinitis (RS)

Retinitis (RS) is the inflammation of the retina and an acute RS can cause vision loss. RS is caused by microbes such as *Toxoplasma*, *Cytomegalovirus*, *Herpes Zoster*, *Herpes Simplex*, and *Candida*. RS can be diagnosed by detecting some lesions that appear in the fundus image such as white or slightly yellow lesions in the posterior pole or pe-

riphery depending upon the aetiological agent. RS may be associated with vasculitis, hemorrhage, vitreous inflammation, and vascular occlusions [Gupta et al., 2021].

2.1.4.23/ Chorioretinitis (CRS)

Chorioretinitis (CRS) is the uveitis or inflammation of the choroid and the retina. It can occur in the iris, ciliary body, choroid, retina, retinal vessels, vitreous, optic nerve head, and sclera. Infections are the common cause of CRS and *Toxoplasma Gondii* is the most common infection. CRS is characterized by diffuse or focal inflammatory infiltrates which appear in the retina and choroid while the parasites may appear as free tachyzoites, or tissue cysts with mononuclear inflammatory infiltrates surrounding retinal blood vessels [Geetha et al., 2021].

2.1.4.24/ Exudation (EDN)

Exudation (EDN) appears as white or yellowish lipid deposits with sharp edges. EDN is a sign of several eye diseases, for instance, circular exudates surrounding the macular area are the sign of circinate retinopathy; mass exudates in the macular region are the sign of exudative maculopathy, and star-like exudates in the macular region are the sign of the macular star [Pachade et al., 2021].

2.1.4.25/ Retinal Pigment Epithelium Changes (RPEC)

Retinal pigment epithelium changes (RPEC) are the changes in the condition of RPE such as ages, a number of structures, increase in the density of residual bodies, accumulation of lipofuscin, accumulation of basal deposits on or within Bruch's membrane, formation of drusen, thickening of Bruch's membrane, microvilli atrophy, and disorganization of the basal infoldings. RPE transports the essential nutrients for the homeostasis of the neural retina and also helps in regeneration [Bonilha, 2008].

2.1.4.26/ Macular Hole (MHL)

Macular hole (MHL) is a full-thickness defect of retinal tissue involving the anatomic fovea [Ho et al., 1998]. MHL is associated with laser treatment, retinal vascular disorders, retinal detachment, and cystoid macular edema [Maggon et al., 2009].

2.1.4.27/ Retinitis Pigmentosa (RP)

Retinitis pigmentosa (RP) is an inherited retinal dystrophy caused by the loss of photoreceptors. The most common form of RP is the progressive loss in the peripheral visual field in daylight and also rod-cony dystrophy that is characterized by night blindness. RP is diagnosed by observing the presence of night blindness and peripheral visual field defects, hypo-volted electroretinogram traces, progressive worsening of these signs, and also lesions in the fundus. The characteristics of RP that are visible in fundus image are pigmentary deposits resembling bone spicules (initially in the peripheral retina), attenuation of the retinal vessels, waxy pallor of the optic disc, and various degrees of retinal atrophy [Hamel, 2006].

2.1.4.28/ Cotton-Wool Spots (CWS)

Cotton-wool spots (CWS) lie superficially as opaque swellings in the retina, occurring as acute lesions. The occurrence of CWSs is a sign of serious vascular damage [Schmidt, 2008].

2.1.4.29/ Coloboma (CB)

Ocular coloboma can be seen in isolation and in an impressive number of multisystem syndromes. Ocular CB is caused by some factors such as inheritance and environmental influences [Gregory-Evans et al., 2004]. Clinically, ocular CB is caused by defective closure of the embryonal fissure. The ocular CB which is caused by chromosomal abnormalities is usually associated with systemic abnormalities. There are several effects of the ocular CB. There are several effects of ocular CB. In the lightest case, the structure and functions of the eye are hardly affected, whereas in the worst case, the eye can become non-functional because of compression by an orbital cyst. In the fundus image, the condition of ocular CB is found based on the size variability of the coloboma (anteroposterior and transverse extent) and the involvement of the optic disc and also fovea [Lingam et al., 2021].

2.1.4.30/ Optic Disc Pit Maculopathy (ODPM)

Optic disc pit maculopathy (ODPM) is characterized by intraretinal and subretinal fluid accumulation in the macula which causes visual deterioration. The source of this fluid may be vitreous fluid, cerebrospinal fluid, leakage from blood vessels at the base of the pit, or leakage from the choroid. ODPM is categorized as a rare diagnosis. The trigger of

its development is still unclear and there are only limited studies in its treatment guidance [Moisseiev et al., 2015].

2.1.4.31/ Preretinal Hemorrhage (PRH)

Preretinal hemorrhage (PRH) may occur after the rupture of retinal vessels. PRH is associated with excessive physical exercises and increased venous pressure (Valsalva retinopathy) or retinal vascular alterations (macroaneurysms and proliferative diabetic retinopathy). The most common site for hemorrhage in Valsalva retinopathy is the posterior pole because of pre-existing anatomical space. However, the term PRH is defined because of the difficulties in the indication of the accurate site of hemorrhage (between the retina and internal limiting membrane, or internal limiting membrane and hyaloid) [Felippe et al., 2004].

2.1.4.32/ Myelinated Nerve Fibers (MNF)

Myelinated nerve fibers (MNF) are developmental anomalies that are present in approximately 1% of all eyes. It is associated with ipsilateral high myopia, amblyopia, and various ocular and systemic abnormalities. MNF is identified by the presence of ectopic oligodendrocytes [Tarabishy et al., 2007]. The lesions of MNF appear as white or gray-white striated patches corresponding in shape to the distribution of retinal nerve fibers and demonstrating frayed borders [Straatsma et al., 1981].

2.1.4.33/ Hemorrhagic Retinopathy (HR)

Hemorrhagic retinopathy (HR) is a form of retinopathy that is hemorrhagic in nature. HR is associated with diabetes, hypertension, and occlusion of the central vein [Pachade et al., 2021].

2.1.4.34/ Central Retinal Artery Occlusion (CRAO)

Central retinal artery occlusion (CRAO) is an acute stroke of the eye and is categorized as an ophthalmic emergency. CRAO is a sign of end-organ ischemia and the underlying atherosclerotic disease that may become a risk of future cerebral stroke and ischaemic heart disease. The most common cause of CRAO is embolism usually due to atherosclerotic plaques. The signs that are found in the fundus image of the CRAO are retinal opacity in the posterior pole, cherry-red spot, cattle trucking, retinal arterial attenuation, optic disc edema, and pallor [Varma et al., 2013].

2.1.4.35/ Tilted Disc (TD)

The tilted disc (TD) is a congenital anomaly of the optic nerve head [Dorrell, 1978]. Visual field deficits, retinal pigment epithelial and choroidal hypoplasia, electro-functional abnormalities, choroidal neovascular development, and refractive error are examined as the features of TD [Sowka et al., 1999].

2.1.4.36/ Cystoid Macular Edema (CME)

Cystoid macular edema (CME) is a common pathologic sequel of the retina which occurs as the result of the cystic accumulation of extracellular intraretinal fluid in the outer plexiform and inner nuclear layers of the retina, as a result of the breakdown of the blood-retinal barrier [Quinn, 1996]. CME may be associated with a wide variety of ocular conditions such as central or branch retinal vein occlusion, diabetic retinopathy, and cataract [Rotsos et al., 2008]. CME is also a major cause of vision loss for patients who are infected by HIV with immune recovery uveitis [Rothova, 2007]. CME is often detected by analyzing the fluid which is visible in the macula area [Rotsos et al., 2008].

2.1.4.37/ Post-Traumatic Choroidal Rupture (PTCR)

Post-traumatic Choroidal Rupture (PTCR) are breaks in the choroid, Bruch's membrane, and RPE which occur as an effect of blunt ocular trauma that can lead to serious macular or hemorrhagic detachment [Pachade et al., 2021].

2.1.4.38/ Choroidal Folds (CF)

Choroidal folds (CF) are clinically detected as dark or light streaks on ophthalmoscopy. CF is associated with other pathological conditions such as tumors, central serous retinopathy, papilloedema, and choroidal naevi. CF occurs because of the combination of an anatomical attachment of Bruch's membrane to the underlying choriocapillaris and congestion of the choriocapillaris [Jaworski et al., 1999].

2.1.4.39/ Vitreous Hemorrhage (VH)

The cause of Vitreous Hemorrhage (VH) is unknown except if there is a clinical sign which clearly points in another direction such as a retinal tear, diabetic retinopathy, and retinal vein occlusion [Lindgren et al., 1996]. It can be also caused by the pathologic mechanisms of disruption of normal retinal vessels, bleeding from diseased retinal vessels or abnormal new vessels, and extension of hemorrhage through the retina from other sources

[Spraul et al., 1997]. There are some other ocular abnormalities that cause VH infrequently such as hemoglobinopathies, age-related macular degeneration, retinal arterial macroaneurysm, intraocular tumors, and vascular occlusive disease [Goff et al., 2006] [Lindgren et al., 1996].

2.1.4.40/ MACROANEURYSM (MCA)

Macroaneurysm (MCA) is an uncommon entity affecting the posterior segment of the eye. MCA is a fusiform or round dilation of the retinal arterioles that occurs in the temporal retina [Pachade et al., 2021]. MCA results in vision loss because of the presence of edema, hemorrhage, exudates, or other secondary complications [Singh et al., 2021].

2.1.4.41/ VASCULITIS (VS)

Vasculitis (VS) is classified based on the size of the vessel, the location, and the associated histological changes. VS is a condition in which the vessel wall is damaged. Clinically, VS is characterized by intraretinal hemorrhage and cotton wool. VS is caused by the infections of the virus such as *Cytomegalo*, *Herpes Simplex*, *Syphilis*, and *Tuberculosis* [Rosenbaum et al., 2016].

2.1.4.42/ BRANCH RETINAL ARTERY OCCLUSION (BRAO)

Branch retinal artery occlusion (BRAO) is a decrement of arterial blood flow to the retina that leads to ischemic damage. The vision is affected due to vascular occlusion. BRAO is caused by the perfusion decrement in a branch retinal artery that comes from emboli to a branch of the central retinal artery. Carotid, hypertension, and hypercholesterolemia are some of the risk factors of BRAO. BRAO is detected by the presence of retinal whitening following the course of a branch artery [Santos et al., 2021].

2.1.4.43/ PLAQUE (PLQ)

Plaque (PLQ) detection is an important sign of atherosclerosis. PLQ consists of embolic material from atheromatous lesions in the aorta or carotid arteries and it can cause occlusion of the arteriole. PLQ appears as a bright orange color spot. It is observed at bifurcations of the retinal arterioles with the occlusive disease within the carotid arterial system and the vertebral-basilar arterial system [Hollenhorst, 1961].

2.1.4.44/ Hemorrhagic Pigment Epithelial Detachment (HPED)

Hemorrhagic pigment epithelial detachment (HPED) is a retinal condition in which the pigment epithelium is separated from the Bruch's membrane because of hemorrhage. HPED is associated with polypoidal choroidal vasculopathy or choroidal neovascularization [Pachade et al., 2021].

2.1.4.45/ Collateral (CL)

Collateral (CL) is the existence of the new retinal vessels which are developed within the existing vessel network. The risk factors of CL are vascular occlusion, glaucoma, and/or optic nerve sheath meningioma [Sowka et al., 2014]. It is associated with other diseases such as optic disc drusen, high myopia, and diabetes [Pachade et al., 2021].

2.1.5/ Fundus Image Datasets

Fundus image datasets are the vital key to design a computer-aided-diagnosis (CAD) system to identify ocular abnormalities. Alongside computer vision, informative analyses of the fundus images are provided to assist the ophthalmologist to conclude the diagnosis. Fundus image datasets provide color fundus images and specific information of the corresponding fundus image as ground-truth data such as lesions, diseases, or parts of the retina. All fundus image dataset used in this research are publicly available online.

They are vary depending on the types of data provided:

- *E-Ophtha* and *IDRiD* are used in microaneurysms detection since they provide the lesions of DR overlaid in each color fundus image.
- *DRIVE* and *STARE* are used in vessel segmentation as they provide the retinal vessels annotation of color fundus images.
- *MESSIDOR* and *IDRiD* are used in OD and Fovea localization as they provide the OD and fovea annotation of color fundus images.
- *RFMiD* is annotated with various eye diseases and is used in multi-disease detection.

2.1.5.1/ E-Ophtha

E-Ophtha [Decenciere et al., 2013] is a public dataset of color fundus images that are collected specially for research in DR. The dataset was generated from the OPH-

DIAT Telemedical network for DR screening, in the framework of the ANR-TECSAN-TELEOPHTA project funded by the French Research Agency (ANR). The fundus images in the E-Ophtha dataset are annotated by expert ophthalmologists. The E-Ophtha dataset provides the annotation of two lesions of DR, microaneurysms (E-Ophtha-MA) and exudates (E-Ophtha-EX). E-Ophtha-MA consists of 148 images with microaneurysms or small hemorrhages and 233 images with no lesions, and E-Ophta-EX consists of 47 images with exudates and 35 images without any lesions. The resolutions of the image are 1440x960 and 2544x1696 pixels. Data was acquired by Canon CR-DGI and Topcon TRC-NW6 cameras with 45° FOV.

2.1.5.2/ Indian Diabetic Retinopathy Image Dataset (IDRiD)

IDRiD dataset [Porwal et al., 2018] is organized in conjunction with IEEE International Symposium on Biomedical Imaging 2018, Washington D.C. Fundus images are obtained from an Eye Clinic that is located in Nanded, India. The images with the resolution of 4288×2848 pixels are acquired using a digital fundus camera (Kowa $VX - 10\alpha$) with 50° FOV.

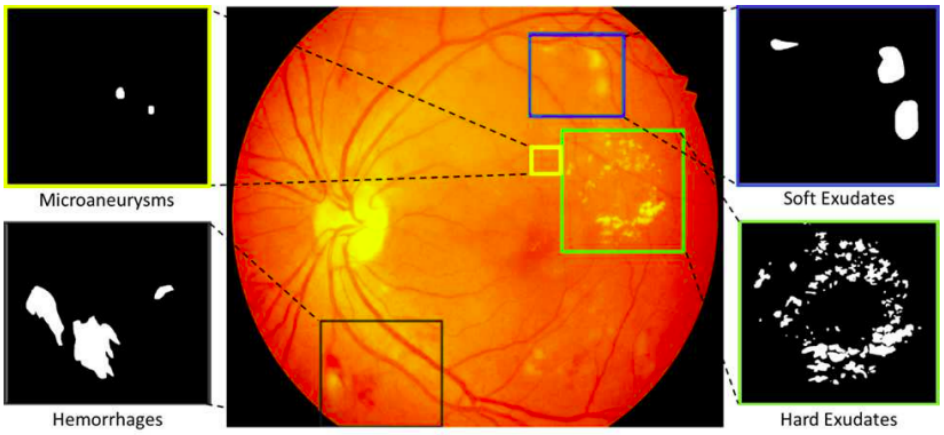


Figure 2.8: Color fundus image and its annotations (lesions of DR) [Porwal et al., 2018].

IDRiD dataset provides color fundus images with three types of data annotation, annotation of lesions in pixel level (see Fig. 2.8), DR and Diabetic Macular Edema (DME) disease grading, and center pixel location of Optic Disc (OD) and fovea. The dataset contains 81 color fundus images with signs of DR and 164 images of healthy eyes. The 81 images with signs of DR are divided into 54 training data and 27 testing data. IDRiD dataset also provides 516 color fundus images with OD and fovea locations that are divided into 413 training data and 103 testing data. DR and DME grading are provided for 516 images.

2.1.5.3/ Digital Retinal Images for Vessel Extraction (DRIVE)

Fundus images in the DRIVE dataset are obtained using a Canon $CR5$ non-mydriatic 3CCD camera with a 45° FOV. Each image has three channels (RGB) with resolutions of 565×584 and 768×584 pixels. DRIVE dataset contains 40 color fundus images with 7 abnormal pathology cases. The images are split into 20 images equally for the training and testing set. The training set consists of a color fundus image with one manual retinal vessel segmentation (see Fig. 2.9) that was annotated by an ophthalmologist expert while the testing set consists of color fundus images with two different observations, the first observation is accepted as ground-truth [Staal et al., 2004].

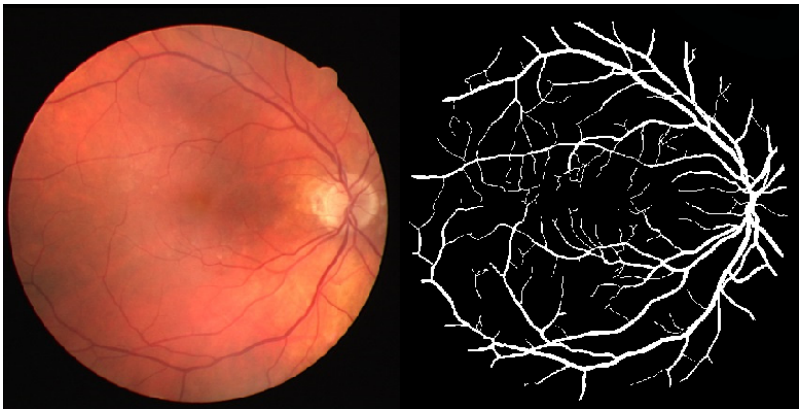


Figure 2.9: Color fundus image and retinal vessel annotation [Iqbal et al., 2018].

2.1.5.4/ STructured Analysis of the Retina (STARE)

STARE project was initiated in 1975 by Michael Goldbaum, M.D., at the University of California, San Diego. The project was funded by the U.S. National Institute of Health. Fundus images in the STARE dataset were obtained from the Shiley Eye Center at the University of California, San Diego, and from the Veterans Administration Medical Center in San Diego. STARE dataset contains around 40 color fundus images with two retinal vessels annotation. This dataset also provides multi-label with the features (402 images), artery/vein (10 images), and optic nerve annotations (80 images). The resolution of the image in this dataset is 700×605 pixels. The data was taken by Topcon TRV-50 camera with 35° FOV [Hoover et al., 2000].

2.1.5.5/ Retinal Fundus Multi-Disease Image Dataset (RFMiD)

RFMiD is a public dataset that provides different information on color fundus images. The fundus images were acquired using three different digital fundus cameras, TOPCON $3D\ OCT - 2000$, Kowa $VX - 10\alpha$, and TOPCON $TRC - NW300$ with 45° and 50°

FOV. The images are centered either on the macula or on the optical disk. They provide images with 2144×1424, 4288×2848, and 2048×1536 resolutions. RFMiD dataset consists of color fundus images with label annotations of the presence of the diseases. The annotations fall into two categories, the classification of normal and abnormal retinal images and the multi-label of 45 different ocular abnormalities (explained in Section 2.1.4). RFMiD dataset contains 3200 color fundus images that are divided into 1920 images for the training set, 640 images for the evaluation set, and 640 for the test set [Pachade et al., 2021].

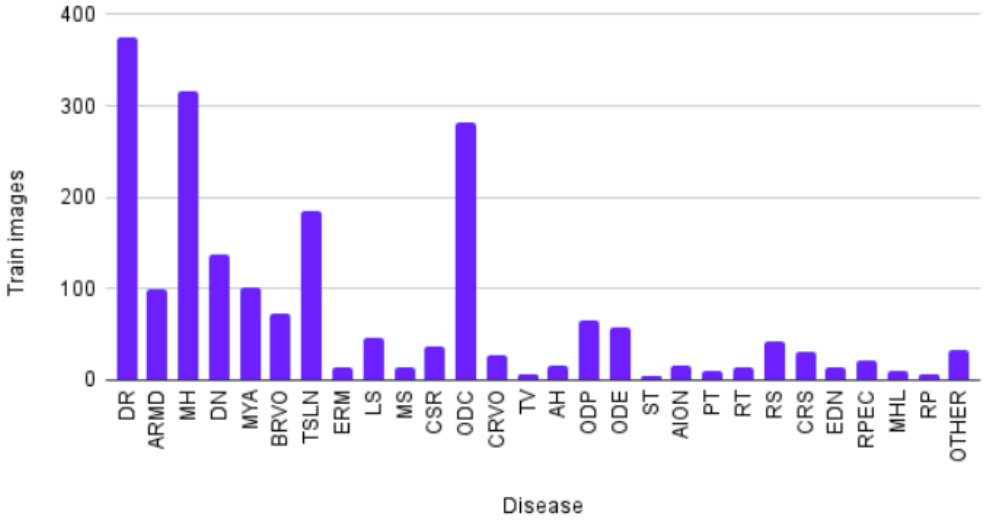


Figure 2.10: Class distribution of RFMiD dataset.

RFMiD is a dataset that provides multi-label annotation for both frequent and rare ocular abnormalities. As figured in Fig. 2.10, the highest total number of images for a class is 376 and the lowest is 6 images for a class. Due to the less acquired data for some rare eye diseases, another class named 'OTHER' is created. It consists of 19 classes that have less than 10 color fundus images. These data are split in the same data distribution into the training set, evaluation set, and testing.

2.1.5.6/ MESSIDOR

Messidor dataset stands for Methods to Evaluate Segmentation and Indexing Techniques in the field of Retinal Ophthalmology. Messidor was a research program funded by the French Ministry of Research and Defense within a 2004 TECHNO-VISION program. This dataset contains 1200 color fundus images, 800 images with pupil dilation (one drop of Tropicamide at 0.5%), and 400 images without dilation. The images were acquired by three ophthalmologic departments using a color video 3CCD camera mounted on a Topcon TRC NW6 non-mydriatic retinograph with a 45° FOV. The color fundus images have

different resolutions, 1440×960, 2240×1488, and 2304×1536. Messidor dataset provides information on DR grades and risk of macular edema diagnosis [Decencière et al., 2014].

Some other datasets are publicly available and could be used to conduct this research. Table 2.1 presents other datasets that also provide color fundus photography with the necessary annotations; microaneurysms, vessels segmentation, multi-label, OD, and Fovea localization.

Table 2.1: Other public color fundus image dataset. H indicates healthy, R indicates right fundus, and L indicates left fundus.

Dataset	Data Provided	Total Images	Image Size	FOV	Camera
ARIA [Farnell et al., 2008]	AMD DR Healthy Vessels OD and Fovea	92 59 61 212 120	768 x 576	50	Zeiss FF450+
DIARETDB1 [Kauppi et al., 2007]	Microaneurysms	84DR+5H	1500 x 1152	50	Zeiss FF450+
ROC [Niemeijer et al., 2009]	Microaneurysms	100	768 x 576, 1058 x 1061, 1389 x 1383	45	Topcon NW100, Topcon NW200, Canon CR5-45NM
ODIR [Challenge, 2019]	Multi-label	5000R+5000L	variety	-	Kowa, Zeiss, and Cannon
UoA-DR [Chalakkal et al., 2017]	Vessels, OD, and Fovea	200	2124 x 2056	45	Zeiss VISUCAM 500 lenses

2.2/ THEORETICAL BACKGROUND

2.2.1/ DEEP LEARNING IN MEDICINE

The necessity of computer-aided diagnosis (CAD) in medicine is due to several factors such as the complexity of the medical diagnosis system, the large amounts of diagnostic knowledge, and the availability of large amounts of complex clinical data relevant to many diseases and conditions [Yanase et al., 2019]. The main objective of CAD in medicine is to increase the efficiency of the diagnosis (both computational cost and time) along with a high performance of diagnosis [Abràmoff et al., 2010]. In particular conditions, CAD systems can also prevent the error caused by humans that may lead to misdiagnosis. It is stated in a study proposed by Bejnordi et al. [Bejnordi et al., 2017] that the performance of the CAD system for the detection of lymph node metastasis of breast cancer is higher than the diagnosis made by 11 experienced pathologists in terms of cost, time, and accuracy. However, the diagnostic capabilities of the CAD systems are not completely exploited due to a lack of research that directly compares artificial intelligence (AI) and physicians' interpretation or validates AI performance in a real clinical environment [Fujita, 2020].

The advances in machine learning techniques gave an impact on transforming CAD

systems from aided to automated. Machine learning and computer vision algorithms allow researchers to analyze the input imaging and/or non-imaging data by extracting and interpreting the information of the data to predict the outcome for a given task [Chan et al., 2020]. However, the transition from traditional CAD systems (conventional image processing and machine learning) to deep learning-based CAD systems is needed due to limitations such as high development cost, high False Positive (FP) rate, effectiveness, and limited specific features [Fujita, 2020].

One of the significant differences between deep learning and machine learning is feature extraction. Instead of having a need for hand-crafted features as in the case of machine learning, deep learning focuses to learn the domain knowledge so that the machine can learn the relevant features needed to interpret the output by correlating with the input data. Deep learning, also called a deep neural network, is a neural network with multiple layers. The success of deep learning to adapt to various applications such as speech recognition, natural language processing, and face recognition makes it interesting for researchers to use it in medical applications [Chan et al., 2020]. In general, deep learning consists of the input layer, hidden layer(s), and output layer. Deep learning frameworks are classified depending on the task and the type of data. The types of deep learning frameworks applied in this research are the Convolutional Neural Network (CNN), the Encoder-Decoder, and the Generative Adversarial Network (GAN).

2.2.1.1/ Convolutional Neural Network

Convolutional neural network (CNN) is the most commonly used deep learning frameworks for computer vision and pattern recognition tasks. CNN has the ability to encode image features into the network architecture and makes it to be suitable for image-related tasks [O'Shea et al., 2015]. CNN is constructed by stacking several types of layers such as convolution, pooling, and fully connected layers as shown in Figure 2.11 [Yamashita et al., 2018]. The CNN-based CAD systems for medical imaging data are widely implemented to perform several tasks such as eye disease detection [Gulshan et al., 2016], cancer diagnosis [Esteva et al., 2017] [Wu et al., 2019], and tumor segmentation [Havaei et al., 2017].

Deeper layers in the network architecture allow the network to learn more object features but with a complexity of increased computational time and more training data. Fine-tuning is a method to reduce the necessity of these requirements by extracting the object features that were learned by a pre-trained network. It is implemented by freezing the initial layers of the pre-trained network and training only the last layer (or last few layers) to learn more features that are adaptable to the current dataset.

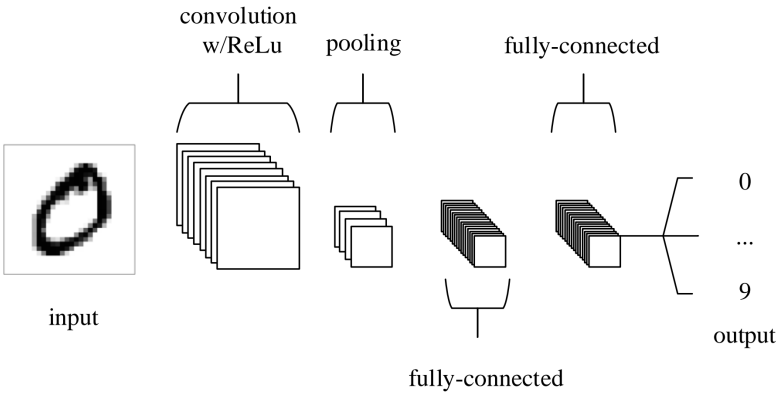


Figure 2.11: Simple CNN architecture [O'Shea et al., 2015].

A pre-trained network is a network trained using other datasets for some specific task. EfficientNet [Tan et al., 2019] and ResNet [He et al., 2016] are the pre-trained networks that are used in this research considering its applauding performance in the classification task.

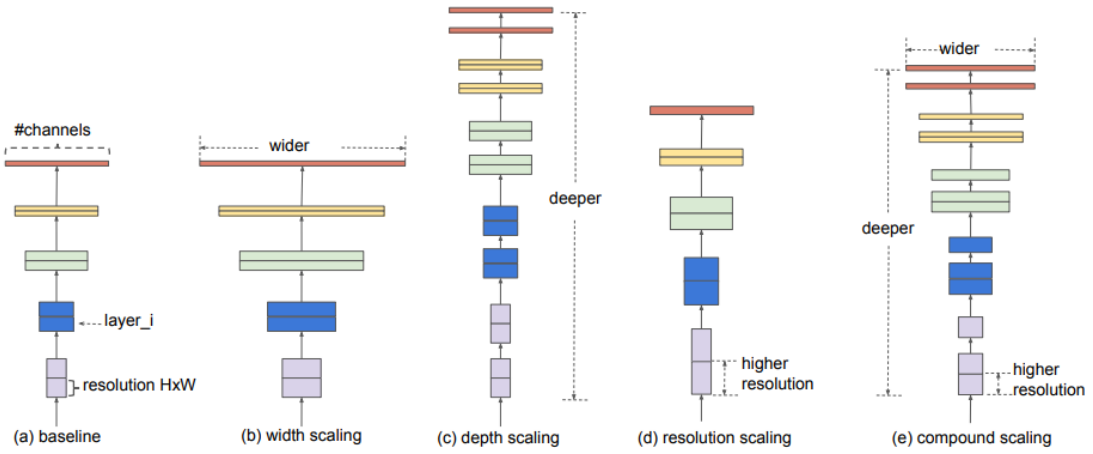


Figure 2.12: Model scaling. (a) is a baseline network example; (b) is scaling the dimension of width only; (c) is scaling the dimension of depth only, (d) is scaling the dimension of resolution only; (e) is the compound scaling method that uniformly scales width, depth, and resolution with a fixed ratio [Tan et al., 2019].

EfficientNet. EfficientNet [Tan et al., 2019] is a CNN architecture with a scaling method that scales the dimensions of depth, width, and resolution uniformly using a compound coefficient (see details in Figure 2.12). The intuition behind the compound scaling method is that the bigger the size of an input image, the higher the number of layers and channels that are required to increase the receptive field and to capture a more fine-grained pattern. The base architecture of EfficientNet is similar to the inverted bottleneck residual blocks of MobileNetV2 [Sandler et al., 2018]. This baseline model is scaled to obtain a variety of EfficientNet (EfficientNet-B0, EfficientNet-B1, EfficientNet-B2, EfficientNet-B3, EfficientNet-B4, EfficientNet-B5, EfficientNet-B6, and EfficientNet-B7). A higher version

in the EfficientNet family has a more complex architecture that can handle the input image with larger resolutions. EfficientNet is fine-tuned for acquiring maximum accuracy at the expense of more computation and prediction time. The EfficientNets achieve state-of-the-art 91.7% accuracy on the CIFAR-100 dataset, 98.8% accuracy for the Flowers dataset, and other transfer learning datasets with a fewer number of parameters.

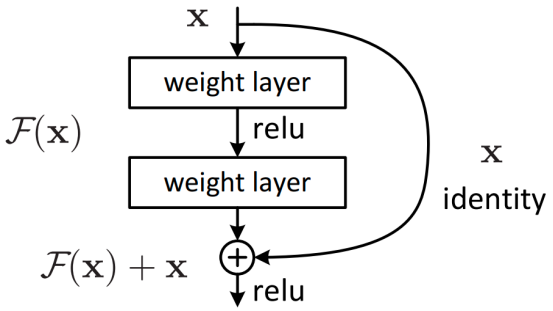


Figure 2.13: Residual learning: a building block [He et al., 2016].

ResNet. ResNet (Residual Network) [He et al., 2016] is a CNN architecture that is able to deal with vanishing gradient issue in training a deep network with large number of layers. ResNet architecture introduces the identity shortcut connection concept between layers as illustrated in Figure 2.13. Stacking the identity mappings in the network keeps the network performance without any degradation. The main idea behind the residual block is to let the network learns a residual function ($F(x)$) of the previous layers, the difference between the input (x) and output layer ($F(x) + x$), by adding an input identity to the output activation layer.

ResNet architectures are built by stacking the residual blocks together. ResNet is introduced in variety deeper levels such as ResNet34, ResNet50, ResNet101, ResNet152, and ResNet1202. A deeper ResNet has more complexity. Though deeper ResNet has no optimization difficulty, the training may lead to overfitting depending on the quantity of the training dataset. And, shallower ResNet does not always give the highest performance. ResNet architecture won the first places in several tracks in ILSVRC and COCO 2015 competitions: ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation.

2.2.1.2/ ENCODER-DECODER

Inspired by neural machine translation, the encoder-decoder neural network is composed of an encoder and a decoder. Figure 2.14 shows the general concept of encoder-decoder in which an encoder is a process to encode the input to the specific state and a decoder is a process to decode that specific state to become an output. In an encoder-decoder neural network, the encoder is a stacking of layers that extracts a higher dimensional

representation of the input data and the decoder is a stacking of layers that uses this higher dimensional representation to generate the specific-lower dimensional output data as expected for that specific task [Cho et al., 2014]. In a CNN-based encoder-decoder, the input image is encoded into a higher dimensional feature map and a new feature map is generated by the decoder to have the same resolution as the input image to perform pixel-wise classification [Badrinarayanan et al., 2017]. Encoder-decoder neural network is implemented to perform several tasks where the output is expected to be in a specific dimension such as image segmentation in computer vision, and text summarization and question answering in natural language processing (NLP).

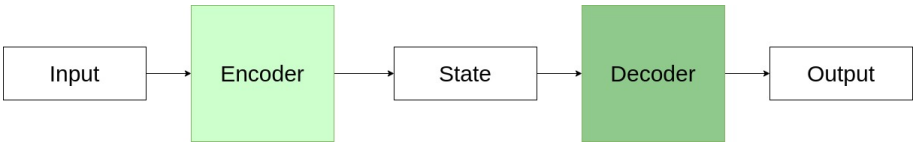


Figure 2.14: The architecture of encoder-decoder.

2.2.1.3/ GENERATIVE ADVERSARIAL NETWORK (GAN)

A generative adversarial network (GAN) is designed to handle the issue of limited labeled data which is a major challenge in deep learning. GAN is a neural network that consists of a generator and a discriminator which are trained under the adversarial learning scheme [Wang et al., 2017]. The overview of the computation procedure and the structure of GAN is described in Figure 2.15.

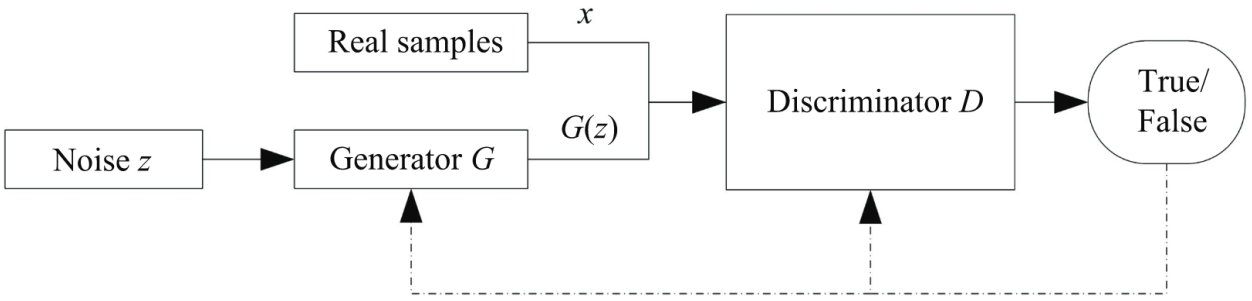


Figure 2.15: Computation procedure and structure of GAN [Wang et al., 2017].

The generator (G) is typically a decoder that captures the distribution of the real samples data (x) from the noise (z) to generate the new data ($G(z)$), and the discriminator (D) is an encoder that discriminates the newly generated data ($G(z)$) with the real samples (x) as accurate as possible [Gui et al., 2021]. The generator and discriminator learn simultaneously. The generator is continuously trained to lower the accuracy of the discriminator while the discriminator may be frozen when it has already reached the optimal

performance. The performance of GAN is optimal when the generator can match the real data distribution that confuses the discriminator maximally (predicts 0.5 for all input) [Creswell et al., 2018]. GAN uses the minimax optimization method where the goal is to reach Nash equilibrium. GAN has been implemented in various applications such as computer vision, speech processing, and NLP.

2.2.1.4/ TRANSFORMER

Transformer [Vaswani et al., 2017] is a deep learning model that transforms a sequence of elements into another sequence by adopting self-attention mechanisms. The attention mechanism is a mapping of the query and a set of key-value pairs to an output.

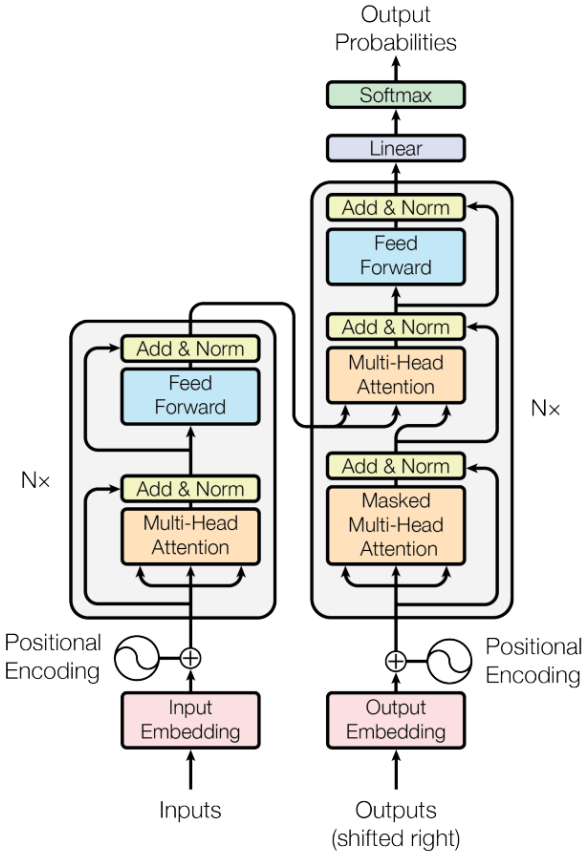


Figure 2.16: Model architecture of Transformer [Vaswani et al., 2017].

As seen in Figure 2.16, transformer consists of Encoder and Decoder parts. The encoder is on the left part of the architecture and the decoder is on the right one. The encoder takes a sequence input to a continuous representation. An encoder contains two sub-layers, a multi-head attention sub-layer with self-attention, and a feed-forward sub-layer (FFN). As seen in Figure 2.17, multi-head attention is a stacked scaled-dot product attention that are concatenated.

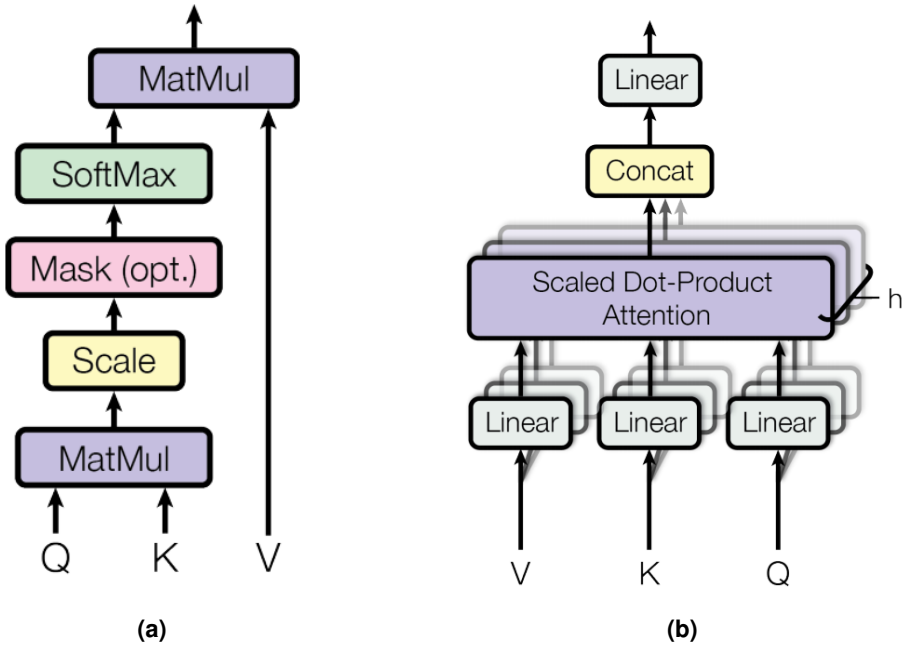


Figure 2.17: (a) Scaled-dot product attention, (b) Multi-head attention [Vaswani et al., 2017].

As described in Eq. 2.1, the attention is computed by softmax of a dot product between query (Q) and key vector (K) that is multiplied with value vector (V). This dot product equation is scaled by $\frac{1}{\sqrt{d_k}}$ where d_k is the dimension of the key vector.

$$Attention\ (Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.1}$$

Feed-forward network (FFN) is a sub-layer that consists of two linear transformations with a ReLU activation in between (see Eq. 2.2). Residual learning and normalization are applied in each sub-layer. Positional encoding (PE) is also summed to the input and initial outputs to capture the information about the relative positions.

$$FFN(x) = ReLU\ (xW_1 + b_1)\ W_2 + b_2 \tag{2.2}$$

The dimension of the positional encoding is the same as the dimension of the model. Positional encoding can be fixed and learned. Fixed positional encoding is generated from the sine and cosine functions of different frequencies as described in Eq. 2.3 where pos is the position and i is the dimension. Each dimension of the positional encoding corresponds to a sinusoid. The wavelengths form a geometric progression from 2π to 10000 x 2π.

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{\text{model}}}\right)$$
$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d_{\text{model}}}\right) \tag{2.3}$$

The decoder generates the output probabilities from the output of the encoder and the initial decoder outputs. A decoder contains multi-head attention with self-attention, multi-head attention with cross-attention, and a FFN sub-layer. Multi-head attention with cross-attention is the multi-head that receives the key (K) and value (V) from the encoder output and query (Q) from the decoder self-attention. The last step of the decoder is to linearly project the Q, K, and V to produce the output probabilities.

Transformer with Image. Transformer model is adopted widely to solve tasks in different fields such as NLP and computer vision such as DETR [Carion et al., 2020] and ViT [Dosovitskiy et al., 2020].

DETR (DEtection TRansformer) [Carion et al., 2020] is one of applications in object detection that applies transformer approach. The goal of object detection in DETR is to predict a set of bounding boxes and category labels for each object of interest. The architecture of DETR model is described in Figure 2.18. A set of features is extracted from pre-trained CNN to represent the information in the image. Since transformer encoder needs a sequence input, this spatial information is flattened into one dimension. With the positional encoding addition, these information are fed into transformer encoder to encode it into a continuous features. The positional encoding is adopted from the original transformer [Vaswani et al., 2017] to the 2D case. The object queries are the initial output of decoder to generate a mapping of key and value from the encoder output with the corresponding object queries. The object queries consists of a set of object category probabilities and the bounding box of the object location.

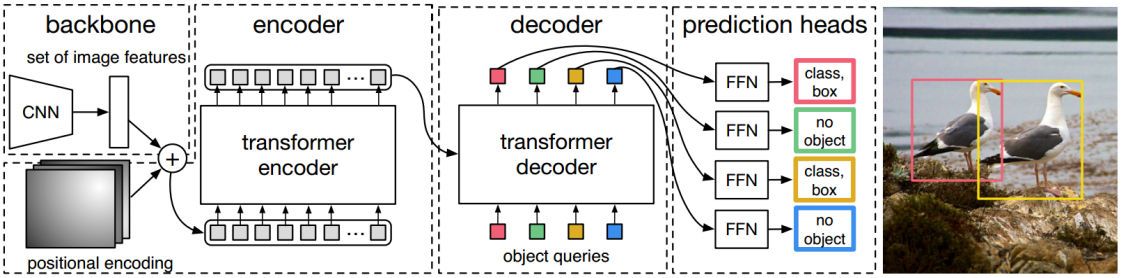


Figure 2.18: Model architecture of DETR [Carion et al., 2020].

ViT (Vision Transformer) [Dosovitskiy et al., 2020] is another application for image recognition using transformer approach in their method. Different from DETR, ViT applies only transformer encoder to recognize the image. The goal of the ViT is to classify the image. Transformer encoder requires a sequence as the input. Hence, ViT divides the input image into several patches that are flattened. As seen in Figure 2.19 about the model

architecture of ViT, these patches are fed into the transformer encoder added with the positional embedding as the input. The positional embedding is a 1-dimensional vector by considering the inputs as a sequence of patches in the raster order. The output of the transformer encoder is transformed into a class prediction. This class prediction is a small multi-layer perceptron (MLP) with tanh activation in the single hidden layer. The initial class prediction is added as an extra learnable class embedding to the input of the transformer encoder.

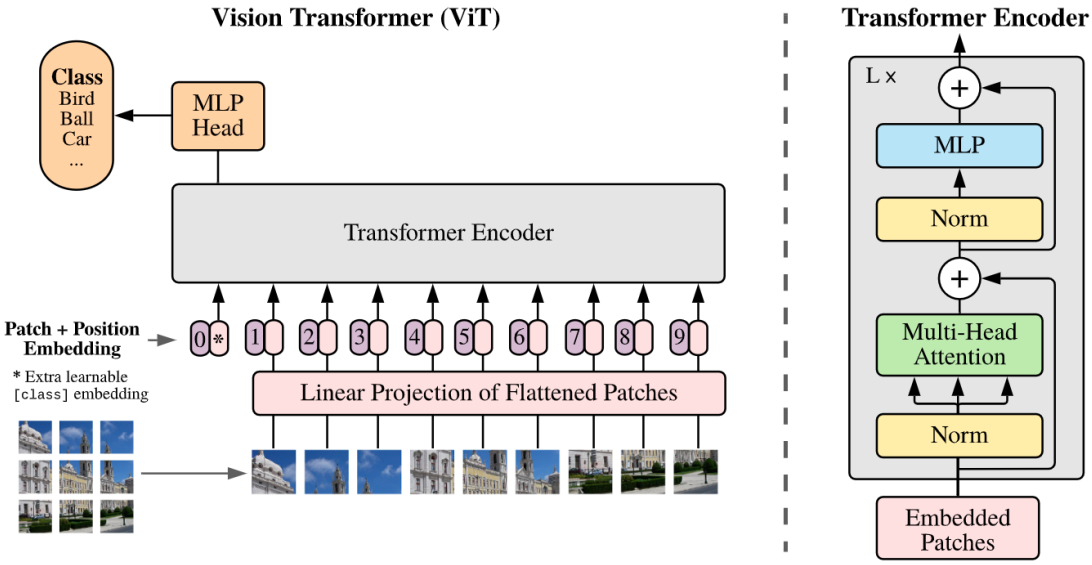


Figure 2.19: Model architecture of ViT [Dosovitskiy et al., 2020].

2.2.2/ WORD EMBEDDING

NLP is the intersection of artificial intelligence and linguistics. NLP is a study to process and analyze natural language data. Word embedding is one of the applications of NLP. Word embeddings are fixed-length vector representations for words that are built based on the distributional hypothesis [Almeida et al., 2019]. It aims to map words or phrases into a low-dimensional continuous space accurately in both syntactic and semantic information. Syntactic information represents the structural roles of the words and semantic information represents the meaning of the words [Li et al., 2018]. The conventional method of word embedding consists in taking a one-hot representation of each word, a vocabulary-size vector with only one non-zero entry. However, this method faces problems with sparse data and also lacks semantic information of words. The neural network-based word embedding produces a better model that can overcome these drawbacks [Wang et al., 2020]. The neural network language model (NNLM) is a word embedding method using a neural network proposed by Bengio et al. [Bengio, 2008]. Several word embedding techniques have been developed by adapting the NNLM and Transformer

methods to increase the efficiency and the quality of the vectors. Word2Vec and FastText are examples of NNLM-based word embedding models, which use neural network language models to generate word vectors. BERT, on the other hand, is an example of a Transformer-based word embedding model, which uses the transformer architecture to generate more contextually-rich word vectors.

2.2.2.1/ WORD2VEC

Word2Vec [Mikolov et al., 2013] is the most representative method of the NNLM-based approach [Li et al., 2018]. The basic assumption behind Word2Vec is that words with similar contexts are supposed to have similar meanings. Word2Vec model is capable of capturing the semantic information of the words. This model implements continuous bag-of-words (CBOW) and skip-gram models to vectorize the words.

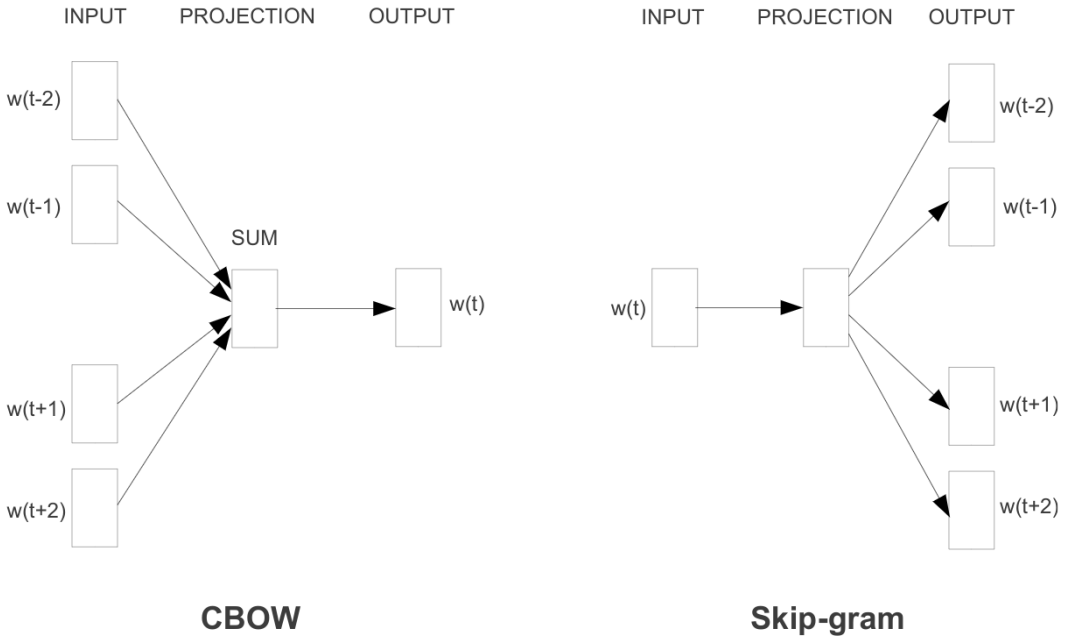


Figure 2.20: Word2Vec model architecture (window size = 2) [Mikolov et al., 2013].

As seen in Figure 2.20, CBOW model is trained to predict the center word $(w(t))$ as the output based on its context $(w(t-2), w(t-1), w(t+1), w(t+2))$ as the input. The goal of the CBOW model is to find the word representation (projection) that is useful to predict the target word by its context words, whereas the skip-gram model is the reverse of the CBOW model. Skip-gram model is trained to predict the context words $(w(t-2), w(t-1), w(t+1), w(t+2))$ as the output given the center word $(w(t))$ as the input. Projection contains a weight matrix with $V \times N$ size, where V is the number of unique words and N is the number of neurons in the hidden layer.

The center word and context words are defined by a word window size. For instance, in the sentence "*I saw a new book store near the park*", the center word "*store*" has {"*new*", "*book*", "*near*", and "*the*"} as its context words with window size '2'. Before feeding it to the network, in pre-processing, each unique word is converted into a one-hot encoding form. Consequently, the negative data are presented more than the positive data. This condition pushes the network to be more sensitive to overfitting and also increases the number of data to be trained. Negative sampling is applied in Word2Vec to overcome this issue by subsampling the negative data based on the distribution of the frequency of words. Instead of using the network for predicting the output, the projection (the hidden layer) is used as the word embedding [Almeida et al., 2019].

2.2.2.2/ FASTTEXT

The FastText model [Mikolov et al., 2018] was proposed by the Facebook AI Research group. FastText is an advancement of the Word2Vec model which is also built based on CBOW and skip-gram models with negative sampling. The main difference between FastText and Word2Vec is the vector assignment. Word2Vec assigns the vectors only based on words while FastText assigns them based on also the characters (called atomic tokens). FastText applied subword n-gram information to acquire the relationship between characters to capture the internal semantics of word information [Wang et al., 2020]. FastText generates character n-grams of specific window sizes to represent the center word. For instance, the sentence "*I saw a new book store near the park*" is divided into words and a list of character n-gram is generated for each center word. For the center word "*store*", a list of character n-gram $(n = 3)$ {"*sto*", "*tor*", "*ore*"} is generated. This concept adds the advantage to the model to predict the unseen word (out-of-vocabulary) which is not the case in Word2Vec. FastText model is reportedly acquiring better results than Word2Vec, especially in the languages that have a heavy morphology and compositional word-building like French, German, and Spanish [Almeida et al., 2019].

2.2.2.3/ BERT

Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al., 2018] adopts the transformer model to train a language model. The architecture of the BERT model is a multi-layer bidirectional Transformer encoder. BERT has two different versions based on the model's complexity. BERT-base is a transformer encoder model with 12 layers, 768 sizes for the hidden layer, and 12 self-attentions heads, and BERT-large consists of 24 layers, 1024 size for the hidden layer, and 16 self-attention heads. The representation of the input in the BERT model is illustrated in Figure 2.21. In the illustration, the input consists of multiple sentences. Firstly, the token [CLS] and [SEP] are

added to the beginning and the end of the data as the symbol of the beginning of the data and the end of the sentences. Then each input element is tokenized with the addition of the position embedding, the position information of all data, and segment embedding, the segmentation between each sentence.

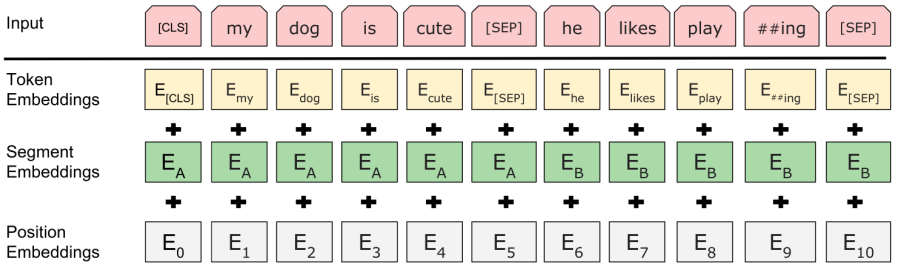


Figure 2.21: BERT input representation [Devlin et al., 2018].

Clinical BERT. Clinical BERT is a clinically oriented BERT model [Alsentzer et al., 2019]. The model was initialized with BioBERT [Lee et al., 2020] (BERT-base that was trained with PubMed article abstracts and PubMed Central article full texts) and trained on either all MIMIC notes or only discharge summaries. MIMIC database contains electronic health records from ICU patients at the Beth Israel Hospital in Boston, MA.

2.2.2.4/ Out-of-Vocabulary (OOV)

Out-of-vocabulary (OOV) is one of the major challenges in a word embedding. OOV words correspond to words that are not trained in the network because of the limited number of such words available in the training datasets. Wang et al. [Wang et al., 2020] categorized the type of OOV into four:

- Professional nouns and research field names
- Emerging common vocabulary e.g. online terms
- Proper names e.g. name of places, name of people, name of the organization
- Other terminology e.g. name of the products, title of books, the title of movies

Word embedding models like Word2Vec for instance can vectorize the words by learning the semantic representations of the words on large datasets. However, these models have no ability to embed the words which belong to the OOV category since these words produce zero-vectors or no output that degrades the quality of the model [Lochter et al., 2020]. There are several solutions that are commonly applied to deal with

OOV such as re-training, ignoring, or replacing with other vectors. Among these solutions, re-training is the most preferred solution as it can expand the vocabulary to cover all needed words [Wang et al., 2020]. However, the vector representation of OOV words can only be generated from the model re-training if it is trained with enough data. The traditional solution to deal with OOV is to ignore the word but it restricts the model to understand the meaning of the sentence [Adams et al., 2017].

The solution to replace the OOV output with a new representative vector has been actively studied such as replacing it with the vector's average of the closest words of an OOV word in the sample [Khodak et al., 2018], and generating a new vector based on the morphological structure. Though the method proposed by Khodak et al. [Khodak et al., 2018] is able to generate a new vector for the OOV, it lacks to capture the complex semantic relationship. The new OOV vector based on morphological structure generates a good enough representative vector to enhance the model performance. FastText is one of the word embedding methods that performs the morphological structure (subword) analysis to deal with OOV. It finds the meaning of the OOV subword by taking the meaning of the other subwords which are captured from other words in the vocabulary. However, the main drawback of this approach is that it is incapable of handling words that have different meanings in a different context. The morphological information approach is capable of generating only the same representation of the OOV words regardless of the context in which it appears [Lochter et al., 2020].

2.3/ Evaluation Metrics

Evaluation metrics computes the performance of the model predictions. In this research, Receiver Operating Characteristic (ROC), Free-Receiver Operating Characteristic (FROC), Precision-Recall (PR), Average Precision (AP), and Euclidean Distance (ED) metrics are utilized.

2.3.1/ Receiver Operating Characteristic (ROC)

Receiver operating characteristic (ROC) is an effective method to evaluate the performance of a diagnostic model [Kumar et al., 2011]. ROC plots the sensitivity or true positive rate (TPR) on the Y-axis against the false positive rate (FPR) on the X-axis with different thresholds. TPR and FPR are calculated as follows:

$$FPR = \frac{FP}{FP + TN} \tag{2.4}$$

$$TPR \textbf{ or } Sensitivity \textbf{ or } Recall = \frac{TP}{TP + FN} \tag{2.5}$$

True Positive (TP) indicates that the result predicts the presence of the label correctly and True Negative (TN) indicates that the result predicts the absence of the label correctly. While False Positive (FP) indicates that the result is wrongly predicted to be present and False Negative (FN) indicates that the result is wrongly predicted to be absent.

The **a**rea **u**nder the **c**urve (AUC) of ROC-curve is a global measurement of a classification model to show its ability to distinguish the classes. An ideal model is represented by the ROC-curve which is closer to the top-left part of the curve and has an AUC value of almost 1.0, while a model that has no ability to discriminate the classes is represented as a perfect diagonal ROC curve with an AUC value of almost 0.5 [Hoo et al., 2017].

2.3.2/ Free-Receiver Operating Characteristic (FROC)

Compared to ROC-curve, FROC-curve is more suitable for evaluating the model with abnormal behavior [Bandos et al., 2009]. FROC-curve provides meaningful information with respect to class imbalance which is vital in many medical applications. This is because the FROC-curve directly focuses on the number of false-positive (FP) without having any impact on the true negative (TN) (as in the case of FPR) that may diverge the analysis of the model performance (see Eq. 2.4). FROC-curve plots per-lesion sensitivity as calculated in Eq.2.5 against the average number of FP detected per image (FPI) for all classification thresholds (see Figure 2.22). It calculates a sensitivity score at some average FPI values in a small range between 0 to 8 (1/8, 1/4, 1/2, 1, 2, 4, 8). An ideal classification model has high sensitivity scores in all average FPI ranges.

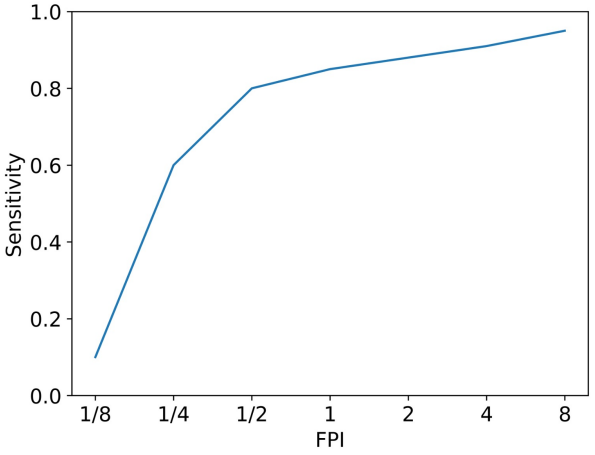


Figure 2.22: Illustration of FROC Curve.

2.3.3/ PRECISION-RECALL (PR)

Precision-Recall (PR) curve plots the precision values (see Eq. 2.6) on the Y-axis against the recall values (see Eq. 2.5) on the X-axis. The **a**rea **u**nder the PR curve (AUPR) represents the performance of a classification model. An ideal model will have high values for both precision and recall and an AUPR value of 1.0 (the highest value).

$$Precision = \frac{TP}{TP + FP} \tag{2.6}$$

2.3.4/ AVERAGE PRECISION (AP)

Average precision (AP) is calculated to evaluate the performance of classification models ideally used for multi-label classification tasks. It summarizes the precision-recall curve as the weighted mean of precisions achieved at each threshold (R_n), with an increase in recall from the previous threshold (P_n) as expressed in Eq.2.7.

$$\text{AP} = \sum_{n} (R_n - R_{n-1})\, P_n \tag{2.7}$$

2.3.5/ EUCLIDEAN DISTANCE (ED)

Euclidean distance (ED) is a distance between two points in Euclidean space. The lower ED value indicates better performance. ED calculation is expressed in Eq. 2.8. $ED(\hat{c}, c)$ is the average of ED with N total number of images, $\hat{c}$ is the ground-truth that contains the pixel coordinates $(\hat{c_x}, \hat{c_y})$, and c is the prediction with (c_x, c_y) pixel coordinates.

$$ED(\hat{c}, c) = \frac{1}{N} \sum^{N} \sqrt{(\hat{c_x} - c_x)^2 + \left(\hat{c_y} - c_y\right)^2} \tag{2.8}$$

II

CONTRIBUTION

3

MICROANEURYSMS DETECTION

3.1/ INTRODUCTION

Early treatment is the most effective way to reduce the risk of vision loss caused by Diabetic Retinopathy (DR) [Faust et al., 2012]. Thus, regular eye screening is an important activity for early DR detection [Stitt et al., 2016]. As mentioned in Section 2.1.4.1, DR is diagnosed by analyzing the presence of some visible lesions in color fundus images obtained from an eye examination [Yau et al., 2012]. Among DR lesions, microaneurysm (MA) is the first symptom of DR that appears as a tiny red spot in the retina image (see Figure 3.1). MA is the swelling of tiny blood vessels caused by a weakening of the vascular walls [Klein et al., 1984] that has 15 to 60 μm in diameter [Meyerle et al., 2008] and seldom exceeds 125 μm [Imani et al., 2015].

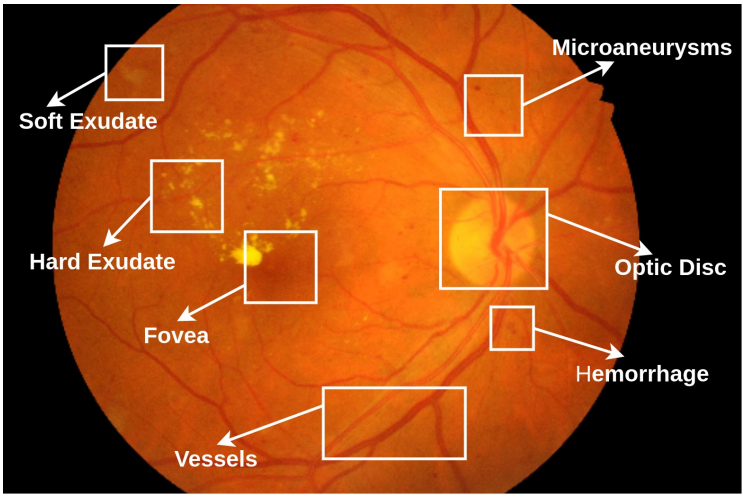


Figure 3.1: Color fundus image with DR lesions.

Computer-aided detection (CAD) for MAs detection is expected to have high accuracy and efficiency in terms of computational cost and time [Abràmoff et al., 2010]. Automatic MAs detection aims to reduce the work of ophthalmologists in examining eye screening by

providing reduced numbers of False Positives (FPs) in the results. However, having a high sensitivity with a small number of False Positive per Image (FPI) has been a challenge in automatic MAs detection. The major causes of this issue is a severe imbalanced data and poor object features. The information of MAs in a color fundus image is much lesser than the non-MAs which in turn affects the learning process in a machine learning-based approach. The poor performance of MAs detection for both supervised and unsupervised approaches shows that, so far, the extracted object features have not represented the object's characteristics accurately.

3.2/ Objective and Constraint

The main objective of the research in this chapter is to build an automatic MAs detection from a color fundus image using CNN as the classifier. The MAs detection is expected to have high sensitivity with a reduced number of FPI. The model is evaluated in individual datasets, so the model should be able to learn from a limited number of training data.

3.3/ Outline of The Chapter

A brief review of the related works is presented in Section 3.4. Section 3.5 describes the proposed methodology applied in this research to solve the problem that is mentioned in Section 3.2. The implementation details of the proposed method is described in Section 3.6, and the results of the experiments are discussed in Section 3.7. Lastly, the conclusion of the experiments that are conducted for MAs detection is explained in Section 3.8.

3.4/ Related Works

As illustrated in Figure 3.2, the general pipeline of MAs detection is divided into three steps: preprocessing, MAs candidate extraction, and MAs detection. All three steps have a significant impact on the results of MAs detection.

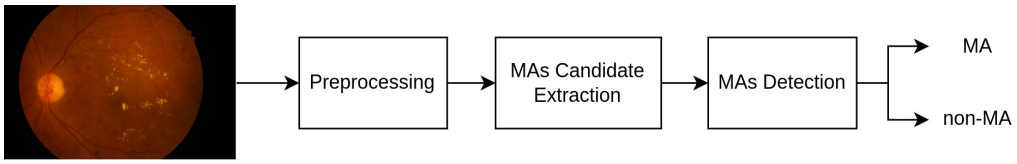
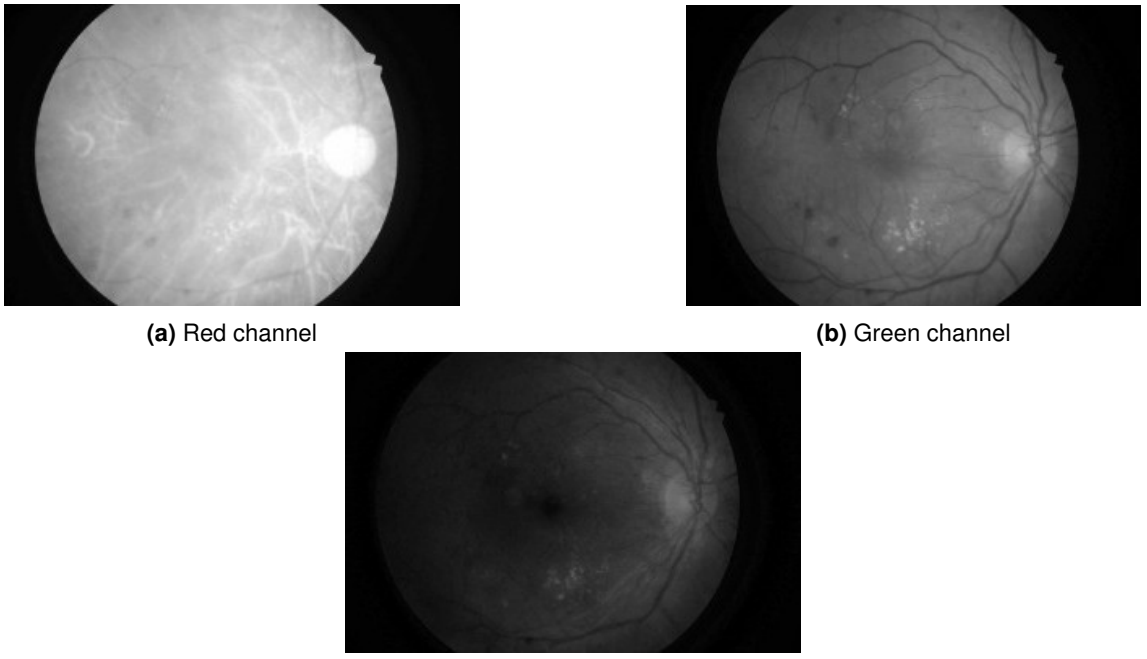


Figure 3.2: General workflow of MAs detection.

Preprocessing is the initial step in MAs detection. It aims to increase the quality of the

information in the input image. As seen in Figure 3.3, each channel of the input image contains different information. The green channel is extracted from the color fundus image and enhanced using CLAHE (Contrast Limited Adaptive Histogram Equalization) [Pizer et al., 1987] since the contrast difference between the background and the lesions is comparatively more important in the green channel.



(a) Red channel **(b)** Green channel

(c) Blue channel

Figure 3.3: Fundus image channels, (a) is the red channel, (b) is the green channel, (c) is the blue channel.

In the second step, MAs candidate extraction is crucial as the sensitivity can decrease significantly due to missed-real-MAs in MAs detection results. Thus, Budak et al. [Budak et al., 2017] and Zhang et al. [Zhang et al., 2021] did not ignore any information from the image and selected all pixels in the image as MAs candidates. Even though this approach does not have any missed-real-MA, it increases the computation cost. On the other hand, Eftekhari et al. [Eftekhari et al., 2019], Long et al. [Long et al., 2020], and Xia et al. [Xia et al., 2021] select the MAs candidates only from the objects that are highly potential to be MA. This approach is computationally efficient but with a higher risk of having missed-real-MA. However, the risk of decrement in sensitivity can be reduced, thanks to the accuracy of the MAs candidate extraction. The MAs candidate can be extracted both with a supervised and unsupervised approach. Eftekhari et al. [Eftekhari et al., 2019] and Xia et al. [Xia et al., 2021] extract the MAs candidates in a supervised approach by generating a probability map using a deep neural network, whereas Dai et al. [Dai et al., 2018], Adal et al. [Adal et al., 2014], Long et al. [Long et al., 2020], and Zhang et al. [Zhang et al., 2020] extract the MAs candidates using an unsupervised

approach by applying hand-crafted feature extraction techniques such as top-hat morphological transformation, hessian matrix, and shape-based features.

Due to the size of the MA, the MAs candidate image is split into patches with one MA candidate present in the center of the patch. The MAs candidate patches are the input of the MAs detection model which is a classifier built to detect MAs by applying machine learning algorithms. In machine learning, the features that differentiate MAs and non-MAs are extracted based on local cross-section [Du et al., 2020], Sliding-Band-Filter [Melo et al., 2020], scale-space, SURF, and Radon-transform [Adal et al., 2014][Giancardo et al., 2011][Giancardo et al., 2010]. The results of hand-crafted features combined with machine learning as a classifier are competitively low and often fail in undefined MAs conditions such as blurry edges and subtle contrast.

In deep learning, the model has the ability to learn the object features based on the given training sets. However, the training process in deep learning still has many issues, such as imbalanced data and overfitting. Dai et al. [Dai et al., 2018] tackled the imbalanced data by applying a cascade learning technique. Cascade learning enables the model to learn more features from non-MAs data with balanced samples to distinguish MAs and non-MAs. Cascading also helps to reduce the FP results by building multiple stages of network filtering. Dai et al. [Dai et al., 2018] also added clinical reports to enrich the object features. AlexNet architecture [Krizhevsky et al., 2012] is used in order to prevent overfitting. The network is trained with a private dataset containing 645 images. Instead of re-training the network with FP data, Budak et al. [Budak et al., 2017] proposed to re-train only the low-performance batch to increase the detection results. This model is built with a simple Deep Convolutional Neural Network (DCNN) and trained with 50 images. The sensitivity of their performance is still lesser than 0.5. Orlando et al. [Orlando et al., 2018] designed their classifier with LeNet architecture [LeCun et al., 1998] to reduce overfitting. They feed the network with the ensemble vector of descriptors and apply the Random Forest as a classifier to refine the prediction results.

Unlike others, Zhang et al. [Zhang et al., 2021] built a stacked fully connected sparse autoencoder to extract the object features and use the last softmax layer to detect the MAs. Background suppression based on dissimilarity features is introduced in the input channel. However, this method requires more training images as it needs to be trained from scratch. Another strategy to deal with imbalanced data is data augmentation. Since general data augmentation such as rotating, shifting, and flipping leads to overfitting due to the small less-texture patch, Xia et al. [Xia et al., 2021] introduced a specific data augmentation for MAs detection by shifting the location of the MA in the patch until it is close to the boundary. They used EfficientNet which has multi kernels to detect the tiny objects. This augmentation technique has less imbalanced data between MAs and non-MAs, but it fails for the patches having more than one MA.

Meanwhile, Kou et al. [Kou et al., 2019], Xie et al. [Xie et al., 2020], and Chen et al. [Chen et al., 2018] detected the MAs using a pixel-based approach. Kou et al. [Kou et al., 2019] implemented different types of encoder-decoder architectures to segment the MAs. The networks are basically modified U-Net-based architecture such as U-Net combined with ResNet blocks (Res-UNet), applied recurrent-convolution blocks into U-Net (RNN-UNet), and a combination of ResNet and recurrent blocks in U-Net architecture. Even though it has high accuracy, this approach still produces many FP predictions without stating any number in detail. Xie et al. [Xie et al., 2020] and Chen et al. [Chen et al., 2018] proposed the segmentation for MAs detection with the refinement of segmentation results by re-segmentation and effective decoder modules to reduce the number of FPs. But this refinement requires a large number of training data. The methods of related works in lesion-based MAs detection are summarized in Table 3.1.

Table 3.1: Methods of related works in MAs detection.

Author	Methods	
	MAs Candidate Extraction	**MAs Detection**
[Long et al., 2020]	Potential pixels (unsupervised)	Machine learning with directional local contrast
[Dashtbozorg et al., 2018]	Potential pixels (unsupervised)	Machine learning with local convergence
[Eftekhari et al., 2019]	Potential pixels (supervised)	Deep learning in two-steps
[Orlando et al., 2018]	Potential pixels (unsupervised)	Machine learning with ensemble features
[Xia et al., 2021]	Potential pixels (supervised)	Deep learning with custom augmentation
[Zhang et al., 2020]	Potential pixels (unsupervised)	Image processing with shape-based filter
[Budak et al., 2017]	All pixels	Deep learning with batch re-training
[Zhang et al., 2021]	All pixels	Stacked fully connected sparse autoencoder
[Dai et al., 2018]	Potential pixels (unsupervised)	Deep learning with cascade learning

The performance details of the related works in lesion-based MAs detection are presented in Table 3.2.

Table 3.2: Performance of related works in MAs detection.

Method	Testing dataset	Training images	FPI						
			1/8	**1/4**	**1/2**	**1**	**2**	**4**	**8**
[Long et al., 2020]	E-Ophta	74	0.075	0.154	0.267	0.358	0.472	0.594	0.699
[Dashtbozorg et al., 2018]	E-Ophta	148	0.358	0.417	0.471	0.522	0.558	0.605	0.638
[Eftekhari et al., 2019]	E-Ophta	198	0.091	0.258	0.401	0.534	0.579	0.667	0.771
[Orlando et al., 2018]	E-Ophta	134	0.14	0.2	0.23	0.37	0.45	0.52	0.62
[Xia et al., 2021]	E-Ophta	452	0.668	0.701	0.71	0.718	0.72	0.733	0.74
[Xia et al., 2021]	IDRiD	452	0.561	0.563	0.565	0.568	0.575	0.601	0.634
[Zhang et al., 2020]	ROC	330	0.12	0.284	0.313	0.391	0.492	0.641	0.654
[Budak et al., 2017]	ROC	50	0.039	0.061	0.121	0.22	0.338	0.372	0.394
[Zhang et al., 2021]	E-Ophta	621	0.885	0.866	0.909	0.986	0.999	1	1
[Zhang et al., 2021]	DIARETDB1	621	0.987	0.984	1	1	0.996	1	1
[Dai et al., 2018]	DIARETDB1	735	0.933	0.951	0.972	0.978	0.984	0.99	0.992

Hence, we proposed two techniques for MAs detection using deep learning-based classifiers instead of segmentation to have more comparisons in the evaluation. The proposed methods tackle the challenges in MAs detection with four factors:

1. Using a better enhancement image technique for the green image channel.
2. Implementing an unsupervised method for MAs candidate extraction to be indepen-

dent of a large required data in MAs candidate extraction since a supervised method needs a large training dataset to have less FPs results in MAs candidate extraction.

3. Adding a background suppression image to the classifier to enhance the features.

4. Applying two different training strategies to the classifier (ensemble and cascading) to reduce the number of FPs.

3.5/ METHODOLOGY

Figure 3.4 illustrates the proposed pipeline to detect the MAs. It starts with preprocessing of the RGB fundus images by extracting and enhancing the green channel ($EnhGreen$). Small patches (so-called MA candidates) are generated from the processed image with the help of vessel segmentation results. Finally, MAs are detected with two different MAs classification strategies, ensemble and cascade learning. Overall, the pipeline can be divided into five parts: Preprocessing, Vessel Segmentation, MAs Candidate Extraction, Patch Generator, and MAs Classification.

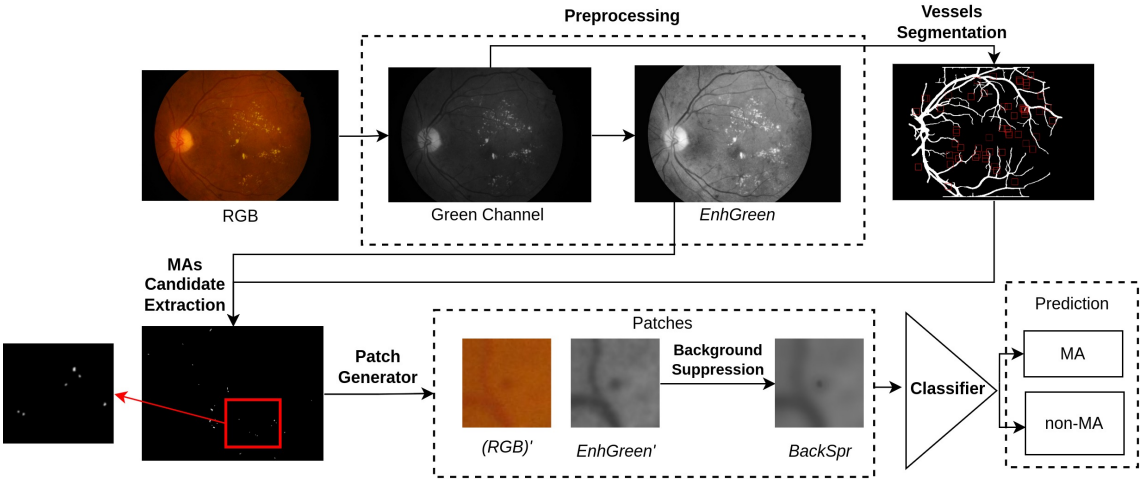


Figure 3.4: The main pipeline of MAs detection.

3.5.1/ PREPROCESSING

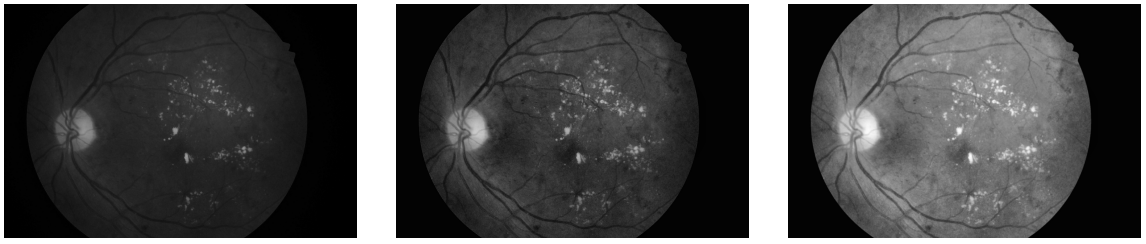
The color fundus image is resized into half to reduce the computation cost while preserving the MAs information. The green channel is extracted as it has the highest contrast between the background and the lesions. The fundus image background is removed to focus on the retinal field-of-view (FOV) by applying the threshold to the result of the median filter (kernel size is 31), followed by the Gaussian filter (kernel size is 31) to remove the noise in the image. The threshold value is defined manually (15), with 0 for the background and 1 for the fundus. Lastly, it is enhanced by r-polynomial transformation

[Walter et al., 2007] to prevent the hidden-potential lesions due to the uneven illumination, followed by CLAHE to correct over-amplification of the image contrast.

As explained in Eq. 3.1, the r-polynomial transformation T is mapping $G \rightarrow U$, where $G = \{g_{\min}, \ldots, g_{\max}\}$ is the green channel of the input image, and $U = \{u_{\min}, \ldots, u_{\max}\}$ is the transformed gray image. The transformation is defined by parameters μ_f, r, u_{min}, and u_{max}. μ_f is calculated as the global intensity mean of the FOV in G, whereas, $r = 1$, $u_{min} = 0$, and $u_{max} = 1$ are selected heuristically.

$$T(g) = \begin{cases} \frac{\frac{1}{2}(u_{\max} - u_{\min})}{(\mu_f - g_{\min})^r} \cdot (g - g_{\min})^r + u_{\min}, & if\ g \leqslant \mu_f \\ \\ \frac{-\frac{1}{2}(u_{\max} - u_{\min})}{(\mu_f - g_{\max})^r} \cdot (g - g_{\max})^r + u_{\max}, & if\ g > \mu_f \end{cases} \tag{3.1}$$

As seen in Figure 3.5, the combination of r-polynomial transformation and CLAHE ($EnhGreen$) can enhance the image much better than using CLAHE alone, preserving the pixels' information, especially the dark-red-lesions.



(a) Green channel **(b)** CLAHE **(c)** R-polynomial and CLAHE

Figure 3.5: Preprocessing methods, (a) is the original image, (b) and (c) are the results of CLAHE and the proposed enhancement techniques.

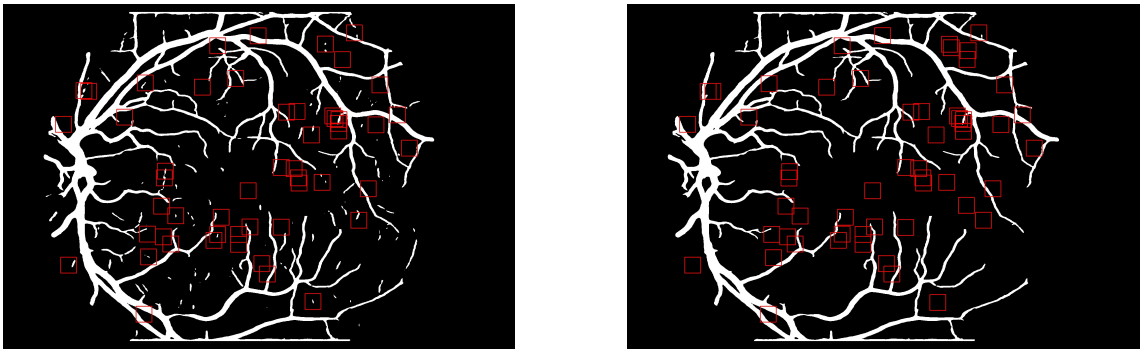
3.5.2/ Vessel Segmentation

The abnormalities in the retinal vessels (vessel branching pattern, vessel width, tortuosity, and vessel density) can indicate many eye diseases due to the complication of hypertension, arteriosclerosis, cardiovascular disease, glaucoma, and stroke [Roychowdhury et al., 2014]. Another usage of retinal vessel segmentation is to remove the vessels from the image in some retinal abnormalities detection such as hemorrhages and in optic nerve localization [Gross et al., 2013]. Many studies focus on vessel segmentation, both in supervised and unsupervised methods. The performance of vessel segmentation with supervised deep learning outperforms the unsupervised methods. Among deep learning methods, U-Net and multi-models are the leading architectures [Chen et al., 2021a]. However, vessel segmentation is not the focus of this research. An

existing vessel segmentation method is applied to generate vessel segmentation for the fundus images that are used in our MAs detection.

Retinal vessel segmentation is required in MAs candidate extraction to focus on the other dark-red objects without interference from the intensities of the vessel. The vessel segmentation is implemented based on the method proposed by Son et al. [Son et al., 2017] to generate the probability map of the retinal vessel. Its performance achieved 0.9803 of ROC, 0.9149 of Precision-Recall, and 0.829 of Dice Coefficient for the DRIVE dataset and 0.9838 of ROC, 0.9167 of Precision-Recall, and 0.834 of Dice Coefficient for STARE dataset.

The vessels are segmented with a GAN-based method to have more realistic results. The full-size color fundus image is provided to the generator instead of patches. The generator is a U-Net architecture which generates the probability map of the retinal vessels. The discriminator determines whether the image provided is the image segmented by the human expert (ground-truth) or the result generated from the generator. The objective function of this network is the minimax function and a loss function that penalizes the distance between the ground-truth and generator result.



(a) Vessel segmentation after thresholding **(b)** Final vessel segmentation

Figure 3.6: Post-processing of retinal vessel segmentation with the red boxes indicates the location of ground-truth MAs.

The probability map of the vessel for the color fundus images in this research is predicted from the pre-trained model on the DRIVE dataset. The threshold value of the probability map is heuristically determined so that the pixel with the highest probability is selected as the main vessel. However, some MAs are miss-segmented due to their appearance and location (close to the vessel). As shown in Figure 3.6, most miss-segmented MAs are disconnected from the main vessel, so post-processing using a connected-component algorithm is applied to remove them based on the size of the objects.

3.5.3/ MAs Candidate Extraction

MAs candidates are extracted in an unsupervised approach to reduce the necessity of a large quantity of training images. Vessel removal is applied to the image by simply changing all intensities that overlapped with the segmentation result to zero. Background estimation is applied to the image to separate the image background (outside FOV and retina background) and the foreground objects. The image background is estimated by applying a median filter (kernel size is 43) to the image and normalizing it with morphological reconstruction [Giancardo et al., 2011].

Inspired by the method proposed by Adal et al. [Adal et al., 2014], the MAs candidates are extracted based on the Gaussian curvature retrieved from the Hessian matrix. Hessian matrix is defined to hold the information about the second-order derivative of the image I (see Eq. 3.2).

$$H = \begin{pmatrix} \frac{\partial^2 I}{\partial^2 x^2} & \frac{\partial^2 I}{\partial x \partial y} \\ \frac{\partial^2 I}{\partial y \partial x} & \frac{\partial^2 I}{\partial^2 y^2} \end{pmatrix} \tag{3.2}$$

The curvature of the overlaid object in the image is analyzed by the eigenvalues of its Hessian matrix (λ_1 and λ_2). As explained in Eq. 3.3, circular-like objects can be detected by calculating the determinant of the Hessian matrix by multiplying its eigenvalues.

$$det(H) = \lambda_1 \cdot \lambda_2 \tag{3.3}$$

A circular-like object has large values for both eigenvalues, so the MAs candidate is found by looking for the local maxima of the $det(H)$. However, other circular-like lesions and objects with a linear-like structure also have a large value in $det(H)$, such as the border of FOV and retinal vessels. Hence, the elimination steps based on the value and sign of the object are applied to $det(H)$ to have only the MAs candidate remaining.

The binary image I_t is produced by thresholding $det(H)$ with ρ value to eliminate the objects based on its value. The potential candidates are indicated as 1 in the threshold results. The selection of threshold value ρ is shown in Eq. 3.4.

$$\rho = \frac{1}{N} \sum^{N} det(H), \quad \forall\ det(H) > \mu \tag{3.4}$$

The value ρ is the truncated mean value of $det(H)$ distribution with mean μ. The object is also eliminated based on the sign of the eigenvalues of its hessian matrix as it defines the curvature of the object.

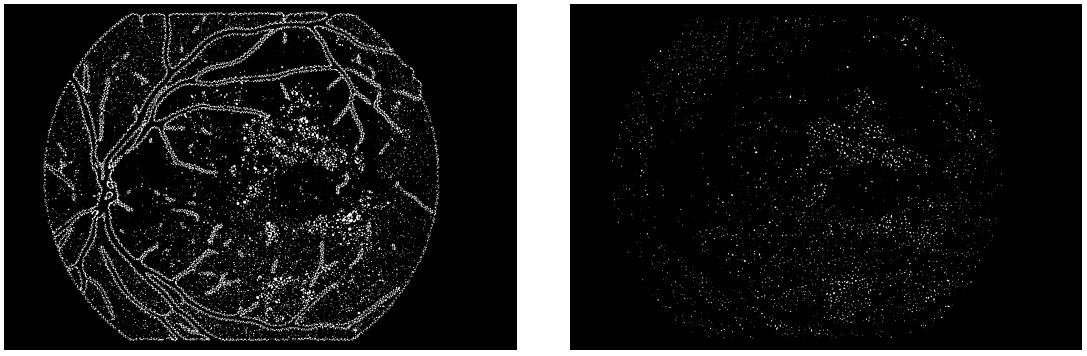
- $|\lambda_1^+| \ll |\lambda_2^-|$ represents a point located in the bright linear-like structure.

- $|\lambda_1^+| \ll |\lambda_2^+|$ represents a point located in the dark linear-like structure.
- $|\lambda_1^-| \approx |\lambda_2^-|$ represents a point located in the bright circular-like structure.
- $|\lambda_1^+| \approx |\lambda_2^+|$ represents a point located in the dark circular-like structure.

λ^+ indicates the eigenvalues having a positive sign and λ^- indicates the eigenvalues having a negative sign. As described in Eq. 3.5, MAs candidates are assigned to objects in the image with positive signed-eigenvalues as it represents the objects with a dark circular-like or linear-dark circular-like structure for all $I_t = 1$.

$$I_{res} = \begin{cases} 1, & I_t > 0 \; and \; \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} > 0 \\ \\ 0, & otherwise \end{cases} \tag{3.5}$$

Figure 3.7 shows the result of value-base elimination followed by sign-based elimination as the output of MAs candidates extraction image I_{res} that contains only MAs candidates.



(a) Value-based elimination **(b)** Sign-based elimination

Figure 3.7: Results of MAs candidate extraction.

3.5.4/ Patch Generator

Patch generation aims to minimize the object's ROI and reduce the computation cost in the classification step. There are five different patches generated: Red', $Green'$, $Blue'$, $EnhGreen'$, and $BackS\,pr$. The Red', $Green'$, $Blue'$ patches are generated by cropping the RGB fundus image and $EnhGreen'$ patches are generated by cropping $EnhGreen$, with one MA candidate located in the center of the patch. $BackS\,pr$ patch generation is explained in Section 3.5.4.1. The patch's size n is selected heuristically so that the MA is covered completely inside the patch to contain more surrounding information.

3.5.4.1/ BACKGROUND SUPPRESSION

The goal of background suppression is to remove the background noise by blurring it so that the only remaining information present is the pixels that belong to the MAs candidate (see Figure 3.9). Inspired by Zhang et al. [Zhang et al., 2021], the object features are extracted based on the local Hessian matrix as explained in Eq. (3.2) because of its effectiveness in representing the structure of objects. The object features are extracted by calculating the determinant of its Hessian matrix values (see Eq. (3.3)) and differentiate the MAs from the other objects based on γ_{bs} threshold value. The algorithm 1 describes the background suppression algorithm.

Three regions of interest (ROIs) are selected from the subset of the image to compare the values of a determinant matrix in order to choose the appropriate γ_{bs} value to eliminate the background objects. First ROI roi_1 is located in the center of the image with the size of its tolerance value (K=5) to directly point to the region where MAs are surely located. The mean value of this ROI μ_1 is computed to be its representation value.

Algorithm 1 Background Suppression

Input: A patch of green enhancement result ($EnhGreen'$)

Output: Result of background suppression ($BackS\,pr$)

1: Extract Hessian matrix as explained in Eq. 3.2 (I_{hes})
2: Find the eigenvalues (λ_1, λ_2)
3: Calculate determinant of I_{hes} as explained in Eq. 3.3 (I_{det})
4: Select ROI (K x K) in the center of the patch (roi_1)
5: Calculate mean of roi_1 (μ_1)
6: Select ROI (2K x 2K) in the center of the patch (roi_2)
7: Calculate mean of roi_2 (μ_2)
8: Calculate mean of I_{det} (μ_3)
9: $\gamma_{bs} = max(\mu_1, \mu_2, \mu_3)$
10: Threshold $EnhGreen'$ image with γ_{bs} ($BackS\,pr$)
11: $BackS\,pr = BackS\,pr \cap \lambda_1^+ \cap \lambda_2^+$
12: Return $BackS\,pr$

The second ROI roi_2 has the same central location as roi_1 but double in size, to be able to learn the contrast difference between the object and nearest-background surroundings. The contrast difference of the smaller MAs is higher than the bigger MAs since the information of the background for the small-sized MA could already be found in the roi_2. These ROI selections are illustrated in Figure 3.8.

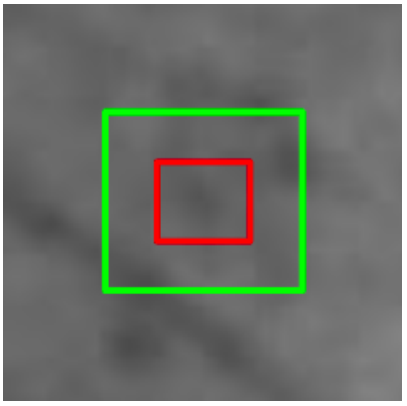


Figure 3.8: Illustration of ROI selections in background suppression. Red box indicates roi_1 and green box indicates roi_2.

However, there is no background information found in the case of bigger MAs, so the mean value of a full image is also computed μ_3. The highest mean value among these regions, γ_{bs} is selected to differentiate the foreground and background. Finally, the results which have negative eigenvalues are also discarded as the local maximum is not the only information contributing to the determinant of the Hessian matrix [Adal et al., 2014]. The final result of background suppression in a green patch is shown in Figure 3.9.

(a) Green patch **(b)** Threshold result **(c)** Final result

Figure 3.9: Background suppression, (a) is the input image, (b) is the mask of foreground (objects that look like MAs), and (c) is the results of background suppression.

3.5.5/ CLASSIFIER

As mentioned in Section 3.4, a classifier for MAs requires a simple network architecture due to the limitation of data and object features. The network architecture implemented in this research is built with EfficientNet-B0 as a base model. Though EfficientNet-B0 is a more complex and deeper network, it can be trained with smaller-size training images with less risk of overfitting because of the compound scaling technique. The fine-tuning of the EfficientNet-B0 that was trained with the ImageNet dataset is applied to reduce the total number of training images required. However, the number of channels of the input image needs to be three. To also add a $BackS\,pr$ image to the classifier, the experiments in this research have two different network strategies, ensemble and cascade learning.

Ensemble learning aims to find the best image channel combination for MAs detection and cascade learning aims to focus to reduce the number of FPs detected.

3.5.5.1/ ENSEMBLE LEARNING

Unlike Orlando et al. [Orlando et al., 2018] who applied the ensemble vector for the descriptors, this experiment implements ensemble learning in the prediction result.

In this experiment, the network learns the object features that are benefited from hand-crafted features. Instead of combining the features from the hand-crafted and neural network and giving it directly to the network as the method proposed by Orlando et al. [Orlando et al., 2018], the hand-crafted features are analyzed beforehand to reduce the complexity of the background surrounding the MA (see Section 3.5.4.1). This background-suppressed image is fed into the network as an additional channel that gives more features to classify the MAs.

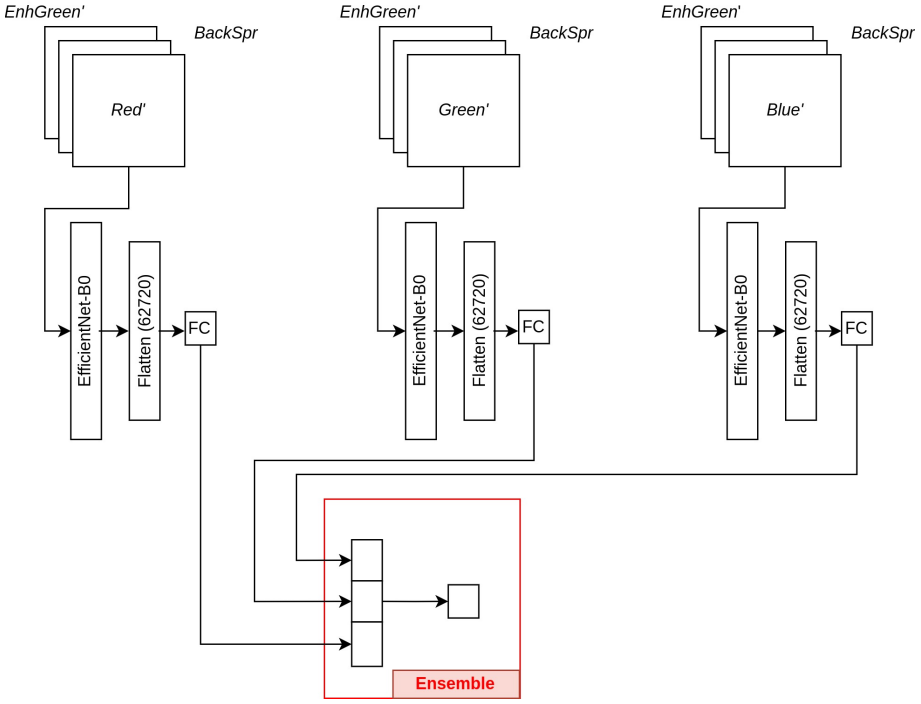


Figure 3.10: Illustration of Ensemble learning. The model consists of three units of identical networks with different input channels. The final prediction is selected by applying ensemble learning to all predictions from all unit networks.

The illustration of the ensemble learning strategy is shown in Figure 3.10. The model consists of three identical unit networks with different input channels followed by an ensemble learning vector as a decision maker for the final prediction. Each unit network has three input image channels. The input image consists of the patches that are generated from the patch generator (see Section 3.5.4). Inspired from [Zhang et al., 2021],

the combination of input image channels is $EnhGreen'$, $BackS\,pr$, and Red' for the first unit network. The Red' is replaced by $Green'$ in the second unit network and $Blue'$ in the third unit network.

The training of each unit network is independent in order to analyze the impact of each unit network on the final prediction. The background suppression image ($BackS\,pr$) is an additional channel to the network. The Red', $Green'$, $Blue'$ images are fed into each unit network separately by considering the information in each color image channel.

The imbalanced dataset is down-sampled randomly to be in an equal number of positive and negative classes without any data augmentation. Each unit network is trained with a different down-sampled dataset with binary cross-entropy as the loss function. Instead of giving more weight to the green channel [Dai et al., 2018][Kou et al., 2019], the weight for each unit network is equal since the information from the green channel is already distributed. The final prediction is decided by voting the ensemble vector. The average probability of the majority unit network is the probability of the final prediction.

3.5.5.2/ CASCADE LEARNING

Cascade learning is the second training strategy implemented in order to reduce the number of FPI. This experiment is inspired by Dai et al. [Dai et al., 2018] and Zhang et al. [Zhang et al., 2021]. The input for this network scheme is $Blue'$, $EnhGreen'$, and $BackS\,pr$ patches. This combination of input image channels is selected based on the outcome of the ensemble learning experiment (see Section 3.5.5.1).

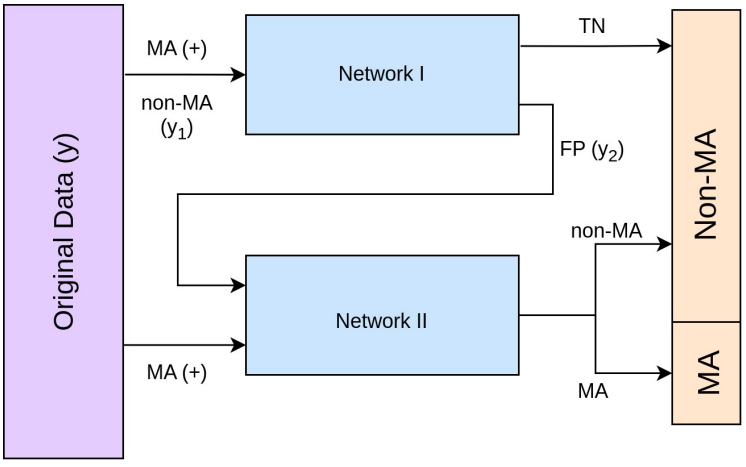


Figure 3.11: Cascade learning.

The illustration of cascade learning in this experiment is described in Figure 3.11. The model of this MAs classifier consists of two-unit networks (Network I and Network II). Each unit network has one neuron in the last layer with a sigmoid activation function. The image

patches y consist of MA (+) and non-MAs data with a severe imbalance distribution. The input of Network I consists of all MAs patches in y and non-MAs patches that are down-sampled into the same number as MAs patches (y_1). The Network I is trained with the ImageNet weight of EfficientNet-B0 as initial weights without any freezing layer applied. After training, Network I predicts the original data (y) to separate the type of data that is able to be classified correctly. The MAs prediction is thresholded with 0.5, greater than 0.5 as MA, and less than 0.5 as non-MA. The miss-predicted outputs, which are assumed to be hard cases, are fed to Network II to learn again. The input of Network II is the FP output from Network I and MA (+), the same MAs data as Network I. The negative data (FP) is also down-sampled (y_2) to match the total of MAs data.

Network II has the same architecture as Network I but with different weight initialization. Network I and Network II are trained sequentially. Network II is trained with the final model of Network I as the initial weights. In testing, the final prediction is decided based on the prediction of all unit networks since each network learns different features. Only the patches predicted as MAs in both networks are classified as MAs.

3.6/ Implementation Details

Computations have been performed on the supercomputer facilities in "Mésocentre de Calcul de Franche-Comté". The experiments of MAs classification are implemented on the Tensorflow platform with python programming language in Ubuntu-based server version 18.04.5 LTS with 252 GB RAM and GPU Tesla V100 DGXS 32 GB. The input image size is 2000×1328, and the patch size is 64x64. The ratio of MA and non-MA patches that are generated is 1:46. The optimizer for the ensemble classifier is Adam, and it is trained with 100 epochs for each unit network. While in the cascade classifier, Network I is trained with Adam optimizer with a 0.001 learning rate and SGD for Network II with the same learning rate. Each network is trained in 600 epochs with a 0.5 reduction in the learning rate every two increased validation loss. The hyperparameters of all networks are selected heuristically. The training and testing for experiments in MAs classification are independent for each dataset.

3.7/ Experimental Results

This section contains the results of the MAs detection with both ensemble learning as well as cascade learning. The first experiment is trained and tested with the IDRiD dataset. The second experiment is trained and tested with IDRiD and E-Ophta dataset. The evaluation metrics that are applied to evaluate the performance of MAs detection are FROC

and PR Curve.

3.7.1/ Ensemble Learning

The performance of MAs detection with ensemble learning is categorized into two, global prediction and local prediction. Global prediction is the performance of the model for a whole fundus image, and local prediction is the performance of the model for the ROI that is located close to the fovea. Local prediction is also evaluated since the MAs that are located close to the fovea have a higher priority in practice. As described in Section 2.1, the fovea is a tiny pit located in the macula of the retina that provides the clearest vision. The failure of recognizing the MAs that are isolated in one to two disc diameters (DD) of fovea often resulted in arbitration for the final gold standard grading [Boucher et al., 2020].

3.7.1.1/ Global Prediction

The performance of this experiment is compared with the other reported state-of-the-art models that focus on the same IDRiD dataset and the same evaluation indicators. The performance is evaluated on each unit network separately with the $Proposed\ method\ (Blue)$ which is the unit network with $Blue'$, the $Proposed\ method\ (Red)$ which is the unit network with Red', and the $Proposed\ method\ (Green)$ which is the unit network with $Green'$. The performance of the final prediction is represented as the $Proposed\ method\ (Final)$.

As described in Table 3.3, the performance of all the proposed methods in the first experiment is lower than MS-EfficientNet [Xia et al., 2021] which has the highest performance in the IDRiD dataset. The significant difference in performance between the proposed method and MS-EfficientNet [Xia et al., 2021] is the number of data used for training of both MAs segmentation (for MAs candidates extraction) and MAs classification. They extracted the MAs candidates and classified the MAs in a supervised approach. They used the cross dataset, a combination of multiple datasets that consists of around 400 images for training, while the proposed method uses only 54 images for IDRiD. The key issue for the classification and segmentation of MAs with a deep learning-based approach is the limited availability of positive data which causes a severe imbalance in the dataset. The proposed method tackles this issue in two steps. One is by applying an unsupervised method for MAs candidates extraction which minimizes the necessity of more data in the training phase and the second is by implementing an ensemble learning scheme with background suppression information to enhance the features even from less positive data. It has to be noted that, SESV-DLab [Xie et al., 2020] and DeepLabv3 [Chen et al., 2018] are evaluated at the pixel level while the proposed methods and Xia et al. are evaluated at the lesion level.

Table 3.3: Comparison of the proposed method performance (patch size 64) with other MAs detection algorithms. Bold values indicate the highest scores among all algorithms. Italic values indicate the highest scores among internal algorithms.

Algorithms	Dataset		FPI							AUC (PR)
	Train	Test	1/8	1/4	1/2	1	2	4	8	
MS-EfficientNet [Xia et al., 2021]	Cross dataset	IDRiD	**0.561**	**0.563**	**0.565**	**0.568**	**0.575**	**0.601**	**0.634**	**0.779**
SESV-DLab. [Xie et al., 2020]	Cross dataset	IDRiD	-	-	-	-	-	-	-	0.51
DeepLabv3 [Chen et al., 2018]	Cross dataset	IDRiD	-	-	-	-	-	-	-	0.487
Proposed method (Blue)	IDRiD	IDRiD	*0.025*	*0.045*	*0.065*	0.103	0.174	0.246	0.376	*0.503*
Proposed method (Final)	IDRiD	IDRiD	0.019	0.036	0.056	*0.105*	*0.189*	*0.287*	*0.41*	0.49
Proposed method (Red)	IDRiD	IDRiD	0.009	0.019	0.036	0.072	0.144	0.247	0.337	0.486
Proposed method (Green)	IDRiD	IDRiD	0.006	0.013	0.026	0.052	0.103	0.207	0.315	0.470

As for the internal performance, the results in Table 3.3 show that the *Proposed method (Blue)* achieved a 0.503 score for AUC (PR) which is the highest score among the other proposed methods. It means that the *Proposed method (Blue)* produces almost the same quantity of FPs and FNs in the predictions. Meanwhile, the AUC (PR) of the *Proposed method (Red)* is less than the *Proposed method (Final)* and *Proposed method (Blue)*. The *Proposed method (Green)* has the lowest performance in FROC and AUC (PR) among other proposed methods. Even though the *Proposed method (Blue)* has the highest AUC (PR), the *Proposed method (Final)* achieved the highest sensitivity for 1, 2, 4, and 8 FPIs. It shows that applying ensemble learning for the final predictions gives a positive impact to reduce the FPs and increase sensitivity. The low sensitivity in the *Proposed method (Final)* for 1/8, 1/4, and 1/2 FPI is caused by the significant sensitivity difference between the *Proposed method (Blue)* with the *Proposed method (Red)* and the *Proposed method (Green)*.

Thus, it can be inferred that the ensemble learning of all color image channels can reduce the number of FPs and increase its sensitivity but the combination of the blue channel, enhanced-green and background suppression holds the highest information to detect the MAs alongside enhanced-green and its background suppression.

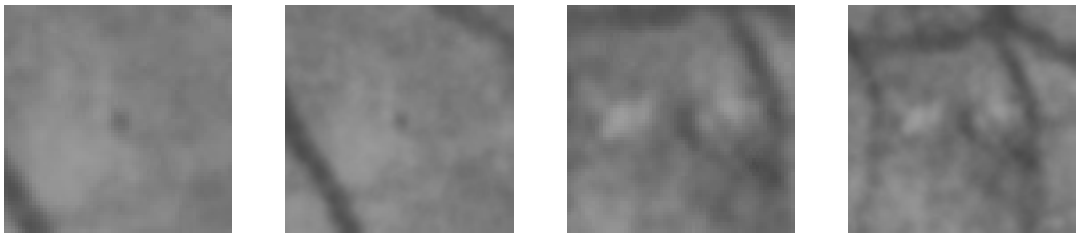
Table 3.4: The performance of the Proposed methods in different patch sizes. Bold values indicate the highest score among patch sizes. Italic values indicate the highest score in each patch size.

Patch Size	Methods	FPI							AUC (PR)
		1/8	1/4	1/2	1	2	4	8	
64	*Proposed method (Blue)*	*0.025*	*0.045*	*0.065*	0.103	0.174	0.246	0.376	*0.503*
64	*Proposed method (Green)*	0.006	0.013	0.026	0.052	0.103	0.207	0.315	0.470
64	*Proposed method (Red)*	0.009	0.019	0.036	0.072	0.144	0.247	0.337	0.486
64	*Proposed method (Final)*	0.019	0.036	0.056	*0.105*	*0.189*	*0.287*	***0.41***	0.49
86	*Proposed method (Blue)*	***0.054***	***0.079***	***0.096***	***0.117***	0.176	0.239	0.334	0.505
86	*Proposed method (Green)*	0.012	0.025	0.051	0.101	***0.203***	***0.295***	*0.379*	***0.526***
86	*Proposed method (Red)*	0.014	0.029	0.054	0.098	0.155	0.218	0.312	0.44
86	*Proposed method (Final)*	0.01	0.019	0.038	0.077	0.153	0.237	0.344	0.475

The global MAs detection with different patch sizes has also been implemented to analyze the impact of patch size with each image color channel. The performance of the Proposed methods for patch sizes 64 and 86 are provided in Table 3.4. Among the patch sizes, the *Proposed method* (*Green*) in patch size 86 achieves the highest score in AUC (PR). It is contradictive to the *Proposed method* (*Green*) which has the lowest AUC (PR) in patch size 64. The *Proposed method* (*Green*) with patch size 86 also achieves the highest sensitivity in 2 and 4 FPI while the *Proposed method* (*Blue*) with the same patch size has the highest sensitivity in FPI<2. The highest sensitivity in 8 FPI is still achieved by the *Proposed method* (*Final*) with patch size 64.

Increasing the patch size means enlarging the ROI in the MA's surroundings. The selection of the patch size is critical in MAs detection because of the size and the structure of the MA. Decreasing the patch size of MA can give direct focus to the network to learn MA's features, but it may lead to a lack of background information when it comes to blurry edges. Whereas increasing the patch size of MA can give more significant features between MA and the background image, it may miss-lead the network to learn more features of the background that has more structures compared to a tiny MA that is located in the center of the patch (see Figure 3.12 as the examples). In this experiment, increasing the patch size to 86 pixels helps the network to distinguish the MA and non-MA better for all unit networks as shown in Table 3.4 especially for the green color channel as it gives the most contrast information for dark objects. The lower AUC (PR) of the *Proposed method* (*Red*) with patch size 86 shows that increasing the patch size leads to an increase in the number of wrong prediction results.

In patch size 86, the *Proposed method* (*Red*) has the lowest score of AUC (PR) and lowest sensitivity for 4 and 8 FPI while the lowest sensitivity for FPI<4 is held by *Proposed method* (*Final*). In general, the main factor of the low performance in *Proposed method* (*Final*) is because of the voting scheme to predict the MAs from the ensemble vectors. In patch size 64, the sensitivity for all the FPIs for *Proposed method* (*Final*) never has the lowest sensitivity compared to other color channels. It shows that most of the MA predictions made by *Proposed method* (*Blue*) which has the highest performance among other unit networks, is matching with the predictions made by *Proposed method* (*Red*) or *Proposed method* (*Green*).



(a) Patch 1 (size: 64) **(b)** Patch 1 (size: 86) **(c)** Patch 2 (size: 64) **(d)** Patch 2 (size: 86)

Figure 3.12: Green Patches in different patch sizes.

So in patch size 64, the voting scheme from the ensemble vectors benefits the final prediction. However, the voting scheme gives different impacts for patch size 86. In the patch size 86, the sensitivity of most of FPIs for the $Proposed\ method\ (Final)$ is the lowest compared to other unit networks while most of the sensitivity for all the proposed methods (Blue, Red, and Green) with patch size 86 is higher than the ones with patch size 64. It shows that the number of majority prediction vectors in $Proposed\ method\ (Final)$ with patch size 86 are from the unit networks that have wrong predictions with high confidence.

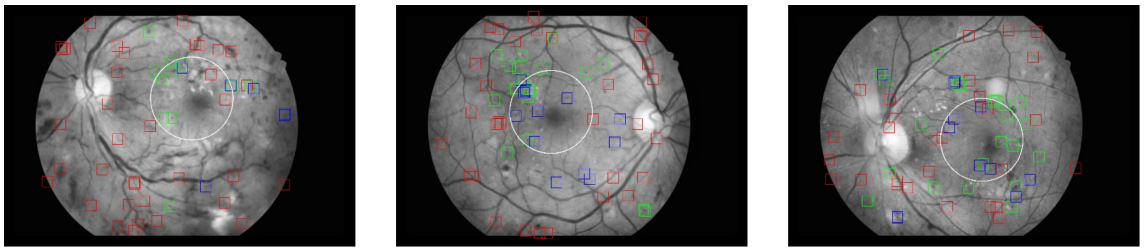
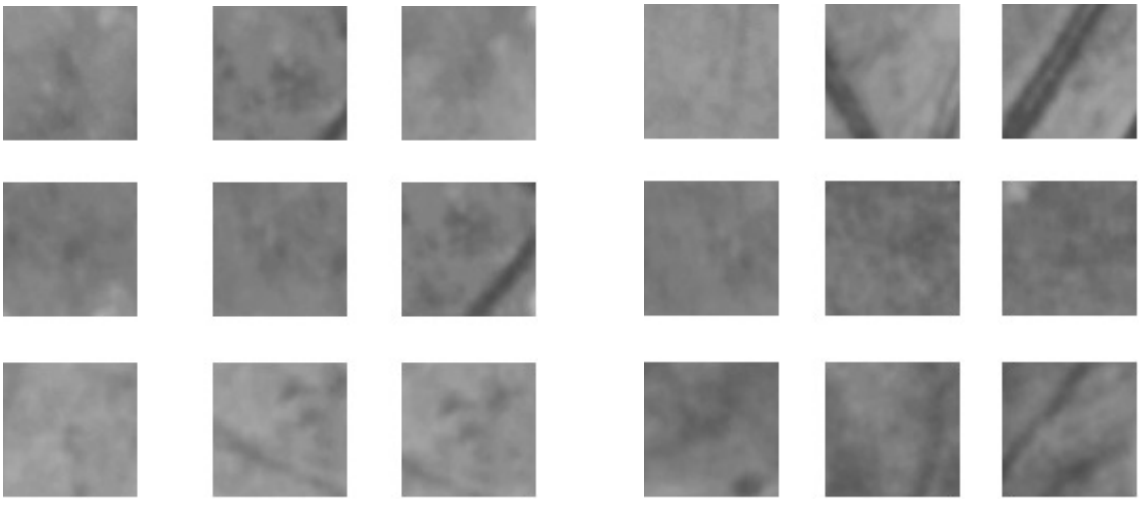


Figure 3.13: Global prediction results for $Proposed\ method\ (Blue)$ with patch size 64. The red, green, and blue bounding box indicates FP, TP, and FN (because of classifier). The white circle indicates the fovea's ROI.

As seen in Figure 3.13, the prediction results of the $Proposed\ method\ (Blue)$ which gives the highest and most stable color impacts in ensemble learning, the number of TP, FP, and FN in all images are almost equal. However, the location of some MAs is often centralized close to the fovea, and the FPs results of the $Proposed\ method\ (Blue)$ are not located in that area. The performance of the MAs detection in this area is explained in Section 3.7.1.2.



(a) False negative (FN) **(b)** False positive (FP)

Figure 3.14: False predictions in IDRiD dataset.

The low performance of proposed methods in FROC and AUC (PR) compared to other algorithms shown in the Table 3.3 describes that the proposed methods are still weak to

predict the MAs, and predicting FPs and FNs with a high probability. This failure occurs in objects that have a similar appearance to MA. As seen in Figure 3.14, the proposed methods fail to distinguish the object in the center of a patch that has the same size as MA and moderate blurry edges.

3.7.1.2/ LOCAL PREDICTION

This section provides the evaluation of MAs detection in the R radius from the center location of the fovea. The need for local prediction is due to the fact that failure in the detection of MAs which are present within one to two disc diameters (DD) of fovea significantly affects the final gold standard grading [Boucher et al., 2020].

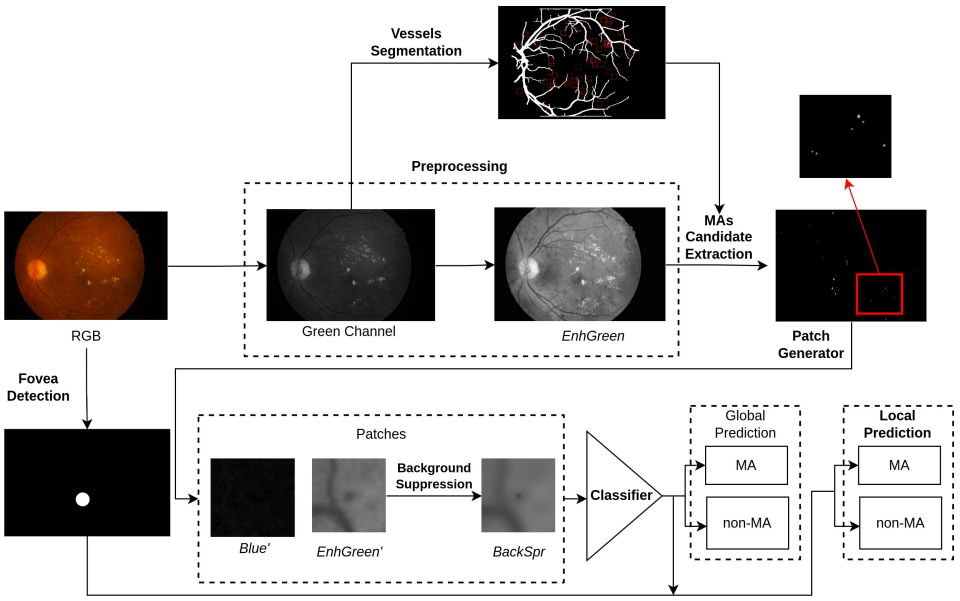


Figure 3.15: Pipeline for local prediction.

The pipeline of the local prediction shown in Figure 3.15 is the modified one from the main pipeline (see Figure 3.4) which includes fovea detection.

MAs local prediction requires the location of the fovea in the fundus image. The method implemented for fovea localization is adapted from the method proposed by Meyer et al. [Meyer et al., 2018], which has the highest evaluation rank of automatic fovea localization in the Messidor dataset. The fovea is localized by a pixel-wise Multi-Task-Learning (MTL)-like strategy by reformulating the problem as regressing the distance from each image location to the closest of both retinal landmarks of interest (OD and fovea). The aim of this network is to also find the location of the optic disc (OD) and learn the fovea localization based on the location of OD.

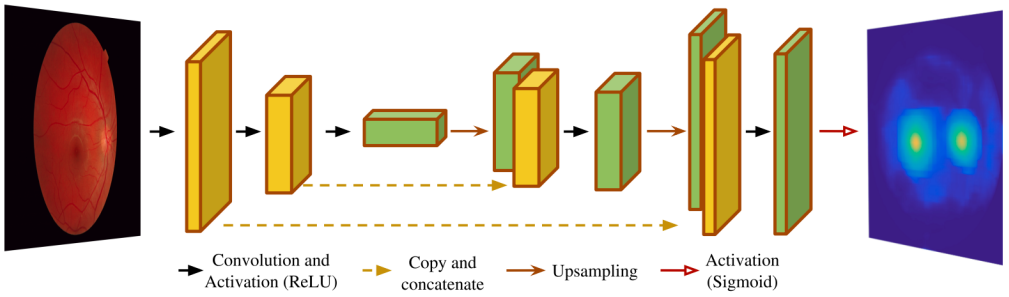


Figure 3.16: U-Net architecture for fovea localization [Meyer et al., 2018].

This network is trained and tested in the Localization set of the IDRiD dataset. The fovea and OD are localized by a segmentation model which is built with U-Net architecture (see Figure 3.16). Due to a difference in the representation of the ground truth dataset between Messidor and IDRiD dataset, the proposed method generates two solid spherical objects with γ_{spr} radius from the center coordinates of the fovea and OD as the ground truth data instead of Gaussian spherical objects. The ground truth of OD and fovea localization for Messidor dataset is annotated as the Gaussian spherical objects. This data is provided by Gegundez et al. [Gegundez-Arias et al., 2013] since Messidor dataset does not provide the annotation of OD and Fovea localization (see Section 2.1.5.6).

Optimized-Mean Square Error (MSE) is applied as a loss function ($\mathcal{L}_{reg}(\theta)$) to calculate the error distance between ground-truth image ($\mathcal{U}_\theta(x,y)$) with its normalized bi-distance map ($\mathcal{B}^N(x,y)$), for fovea location (x_{fov}, y_{fov}) and OD location (x_{od}, y_{od}) as follows:

$$\mathcal{L}_{\text{reg}}(\theta) = \frac{1}{M} \sum_{x,y \in \Omega} \left\| \mathcal{U}_\theta(x,y) - \mathcal{B}^N(x,y) \right\|^2 \tag{3.6}$$

$$\mathcal{B}^N(x,y) = \left(1 - \frac{\mathcal{B}(x,y)}{\max_\Omega \mathcal{B}(x,y)} \right)^\gamma \tag{3.7}$$

$$\mathcal{B}(x,y) = \min \left(\sqrt{(x - x_{od})^2 + (y - y_{od})^2}, \sqrt{\left(x - x_{fov}\right)^2 + \left(y - y_{fov}\right)^2} \right) \tag{3.8}$$

Normalized bi-distance ($\mathcal{B}^N(x,y)$) is a normalized form of bi-distance map ($\mathcal{B}(x,y)$) for each pixel location $(x,y) \in \Omega$, being $\Omega \subset \mathbb{R}^2$ the retinal image $I(x,y)$ domain with γ as parameter governing the spread across the image domain. Post-processing is applied to the segmentation result to classify the objects. Fovea and OD are assigned to the two-highest probability scores. The objects are identified based on their intensities, OD has a bright intensity and fovea has a dark intensity.

Table 3.5: Euclidean Distance Comparison of Fovea Localization in Localization Set of IDRiD Dataset. The bold value in ED indicates the highest scores.

Methods	ED (in pixels)
DeepDR [Porwal et al., 2020]	**64.492**
VRT [Porwal et al., 2020]	68.466
SDNU [Porwal et al., 2020]	85.4
ZJU-BII-SGEX [Porwal et al., 2020]	570.133
Regression [Meyer et al., 2018]	335.868
Proposed method	175.07
Proposed method (without false detection)	48.97

Table 3.5 shows the evaluation of the fovea localization method that is compared with the methods that are reported in the result of fovea localization in the IDRiD challenge (DeepDR, VRT, SDNU, and ZJU-BII-SGEX) [Porwal et al., 2020] and Regression [Meyer et al., 2018]. The performance is evaluated by calculating the Euclidean Distance (ED) between pixel coordinates of the prediction and ground truth in the original resolution.

As seen in Table 3.5, the highest performance is achieved by the winner of the IDRiD challenge, DeepDR, followed by VRT and SDNU. The low ED values of these three methods indicate fovea in all testing images is localized correctly with minor coordinates difference whereas the methods that have ED values for more than 200 pixels such as ZJU-BII-SGEX [Porwal et al., 2020] and Regression [Meyer et al., 2018] show the miss-detected fovea (false detection) in a few testing images.

The proposed method of fovea localization has lower ED than Meyer et al. but it is higher than other methods except for ZJU-BII-SGEX. It is caused by some miss-detected fovea in the testing images. The proposed method fails to detect fovea in 12 images from 103 testing images. The ED of the proposed method to locate the fovea in true detection (Proposed method without false detection) could reach 48.97 pixels. Finally, the proposed method of fovea localization is applied in the same set as the MAs detection dataset (Segmentation Set of IDRiD dataset). As seen in Figure 3.17, the proposed method succeeds to detect and localize the fovea in all testing images.

Local MAs detection is centralized to the area with radius R, from the center of the fovea (see Figure 3.18). The internal evaluation of the local MAs detection is described in Table 3.6. $Proposed\ method\ (Blue)^*$, $Proposed\ method\ (Red)^*$, $Proposed\ method\ (Green)^*$, and $Proposed\ method\ (Final)^*$ are the local MAs detection of $Proposed\ method\ (Blue)$, $Proposed\ method\ (Red)$, $Proposed\ method\ (Green)$, and $Proposed\ method\ (Final)$ respectively, with patch size 64.

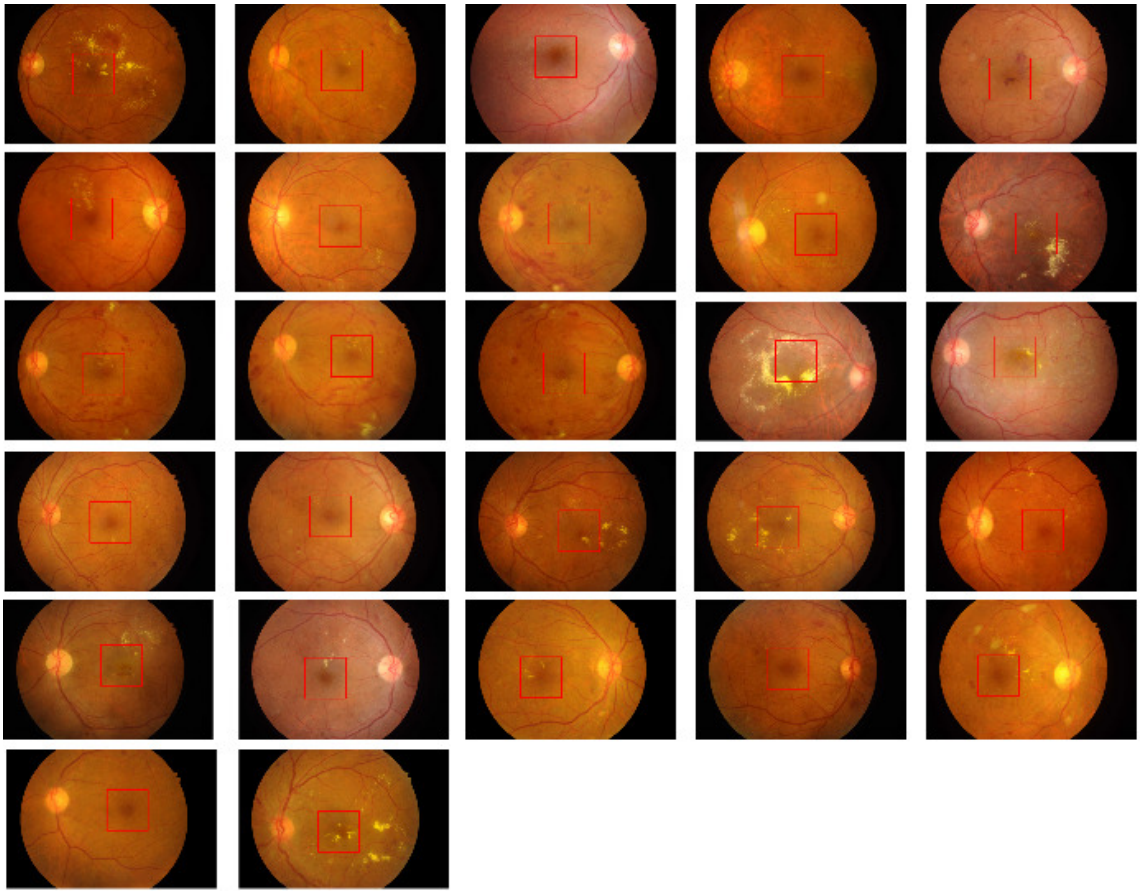


Figure 3.17: Fovea localization results of MAs testing images in Segmentation Set of IDRiD dataset. The red box indicates the location of the fovea.

As seen in Table 3.6, *Proposed method (Blue)** still has the highest AUC (PR) compared to other proposed methods. Unlike the internal evaluation in global prediction, *Proposed method (Blue)** achieves the highest sensitivity for 1/4, 1, 2, 4, and 8 FPI while *Proposed method (Final)** achieves the highest sensitivity only in 1/2 FPI, and *Proposed method (Red)** achieves the highest sensitivity in 1/8 FPI. In local prediction, the blue channel has more dominant information towards the final prediction followed by the red channel and green channel. The low sensitivity of the *Proposed method (Final)** is caused by a significant sensitivity difference between *Proposed method(Green)** with *Proposed method (Blue)** as well as *Proposed method (Red)**.

Table 3.6: Internal comparison of FROC and AUC-PR for local MAs detection. Bold values indicate the highest scores.

Dataset	Methods	FPI							AUC (PR)
		1/8	1/4	1/2	1	2	4	8	
IDRiD Dataset	*Proposed method (Blue)**	0.0981	**0.23**	0.28	**0.395**	**0.522**	**0.6277**	**0.725**	**0.574**
	Proposed method (Final)*	0.11	0.1837	**0.2975**	0.375	0.45	0.55	0.655	0.518
	Proposed method (Green)*	0.0475	0.1175	0.17	0.245	0.3283	0.395	0.52	0.372
	Proposed method (Red)*	**0.12**	0.155	0.215	0.31	0.3825	0.5675	0.635	0.4765

By classifying the MAs inside the radius R from the center coordinates of the fovea reduces the number of FP results. As seen in Figure 3.13, the predictions often fail to classify the MAs (FP) which are located close to the vessels, especially at the end of the vessels. MAs local prediction could detect MAs correctly (TP) with lesser FNs and FPs inside the fovea boundaries (R) since there are fewer vessels in the ROI (see Figure 3.18).

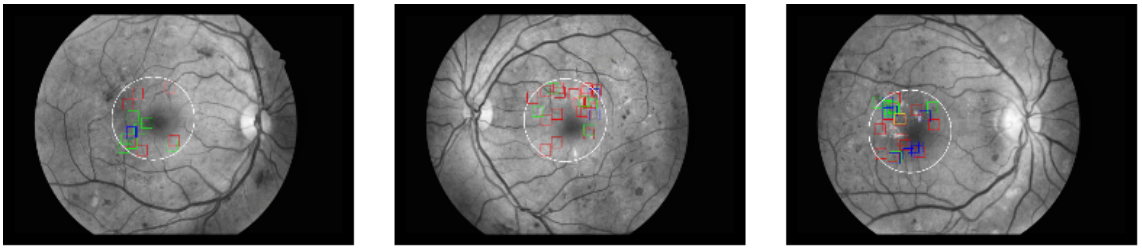


Figure 3.18: Local MAs Prediction Results. The red, green, blue, and yellow bounding box indicates FP, TP, FN (because of classifier), and FN (because of MAs candidates selection). The white circle indicates the fovea's ROI.

3.7.2/ Cascade Learning

The performance of the second experiment for MAs detection, *Proposed method (cascade)*, is compared with the other reported state-of-the-arts that focus on the same dataset and evaluation metric (see Table 3.7). The *Proposed method (cascade)* is compared with the algorithms based on Directional Local Contrast (DLC) [Long et al., 2020], local convergence [Dashtbozorg et al., 2018], two-step CNN [Eftekhari et al., 2019], ensemble features [Orlando et al., 2018], and MS-EfficientNet [Xia et al., 2021] for an evaluation conducted in the E-Ophta-MA dataset, and MS-EfficientNet [Xia et al., 2021], SESV-DLab [Xie et al., 2020] and DeepLabv3 algorithm [Chen et al., 2018] for the IDRiD dataset.

Table 3.7: Comparison of the proposed method performance with other MAs detection algorithms.

Algorithms	Dataset		FPI							AUC (PR)
	Train	Test	1/8	1/4	1/2	1	2	4	8	
DLC [Long et al., 2020]	E-Ophta	E-Ophta	0.075	0.154	0.267	0.358	0.472	0.594	0.699	-
Local convergence [Dashtbozorg et al., 2018]	E-Ophta	E-Ophta	0.358	0.417	0.471	0.522	0.558	0.605	0.638	-
Two-stages-CNN [Eftekhari et al., 2019]	Cross dataset	E-Ophta	0.091	0.258	0.401	0.534	0.579	0.667	0.771	-
Ensemble-features [Orlando et al., 2018]	Cross dataset	E-Ophta	0.14	0.2	0.23	0.37	0.45	0.52	0.62	-
MS-EfficientNet [Xia et al., 2021]	Cross dataset	E-Ophta	**0.688**	**0.701**	**0.71**	**0.718**	**0.720**	**0.733**	0.74	**0.615**
Proposed method (cascade)	E-Ophta	E-Ophta	0.266	0.299	0.344	0.408	0.538	0.669	**0.792**	0.576
MS-EfficientNet [Xia et al., 2021]	Cross dataset	IDRiD	**0.561**	**0.563**	**0.565**	**0.568**	**0.575**	**0.601**	**0.634**	**0.779**
SESV-DLab. [Xie et al., 2020]	Cross dataset	IDRiD	-	-	-	-	-	-	-	0.51
DeepLabv3 [Chen et al., 2018]	Cross dataset	IDRiD	-	-	-	-	-	-	-	0.487
Proposed method (cascade)	IDRiD	IDRiD	0.113	0.121	0.137	0.168	0.23	0.356	0.463	0.424
Proposed method (cascade)	E-Ophta	IDRiD	0.091	0.099	0.115	0.146	0.208	0.291	0.411	0.448
Proposed method (cascade)	IDRiD	E-Ophta	0.091	0.135	0.233	0.343	0.477	0.627	0.758	0.498

As shown in Table 3.7, the sensitivity of the *Proposed method (cascade)* in the E-Ophta-MA dataset reaches 0.792 for 8 FPI which is the highest sensitivity among the

other algorithms that are reported. It surpasses the performance of MS-EfficientNet [Xia et al., 2021] in the 8 number of FPI while they have the highest sensitivity for the smaller FPIs. The sensitivity of the *Proposed method* (*cascade*) in 4 FPI is higher than all the algorithms compared in Table 3.7 except MS-EfficientNet with a 0.064 difference. *Proposed method* (*cascade*) overperformed the sensitivity of DLC [Long et al., 2020] and Ensemble-features [Orlando et al., 2018] in all FPI. It is also competing tightly with local converge-based [Dashtbozorg et al., 2018], and two-stages-CNN [Eftekhari et al., 2019] in FPI > 1.

DLC [Long et al., 2020] is trained and evaluated in E-Ophta-MA without any additional dataset. DLC algorithm was trained with 74 training images from the E-Ophta-MA dataset and tested with another 74 images from the same dataset while the *Proposed method* (*cascade*) was trained in k-fold since there is no separation for training and testing set in E-Ophta-MA dataset. The images are divided into 80% as training set and 20% as validation set under 10 folds cross-validation procedure. The results shown in Table 3.7 are the average of the testing results for all values of 10-folds.

DLC [Long et al., 2020] detects the MAs by extracting 44 local hand-crafted features from an image patch (patch size 25). There are seven types of features extracted including color, grayscale, DLC, shape, texture, Gaussian filter-based, and gradient. It is trained with multiple machine learning techniques to distinguish MA and non-MA. The main factors that contribute to the failure of this method are image noise, irregular shape, and the location of the MA. Image noise is the defect from the camera that can be removed manually using post-processing. Since the main contribution of this method is DLC-based features that focus on the pattern of MA's shape, irregular shape and larger MA cause miss-classified between MA and hemorrhage. The location of MA also increases the number of FN predictions. Some true-MAs are removed in the MAs candidate extraction step, especially the MAs which are located close to the vessels.

The performance of *Proposed method* (*cascade*) that is trained and evaluated in E-Ophta-MA, is higher than the algorithm proposed by Orlando et al. [Orlando et al., 2018] where the model is trained with cross-dataset. They trained the classifier using DIARETDB1 and ROC datasets and evaluated it in the E-Ophta-MA dataset. As seen in Table 3.7, the performance of *Proposed method* (*cascade*) that is trained in the different datasets (IDRiD in this experiment) and evaluated in E-Ophta-MA, also still achieves better sensitivity in FPI>1/4. In the method proposed by Orlando et al., the ensemble technique is applied to the features of the MAs candidates. The features are extracted by both hand-crafted feature extractions and CNN-based features. It has 63 hand-craft features based on its intensity and shape. Finally, a random forest classifier is applied to detect the MAs from MAs candidates.

The *Proposed method* (*cascade*) achieves higher sensitivity in 1/8, 1/4, 4, and 8 FPI when

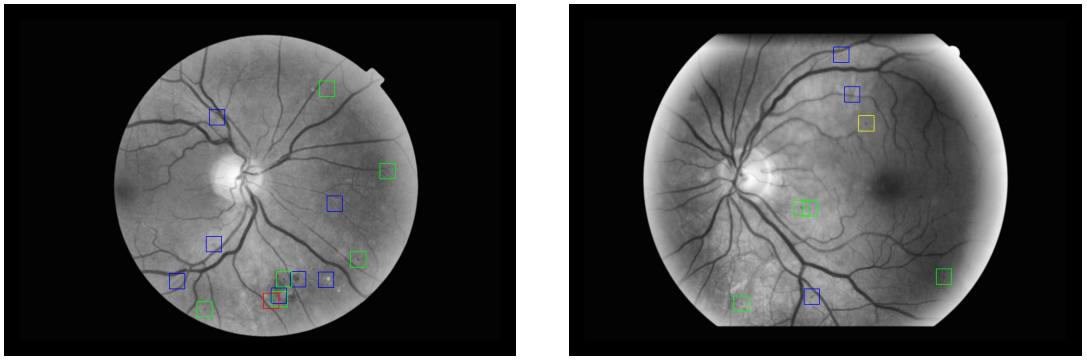
compared to the two-stages CNN algorithm [Eftekhari et al., 2019] but yields comparatively lower sensitivity in 1/2, 1, and 2 FPI. The classifier used by [Eftekhari et al., 2019] is trained and tested with E-Ophta-MA and ROC datasets. Eftekhari et al. extracted MAs candidates through segmentation using the first CNN to reduce the number of non-MA and applied the second CNN encoder to classify the MA and non-MA. In the testing scheme, MAs candidates are assigned to the objects that have a minimum of 0.6 confidence. Applying CNN for both MAs candidate extraction and classifier requires a large quantity of image training data. The MAs candidate extraction with image segmentation usually produces more wrong predictions but there is no information explained related to the performance of the first stage of the CNN in their paper.

Compared to the local-convergence algorithm [Dashtbozorg et al., 2018], the performance of the *Proposed method* (*cascade*) has higher sensitivity only in 4 and 8 FPI with significant differences but a slight-lower sensitivity in other FPIs. Dashtbozorg et al. trained and tested the model with repeated 10-folds cross-validation of the E-Ophta-MA dataset. They extract the 29 features from MAs candidates, 7 intensity-based features, 10 shape-based descriptors, and 12 local convergence-based features. The MA and non-MA are classified using the Random UnderSampling Boost classifier (RUSBoost) to deal with imbalanced data issues. The performance of this algorithm is affected by the FN produced in MAs candidate extraction. The sensitivity of the MAs candidate extraction step in this method for the E-Ophta-MA dataset is 0.95.

Proposed method (*cascade*) achieves the highest sensitivity in 8 FPI compared to other algorithms reported in Table 3.7 including MS-EfficientNet [Xia et al., 2021]. As mentioned in Section 3.7.1.1, the significant difference in performance between the proposed method and Xia et al. is the number of data used for training of both MAs segmentation (for MAs candidates extraction) and MAs classification. Xia et al. used a cross dataset (around 400 images) for training, while the proposed method uses only two specific datasets individually (148 images for E-Ophta and 54 images for IDRiD).

Proposed method (*cascade*) is also evaluated based on the PR curve as shown in Table 3.7. In the E-Ophta-MA dataset, *Proposed method* (*cascade*) and MS-EfficientNet [Xia et al., 2021] are the only algorithms that are evaluated in AUC (PR) metric. The AUC (PR) of Xia et al. is still higher than the *Proposed method* (*cascade*).

In the IDRiD dataset, the *Proposed method* (*cascade*) is compared with MS-EfficientNet [Xia et al., 2021] in FROC and AUC (PR) metrics evaluation. Both sensitivities of FPI and AUC (PR) score for the *Proposed method* (*cascade*) are lower than MS-EfficientNet for the same reasons as in the case of E-Ophta-MA. Even though the AUC (PR) of the *Proposed method* (*cascade*) is lower than SESV-DLab [Xie et al., 2020] and DeepLabV3 [Chen et al., 2018], the performance can not be compared since they are evaluated at pixel-level while *Proposed method* (*cascade*) is evaluated at lesion-level.



(a) Output 1 **(b)** Output 2

Figure 3.19: The results of MAs detection in the E-Ophta dataset. Red, green, and blue bounding boxes indicate FN, TP, and FP. Yellow bounding boxes indicate missed-real-MA from preprocessing.

The results of MAs detection from the E-Ophta-MA dataset are shown in Figure 3.19. It shows that the model of the *Proposed method (cascade)* produces more FNs compared to FPs since the cascade learning only focuses to eliminate the FPs. In *Proposed method (cascade)*, the FP prediction occurs on the objects that have dark dots at the end of the small vessels and in the region which is surrounded by the brighter objects.

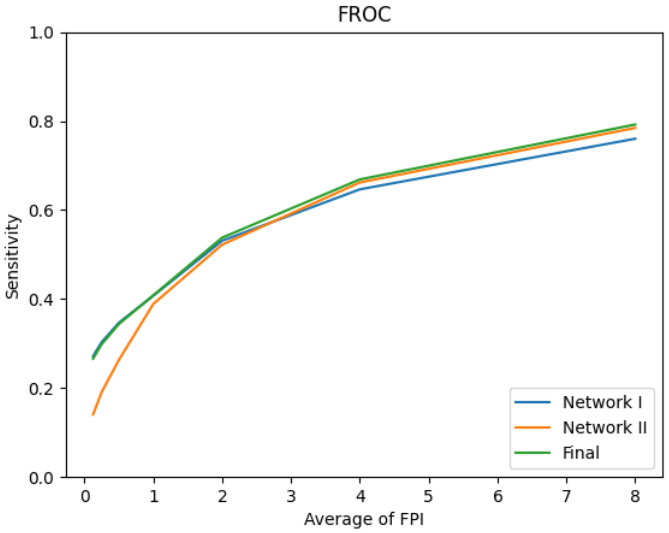


Figure 3.20: FROC curve of E-Ophta-MA dataset.

The *Proposed method (cascade)* is also tested in different datasets to find out whether it is generalized. As seen in the last block of rows in the Table 3.7, the *Proposed method (cascade)* achieves competitive results without any fine-tuning, so it generalizes quite well. As seen in Table 3.7, the results of the *Proposed method (cascade)* that is tested in IDRiD but trained with E-Ophta tend to have a lower performance, while the *Proposed method (cascade)* that is tested in E-Ophta but trained with IDRiD tends to have a higher performance. It can be inferred that the features extracted from the train-

ing images from only one dataset may not be enough to distinguish the MA and non-MA in IDRiD testing images, while IDRiD training images can give good enough features to classify the MA and non-MA in the E-Ophta dataset.

The evaluation metric of the internal performance of the *Proposed method (cascade)* that is conducted on the E-Ophta-MA dataset is described in Figure 3.20. The sensitivity of Network I is higher than Network II for FPI<2 and it keeps going below the sensitivity of Network II for FPI>2. On the other hand, the sensitivity of Network II is low for FPI<2 but it increases and surpasses Network I for the FPI>2. As shown in Figure 3.20, the performance of Final has a sensitivity almost close to Network I for FPI<2 and has a sensitivity higher than Network II for FPI>2. This shows the sensitivity of the cascade learning scheme is higher than the individual networks (Network I and Network II).

3.7.3/ Discussion Across Methodologies

This section explains the comparison between MAs detection with ensemble learning and cascade learning results. The comparison is described in Table 3.8. *Proposed method (Blue)* 64 and *Proposed method (Final)* 64 are the best performances in MAs detection with ensemble learning with patch size 64, *Proposed method (Blue)* 86 and *Proposed method (Green)* 86 are the best performances in MAs detection with ensemble learning with patch size 86.

Table 3.8: Comparison of the performance of the proposed method with ensemble learning and cascade learning. Bold values indicate the highest scores among all algorithms. Italic values indicate the highest scores among internal algorithms.

Algorithms	Dataset		FPI							AUC (PR)
	Train	Test	1/8	1/4	1/2	1	2	4	8	
Proposed method (Blue) 64	IDRiD	IDRiD	0.025	0.045	0.065	0.103	0.174	0.246	0.376	0.503
Proposed method (Final) 64	IDRiD	IDRiD	0.019	0.036	0.056	0.105	0.189	0.287	*0.41*	0.49
Proposed method (Blue) 86	IDRiD	IDRiD	*0.054*	*0.079*	*0.096*	*0.117*	0.176	0.239	0.334	0.505
Proposed method (Green) 86	IDRiD	IDRiD	0.012	0.025	0.051	0.101	*0.203*	*0.295*	0.379	***0.526***
Proposed method (cascade)	IDRiD	IDRiD	***0.113***	***0.121***	***0.137***	***0.168***	***0.23***	***0.356***	***0.463***	0.424

As seen in Table 3.8, *Proposed method (cascade)* achieves the highest sensitivity in all FPI compared to all other proposed methods with ensemble learning, but it has the lowest AUC (PR) score. Though *Proposed method (Green)* 86 is the highest AUC (PR), the score is slightly higher than 0.5. It implies that the MAs detections with both ensemble and cascade are still weak. However, the cascade learning with Blue-enhanced green-background suppression input can increase the sensitivity in a small number of FPIs.

All performance of Proposed methods for ensemble learning and cascade learning are computed including the FN results that are produced from the MAs candidate extraction process. The MAs candidates extraction gives an average of 95% sensitivity per image.

3.8/ CONCLUSION

We have proposed methods to answer the objective of this research related to MAs detection with two strategies. One is by applying an unsupervised method for MAs candidate extraction which minimizes the necessity of a large number of data in the training phase, and the second is by implementing both ensemble and cascade learning with background suppression information to enhance the image features.

The MAs detection with ensemble learning is built with the help of three identical fine-tuned networks with different input channels and an ensemble vector to decide the final decision by the voting scheme. This method provides an overall analysis of the impact of each color channel in MAs detection. Different color channels have different impacts on MAs detection depending on the patch size of the image. The combination of input channels (green, enhanced green, and background suppressed image) with patch size 86 achieves the highest AUC (PR) score. In contrast, the combination of input channels (blue, enhanced green, and background suppressed image) has a more stable performance irrespective of the patch size. However, this method fails to classify objects with blurry edges.

Combining the results of the previous method with cascade learning, the sensitivity of the MAs detection is increased in smaller numbers of FPI. Cascade learning gives the classifier the ability to re-learn the features of the object which fails to be learned in the previous unit network. The final prediction of this method can reduce the number of FPs per image without a significant reduction in the sensitivity since it is dependent on both unit networks. With a significantly different quantity of training data, the performance of the proposed network reaches the highest sensitivity for the 8 FPI and competing sensitivity values for FPI<8 in E-Ophta dataset. Yet, this method still produces the FPs for dot-like objects which are located at the end of the small vessels and between some brighter objects.

3.9/ FUTURE WORKS

Furthermore, the classifier can be improved to focus to reduce not only FP but also FN, since cascade learning reduces the FP but increases the FN in the end results. The patch size contributes significantly to the results of MAs detection. Choosing the appropriate patch size might enhance the results.

4

Multi-Label Ocular Abnormalities Detection

4.1/ Introduction

Persistent ocular diseases give a high risk of retinal damage which can lead to blindness [Congdon et al., 2004]. One major challenge of this situation is the lack of quality diagnosis and prevention [Pachade et al., 2021]. Early detection and timely diagnosis of ocular pathologies are effective ways to prevent this visual impairment. The detection of ocular diseases with the help of computer-aided diagnosis (CAD) is facilitated by the availability of retinal image modalities such as optical coherence tomography (OCT), fluorescein angiography (FA), and color fundus photography (CFP). Among other retinal image modalities, CFP is the most cost-effective and simple approach for eye screening. Because of its effectiveness and efficiency, multi-label detection of the CFP is vital considering the possibility of a patient getting affected by more than one ocular disease. However, one of the challenges of multi-label detection in ocular diseases is to detect both frequent and rare ocular abnormalities. Rare ocular abnormalities are usually ignored in the detection because of the limitation in training data quantity [Pachade et al., 2021].

There are limited studies that address the multi-label detection of ocular abnormalities from a single color fundus image. The common approach to solve this problem is based on only spatial relationship learning and ignores the co-occurrence dependency issue which is a crucial problem in multi-label detection [Zhao et al., 2021b]. The interconnection between classes (semantic information) can be acquired either only from a visual modality or from multi-modalities (visual and linguistic modalities). The model of co-occurrence dependency which is built based on the visual and linguistic modality is widely studied in general multi-label detection tasks but not in the medical application. The linguistic modality in medical applications may contain uncorrelated labels and semantic information that becomes the noise that hinders the detector to learn powerful features [Guan et al., 2020]. However, the label semantic information can be a complement for

the visual features to build the model of co-occurrence dependency in multi-label detection for both frequent and rare ocular abnormalities considering the image availability. In the proposed methods, the uncorrelated label semantic information in medical application issues is prevented by building the semantic dictionary learning to avoid the irrelevant correlation between the label and visual features and handling the out-of-vocabulary (OOV) in the linguistic modality that causes irrelevant word features.

4.2/ Objectives

The main objective of the research in this chapter is to analyze the impact of the linguistic modality as the complement to the visual modality in the multi-label detection of frequent and rare ocular abnormalities from a color fundus image. The model is trained and tested in RFMiD dataset (see Section 2.1.5.5), so it is designed to detect 28 labels that consist of 45 ocular abnormalities.

4.3/ Outline of The Chapter

A brief review of the related works is presented in Section 4.4. Section 4.5 describes the proposed methodologies applied in this research to solve the problem that is mentioned in Section 4.2. The implementation details of the proposed methods are described in Section 4.6, and the results of the experiments are discussed in Section 4.7. Lastly, the conclusion of the experiments that are conducted for multi-label detection is explained in Section 4.8.

4.4/ Related Works

There are limited numbers of studies that focus on multi-label detection for ocular abnormalities, especially from a single color fundus image. Table 4.1 summarizes the earlier studies related to multi-label detection that is conducted in the RFMiD dataset. The multi-label detection models of KAMATALAB, Schulich Applied Computing, BNUAA, Nekaravuru, WWW, IGSTfencing, MISIT, and Chizu & Arai & Okada [Pachade et al., 2023] are reported as the highest model performances in the RFMiD dataset challenge, while other studies shown in Table 4.1 are reported in separate publications. Most of the methods applied to detect multi ocular abnormalities are ensemble learning, combining multiple CNNs as the backbone to extract the features. Even though this method can deal with limited quantity of training data, it requires higher resources.

Table 4.1: Related works in multi-label detection with RFMiD dataset.

Authors	Method	Labels	Training Dataset	Performance
KAMATALAB [Pachade et al., 2023]	Ensembling 3 backbones. The task is divided into 3 sub-tasks.	28	RFMiD	Final score in Evaluation set is 0.802 and 0.782 in Test set.*
Schulich Applied Computing [Pachade et al., 2023]	Ensembling 5 backbones. Trained with Asymmetric Loss.	28	RFMiD	Final score in Evaluation set is 0.786 and 0.778 in Test set.*
BNUAA [Pachade et al., 2023]	Ensembling 2 backbones with a heavy image augmentation.	28	RFMiD	Final score in Evaluation set is 0.782 and 0.758 in Test set.*
Nekaravuru [Pachade et al., 2023]	Ensembling 3 backbones. The task is divided into 2 sub-tasks.	28	RFMiD	Final score in Evaluation set is 0.757 and 0.749 in Test set.*
WWW [Pachade et al., 2023]	Applying CLAHE in pre-processing. Ensembling 6 backbones.	28	RFMiD	Final score in Evaluation set is 0.7584 and 0.738 in Test set.*
IGSTfencing [Pachade et al., 2023]	The model is semi-supervised siamese graph CNN. The relation between labels is learnt by GCN.	28	RFMiD and ODIR	Final score in Evaluation set is 0.791 and 0.71 in Test set.*
MISIT [Pachade et al., 2023]	Upsampling the data. Ensembling 4 backbones.	28	RFMiD	Final score in Evaluation set is 0.7027 and 0.6893 in Test set.*
Chizu & Arai & Okada [Pachade et al., 2023]	Pre-processed the image to have the same aspect ratio alignment. Ensembling multiple backbones	28	RFMiD	Final score in Evaluation set is 0.738 and 0.78 in Test set.*
[Rodriguez et al., 2022]	The model is based on C-Tran, transformer encoder that is trained with partial labels.	20	RFMiD, ARIA, and STARE	For 15 labels, final score is 0.824, F1 is 0.573, mAP is 0.685, and AUC is 0.962.
[Sun et al., 2022a]	The features are extracted by EfficientNet with adding a spatial attention layer.	7	RFMiD	Accuracy is 0.883 and F1 score is 0.762.
Multi-Expert [Lyu et al., 2022]	Ensembling multiple backbones with heuristic stacking to decide the final predictions.	28	RFMiD	Final score in Test set is 0.778.

* indicates the methods presented in the RFMiD challenge.

The imbalanced data issue is tackled by selecting the loss function that focuses on the imbalance data, such as asymmetric loss, and by dividing the model into multiple subtasks. In task division, each sub-tasks learn to predict some group of labels that has less-imbalance distribution. The methods proposed by IGSTfencing [Pachade et al., 2021]

and Rodriguez et al. [Rodriguez et al., 2022] are the only methods that also learn the correlation between labels (implicitly or explicitly) to detect ocular abnormalities. However, all the methods use only spatial information as the input of their models. IGSTfencing [Pachade et al., 2023] computes the labels correlation based on the co-occurrence of the labels. It needs large training data to build an accurate correlation. Rodriguez et al. [Rodriguez et al., 2022] learn the correlation between co-occurrence labels and spatial information implicitly by using C-Tran architecture that masks some labels in the training phase so that the model can learn many possible labels. This model is trained in a custom dataset that consists of RFMiD, ARIA, and STARE datasets. However, the model is designed to predict only 20 labels: 15 labels from RFMiD dataset and 5 labels from ARIA and STARE datasets. The performance of this model in the RFMiD dataset is reported with the same evaluation metrics as the RFMiD challenge. However, it is not mentioned if the model is tested in the Evaluation set or Test set. The same occurred in the model performance of Sun et al. [Sun et al., 2022a]. The model is evaluated in the RFMiD dataset but has no information if it is tested in the Evaluation set or Tested set. The evaluation metrics that they use are accuracy, F1 score, complete match, and index comparison. Though all the evaluation metrics are different from other studies, they provide the AUC of the ROC curve for each ocular abnormality in testing phase. The model is designed to predict 7 ocular abnormalities.

Due to the limited number of studies that focus on the multi-label detection of ocular abnormalities from a single color fundus image, other related works that study the multi-label detection of ocular abnormalities from binocular color fundus images are also reviewed. In addition, earlier studies related to the multi-label detection of natural images are explored to learn the correlation between spatial input and label co-occurrence dependency. These studies are summarized in Table 4.2.

Table 4.2: Other related works in multi-label detection.

Authors	Input type	Modality	Method
[He et al., 2021]	Binocular CFP	1	CNN with pixel-wise correlation.
[Gour et al., 2021]	Binocular CFP	1	CNN with joint features.
[Sun et al., 2022b]	Binocular CFP	2	GCN with LightGBM classifier.
[Zhu et al., 2017]	Natural image	1	CNN with attention layer.
[Chen et al., 2019]	Natural image	1	Multi GCNs.
[Wen et al., 2020]	Natural image	1	Projection of visual and label features.
[Chen et al., 2021b]	Natural image	2	C-GCN and P-GCN.
[Nguyen et al., 2021]	Natural image	2	Stacked GCN and n-modules CNN.
[Zhao et al., 2021a]	Natural image	2	Learning the structural relation graph.
[Zhao et al., 2021b]	Natural image	3	Transformer with LGE.
[Zhou et al., 2021]	Natural image	2	CNN with dictionary learning.
[Liu et al., 2021a]	Natural image	1	Query-based transformer.

The interference of label information in multi ocular abnormalities detection is not yet studied widely, either from a single CFP or binocular CFPs. More studies related to multi

ocular abnormalities detection are found from a pair of color fundus photography images (CFP) such as He et al. [He et al., 2021], Gour et al. [Gour et al., 2021], and Sun et al. [Sun et al., 2022b]. Like other common approaches, He et al. [He et al., 2021] and Gour et al. [Gour et al., 2021] use only the spatial information as the input for their detection model. He et al. [He et al., 2021] learn the pixel-wise correlation between a pair of CFPs by Spatial Correlation Module (SCM), and Gour et al. [Gour et al., 2021] learn the features of individual CFPs and concatenate these features directly to classify the ocular diseases. Unlike other studies with binocular CFP images, Sun et al. [Sun et al., 2022b] proposed a multi-label detection by considering the co-occurrence dependency. They build the graph convolutional network (GCN) to model the correlation between labels and classify the ocular diseases using LightGBM. This method is designed to detect seven frequent ocular diseases from binocular CFP images.

In a general multi-label detection task, the co-occurrence dependency can be modeled from a single visual modality and multi-modality (visual and linguistic modalities). Some studies model the co-occurrence dependency from a single visual modality (image) by learning the implicit spatial relations with self-attention mechanism [Zhu et al., 2017] and enforcing the visual consistency to the attention regions under different spatial transformations. The model of the co-occurrence dependency is built explicitly by graph models such as Chen et al. [Chen et al., 2019] that generate the graph models of the co-occurrence dependency between labels from the label frequency, and Wen et al. [Wen et al., 2020] that learn the correlation between labels from the projection of the visual features to the label mapping. The features with the same labels are closer to each other in the projection space.

More studies focus on multi-modality (visual and linguistic modality) as another approach to model the co-occurrence dependency in the multi-label detection task. The common method to model co-occurrence dependency from labels is by word embedding. Chen et al. [Chen et al., 2021b] explicitly model the label semantic information by graph convolutional network (GCN). They proposed a classifier GCN (C-GCN) and a prediction GCN (P-GCN) model. C-GCN learns to map the label semantic information into an interdependent classifier that is shared with all images while P-GCN learns to disentangle the visual features into class-relevant features. Nguyen et al. [Nguyen et al., 2021] focus to apply a graph approach to learn the label semantic information and its topology information. They proposed a divide-and-conquer technique to learn the visual features of the separate objects using n-modules of CNN. The label embedding and the topology structure of the labels are fed to the stacked GCN to support the visual recognition to classify the labels. Instead of learning the visual features in high dimension, Zhao et al. [Zhao et al., 2021a] learn the structure relation graph of the visual features that are extracted from the CNN. The co-occurrence dependency of the labels is modeled using a GCN. This label semantic information is added to the structural relation graph of the

visual features information to classify the labels. However, modeling the label features adds another complexity to the network. The label semantic information using GCN may learn spurious correlation when the label statistics are not enough [Liu et al., 2021a].

The visual representation of the correlation between visual and linguistic features is important to study in the multi-label detection task to prevent the irrelevant correlation between visual and linguistic features. Zhao et al. [Zhao et al., 2021b] proposed the linguistic guided enhancement module (LGE) to enhance the representation across modalities. The visual modality, semantic modality, and linguistic modality are the inputs of this model. The visual modality is extracted by linear projection on a sequence of image patches. Semantic modality is extracted from a pre-trained CNN, and linguistic modality is extracted from the label embedding. The features of all modalities are trained using a multi-modal transformer. Zhou et al. [Zhou et al., 2021] proposed a semantic dictionary learning as the visual representation of the visual features and label embedding. This model is constrained by the consistency of the label embedding features and visual features. Liu et al. [Liu et al., 2021a] proposed a query-based transformer to predict the multi-label from an image. Instead of using the label embedding generated from the pre-trained network, they choose to learn the label features vector independently. This model is an improved DETR model [Carion et al., 2020] where each query corresponds to one label class.

Inspired by Zhou et al. [Zhou et al., 2021] and Liu et al. [Liu et al., 2021a], we proposed a multi-label detection for frequent and rare ocular abnormalities using semantic dictionary learning since it is an effective way to exploit the correlation between labels and visual features to avoid uncorrelated information between labels and visual features. In multi-label detection for medical application, the label features vector may contain out-of-vocabulary (OOV) vector because of the category difference, so the method is also designed to be adaptable with OOV words to reduce the irrelevant word features.

4.5/ METHODOLOGY

The methodology in this chapter is divided into pre-processing section and two different proposed methods using semantic dictionary learning. Pre-processing is applied to all proposed methods as the initial process of multi-label detection. The first proposed method is a multi-label detection that learns the semantic dictionary learning with visual and labels embedding consistency. The model is also divided into the OOV model and the non-OOV model. The second proposed method is a multi-label detection that learns the semantic dictionary learning with the transformer decoders.

4.5.1/ PRE-PROCESSING

Pre-processing is an important initial process to uniformize the image size and to select the ROI of the fundus area. The image size is resized into 224 x 224 pixels for the first proposed method and 448 x 448 pixels for the second proposed method. The background of the color fundus image is cropped to have a direct focus on the FOV of the fundus. The background cropping is described in Algorithm 2.

Algorithm 2 Background Cropping

Input: A color fundus image (I)
Output: Result of background cropping (I')
1: I_{gray} = CONVERT_RGB_TO_GRAY(I)
2: I_{thresh} = OTSU_THRESHOLDING(I_{gray})
3: $I_{contours}$ = FIND_CONTOURS(I_{thresh})
4: $I_{contours}$ = SORTED_CONTOURS_AREA($I_{contours}$)
5: $I_{contour} = I_{contours}[0]$
6: $x, y, w, h = I_{contour}$
7: $I' = I[y : y + h, x : x + w]$
8: Return I'

4.5.2/ CNN-BASED SEMANTIC DICTIONARY LEARNING

4.5.2.1/ DATA PREPARATION

The experiments of the data preparation are divided into sampling and non-sampling categories:

- **Sampling Experiment**. The sampling experiment is a strategy applied by MISIT [Pachade et al., 2023] to solve the problem of limited data in a multi-label task for the RFMiD dataset. They applied a novel distributed upsampling that was introduced by Müller et al. [Müller et al., 2021b] to increase the number of images that contain the minority labels-combination in order to balance the data.

  1. **Distributed Upsampling**. Data is upsampled to balance the label distribution by augmenting the image (flipping, rotation, and color jittering) [Müller et al., 2021a]. Upsampling is applied in different quantities to each label depending on its distribution. The quantity of the upsampled data is increased up to three times from the original data with each label occurring at least 100 times. This upsampling technique can reduce the imbalance issue in the RFMiD dataset that has a critical imbalance condition (see Figure 2.10).
  2. **Global Upsampling.** Distributed sampling gives equivalent weight for each label in the upsampled data. It may cause an increment performance for rare oc-

ular abnormalities, but also decrement performance in frequent labels. Hence, an experiment with global upsampling is done to compare the model performance with distributed upsampling. The global upsampling also utilizes image augmentation techniques to generate additional images. The upsampling is applied to each label equally. In this experiment, the original data is upsampled three times.

The upsampled data is split into training and validation data. The intensities of the image, in training and validation data, are converted into a range [0,1].

- **Non-sampling Experiment**. The non-sampling experiment consists of data splitting, image augmentation, and image conversion. The data is split into training and validation data without considering the distribution of the labels. The image augmentation (color jitter, horizontal flip, rotation) is applied only to the training data. The intensities of the image are converted into a range [0,1] for both training and validation data. Due to the imbalance data issue, the non-sampling experiment is applied in two different scenarios:

  1. **Stages.** Inspired by Zhou et al. [Zhou et al., 2022], KAMATALAB [Pachade et al., 2023], and Nekaravaru [Pachade et al., 2023], the prediction of the label is divided into two stages to reduce the data imbalance.

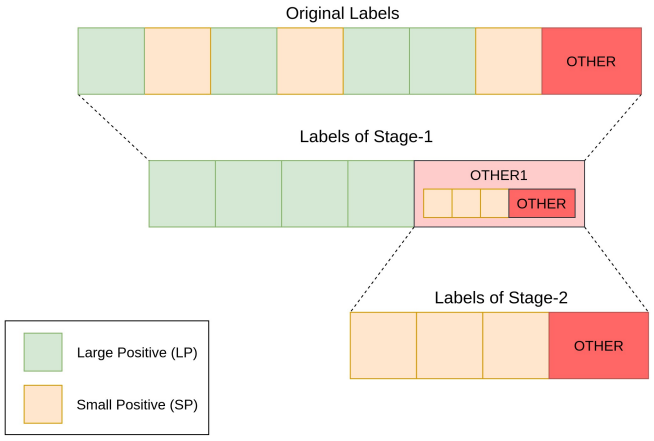


Figure 4.1: Illustration of stages scenario.

As illustrated in Figure 4.1, the first stage focuses to predict only the large positive labels (LP), and the small positive labels (SP) are added to OTHER label (named OTHER1). LP refers to the label that has a minimum of 30 images and SP refers to the label that has less than 30 images. The second stage focuses to predict SP labels and also to differentiate between SP labels and OTHER label. OTHER label is a combination of images with very rare labels. In the

testing scheme, the input of the second stage is the OTHER1 prediction from the first stage.

2. **No-stage.** In this scenario, all labels (27 labels and OTHER label) are predicted in a network without being divided into multiple stages.

4.5.2.2/ CLASSIFICATION

The first proposed method is inspired by Zhou et al. [Zhou et al., 2021]. The semantic dictionary built by Zhou et al. fails to learn the labels that contain irrelevant label features which are commonly caused by OOV words, the words that are not trained in the network because of the limited number of such words available in the training datasets. The main cause of this failure is the zero vector in the label features that represent the OOV words in a word embedding. Thus, we proposed an OOV-adaptive deep semantic dictionary learning that considers the visual consistency between image features which are generated from the color fundus image, and label features which are generated from the linguistic information.

The overview of the first proposed method is described in Figure 4.2. To avoid the learning failure in semantic dictionary, the OOV labels are separated from the original data labels. The classification process of the OOV labels and non-OOV labels are also separated inside the network. The detection of the image that belongs to OOV labels is predicted directly from the image features, while the detection of the non-OOV labels is predicted by the semantic dictionary learning from the image features and label features.

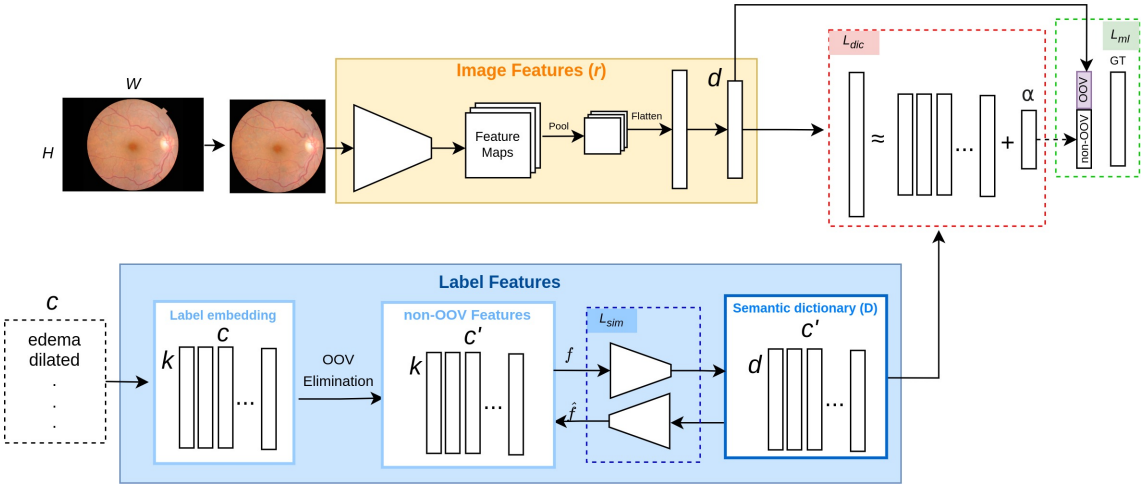


Figure 4.2: Overview of the first proposed method. The background of RGB image is cropped and fed into a classifier to extract the image features (f) while the list of labels is fed into the NLP model to extract the label features. The correlation between these features is described in the semantic dictionary (D).

Image Features Extraction. The input of the classifier is a color fundus image (Img) with H x W x 3 dimension, the image output of the pre-processing step, and the list of labels c. The spatial features are extracted by feeding the image Img into a fine-tuned pre-trained EfficientNet-B4 with the last three trainable layers. The image features have d dimensions. This proposed method is backbone-agnostic. The pre-trained network to extract the image features can be changed into another pre-trained network.

Linguistic Features Extraction. The linguistic features are extracted by vectorizing the labels. However, linguistic features can only be extracted from known labels. Therefore, the OOV labels need to be eliminated from the labels c (OOV Elimination). Initially, all labels c are fed into a pre-trained word embedding network by transfer learning. The non-OOV labels will have non-zero vectors while the OOV labels will have zero vectors or no vectors. The output of the OOV elimination is the non-OOV label features that have c' x k dimension. Each word in the non-OOV labels has k linguistic feature vector. For the labels that consist of multiple words, the average of the vector from each word with k elements is calculated to be the final word vector. This proposed network is also label embedding-agnostic. The pre-trained NLP network to extract the label embedding can be changed into another word embedding extractor.

Semantic Dictionary Learning. The semantic dictionary consists of a learnable metric that represents the correlation between non-OOV label embedding and image features. It is obtained by an auto-encoder network with bidirectional transformation to prevent overfitting. The auto-encoder is also applied to generate a different dimension for the semantic dictionary that matches the dimension of image features d.

The semantic dictionary is restricted by information that is consistent with the label embedding vectors. As seen in Eq. 4.1, this consistency is defined by a cost (L_{sim}) that maximizes the similarity between each vector of re-generated ($\hat{f}$) and original (f) non-OOV features as a cosine similarity function, where, ψ is the set of learnable parameters and n is the number of labels.

$$\arg\max_{\psi} L_{sim} \; := \; \frac{1}{n}\sum_{i=1}^{n} \cos\left(f_i, \hat{f}_i\right) \; = \; \frac{1}{n}\sum_{i=1}^{n} \frac{f_i \cdot \hat{f}_i}{\|f_i\|_2 \times \left\|\hat{f}_i\right\|_2} \tag{4.1}$$

In this network, the L_{dic} is computed to represent the visual consistency. Visual consistency of the image and label features describe their visual correlation. Instead of forcing the assumption on the network that the label features contain the same information as the image features, this network adds the vector α to learn the degree of the correlation between the image features and label features.

$$L_{dic} = \|r - D\alpha\|_2^2 + \lambda\|\alpha\|_2^2 \tag{4.2}$$

The L_{dic} is explained in Eq. 4.2 where r is the image features, D is the semantic dictionary, and λ is the regularization parameter. The vector α is updated in forward propagation as calculated in Eq. 4.3 where I is an identity matrix.

$$\alpha = \left(D^T D + \lambda I\right)^{-1} D^T r \tag{4.3}$$

The prediction output of the OOV and non-OOV labels are concatenated in the last layer that is optimized by minimizing the loss of the multi-label prediction (L_{ml}). OOV labels are predicted directly from the image features while non-OOV labels are predicted from semantic dictionary learning. The semantic dictionary learning is optimized in backward propagation by minimizing the L_{total} as the total loss.

$$L_{total} = \frac{L_{ml} + \beta L_{dic}}{L_{sim}} \tag{4.4}$$

As seen in Eq. 4.4, the total loss L_{total} consists of L_{ml}, β, L_{dic}, and L_{sim} where β is a hyper-parameter to balance the loss and L_{ml} is the multi-label loss. Inspired by the focal loss [Lin et al., 2017b], the multi-label loss L_{ml} (see Eq. 4.5) applied in this method combines the focal loss and balanced binary cross entropy with weight for each class (w_{cls}) and weight for the sparsity of the positive and negative labels (w_{spr}) to deal with the imbalance of labels distribution and the imbalance of label vector.

$$L_{ml} = -\sum_{i=1}^{n} w_{cls}^{i} \, (1 - p_t^i)^{\gamma} \, (w_{spr} \; log(p_t^i)) \tag{4.5}$$

The value of p_t is equal to p if the label is 1 and $(1 - p)$ if the label is equal to 0.

Class weight (w_{cls}) is computed based on the inverse of the square root of a number of samples (ISNS), $w_{cls} = c\frac{w_{nc}}{\sum w_{nc}}$, where $w_{nc} = \frac{1}{\sqrt[2]{N_c}}$ and N_c is the number of samples per class.

Sparsity weight (w_{spr}) is a coefficient that is calculated as $w_{spr} = \frac{N^0}{N^1}$. N^0 and N^1 are the number of samples in class 0 and class 1.

Figure 4.3: Data Illustration: Two image samples with 5 labels.

Given a data illustration as described in Figure 4.3, the illustration of the w_{cls} calculation is explained in Figure 4.4 and the value of w_{spr} is 0.66, N^0=4 and N^1=6.

N_c	1	1	2	1	1
W_{nc}	1	1	0.71	1	1
W_{cls}	1.06	1.06	0.42	1.06	1.06

Figure 4.4: Illustration: The calculation of w_{cls}.

4.5.3/ Transformer-based Semantic Dictionary Learning

4.5.3.1/ Data Preparation

The data preparation for the second proposed method consists of data splitting, image augmentation, intensity conversion, and image normalization. The data is split into training and validation data without considering the distribution of the labels. The random augmentation [Cubuk et al., 2020] is applied to the color fundus image only to the training data. The intensities of the image are converted into a range [0,1] for both training and validation data. Then, the image is normalized with the standard deviation and mean of ImageNet [Russakovsky et al., 2015].

4.5.3.2/ Classification

The second proposed method is inspired by Liu et al. [Liu et al., 2021a] and Zhou et al. [Zhou et al., 2021]. We proposed a semantic dictionary with visual attention that learns to map the visual features that are extracted from the color fundus image with the label features that are generated from the label embedding.

Compared to the first proposed method that constrains the label features to have the same consistency with the image features, the second proposed method focuses to map the image features as the key and value, with the semantic dictionary as the query, to have the same focus to the object region attention.

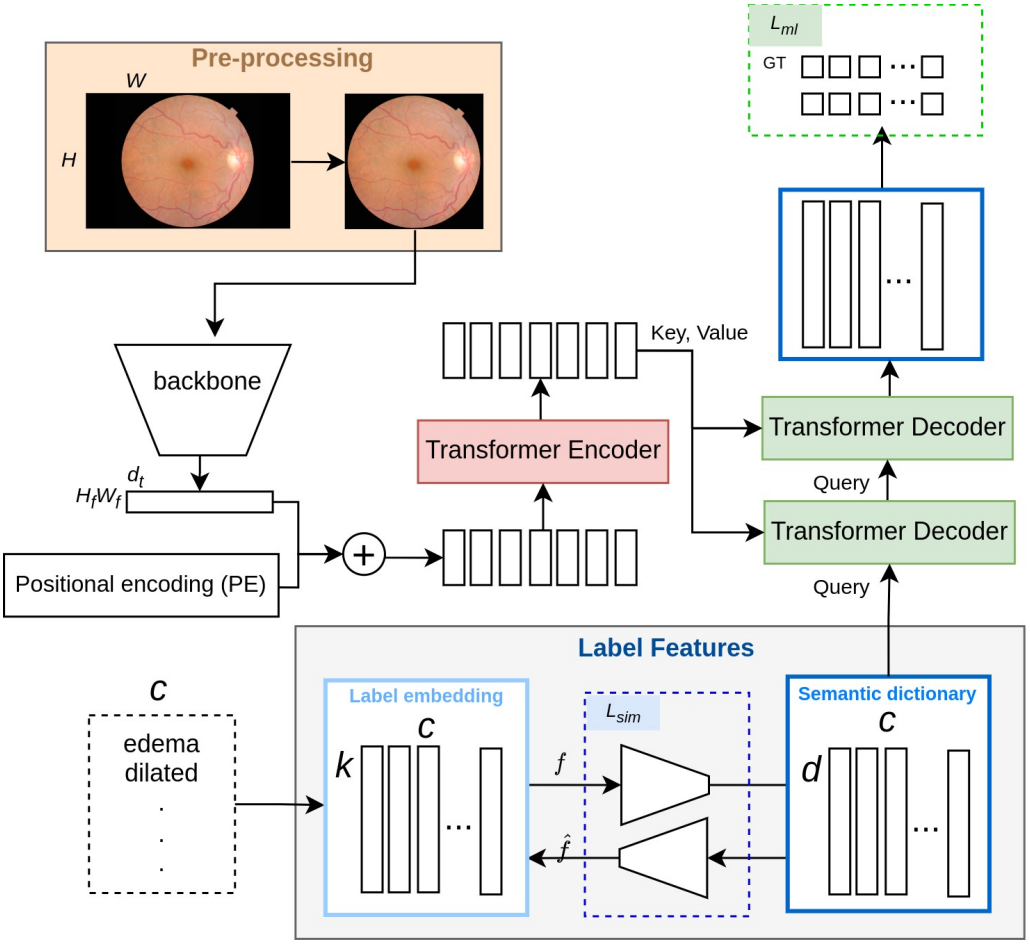


Figure 4.5: Overview of the second proposed method

The overview of this proposed method is seen in Figure 4.5. The main transformer architecture of the second proposed method is mostly adopted from DETR [Carion et al., 2020]. It uses the features from a pre-trained CNN as the input of the model and it detects the objects with transformer decoders. The difference with DETR and the second proposed method lies in the query of the decoder. The query of the decoder in DETR is to predict the presence and location of ROI for all the seen objects while the query of the second proposed method is to build a semantic dictionary. The pre-trained network as a backbone is selected to have a more compact features representation and also to reduce the necessity of large training images as the drawback of ViT [Dosovitskiy et al., 2020].

Image Features Extraction. The classifier aims to detect the presence of c ocular abnormalities given a color fundus image (I) as the input image with H x W x 3 dimension. After cropping the image background in pre-processing, the spatial features of the image with H_f x W_f x d_f are extracted from a backbone and are projected to H_f x W_f x d_t to match with the dimension of the transformer. The dimension of these features are

also reshaped into $H_f W_f$ x d_t to be a sequence input for the transformer encoder. The backbone for the spatial features extraction is a pre-trained CNN, ResNet101. However, this proposed method is backbone-agnostic. The backbone can be changed with any pre-trained image network.

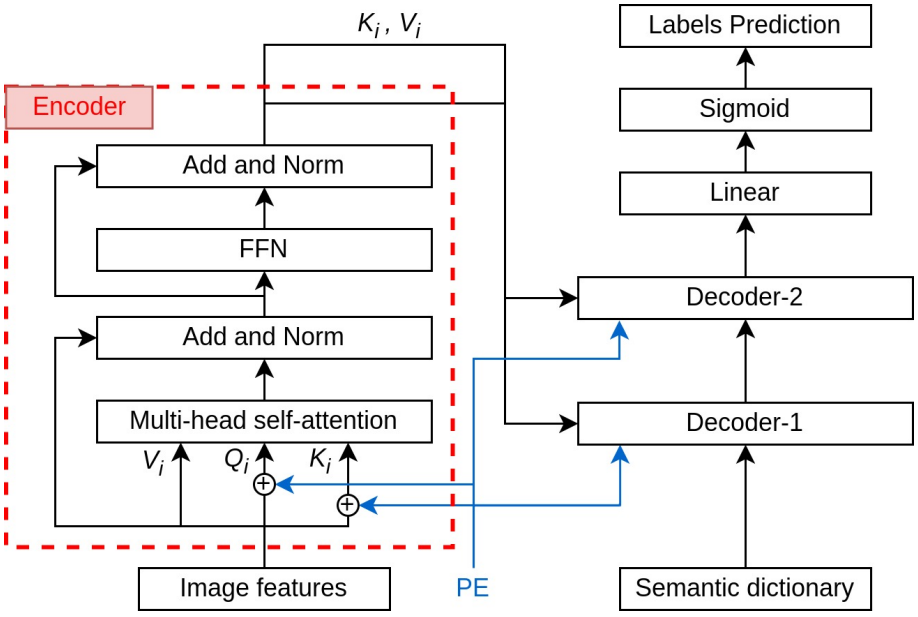


Figure 4.6: Encoder detail architecture.

Spatial positional encoding (PE) in this proposed method is fixed value and it is generated in the same way as the positional encoding generation in DETR. As seen in Figure 4.6, the image features added with positional encoding are fed into one transformer encoder. The transformer encoder performs multi-head self-attention (with 4 heads) from the image features. The residual learning is applied to each Add and Norm block. The key (K_i) and value (V_i) of the image features, as the output of the encoder, are given into two transformer decoders.

Linguistic Features Extraction. Linguistic modality is extracted from the list of labels c and it is used as another input of the classifier. The list of labels c is converted into a label embedding metric before adding it to the model. The label embedding with k x c dimension is generated from a pre-trained network that is trained in text documents. Each label has k vector to represent its linguistic features. The feature vector of the label that consists of multi-words or phrases is finalized by averaging the vector of each word. The OTHER label is also computed by the vector average of all ocular abnormalities that belong to the OTHER label. This proposed method is also agnostic to the label embedding method. The pre-trained network to generate the label embedding can be changed with the network that has a non-zero vector for OOV words.

Semantic Dictionary Learning. The semantic dictionary is a learnable metric that maps the correlation between image features and linguistic features. The value of the semantic dictionary is updated by two transformer decoders that inject the image features and an auto-encoder network that injects the label features. The updates of the semantic dictionary are constrained by linguistic features consistency and the labels ground-truth (GT).

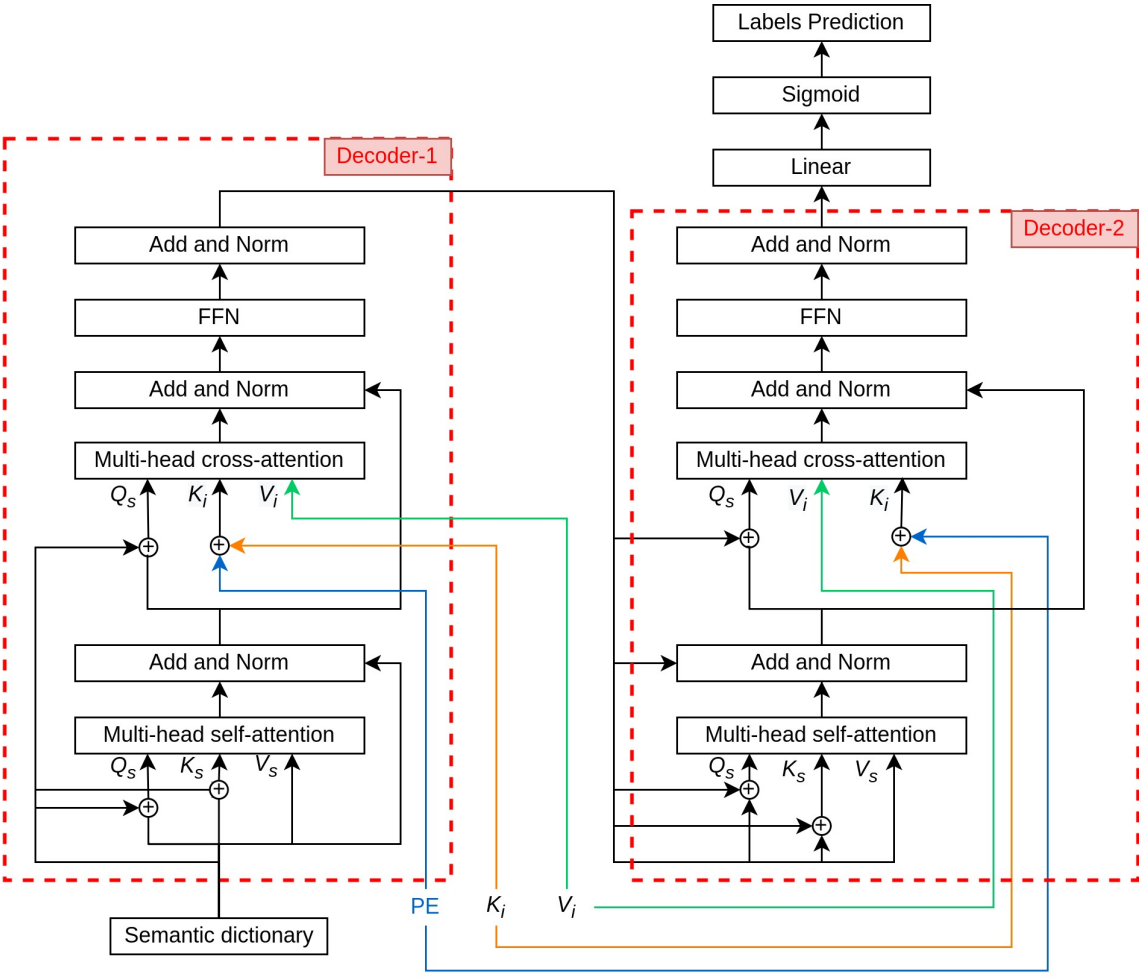


Figure 4.7: Decoders detail architecture.

Similar concept to the first proposed method (see Section 4.5.2.2), the linguistic features consistency (L_{sim}) is also restricted by Eq. 4.1. The auto-encoder generates a d x c semantic dictionary from a k x c label embedding with constant information. As seen in Figure 4.7, the semantic dictionary is given into the first transformer decoder as the query (Q_s) along with K_i and V_i that comes from the encoder output, and spatial positional encoding. As mentioned in Section 2.2.1.4, each transformer decoder consists of a self-attention block, cross-attention block, and FFN block. The four-head self-attention produces intra-attention that computes the similarity weight (V_s) corresponding to the

query (Q_s) and key (K_s) of the semantic dictionary. The residual connection and layer normalization are applied to each decoder block. Similar to DETR, the positional encoding of the transformer decoders is the input of the decoder. It is learnable, not a fixed positional encoding that is applied in the encoder. The input of the decoder can also act as the positional encoding since the value of the semantic dictionary is expected to be unique for each class. In Decoder-1, the semantic dictionary as the input of the first decoder is added to Q_s and K_s in the self-attention block.

The semantic dictionary is also added to the Q_s in four-head cross-attention as the positional encoding while spatial positional encoding (PE) is added to the K_i. In cross-attention block, the query from the semantic dictionary (Q_s) learns to map the value (V_i) from the image based on the image key (K_i) with its positional encoding. The output of the cross-attention goes through the FFN block before being fed into the next decoder block. The query of the output from Decoder-1 (Q_s) is given to Decoder-2 along with key (K_i), value (V_i), and spatial positional encoding (PE) from the image.

All decoders consist of the same blocks sequence. In Decoder-2, the Q_s is added to the Q_s and K_s in four-head self-attention block, and it is added also to the Q_s in four-head cross-attention block. The output of the Decoder-2 has d x c dimension, the same dimension as the semantic dictionary. The output is projected into c vector in Linear layer followed by Sigmoid activation layer to have multi-labels prediction.

The semantic dictionary is optimized by minimizing the multi-label loss (L_{ml}) and maximizing the label features consistency (L_{sim}). The total loss for the second proposed method is described in Eq. 4.6.

$$L_{total} = \frac{L_{ml}}{L_{sim}} \tag{4.6}$$

The multi-label loss (L_{ml}) is defined by asymmetric loss (ASL) [Ridnik et al., 2021], a variant focal loss with different γ for positive and negative. The ASL loss is explained in Eq. 4.7 where p is the probabilities, γ^+ is 0, and γ^- is 2.

$$L_{ml} = \frac{1}{c}\sum_{c=1}^{c}\begin{cases}(1-p)^{\gamma^+}\log(p), & y=1\\ (p)^{\gamma^-}\log(1-p), & y=0\end{cases} \tag{4.7}$$

4.6/ Implementation Details

Computations have been performed on the supercomputer facilities in "Mésocentre de Calcul de Franche-Comté". The experiments of multi-label ocular abnormalities detection are implemented on the PyTorch platform with python programming language in Ubuntu-

based server version 18.04.5 LTS with 252 GB RAM and GPU Tesla V100 DGXS 32 GB. The experiments are conducted in the RFMiD dataset.

For the first proposed method, the optimizer is stochastic gradient descent (SGD) with 0.9 momentum. It is trained with 100 epochs with an early stopper, learning rate 1e-3, and weight decay 1e-4. The hyper-parameters λ, and β are selected heuristically and explained in Section 4.7.

The second proposed method is optimized with AdamW optimizer. It is trained in 80 epochs with an early stopper. The learning rate is 1e-4 and 0.01 weight decay. The dimension of the FFN block is 8192.

4.7/ Experimental Results

All experiments in the first proposed method and second proposed method are conducted on the RFMiD dataset. Training and validation are conducted in the Training set with 3-fold cross-validation, and the model is tested in the Evaluation set by averaging the performances from all folds models. The model has also been tested in the Test set only for the second proposed method. The performance of proposed methods for multi-label ocular abnormalities detection is evaluated both globally (Eq. 4.8) and also per label (Eq. 4.9). Eq. 4.8 calculates the final score globally by considering each element of the label indicator matrix as a label.

$$Final\ Score = \frac{AUC_{global} + mAP_{global}}{2} \tag{4.8}$$

While Eq. 4.9 calculates the average of the final score for each label. This does not take label imbalance into account. C indicates the number of labels.

$$Final\ Score = \frac{1}{C}\sum_{i=1}^{C}\frac{(AUC_i + mAP_i)}{2} \tag{4.9}$$

The final score is calculated based on the values of the area under the Receiver Operating Characteristic (ROC) curve (AUC) and the mean of Average Precision (mAP), which are explained in 2.3.

4.7.1/ CNN-Based Semantic Dictionary Learning

Several experiments have been deployed in this first proposed method. Since the first proposed method is word embedding-agnostic, the experiments are implemented with Word2Vec [Mikolov et al., 2013] and FastText [Mikolov et al., 2018] pre-trained word em-

bedding networks. To analyze the impact of word embedding injection towards the model with image features, the experiments of CNN-based semantic dictionary learning are categorized as described in Table 4.3.

Table 4.3: Experiments Details

ID	Word Embedding	Upsampling	Stages
Experiment-1	Word2Vec	Distributed	no
Experiment-2	Word2Vec	Global	yes
Experiment-3	Word2Vec	-	yes
Experiment-4	Word2Vec	-	no
Experiment-5	FastText	-	yes
Experiment-6	FastText	-	no
Experiment-7	-	-	-

Experiment-1 is carried out with the distributed upsampling to reduce the data imbalance. In Experiment-2, the model is built with global upsampling to increase the number of ocular abnormality images globally and it is also built in two stages to have less data imbalance by separating the large positive (LP) and small positive (SP) data distributions. Experiment-3 and Experiment-4 are deployed to have a more realistic scenario by not applying the data upsampling. Instead, the real-time image augmentation is applied to the training data only. The difference between Experiment-3 and Experiment-4 is the training strategy. Similar to Experiment-2, the model training in Experiment-3 is done in two stages, for LP ocular abnormalities and SP ocular abnormalities. As explained in Section 4.5.2.1, the testing schema is also applied in stages. The first stage predicts the probability of LP and OTHER1 ocular abnormalities, while the second stage predicts SP and OTHER ocular abnormalities from the output of OTHER1. Experiment-4 is carried out without any upsampling or stage scenario to deal with data imbalance. The model predicts directly all ocular abnormalities and OTHER.

The experiments with FastText as word embedding are carried out without upsampling. The training and testing in Experiment-5 are done in two stages with the same scenario as Experiment-3, while Experiment-6 is deployed with no-stage which has a similar scenario as Experiment-4. Different from other experiments, Experiment-7 is executed to be able to compare the impact of two input modalities in multi-ocular abnormalities detection. Experiment-7 has a similar architecture to other experiments but without label features and a semantic dictionary.

4.7.1.1/ Comparison between Experiments

Word2Vec produces zero vectors for OOV words. From 27 labels in the RFMiD dataset, three ocular abnormalities belong to OOV words. Differently, FastText gives no vector value for OOV words. With FastText as word embedding, there is only one word that

belongs to OOV words.

The training in Experiment-1 and Experiment-2 have no issue with overfitting in the training phase since the data splitting is done after upsampling the original data. The model in Experiment-1 is trained with high metrics for both LP and SP ocular abnormalities. However, the testing results are low for LP ocular abnormalities and it is poor for SP ocular abnormalities. It shows that the model is overfitted to the training data. Though true prediction for SP ocular abnormalities is necessary, a high false prediction in LP ocular abnormalities significantly affects the overall prediction performance metrics. Similar case with Experiment-1, the model in Experiment-2 is trained well with high metrics in LP and SP ocular abnormalities. Dividing the model to train a different group of labels in each stage gives higher performance for LP ocular abnormalities, but lower performance for SP labels. However, the performance for LP ocular abnormalities in Experiment-2 is higher than that of Experiment-1.

Table 4.4: Testing results in Evaluation Set for experiments without sampling. The mAP, AUC, and Final scores are calculated globally. Bold values indicate the highest score.

Experiment	Word Embedding	Stage	mAP	AUC	Final Score
Experiment-3	Word2Vec	yes	0.144	0.767	0.456
Experiment-4	Word2Vec	no	0.232	0.829	0.530
Experiment-5	FastText	yes	0.331	0.845	0.588
Experiment-6	FastText	no	**0.397**	0.881	**0.639**
Experiment-7	-	no	0.338	**0.889**	0.614

The other performances of the CNN-based semantic dictionary learning with Word2Vec and FastText experiments without data sampling are described in Table 4.4. Among experiments that are listed in Table 4.4, Experiment-6 has the highest score for final score and mAP values. The mAP and final score of Experiment-6 surpass Experiment-7, the baseline model without label features, with a slight difference in AUC values. The interference of word embedding in the detection architecture can increase the detection performance but also can decrease the detection performance. More OOV words found in the word embedding with non-zero or non-value vectors affect the performance of the multi-label detection. Between stages and non-stage scenarios, experiments that predict the labels without dividing them into multiple stages also have higher performance than the experiments with stages.

Table 4.5: Testing results in Evaluation Set for experiments with stages. The mAP, AUC, and Final scores are calculated globally. Bold values indicate the prediction for complete labels.

ID	Experiment	Word Embedding	Stage	mAP	AUC	Final Score
ID-1	Experiment-3	Word2Vec	Stage-1	0.264	0.800	0.532
ID-2	Experiment-3	Word2Vec	Stage-2	0.243	0.731	0.487
ID-3	**Experiment-3**	**Word2Vec**	**Overall**	**0.144**	**0.767**	**0.456**
ID-4	Experiment-5	FastText	Stage-1	0.349	0.824	0.587
ID-5	Experiment-5	FastText	Stage-2	0.242	0.756	0.499
ID-6	**Experiment-5**	**FastText**	**Overall**	**0.331**	**0.845**	**0.588**

Table 4.5 shows the performance details of the experiments with stages. ID-1 of Table 4.5 is the performance of the Experiment-3 which is tested in Stage-1 to predict LP labels from all input images, ID-2 is the performance of the Experiment-3 which is tested in Stage-2 to predict SP labels and OTHER label from all input images, and ID-3 is the overall performance of the Experiment-3 that is tested to predict LP labels from the model of Stage-1, and predict the OTHER1 labels into the rest of labels (SP labels and OTHER label) from the model of Stage-2. ID-4, ID-5, and ID-6 are similar to ID-1, ID-2, and ID-3 respectively but with different word embedding. Though the performance of ID-1 and ID-2 is quite the same, the overall prediction in ID-3 has a lower performance compared to ID-1 and ID-2. The decrement in the performance of the model that is tested with overall labels is due to the less accurate prediction for SP labels and OTHER label from the Stage-2 model. In contrast, the performance of ID-6 is higher than ID-5 but lower than ID-4 even though it is not much different in ID-5 and ID-2 performance. It is because of the total images that are predicted to have OTHER1 ocular abnormality. The input images for Stage-2 in overall prediction are the images that are predicted to have OTHER1 ocular abnormalities from Stage-1. The more accurate the model prediction of Stage-2 can boost the performance of overall prediction. However, the accuracy of OTHER1 label prediction from Stage-1 is also an important factor since Stage-2 predicts only the images that are predicted to have OTHER1.

Table 4.6 describes the comparison of the performances between Experiment-6 with the other algorithms that are reported in their publications such as KAMATALAB, IGSTfencing, Schulich Applied Computing, BNUAA, WWW, Nekaravuru, Chizu & Arai & Okada, and MISIT that are reported in RFMiD challenge [Pachade et al., 2023]. KAMATALAB [Pachade et al., 2023] has the highest performance final score for multi-label detection. They proposed a detection method that consists of three different paths. Every path is built with a pre-trained CNN and different input image sizes. One model path is trained with 3 sub-label groups (stage) while the other two model paths are trained with complete labels. Post-processing is applied to the model; average blending for major classes only, providing meta-information, replacement with min-max values, and rank voting.

Table 4.6: Comparison of Testing Results in Evaluation Set. Bold values indicate the highest score among all algorithms. Sm refers to Sampling, BBone refers to Backbone, WEmbd refers to Word Embedding, and FScore refers to Final Score.

Algorithm	Sm	Stage	BBone	WEmbd	mAP	AUC	FScore
KAMATALAB [Pachade et al., 2023]	no	3	3	-	-	-	**0.802**
IGSTfencing [Pachade et al., 2023]	no	no	1	-	-	-	0.791
Schulich Applied Computing [Pachade et al., 2023]	no	no	5	-	-	-	0.786
BNUAA [Pachade et al., 2023]	no	no	2	-	-	-	0.782
Multi-Expert [Lyu et al., 2022]	no	no	multiple	-	-	-	0.778
WWW [Pachade et al., 2023]	no	no	6	-	-	-	0.758
Nekaravuru [Pachade et al., 2023]	no	2	3	-	-	-	0.757
Chizu & Arai & Okada [Pachade et al., 2023]	no	no	multiple	-	-	-	0.738
MISIT [Pachade et al., 2023]	yes	no	4	-	-	-	0.703
Experiment-6 (Global calculation)	no	no	1	FastText	**0.397**	0.881	0.639

Nekaravuru [Pachade et al., 2023] applied the multi-stage approach that was trained together with different weight loss in each stage. They also stack three pre-trained networks as the backbone to have more generalized model. However, these methods require larger resources because of the computation and number of models in both training and testing. IGSTfencing [Pachade et al., 2023] is the only recent algorithm that proposed the detection method without a model ensemble. The detection model is built with semi-supervised siamese GCN. The model is trained with label and unlabeled data and it also learns the co-occurrence dependencies between labels by constructing two GCNs. This method is the second-highest performance listed in Table 4.6. Nevertheless, this detection model needs more data in the training phase. It was trained with RFMiD and ODIR datasets.

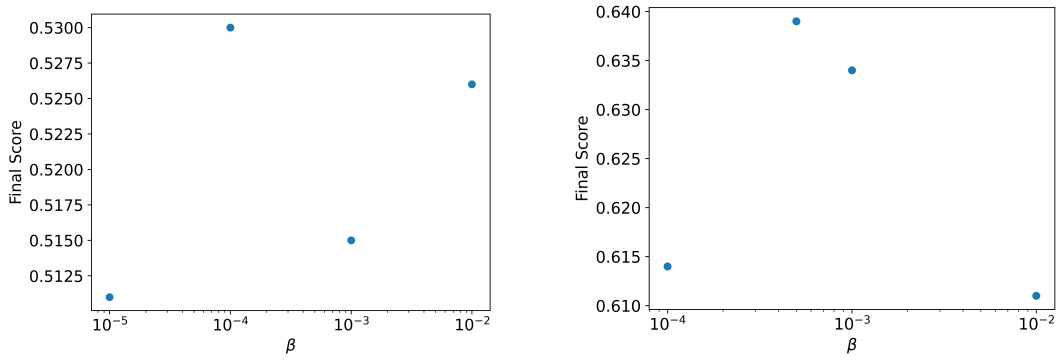
This first proposed method is backbone-agnostic and word embedding-agnostic. Selecting another word embedding technique for linguistic features that have treated the OOV words with a non-zero vector can increase the performance of multi-label detection with semantic dictionary learning.

4.7.1.2/ Hyper-parameter Analysis

There are two special hyper-parameters for CNN-based semantic dictionary learning: Hyper-parameter λ that is explained in Eq. 4.3 as the regularizer, and hyper-parameter β that is explained in Eq. 4.4 as the balance between L_{ml} and L_{dic}. The value of λ and β are selected heuristically. For Experiment-4, the hyper-parameter λ is 10 and β is $1.00\ e^{-04}$. The hyper-parameter λ for Experiment-6 is 1 and β is $5.00\ e^{-04}$. During training, the value of λ affects how the model is learning. A lower value of λ makes the model learn slowly and converge in the non-optimum loss, and a higher value of λ makes the model to be sensitive only in label 0.

Figure 4.8 shows the different final scores corresponding to different values of β. The

other hyper-parameters for Word2Vec are set to be the same as Experiment-4, and the other hyper-parameters for FastText are the same as Experiment-6. As seen in Figure 4.8, the hyper-parameter β is quite sensitive toward the value of the final score. Though a higher value of β can give also a higher value of final score (see Figure 4.8a), it puts the stability of the dictionary loss (L_{dic}) in the risk. In this research, the lowest range for the selection of β value is $1.00\ e^{-05}$ to see the impact of a CNN-based semantic dictionary with visual consistency towards the detection model.



(a) Word2Vec **(b)** FastText

Figure 4.8: Hyper-parameter beta selection for Experiment-4 and Experiment-6.

4.7.2/ Transformer-Based Semantic Dictionary Learning

Though the transformer-based semantic dictionary is also backbone-agnostic and word embedding-agnostic, the experiments that are deployed in this research have an identical backbone and variety of word embeddings. Four experiments are implemented in this proposed method:

1. **Experiment-1** is executed with FastText for the label embedding as the highest performance in CNN-based semantic dictionary learning.
2. **Experiment-2** is executed with BERT [Devlin et al., 2018].
3. **Experiment-3** is executed with Clinical BERT [Alsentzer et al., 2019] considering the source of the training data is the limited clinical reports.
4. **Experiment-4** is executed with spatial learnable word embedding [Liu et al., 2021a]. This model is identical to the model with the second proposed method but without any interference from any word embedding technique. The learnable word embedding that is generated from this model is the other representation of the spatial features. This experiment is the baseline for the second proposed method.

The performance of the transformer-based semantic dictionary learning experiments are described in Table 4.7. The experiments are tested in Evaluation Set and Test Set on the RFMiD dataset. ID-1, ID-2, ID-3, and ID-4 show the performance of Experiment-1, Experiment-2, Experiment-3, and Experiment-4 which are tested on the Evaluation Set while ID-5, ID-6, ID-7, and ID-8 are the experiments that are tested on the Test Set respectively.

Table 4.7: Testing Intra-performances in Evaluation Set and Test Set for multi-label detection model. Evaluation metrics are calculated per label. Bold values indicate the highest score in each Data Testing.

ID	Algorithm	Data Testing	Word Embedding	mAP	AUC	Final Score
ID-1	Experiment-1	Evaluation Set	FastText	0.432	0.861	0.647
ID-2	Experiment-2	Evaluation Set	BERT	0.449	0.856	0.652
ID-3	Experiment-3	Evaluation Set	Clinical BERT	**0.455**	**0.874**	**0.665**
ID-4	Experiment-4	Evaluation Set	Learnable	0.434	0.863	0.648
ID-5	Experiment-1	Test Set	FastText	0.431	0.858	0.645
ID-6	Experiment-2	Test Set	BERT	0.430	0.835	0.632
ID-7	Experiment-3	Test Set	Clinical BERT	0.442	0.842	0.642
ID-8	Experiment-4	Test Set	Learnable	**0.443**	**0.862**	**0.653**

In the testing phase, for Evaluation Set, the model that is trained with Clinical BERT (ID-3) has the highest scores in mAP, AUC, and final score followed by the model trained with BERT (ID-2), while the baseline model (ID-8) has the highest score on the Test Set with a slight different performance with the model trained with FastText (ID-5) followed by the model trained with Clinical BERT (ID-7). The result in Evaluation Set shows that the model that is trained with the word embedding technique can increase the model performance compared with the baseline model. On the other hand, the result in Test Set shows that the word embedding that is learned from the spatial features (baseline model) is enough to achieve good performance. Due to these different inferences, the evaluation metrics that are calculated globally are also provided in Table 4.8.

Table 4.8: Testing Intra-performances in Evaluation Set and Test Set for multi-label detection model. Evaluation metrics are calculated globally. Bold values indicate the highest score in each Data Testing.

ID	Algorithm	Data Testing	Word Embedding	mAP	AUC	Final Score
ID-9	Experiment-1	Evaluation Set	FastText	0.636	**0.929**	0.783
ID-10	Experiment-2	Evaluation Set	BERT	**0.647**	0.927	**0.787**
ID-11	Experiment-3	Evaluation Set	Clinical BERT	0.642	0.920	0.781
ID-12	Experiment-4	Evaluation Set	Learnable	0.638	**0.929**	0.783
ID-13	Experiment-1	Test Set	FastText	0.661	**0.935**	0.798
ID-14	Experiment-2	Test Set	BERT	**0.679**	0.929	**0.804**
ID-15	Experiment-3	Test Set	Clinical BERT	0.671	0.925	0.798
ID-16	Experiment-4	Test Set	Learnable	0.668	0.934	0.801

Table 4.8 shows that the experiments with word embedding BERT (ID-10 and ID-14) achieve the highest final score with also the highest mAP in both Evaluation Set and Test Set. Overall, the model with the word embedding from the linguistic features has better performance than the baseline model in most cases.

Table 4.9: Comparison of Testing Results in Evaluation Set. Bold values indicate the highest score among all algorithms. Sm refers to Sampling, BB refers to Backbone, WEmb refers to Word Embedding, and FScore refers to Final Score.

Algorithm	Sm	Stage	BB	WEmb	mAP	AUC	FScore
KAMATALAB [Pachade et al., 2023]	no	3	3	-	-	-	**0.802**
IGSTfencing [Pachade et al., 2023]	no	no	1	-	-	-	0.791
Schulich Applied Computing [Pachade et al., 2023]	no	no	5	-	-	-	0.786
BNUAA [Pachade et al., 2023]	no	no	2	-	-	-	0.782
WWW [Pachade et al., 2023]	no	no	6	-	-	-	0.758
Nekaravuru [Pachade et al., 2023]	no	2	3	-	-	-	0.757
Chizu & Arai & Okada [Pachade et al., 2023]	no	no	multiple	-	-	-	0.738
MISIT [Pachade et al., 2023]	yes	no	4	-	-	-	0.703
2^{nd} Proposed Method (Experiment-3)	no	no	1	Clinical BERT	**0.455**	**0.874**	0.665

Table 4.9 describes the proposed methods compared to other recent studies in multi-label detection for the RFMiD dataset. All the algorithms reported in Table 4.9 are tested in Evaluation Set. The evaluation metric of all algorithms is calculated by averaging the final score per label. The final score of 2^{nd} Proposed Method (Experiment-3) is 0.038 difference with MISIT [Pachade et al., 2023]. Table 4.10 describes the proposed methods compared to other recent studies in multi-label detection for the RFMiD dataset in Test Set: the highest performances in the RFMiD dataset challenge [Pachade et al., 2023] and Multi-Expert [Lyu et al., 2022]. The final score of 2^{nd} Proposed Method (Experiment-1) is 0.044 difference with MISIT [Pachade et al., 2023].

Table 4.10: Comparison of Testing Results in Test Set. Bold values indicate the highest score among all algorithms. Sm refers to Sampling, BB refers to Backbone, WEmb refers to Word Embedding, and FScore refers to Final Score.

Algorithm	Sm	Stage	BB	WEmb	mAP	AUC	FScore
KAMATALAB [Pachade et al., 2023]	no	3	3	-	-	-	0.782
Schulich Applied Computing [Pachade et al., 2023]	no	no	5	-	-	-	0.778
Multi-Expert [Lyu et al., 2022]	no	no	multiple	-	-	-	0.778
BNUAA [Pachade et al., 2023]	no	no	2	-	-	-	0.758
Nekaravuru [Pachade et al., 2023]	no	2	3	-	-	-	0.749
WWW [Pachade et al., 2023]	no	no	6	-	-	-	0.738
IGSTfencing [Pachade et al., 2023]	no	no	1	-	-	-	0.710
MISIT [Pachade et al., 2023]	yes	no	4	-	-	-	0.689
2^{nd} Proposed Method (Experiment-1)	no	no	1	FastText	**0.431**	**0.862**	0.645

The results of the 2^{nd} Proposed Method (Experiment-2) are also compared with Sun et al. [Sun et al., 2022a] who focus to predict 7 frequent ocular abnormalities in the RFMiD dataset. Table 4.11 presents the AUC-ROC metric of these 7 labels to be compared.

The model of Sun et al. [Sun et al., 2022a] and 2^{nd} Proposed Method (Experiment-2) are tested in Test Set. As shown in Table 4.11, the prediction of the frequent labels from the 2^{nd} Proposed Method (Experiment-2) has competitive results compared to the multi-frequent label detection model proposed by Sun et al. [Sun et al., 2022a].

Table 4.11: The comparison of AUC-ROC metric per labels in Test Set. Bold value indicates the highest value.

Label	Sun et al. [Sun et al., 2022a]	2^{nd} Proposed Method (Experiment-2)
DR	0.97	**0.98**
ARMD	**0.98**	0.95
MH	0.94	**0.96**
DN	0.84	**0.89**
MYA	**0.98**	0.95
TSLN	**0.95**	0.94
ODC	**0.9**	0.89

A complete label-level performance of the 2^{nd} Proposed Method (Experiment-2) is presented in Table 4.12. The Final score and mAP metrics are also provided in Table 4.12 to have a more focused evaluation, especially for the rare ocular abnormalities. It is quite difficult to evaluate the model performance from only AUC-ROC for rare ocular abnormalities. A contrast value between AUC-ROC and mAP metrics in rare ocular abnormalities is caused by a small number of total images that are assigned to be positive.

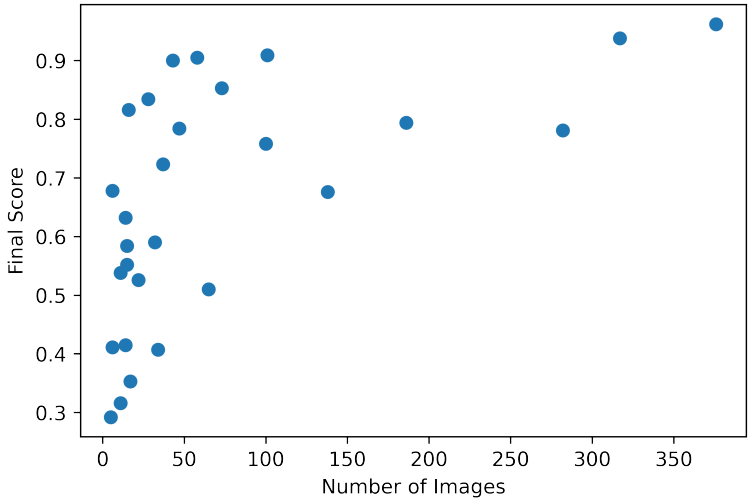


Figure 4.9: Final scores of each label are plotted against the number of training images.

Figure 4.9 describes the plot of the number of training images against the final score for all the labels. It shows that the labels that have large training images (minimum 100 images) tend to have a final score above 0.6. However, the fluctuated performances are

captured between 0 and 100 training images. To analyze the behavior of the model's performance, the labels are divided into two groups (LP and SP) based on the number of training images with a threshold of 30. LP (Large Positives) refers to the labels that have more than 30 training images and SP (Small Positives) refers to the labels that have less than 30 training images.

Table 4.12: Evaluation metrics per labels in Test Set. Bold values indicate the exception labels.

Label	Total Images in Training Set	AUC	mAP	Final score
DR	376	0.982	0.941	0.962
MH	317	0.965	0.911	0.938
ODC	282	0.886	0.676	0.781
TSLN	186	0.940	0.649	0.794
DN	138	0.891	0.461	0.676
MYA	101	0.954	0.864	0.909
ARMD	100	0.947	0.570	0.758
BRVO	73	0.964	0.742	0.853
ODP	**65**	**0.830**	**0.190**	**0.510**
ODE	58	0.955	0.856	0.905
LS	47	0.931	0.636	0.784
RS	43	0.988	0.811	0.900
CSR	37	0.901	0.544	0.723
OTHER	**34**	**0.714**	**0.100**	**0.407**
CRS	**32**	**0.900**	**0.280**	**0.590**
CRVO	**28**	**0.936**	**0.731**	**0.834**
RPEC	22	0.948	0.104	0.526
AION	17	0.677	0.029	0.353
AH	16	0.988	0.644	0.816
EDN	15	0.836	0.332	0.584
MS	15	0.894	0.209	0.552
ERM	14	0.616	0.214	0.415
RT	14	0.955	0.308	0.632
PT	11	0.617	0.015	0.316
MHL	11	0.727	0.349	0.538
TV	6	0.851	0.505	0.678
RP	6	0.810	0.012	0.411
ST	5	0.572	0.013	0.292

The performance of LP labels is shown in Figure 4.10. It shows that most of LP labels achieve final scores greater than 0.6 except for three labels: ODP, CRS, and OTHER.

On the other hand, the performance of SP labels is shown in Figure 4.11. The performance of SP labels is 0.534 final score on average with 0.169 standard deviation. However, the plot of the performance against the final score in SP labels has no pattern. There are some labels that can achieve high final scores up to 0.8, but there are also some other labels that have very low final scores.

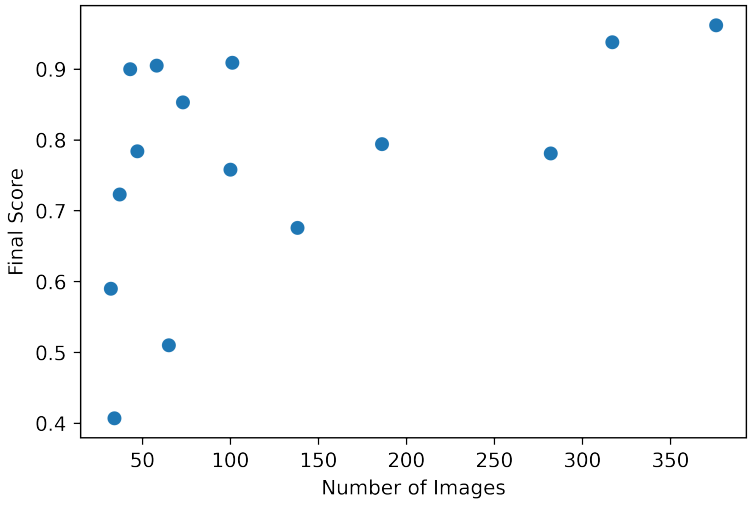


Figure 4.10: Final scores of each LP label plotted against the number of training images.

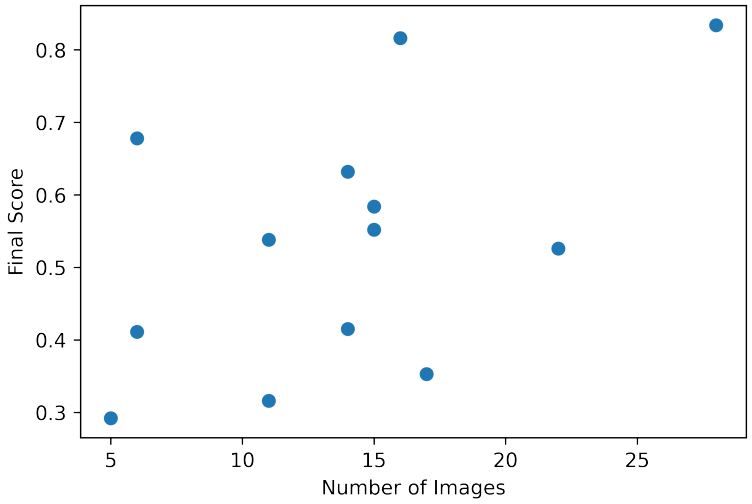


Figure 4.11: Final scores of each SP label plotted against the number of training images.

Figure 4.12 shows the plot between the final score and the number of training images for SP labels with adding the co-occurrence dependency to be the third variable. The 0.0 value in the ratio of co-occurrence indicates that the label has no co-occurrence with any other labels, whereas the value 1.0 indicates that the label is highly co-occurrent with other labels. Based on the data plotted in Figure 4.12, it shows that most of the SP labels that have a ratio of co-occurrence of more than 0.5 can achieve a final score greater than 0.5 while the SP labels that have low co-occurrence tend to have a low final score (less than 0.5), except for CRVO label. Co-occurrence dependency gives a positive impact on

SP labels since most of the co-occurred labels are LP labels.

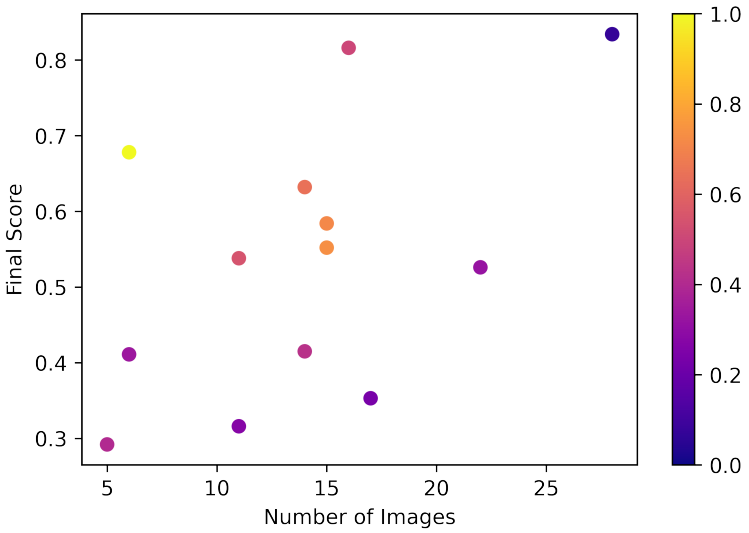
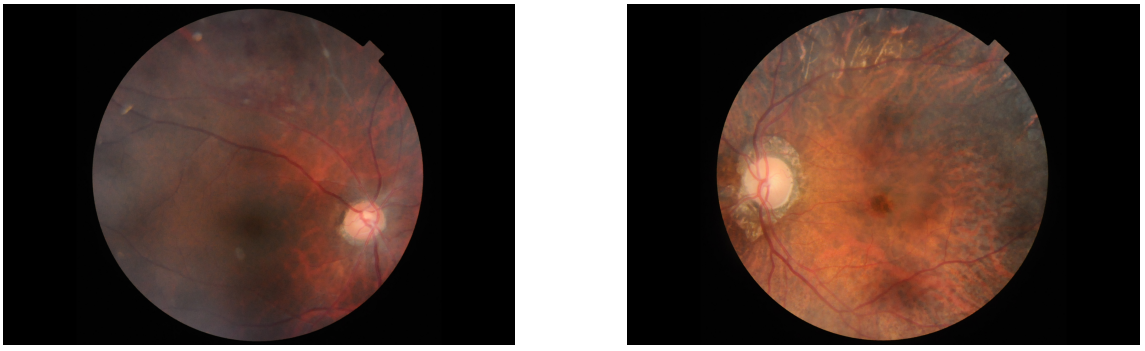


Figure 4.12: Final scores of each SP label plotted against number of training images. The color bar denotes the ratio of co-occurrence of the labels.

However, the type of visible sign in the ocular abnormality also affects the prediction result. The model has difficulties in differentiating labels that have similar visible signs. In some test images, the model miss-predicts ODP (Optic Disc Pallor) to ODC (Optic Disc Cupping) and ODC to ODP. Figure 4.13 shows the false predictions of ODP, 4.13a is the fundus image that is predicted to have ODC when the ground truth is ODP, and 4.13b is the fundus image that is predicted to have ODP when the ground truth is ODC.

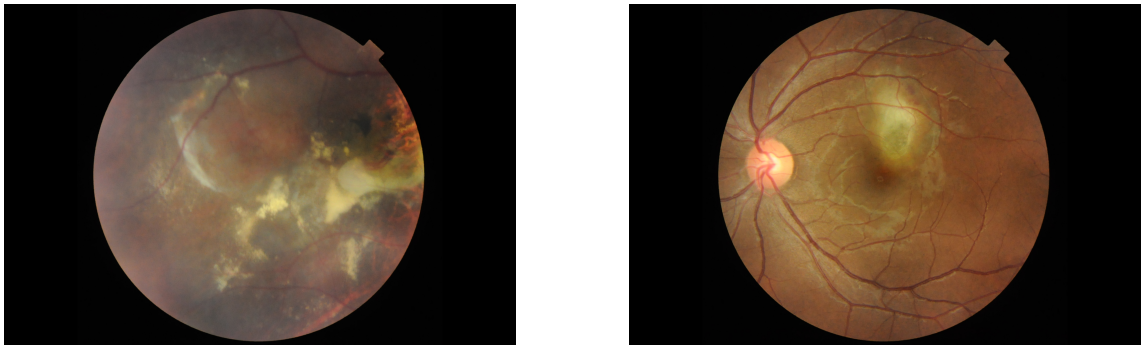


(a) False Positive **(b)** False Negative

Figure 4.13: ODP false prediction.

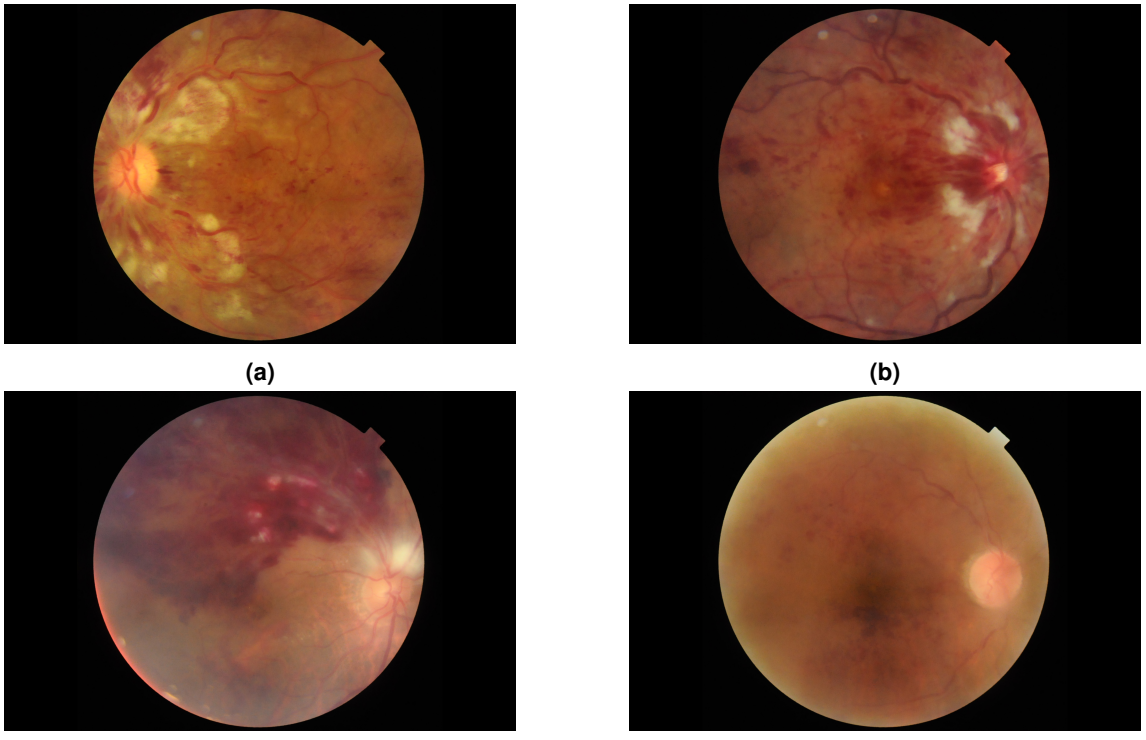
The same issue also occurs with CRS (Chorioretinitis) label. The model has hard difficulty differentiating CRS and ARMD (Age-Related Macular Degeneration). Figure 4.14 shows the false predictions of CRS, 4.14a is the fundus image that is predicted to have

CRS when the ground truth is ARMD, and 4.14b is the fundus image that is predicted to have ARMD when the ground truth is CRS. This issue is the causing factor of 75% false prediction.



(a) False Positive **(b)** False Negative

Figure 4.14: CRS false prediction.



(a) **(b)**

(c) **(d)**

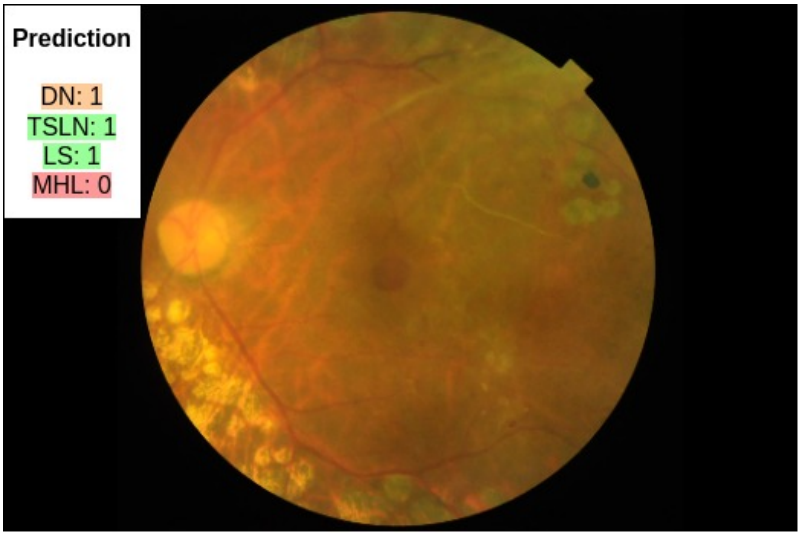
Figure 4.15: CRVO images predictions; (a) and (b) are True Predictions, (c) is False Positive, and (d) is False Negative prediction.

The type of visible sign of an ocular abnormality can also affect the performance increment. As seen in Figure 4.15, the visible sign of CRVO (Central Retinal Vein Occlusion) is clearly visible and it spreads all over the fundus. It causes the model to predict the ocular abnormality correctly. Yet, the false predictions in CRVO are also caused by the model that is still weak to differentiate CRVO and BRVO. In addition, the low performance

of OTHER label shows that the model is having difficulty differentiating the visible signs that belong to multiple ocular abnormalities.

The final prediction results are illustrated in Figure 4.16. Label 1 indicates the existence of an ocular abnormality and 0 is the opposite. The detection model can predict correctly all ocular abnormalities in the CFP in Figure 4.16a. Figure 4.16b illustrates the case that the detection model fails to predict some ocular abnormalities.

(a) Image-1



(b) Image-2

Figure 4.16: The illustration of final prediction results from CFPs in Test Set. Green indicates TP, orange indicates FP, and red indicates FN.

4.7.3/ Discussion Across Methodologies

This section explains the comparison between CNN-based and Transformer-based semantic dictionary learning results. The evaluation metrics are calculated globally to

have a fair comparison. The comparison is described in Table 4.13. It compares the performances of the multi-label detection with a CNN model without word embedding (CNN baseline method), the best-performing model from the 1^{st} proposed method, and the 2^{nd} proposed method (transformer-based) that is trained using the same backbone (EfficienNet-B4) and word embedding (FastText).

Table 4.13: Comparison of performance of CNN-based and Transformer-based semantic dictionary learning in Evaluation Set for multi-label detection model. Evaluation metrics are calculated globally. Bold values indicate the highest score.

Algorithm	Backbone	Word Embedding	mAP	AUC	Final Score
CNN baseline method	EfficientNet-B4	-	0.338	0.889	0.614
1^{st} Proposed method	EfficientNet-B4	FastText	0.397	0.881	0.639
2^{nd} Proposed method	EfficientNet-B4	FastText	**0.575**	**0.924**	**0.750**

As shown in Table 4.13, injecting the word embedding as an additional input modality along with a single CNN can boost the model performance. Both 1^{st} proposed method and 2^{nd} proposed method has higher performances compared to the CNN baseline method with 2^{nd} Proposed method as the highest performance. It also shows that treating the linguistic features as a main part of the detection method (2^{nd} proposed method) can give a higher impact than treating it as a weak factor (1^{st} proposed method).

4.8/ CONCLUSION

In conclusion, providing linguistic modality along with spatial modality with semantic dictionary learning increases the performance of multi-label detection from a single color fundus image for frequent and rare ocular abnormalities, compared to the model detection with only spatial modality. In this chapter, we proposed two different approaches of semantic dictionary learning: CNN-based (1^{st} Proposed Method) and transformer-based (2^{nd} Proposed Method). In CNN-based semantic dictionary learning method, the spatial features are extracted from the pre-trained CNN and linguistic features are extracted from the word embedding network. The correlation between these features is learned by a semantic dictionary that is generated based on the visual features representation module, whereas in Transformer-based semantic dictionary learning, the correlation between spatial and linguistic features is learned based on the attention regions with transformer decoders.

The CNN-based semantic dictionary learning can adapt the detection model with OOV words that come from the word embedding technique. With a simple fine-tuned CNN as backbone, the model with multi-modalities (spatial and linguistic) shows an increment in performance compared with the base-model with only spatial modality. However, more

OOV words that are produced from word embedding give a higher risk of decrement in detection performance.

On the other hand, the performance of transformer-based semantic dictionary learning surpasses the performance of 1^{st} Proposed Method. This 2^{nd} Proposed Method can map the correlation between spatial and linguistic modality explicitly. Both 1^{st} Proposed Method and 2^{nd} Proposed Method are backbone-agnostic and word embedding-agnostic. Changing the backbone and the word embedding technique that has better performance may boost these multi-label detections. Overall, using a semantic dictionary as a features correlation method, it is important to consider linguistic features as one of the strong input factors in multi-label detection for frequent and rare ocular abnormalities.

III

CONCLUSION

5

General Conclusion

5.1/ Summary of The PhD Thesis

This thesis presents CAD models involved in ocular abnormalities detection from a single color fundus photography. The first CAD model is implemented to detect microaneurysms with high sensitivity and a low number of FPI. The main challenge in MAs detection is the limited number of data with MA that causes severe data imbalance. We proposed a MAs detection that consists of three main processes: pre-processing, MAs candidate extraction, and MAs classification. The green channel of a CFP is enhanced by CLAHE followed by r-polynomial transformation. To reduce the need for a large number of training data, the MAs candidates are extracted in an unsupervised approach. The data patches are generated from MAs candidates to be fed into the MAs classifier. The image background of the patches is suppressed to reduce its complexity. In MAs classification, few fine-tuned CNN networks are applied to distinguish MA and non-MA with limited training data. The MAs classifiers are built in two methods: ensemble and cascade learning. MAs classifier with ensemble learning is built to analyze the best input channels. The experiments are conducted on the IDRiD dataset. The results show that the combination of the enhanced-green channel, their background suppression image, and the blue channel gives the highest performance for MAs detection. This combination is then used in the second MAs classifier that is built with cascade learning. The aim of the second MAs classifier is to reduce the number of FPI with high sensitivity. The experiments are conducted in E-Ophta and IDRiD datasets separately. MAs classifier with cascade learning that is trained and tested with the E-Ophta dataset achieves the highest sensitivity in 8 FPI compared to other existing methods in the literature. Also, the competitive performances of MAs classifiers trained and tested with different datasets show that the MAs classifier generalizes well.

The second CAD model proposed in this thesis is for the multi-label detection of 28 ocular abnormalities that consist of frequent and rare abnormalities. Rare ocular abnormalities are usually ignored because of a lack of features. We tackle the condition of rare labels

by adding the co-occurrence dependency factor to the model. Instead of using the hot-encoded label co-occurrence as the second input modality, the linguistic features of the labels are applied to represent the co-occurrence features. The model learns the relation between spatial features and linguistic features represented as a semantic dictionary. Two approaches for multi-label detection with deep learning are proposed in this thesis: CNN-based and Transformer-based semantic dictionary learning. CNN-based semantic dictionary learning focuses to learn the semantic dictionary with the visual representation constraint. The results show that adding the linguistic modality and generating the semantic dictionary can increase the model performance compared with the model that uses only spatial modality. Yet, the semantic dictionary is treated as a weak factor due to the visual representation constraint. Hence, the second proposed method in multi-label detection for ocular abnormalities is built based on transformers. In this approach, the semantic dictionary becomes a crucial part of the model. The semantic dictionary acts as the query while the spatial features are the key and value. The experiments are conducted on the RFMiD dataset. The results show that the proposed method can achieve higher performances compared with the CNN-based semantic dictionary learning method. The transformer-based approach is also able to detect some rare ocular abnormalities with a minimum 0.5 final score for the rare labels that have a high ratio of co-occurrence. Adding linguistic modality along with spatial modality in multi-label detection for frequent and rare ocular abnormalities can boost the detection performance.

5.2/ PERSPECTIVES

5.2.1/ MICROANEURYSMS DETECTION

By the completion of this thesis, we came across some aspects related to MAs detection that need deep analysis. The first aspect is the dataset. Though there are numerous annotated public datasets available, the quality of the MAs annotation in these datasets is still uncertain. It has been proven in the study by Krause et al. [Krause et al., 2018], which shows that there are some disagreements in annotation between the ground-truth dataset and the experts in the hospital, and the most common disagreement is for MAs annotation. High-quality annotation is an important factor in MAs detection since false predictions usually occur in weak-appearance objects such as blurry edges or less-contrast backgrounds.

The second aspect is the method to increase the performance of MAs detection. Other than cascade learning, the method proposed by Galdran et al. [Galdran et al., 2022] can be a potential method to refine the prediction of MAs detection. Cascade learning refines the predictions by training another network that focuses to learn the features of

FP predictions. However, the cascade learning proposed in this thesis is designed only to decrease the number of FPs without any specific refinement step to reduce the FN predictions. On the other hand, Galdran et al. proposed a simpler refinement approach for blood vessel segmentation that has two CNNs, one to segment the vessels from a CFP and the other to refine the segmentation results by providing pseudo-labels as additional input. The performance of their method surpassed the other existing methods that are trained with the combination of multiple datasets (cross-dataset). Considering the results of their method, it looks feasible for it to be adapted for MAs detection.

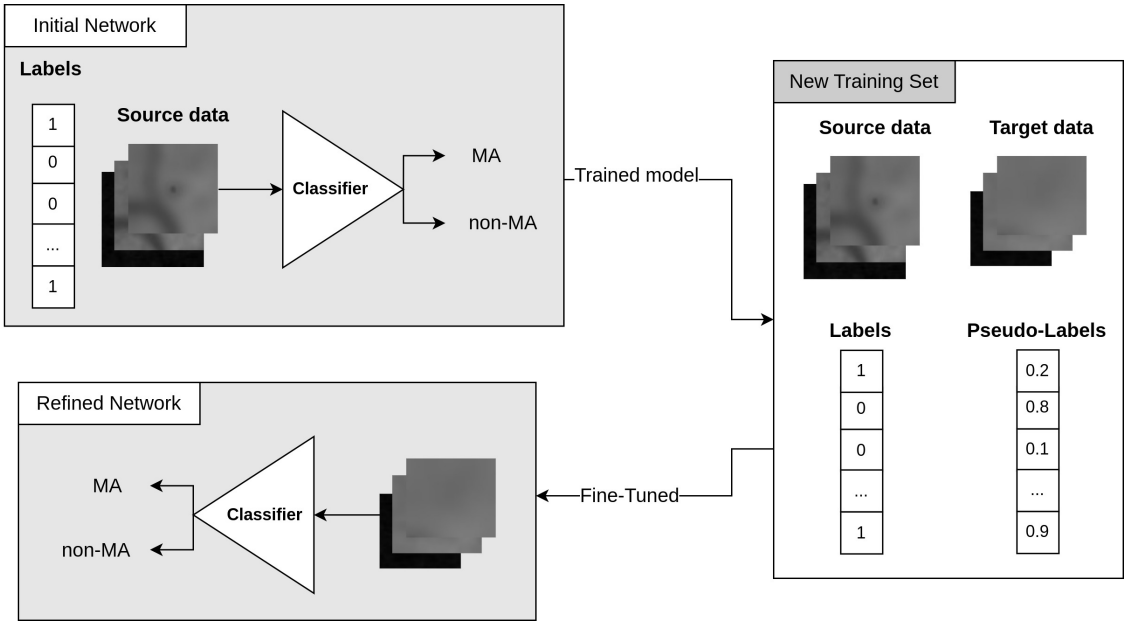


Figure 5.1: An illustration of the potential refinement method for MAs detection adapted from Galdran et al. [Galdran et al., 2022].

The refinement step proposed by Galdran et al. [Galdran et al., 2022] is illustrated in Figure 5.1. Firstly, the training data is split into source data and target data. The initial network is trained with the patches (from MAs candidates extraction) and the labels of the source data. After that, the target data is used to evaluate the performance of the initial network to get the probability of the labels which are called "pseudo-labels". The source data and its label along with target data and its pseudo-labels are used in training the refinement network. The final predictions are based on the results of the refinement network. This method is expected to significantly reduce false (both FP and FN) predictions.

Another interesting approach to increase the performance of MAs detection is Transformers. In the image classification task, the transformers achieve higher performance than CNN which can be of significant advantage in MAs detection tasks [Dosovitskiy et al., 2020]. As for the input image, giving a full fundus image that is divided into bigger patches can be a wise option instead of giving the small patches of MAs candidates extraction that consist of one MA per patch. A bigger patch can consist of one or

multiple MAs in a patch. It can also give more information about the contrast between objects and the background to understand the similarity between objects in the same background. To be suitable with this type of input image, DETR model [Carion et al., 2020] can be adapted to detect multiple MAs from an image.

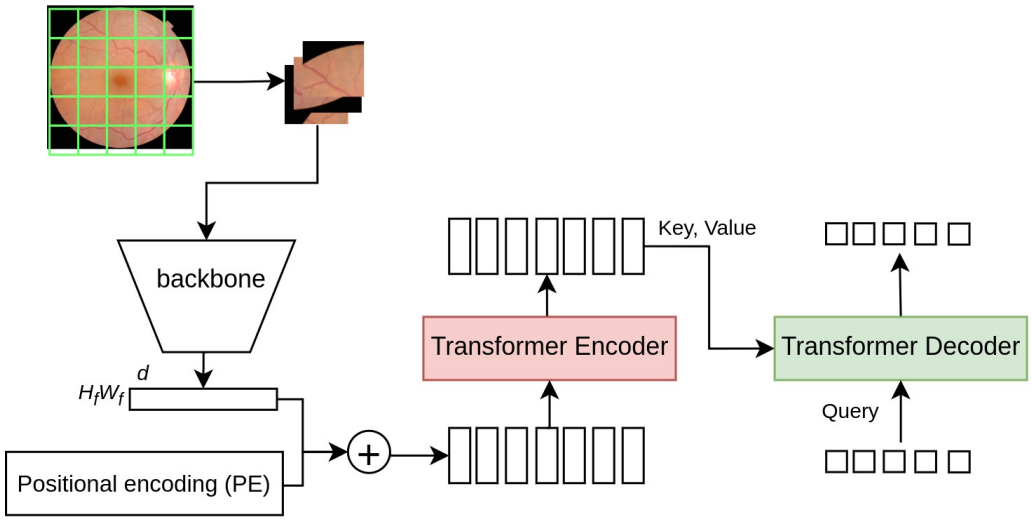


Figure 5.2: MAs detection with Transformer.

The illustration of MAs detection based on DETR is shown in Figure 5.2. The spatial features of the patches are extracted by a backbone network. The spatial features and positional encoding are given to the transformer encoder. The key and value from the transformer encoder is mapped by the object query in the transformer decoder to find the location of the MAs in the patch. The augmentation method proposed by Xia et al. [Xia et al., 2021] can also be adapted to this approach to increase the number of positive data. To adapt the DETR model with tiny objects, ViTDet [Li et al., 2022a] can be applied to give multi-scale features to the ViT backbone.

Overall, a high-performance MAs detection is not only useful to detect the MAs solely but also boost the performance of DR grading since the inefficient detection of MAs can lead to miss-grading of the DR between mild and moderate [Krause et al., 2018].

5.2.2/ Multi-label Detection for Ocular Abnormalities

In terms of model improvement for multi-label detection of ocular abnormalities proposed in this thesis, there are a few main factors that need to be analyzed in depth.

Linguistic Features. The rapid advancement in language modeling paves way for the improvement in the word embedding models representing the linguistic features in multi-label detection. GPT (Generative Pre-trained Transformer) [Radford et al., 2018] and CALM (Confident Adaptive Language Modelling) [Schuster et al., 2022] are language

models that can be applied to extract the linguistic features in multi-label detection because of their performances. GPT [Radford et al., 2018] is the backend of ChatGPT, the recent chatbot that was launched in November 2022. GPT has a variety of versions, the latest version is GPT-3 [Brown et al., 2020] and the next version (GPT-4) is going to be released soon. GPT is built based on general transformer architecture. Instead of fine-tuning, GPT-3 implements zero-shot, one-shot, and few-shot learning in their experiments. GPT-3 was trained in 470 times bigger in parameter size than BERT-Large. CALM [Schuster et al., 2022] is also a large language model that utilizes a transformer-based architecture. CALM is built with encoders that change the text input to the dense representation and decoders to generate the new text. The main advantage of CALM is the ability to predict the next word before all decoders are completed. CALM is capable to maintain high performance with high-quality output and also to increase the speed of text generation.

Although some publicly available language models are trained with medical documents such as ClinicalBERT [Alsentzer et al., 2019] and PubMedGPT [Bolton et al., 2022], retraining the network with eye-related documents as text database might be necessary considering the existence of the rare ocular abnormalities.

Spatial Features. In this thesis, the spatial features are extracted from a single fine-tuned CNN. As mentioned in Section 4, the proposed methods of the multi-label detection are backbone-agnostic so it can be changed with multi-CNNs or with a Transformer-based backbone such as Vision Transformers (ViT) [Dosovitskiy et al., 2020], Shifted Window transformers (Swin) [Liu et al., 2021b], and Multi-scale Vision Transformers (MViT) [Fan et al., 2021]. ViT is the standard transformer model used in image classification while Swin and MViT are the leading adaptations of ViT. Unlike ViT which is implemented with a fixed patch size, the Swin transformer uses shifted window concept similar to CNNs with varying patch sizes. The scale-up Swin transformer model with the model capacity that can receive bigger image resolution is also available in Swin V2 [Liu et al., 2022]. MViT is another variant of ViT that is built by connecting the seminal idea of multiscale feature hierarchies with transformer models. MViTv2 [Li et al., 2022b] is an upgraded version of MViT that incorporates decomposed relative positional embeddings and residual pooling connections. Both these transformers have shown good performance in image classification and object detection tasks. Unlike Swin and MViT which are built in hierarchical architecture, Li et al. [Li et al., 2022a] proposed ViTDet architecture with a single-scale backbone. ViTDet architecture focuses to reserve the backbone to be task-agnostic. Inspired by Feature Pyramid Networks (FPN) [Lin et al., 2017a], ViTDet builds a simple pyramid on the single-scale backbone. The network is trained by Masked Autoencoder (MAE) [He et al., 2022] with non-overlapping windows.

Sub-task Division. The discussion of the multi-label detection of ocular abnormalities

in Section 4.7.2 shows that false predictions are usually related to some ocular abnormalities that have symptoms with similar appearances. Therefore, it is also important to focus on differentiating these confusing abnormalities. The abnormalities with symptoms-similarities can be grouped as follows:

- Optic disc abnormality: ODC, ODE, and ODP
- Vessels abnormality: CRVO and BRVO
- Yellow lesion abnormality: Exudates and drusen
- Red lesion abnormality: Microaneurysm and hemorrhage.

Symptoms-based TSDL (S-TSDL). Although the results of transformer-based semantic dictionary learning (TSDL) are quite promising in multi-label detection, they can still be improved. Instead of a class-based semantic dictionary as proposed in this thesis, the same model can be applied to generate a symptom-based semantic dictionary.

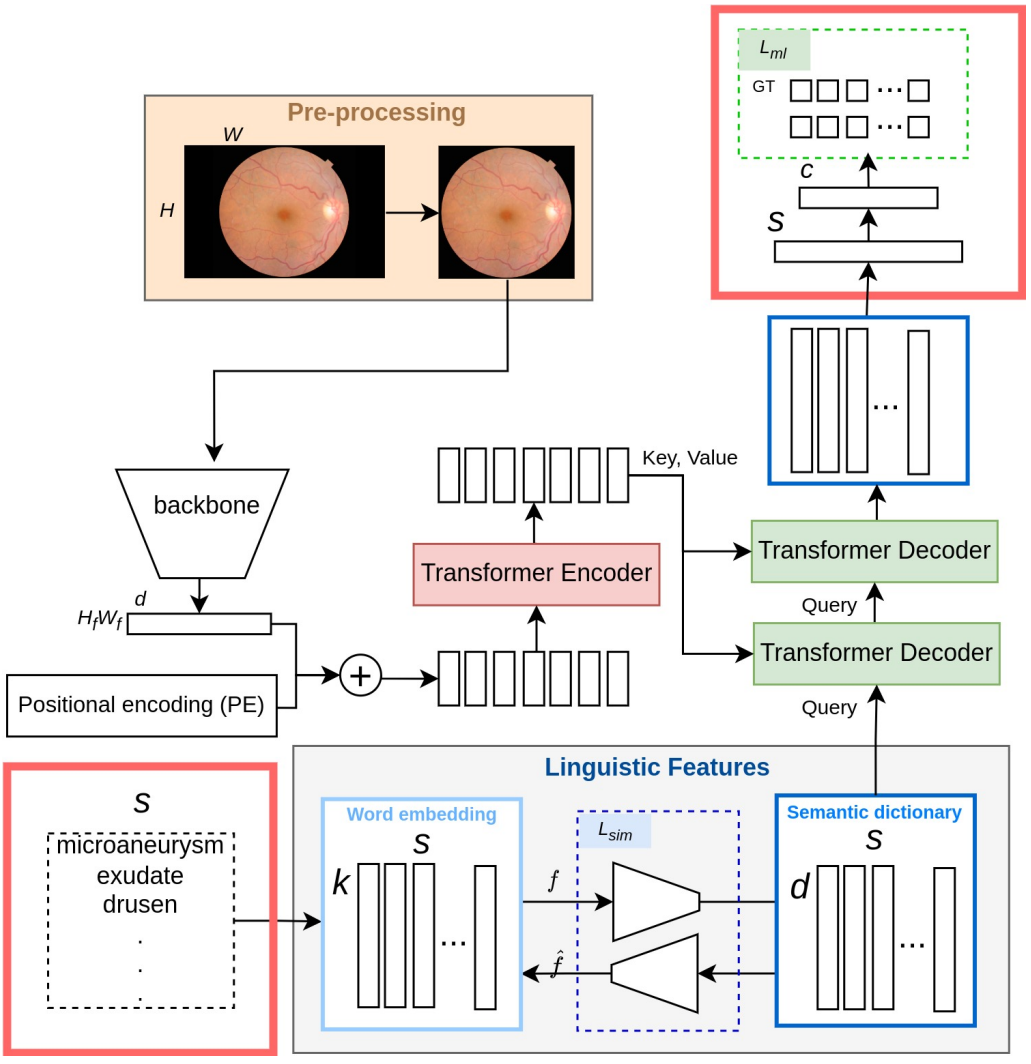


Figure 5.3: The overview illustration of S-TSDL.

Figure 5.3 shows the possible modification of TSDL to get a symptom-based TSDL (S-TSDL). The main difference between TSDL and S-TSDL models is the input for the linguistic features. TSDL requires a list of ocular abnormalities for the input of its linguistic features while S-TSDL requires a list of symptoms of all ocular abnormalities. The other difference in the architecture is the conversion of the semantic dictionary into labels. In TSDL, the updated values of the semantic dictionary are projected directly to classify abnormalities while in S-TSDL the updated values of the semantic dictionary are projected to classify symptoms followed by another Linear module to classify abnormalities.

There are some advantages of generating a symptom-based semantic dictionary:

- Instead of learning the features of the ocular abnormality, the S-TSDL model focuses to learn the features of the symptoms and gives the conclusion of the type of ocular abnormality based on the symptoms.
- The model learns to define the spatial features of all symptoms.
- The word embedding extracted from symptoms contains the co-occurrence between symptoms that can describe the co-occurrence between labels.

Furthermore, providing an additional image with a smaller region of interest (ROI) such as an RGB patch that focuses on OD, can also be considered to extract more detailed features. It can help the model to differentiate some ocular abnormalities that occur in OD such as ODE, ODP, and ODC, which are challenging for the network to learn (see details in Section 4.7.2). The multi-input images can be treated as the multi-modality approach in the Transformer to avoid channel issues. The necessary modification to adapt these multiple input images with the TSDL or S-TSDL model is in the transformer encoder part. There are several ways to design the transformer's encoder for multi-input images as explained in [Xu et al., 2022].

Medical Report Generation. S-TSDL is also a potential method to be adapted for medical report generation. In automatic medical report generation, the input of the model is a medical image and the main concept is an image-to-text approach. Nooralahzadeh et al. [Nooralahzadeh et al., 2021] generated a medical report that represents two corresponding medical images by feeding each image to the visual backbones to extract the visual features before going through the two text generation modules. The first module is to generate the symptoms of the abnormality based on the visual features and the second module is to generate the sentences based on the symptoms to have a human-readable medical report. On the other hand, Kaur et al. [Kaur et al., 2022] implemented a reverse technique of text generation. Here, the medical report is first generated by text generation language modeling and then refined by sentiment analysis language modeling to add sentiment effects to the medical report.

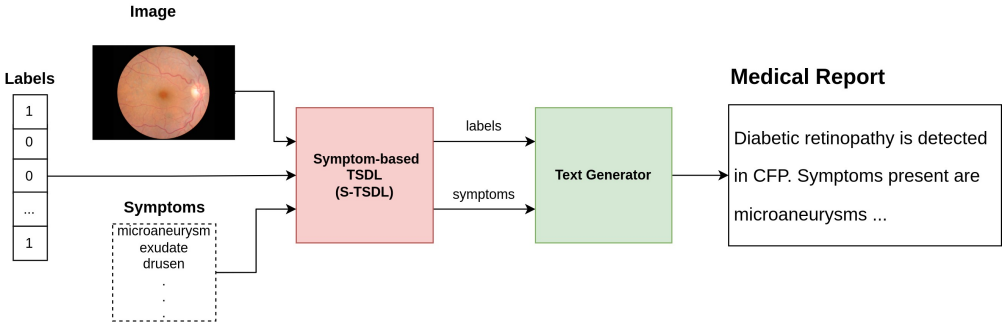


Figure 5.4: Medical report generation with semantic dictionary learning.

The combination of S-TSDL with the method proposed by [Nooralahzadeh et al., 2021] is expected to give better visualization features that can help the text generator to produce a medical report. Figure 5.4 illustrates the main scheme of medical report generation with S-TSDL. The input of the model is a CFP along with labels and symptoms. Symptoms are the knowledge base of the model and it is provided only for training. S-TSDL produces the predicted labels and appeared symptoms that will be provided to the text generator to generate the medical report. However, the dataset that is publicly available for medical report generation is limited to x-ray images, but the adaptation of S-TSDL's concept in medical report generation from other medical image modalities might be possible.

BIBLIOGRAPHY

[Abràmoff et al., 2010] Abràmoff, M. D., Garvin, M. K., and Sonka, M. (2010). **Retinal imaging and image analysis**. *IEEE reviews in biomedical engineering*, 3:169–208.

[Acharya et al., 2008] Acharya, R., Ng, Y. E., and Suri, J. S. (2008). **Image modeling of the human eye**. Artech House.

[Adal et al., 2014] Adal, K. M., Sidibé, D., Ali, S., Chaum, E., Karnowski, T. P., and Mériaudeau, F. (2014). **Automated detection of microaneurysms using scale-adapted blob analysis and semi-supervised learning**. *Computer methods and programs in biomedicine*, 114(1):1–10.

[Adams et al., 2017] Adams, O., Makarucha, A., Neubig, G., Bird, S., and Cohn, T. (2017). **Cross-lingual word embeddings for low-resource language modeling**. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 937–947.

[Allyn,] Allyn, W. https://www.welchallyn.com/en/microsites/iexaminer.html.

[Almeida et al., 2019] Almeida, F., and Xexéo, G. (2019). **Word embeddings: A survey**. *arXiv preprint arXiv:1901.09069*.

[Alsentzer et al., 2019] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., and McDermott, M. (2019). **Publicly available clinical BERT embeddings**. *arXiv preprint arXiv:1904.03323*.

[Badrinarayanan et al., 2017] Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). **Segnet: A deep convolutional encoder-decoder architecture for image segmentation**. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495.

[Bandos et al., 2009] Bandos, A. I., Rockette, H. E., Song, T., and Gur, D. (2009). **Area under the free-response ROC curve (FROC) and a related summary index**. *Biometrics*, 65(1):247–256.

[Baumal, 2018] Baumal, C. R. (2018). **Chapter 4 - Imaging in Diabetic Retinopathy**. In Baumal, C. R., and Duker, J. S., editors, *Current Management of Diabetic Retinopathy*, pages 25–36. Elsevier.

[Bejnordi et al., 2017] Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J. A., Hermsen, M., Manson, Q. F., Balkenhol, M., and others (2017). **Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer**. *Jama*, 318(22):2199–2210.

[Bengio, 2008] Bengio, Y. (2008). **Neural net language models**. *Scholarpedia*, 3(1):3881.

[Besenczi et al., 2016] Besenczi, R., Tóth, J., and Hajdu, A. (2016). **A review on automatic analysis techniques for color fundus photographs**. *Computational and structural biotechnology journal*, 14:371–384.

[Bolton et al., 2022] Bolton, E., Hall, D., Yasunaga, M., Lee, T., Manning, C., and Liang, P. (2022). **Stanford CRFM Introduces PubMedGPT 2.7B**. https://hai.stanford.edu/news/stanford-crfm-introduces-pubmedgpt-27b.

[Bonilha, 2008] Bonilha, V. L. (2008). **Age and disease-related structural changes in the retinal pigment epithelium**. *Clinical Ophthalmology (Auckland, NZ)*, 2(2):413.

[Boucher et al., 2020] Boucher, M. C., Nguyen, M. T. D., and Qian, J. (2020). **Assessment of training outcomes of nurse readers for diabetic retinopathy Telescreening: validation study**. *JMIR diabetes*, 5(2):e17309.

[Brown et al., 2020] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., and others (2020). **Language models are few-shot learners**. *Advances in neural information processing systems*, 33:1877–1901.

[Budak et al., 2017] Budak, U., Şengür, A., Guo, Y., and Akbulut, Y. (2017). **A novel microaneurysms detection approach based on convolutional neural networks with reinforcement sample learning algorithm**. *Health information science and systems*, 5(1):14.

[Carion et al., 2020] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). **End-to-end object detection with transformers**. In *European conference on computer vision*, pages 213–229. Springer.

[Chalakkal et al., 2017] Chalakkal, R. J., Abdulla, W. H., and Sinumol, S. (2017). **Comparative analysis of university of Auckland diabetic retinopathy database**. In *Proceedings of the 9th International Conference on Signal Processing Systems*, pages 235–239.

[Challenge, 2019] Challenge, G. (2019). **Peking University International Competition on Ocular Disease Intelligent Recognition (ODIR-2019)**. https://odir2019.grand-challenge.org/dataset/.

[Chan et al., 2020] Chan, H.-P., Hadjiiski, L. M., and Samala, R. K. (2020). **Computer-aided diagnosis in the era of deep learning**. *Medical physics*, 47(5):e218–e227.

[Chen et al., 2021a] Chen, C., Chuah, J. H., Ali, R., and Wang, Y. (2021a). **Retinal vessel segmentation using deep learning: a review**. *IEEE Access*, 9:111985–112004.

[Chen et al., 2018] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). **Encoder-decoder with atrous separable convolution for semantic image segmentation**. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818.

[Chen et al., 2019] Chen, T., Xu, M., Hui, X., Wu, H., and Lin, L. (2019). **Learning semantic-specific graph representation for multi-label image recognition**. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 522–531.

[Chen et al., 2011] Chen, W. S., Friberg, T. R., Eller, A. W., and Medina, C. (2011). **Advances in retinal imaging of eyes with hazy media: Further Studies**. *Investigative Ophthalmology & Visual Science*, 52(14):4036–4036.

[Chen et al., 2021b] Chen, Z., Wei, X.-S., Wang, P., and Guo, Y. (2021b). **Learning graph convolutional networks for multi-label recognition and applications**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Cho et al., 2014] Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). **On the properties of neural machine translation: Encoder-decoder approaches**. *arXiv preprint arXiv:1409.1259*.

[Congdon et al., 2004] Congdon, N., O'Colmain, B., Klaver, C., Klein, R., Muñoz, B., Friedman, D. S., Kempen, J., Taylor, H. R., Mitchell, P., and others (2004). **Causes and prevalence of visual impairment among adults in the United States.** *Archives of Ophthalmology (Chicago, Ill.: 1960)*, 122(4):477–485.

[Creswell et al., 2018] Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). **Generative adversarial networks: An overview**. *IEEE signal processing magazine*, 35(1):53–65.

[Cubuk et al., 2020] Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. (2020). **RandAugment: Practical automated data augmentation with a reduced search space**. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703.

[Dai et al., 2018] Dai, L., Fang, R., Li, H., Hou, X., Sheng, B., Wu, Q., and Jia, W. (2018). **Clinical report guided retinal microaneurysm detection with multi-sieving deep learning**. *IEEE transactions on medical imaging*, 37(5):1149–1161.

[Dashtbozorg et al., 2018] Dashtbozorg, B., Zhang, J., Huang, F., and ter Haar Romeny, B. M. (2018). **Retinal microaneurysms detection using local convergence index features**. *IEEE Transactions on Image Processing*, 27(7):3300–3315.

[Davis et al., 2005] Davis, M. D., Gangnon, R. E., Lee, L. Y., Hubbard, L. D., Klein, B., Klein, R., Ferris, F. L., Bressler, S. B., Milton, R. C., and others (2005). **The Age-Related Eye Disease Study severity scale for age-related macular degeneration: AREDS report No. 17.** *Archives of ophthalmology (Chicago, Ill.: 1960)*, 123(11):1484–1498.

[Decenciere et al., 2013] Decenciere, E., Cazuguel, G., Zhang, X., Thibault, G., Klein, J.-C., Meyer, F., Marcotegui, B., Quellec, G., Lamard, M., Danno, R., and others (2013). **TeleOphta: Machine learning and image processing methods for teleophthalmology**. *Irbm*, 34(2):196–203.

[Decencière et al., 2014] Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordonez, R., Massin, P., Erginay, A., and others (2014). **Feedback on a publicly distributed image database: the Messidor database**. *Image Analysis & Stereology*, 33(3):231–234.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). **Bert: Pre-training of deep bidirectional transformers for language understanding**. *arXiv preprint arXiv:1810.04805*.

[Dorrell, 1978] Dorrell, D. (1978). **The tilted disc.** *British Journal of Ophthalmology*, 62(1):16–20.

[Dosovitskiy et al., 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., and others (2020). **An image is worth 16x16 words: Transformers for image recognition at scale**. *arXiv preprint arXiv:2010.11929*.

[Du et al., 2020] Du, J., Zou, B., Chen, C., Xu, Z., and Liu, Q. (2020). **Automatic microaneurysm detection in fundus image based on local cross-section transformation and multi-feature fusion**. *Computer Methods and Programs in Biomedicine*, 196:105687.

[Eftekhari et al., 2019] Eftekhari, N., Pourreza, H.-R., Masoudi, M., Ghiasi-Shirazi, K., and Saeedi, E. (2019). **Microaneurysm detection in fundus images using a two-step convolutional neural network**. *Biomedical engineering online*, 18(1):67.

[Esteva et al., 2017] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). **Dermatologist-level classification of skin cancer with deep neural networks**. *nature*, 542(7639):115–118.

[Fan et al., 2021] Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., and Feichtenhofer, C. (2021). **Multiscale vision transformers**. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835.

[Farnell et al., 2008] Farnell, D. J., Hatfield, F. N., Knox, P., Reakes, M., Spencer, S., Parry, D., and Harding, S. P. (2008). **Enhancement of blood vessels in digital fundus photographs via the application of multiscale line operators**. *Journal of the Franklin institute*, 345(7):748–765.

[Faust et al., 2012] Faust, O., Acharya, R., Ng, E. Y.-K., Ng, K.-H., and Suri, J. S. (2012). **Algorithms for the automated detection of diabetic retinopathy using digital fundus images: a review**. *Journal of medical systems*, 36(1):145–157.

[Felippe et al., 2004] Felippe, E., Sales, Á., Soglia, J., Ribeiro, R., Póla, C., Mitre, J., and Carvalho, J. R. (2004). **Preretinal hemorrhage**. *Einstein*, 2(4):314–5.

[Fong et al., 1993] Fong, A. C., and Schatz, H. (1993). **Central retinal vein occlusion in young adults**. *Survey of ophthalmology*, 37(6):393–417.

[Foos, 1974] Foos, R. Y. (1974). **Vitreoretinal juncture—simple epiretinal membranes**. *Albrecht von Graefes Archiv für klinische und experimentelle Ophthalmologie*, 189(4):231–250.

[Frangieh et al., 1982] Frangieh, G. T., Green, W. R., and Fine, S. L. (1982). **A histopathologic study of best's macular dystrophy**. *Archives of Ophthalmology*, 100(7):1115–1121.

[Fujimoto et al., 2000] Fujimoto, J. G., Pitris, C., Boppart, S. A., and Brezinski, M. E. (2000). **Optical coherence tomography: an emerging technology for biomedical imaging and optical biopsy**. *Neoplasia*, 2(1-2):9–25.

[Fujita, 2020] Fujita, H. (2020). **AI-based computer-aided diagnosis (AI-CAD): the latest review to read first**. *Radiological physics and technology*, 13(1):6–19.

[Galdran et al., 2022] Galdran, A., Anjos, A., Dolz, J., Chakor, H., Lombaert, H., and Ayed, I. B. (2022). **State-of-the-art retinal vessel segmentation with minimalistic models**. *Scientific Reports*, 12(1):1–13.

[Galloway et al., 2016] Galloway, N. R., Amoaku, W. M., Galloway, P. H., and Browning, A. C. (2016). **Common eye diseases and their management**. Springer.

[Geetha et al., 2021] Geetha, R., and Tripathy, K. (2021). **Chorioretinitis**. In *StatPearls [Internet]*. StatPearls Publishing.

[Gegundez-Arias et al., 2013] Gegundez-Arias, M. E., Marin, D., Bravo, J. M., and Suero, A. (2013). **Locating the fovea center position in digital fundus images using thresholding and feature extraction techniques**. *Computerized Medical Imaging and Graphics*, 37(5-6):386–393.

[Gella et al., 2016] Gella, L., Raman, R., and Sharma, T. (2016). **Imaging drusens using spectral domain optical coherence tomography**. *Saudi Journal of Ophthalmology*, 30(2):88–91.

[Giancardo et al., 2011] Giancardo, L., Meriaudeau, F., Karnowski, T. P., Li, Y., Tobin, K. W., and Chaum, E. (2011). **Automatic retina exudates segmentation without a manually labelled training set**. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1396–1400. IEEE.

[Giancardo et al., 2010] Giancardo, L., Mériaudeau, F., Karnowski, T. P., Tobin, K. W., Li, Y., and Chaum, E. (2010). **Microaneurysms detection with the radon cliff operator in retinal fundus images**. In *Medical Imaging 2010: Image Processing*, volume 7623, pages 292–299. SPIE.

[Goff et al., 2006] Goff, M. J., McDonald, H. R., Johnson, R. N., Ai, E., Jumper, J. M., and Fu, A. D. (2006). **Causes and treatment of vitreous hemorrhage.** *Comprehensive ophthalmology update*, 7(3):97–111.

[Gour et al., 2021] Gour, N., and Khanna, P. (2021). **Multi-class multi-label ophthalmological disease detection using transfer learning based convolutional neural network**. *Biomedical Signal Processing and Control*, 66:102329.

[Gregory-Evans et al., 2004] Gregory-Evans, C., Williams, M., Halford, S., and Gregory-Evans, K. (2004). **Ocular coloboma: a reassessment in the age of molecular neuroscience**. *Journal of medical genetics*, 41(12):881–891.

[Gross et al., 2013] Gross, S., Klein, M., and Schneider, D. (2013). **Segmentation of blood vessel structures in retinal fundus images with Logarithmic Gabor filters**. *Current Medical Imaging*, 9(2):138–144.

[Guan et al., 2020] Guan, Q., and Huang, Y. (2020). **Multi-label chest X-ray image classification via category-wise residual attention learning**. *Pattern Recognition Letters*, 130:259–266.

[Gui et al., 2021] Gui, J., Sun, Z., Wen, Y., Tao, D., and Ye, J. (2021). **A review on generative adversarial networks: Algorithms, theory, and applications**. *IEEE Transactions on Knowledge and Data Engineering*.

[Gulshan et al., 2016] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., and others (2016). **Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs**. *Jama*, 316(22):2402–2410.

[Gupta et al., 2021] Gupta, A., Paulbuddhe, V. S., and Tripathy, K. (2021). **Exudative retinitis**. In *StatPearls [Internet]*. StatPearls Publishing.

[Hamel, 2006] Hamel, C. (2006). **Retinitis pigmentosa**. *Orphanet journal of rare diseases*, 1(1):1–12.

[Haskes et al., 1995] Haskes, C., and Haskes, L. P. (1995). **Acquired optociliary shunt vessels and their clinical occurrences**. *Clinical Eye and Vision Care*, 7(2):69–77.

[Hassan et al., 2021] Hassan, S. A., Akbar, S., Rehman, A., Saba, T., Kolivand, H., and Bahaj, S. A. (2021). **Recent developments in detection of central serous retinopathy through imaging and artificial intelligence techniques–a review**. *IEEE Access*.

[Havaei et al., 2017] Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., and Larochelle, H. (2017). **Brain tumor segmentation with deep neural networks**. *Medical image analysis*, 35:18–31.

[Hayreh et al., 1981] Hayreh, S. S., and Zahoruk, R. M. (1981). **Anterior ischemic optic neuropathy**. *Ophthalmologica*, 182(1):13–28.

[He et al., 2021] He, J., Li, C., Ye, J., Qiao, Y., and Gu, L. (2021). **Multi-label ocular disease classification with a dense correlation deep neural network**. *Biomedical Signal Processing and Control*, 63:102167.

[He et al., 2022] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). **Masked autoencoders are scalable vision learners**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). **Deep residual learning for image recognition**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Ho et al., 1998] Ho, A. C., Guyer, D. R., and Fine, S. L. (1998). **Macular hole**. *Survey of ophthalmology*, 42(5):393–416.

[Ho et al., 2017] Ho, T., and Nallasamy, S. (2017). **Myopia: Epidemiology and Strategies for Intervention**. *Advances in Ophthalmology and Optometry*, 2(1):63–74.

[Hollenhorst, 1961] Hollenhorst, R. W. (1961). **Significance of bright plaques in the retinal arterioles**. *Jama*, 178(1):23–29.

[Hollow, 2015] Hollow, F. (2015). **Diabetes eye health: A guide for health professionals**. *The Fred Hollows Foundation in partnership with the International Diabetes Federation (IDF)*.

[Hoo et al., 2017] Hoo, Z. H., Candlish, J., and Teare, D. (2017). **What is an ROC curve?**

[Hoover et al., 2000] Hoover, A., Kouznetsova, V., and Goldbaum, M. (2000). **Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response**. *IEEE Transactions on Medical imaging*, 19(3):203–210.

[Imani et al., 2015] Imani, E., Pourreza, H.-R., and Banaee, T. (2015). **Fully automated diabetic retinopathy screening using morphological component analysis**. *Computerized medical imaging and graphics*, 43:78–88.

[Iqbal et al., 2018] Iqbal, T., and Ali, H. (2018). **Generative adversarial network for medical images (MI-GAN)**. *Journal of medical systems*, 42(11):1–11.

[Jager et al., 2008] Jager, R. D., Mieler, W. F., and Miller, J. W. (2008). **Age-related macular degeneration**. *New England Journal of Medicine*, 358(24):2606–2617.

[Jaulim et al., 2013] Jaulim, A., Ahmed, B., Khanam, T., and Chatziralli, I. P. (2013). **Branch retinal vein occlusion: epidemiology, pathogenesis, risk factors, clinical features, diagnosis, and complications. an update of the literature**. *Retina*, 33(5):901–910.

[Jaworski et al., 1999] Jaworski, A., Wolffsohn, J. S., and Napper, G. A. (1999). **Aetiology and management of choroidal folds**. *Clinical and Experimental Optometry*, 82(5):169–176.

[Jogi, 2003] Jogi, R. (2003). **Basic ophthalmology**. Jaypee Brothers, Medical Publishers.

[Kauppi et al., 2007] Kauppi, T., Kalesnykiene, V., Kamarainen, J.-K., Lensu, L., Sorri, I., Raninen, A., Voutilainen, R., Uusitalo, H., Kälviäinen, H., and Pietilä, J. (2007). **The DIARETDB1 diabetic retinopathy database and evaluation protocol.** In *BMVC*, volume 1, pages 1–10.

[Kaur et al., 2022] Kaur, N., and Mittal, A. (2022). **RadioBERT: A deep learning-based system for medical report generation from chest X-ray images using contextual embeddings**. *Journal of Biomedical Informatics*, 135:104220.

[Kertes et al., 2007] Kertes, P. J., and Johnson, T. M. (2007). **Evidence-based eye care**. Lippincott Williams & Wilkins.

[Khodak et al., 2018] Khodak, M., Saunshi, N., Liang, Y., Ma, T., Stewart, B., and Arora, S. (2018). **A la carte embedding: Cheap but effective induction of semantic feature vectors**. *arXiv preprint arXiv:1805.05388*.

[Khoshnevis et al., 2019] Khoshnevis, M., Rosen, S., and Sebag, J. (2019). **Asteroid hyalosis—a comprehensive review**. *Survey of Ophthalmology*, 64(4):452–462.

[Kim et al., 2019] Kim, M., Yun, J., Cho, Y., Shin, K., Jang, R., Bae, H.-j., and Kim, N. (2019). **Deep learning in medical imaging**. *Neurospine*, 16(4):657.

[Klein et al., 1984] Klein, R., Klein, B. E., and Moss, S. E. (1984). **Visual impairment in diabetes**. *Ophthalmology*, 91(1):1–9.

[Kou et al., 2019] Kou, C., Li, W., Liang, W., Yu, Z., and Hao, J. (2019). **Microaneurysms segmentation with a u-net based on recurrent residual convolutional neural network**. *Journal of Medical Imaging*, 6(2):025008.

[Koustenis et al., 2017] Koustenis, A., Harris, A., Gross, J., Januleviciene, I., Shah, A., and Siesky, B. (2017). **Optical coherence tomography angiography: an overview of the technology and an assessment of applications for clinical research**. *British Journal of Ophthalmology*, 101(1):16–20.

[Krause et al., 2018] Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Corrado, G. S., Peng, L., and Webster, D. R. (2018). **Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy**. *Ophthalmology*, 125(8):1264–1272.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). **Imagenet classification with deep convolutional neural networks**. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

[Kumar et al., 2011] Kumar, R., and Indrayan, A. (2011). **Receiver operating characteristic (ROC) curve for medical researchers**. *Indian pediatrics*, 48(4):277–287.

[LeCun et al., 1998] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). **Gradient-based learning applied to document recognition**. *Proceedings of the IEEE*, 86(11):2278–2324.

[Lee et al., 2004] Lee, J., Yap, E., and others (2004). **Optociliary shunt vessels in diabetes mellitus**. *Singapore medical journal*, 45(4):166–169.

[Lee et al., 2020] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). **BioBERT: a pre-trained biomedical language representation model for biomedical text mining**. *Bioinformatics*, 36(4):1234–1240.

[Li et al., 2022a] Li, Y., Mao, H., Girshick, R., and He, K. (2022a). **Exploring plain vision transformer backbones for object detection**. *arXiv preprint arXiv:2203.16527*.

[Li et al., 2022b] Li, Y., Wu, C.-Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., and Feichtenhofer, C. (2022b). **MViTv2: Improved Multiscale Vision Transformers for Classification and Detection**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814.

[Li et al., 2018] Li, Y., and Yang, T. (2018). **Word embedding for understanding natural language: a survey**. In *Guide to big data applications*, pages 83–104. Springer.

[Lin et al., 2017a] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017a). **Feature pyramid networks for object detection**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125.

[Lin et al., 2017b] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017b). **Focal loss for dense object detection**. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

[Lindgren et al., 1996] Lindgren, G., and Lindblom, B. (1996). **Causes of vitreous hemorrhage.** *Current opinion in ophthalmology*, 7(3):13–19.

[Lingam et al., 2021] Lingam, G., Sen, A. C., Lingam, V., Bhende, M., Padhi, T. R., and Xinyi, S. (2021). **Ocular coloboma—a comprehensive review for the clinician**. *Eye*, 35(8):2086–2109.

[Liu et al., 2021a] Liu, S., Zhang, L., Yang, X., Su, H., and Zhu, J. (2021a). **Query2label: A simple transformer way to multi-label classification**. *arXiv preprint arXiv:2107.10834*.

[Liu et al., 2022] Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., and others (2022). **Swin transformer v2: Scaling up capacity and resolution**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019.

[Liu et al., 2021b] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021b). **Swin transformer: Hierarchical vision transformer using shifted windows**. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

[Lochter et al., 2020] Lochter, J. V., Silva, R. M., and Almeida, T. A. (2020). **Deep learning models for representing out-of-vocabulary words**. In *Brazilian Conference on Intelligent Systems*, pages 418–434. Springer.

[Long et al., 2020] Long, S., Chen, J., Hu, A., Liu, H., Chen, Z., and Zheng, D. (2020). **Microaneurysms detection in color fundus images using machine learning based on directional local contrast**. *Biomedical engineering online*, 19(1):1–23.

[Lyu et al., 2022] Lyu, L., Toubal, I. E., and Palaniappan, K. (2022). **Multi-Expert Deep Networks for Multi-Disease Detection in Retinal Fundus Images**. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1818–1822. IEEE.

[Maggon et al., 2009] Maggon, R., Parihar, J., Vats, D., Mathur, V., and Mishra, S. (2009). **Macular hole surgery: Maiden experience**. *Medical Journal, Armed Forces India*, 65(1):77.

[McAllister, 2012] McAllister, I. L. (2012). **Central retinal vein occlusion: a review**. *Clinical & experimental ophthalmology*, 40(1):48–58.

[Melo et al., 2020] Melo, T., Mendonça, A. M., and Campilho, A. (2020). **Microaneurysm detection in color eye fundus images for diabetic retinopathy screening**. *Computers in Biology and Medicine*, 126:103995.

[Meyer et al., 2018] Meyer, M. I., Galdran, A., Mendonça, A. M., and Campilho, A. (2018). **A pixel-wise distance regression approach for joint retinal optical disc and fovea detection**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 39–47. Springer.

[Meyerle et al., 2008] Meyerle, C. B., Chew, E. Y., and Ferris, F. L. (2008). **Nonproliferative diabetic retinopathy**. In *Diabetic retinopathy*, pages 3–27. Springer.

[Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). **Efficient estimation of word representations in vector space**. *arXiv preprint arXiv:1301.3781*.

[Mikolov et al., 2018] Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). **Advances in Pre-Training Distributed Word Representations**. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

[Millay et al., 1986] Millay, R. H., Klein, M. L., Handelman, I. L., and Watzke, R. C. (1986). **Abnormal glucose metabolism and parafoveal telangiectasia**. *American journal of ophthalmology*, 102(3):363–370.

[Mishra et al., 2022] Mishra, C., and Tripathy, K. (2022). **Retinal traction detachment**. In *Statpearls [Internet]*. StatPearls Publishing.

[Moisseiev et al., 2015] Moisseiev, E., Moisseiev, J., and Loewenstein, A. (2015). **Optic disc pit maculopathy: when and how to treat? A review of the pathogenesis and treatment options**. *International journal of retina and vitreous*, 1(1):1–9.

[Müller et al., 2021a] Müller, D., and Kramer, F. (2021a). **MIScnn: a framework for medical image segmentation with convolutional neural networks and deep learning**. *BMC medical imaging*, 21(1):1–11.

[Müller et al., 2021b] Müller, D., Soto-Rey, I., and Kramer, F. (2021b). **Multi-disease detection in retinal imaging based on ensembling heterogeneous deep learning models**. In *German Medical Data Sciences 2021: Digital Medicine: Recognize–Understand–Heal*, pages 23–31. IOS Press.

[Nguyen et al., 2021] Nguyen, H. D., Vu, X.-S., and Le, D.-T. (2021). **Modular graph transformer networks for multi-label image classification**. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9092–9100.

[Niemeijer et al., 2009] Niemeijer, M., Van Ginneken, B., Cree, M. J., Mizutani, A., Quellec, G., Sánchez, C. I., Zhang, B., Hornero, R., Lamard, M., Muramatsu, C., and others (2009). **Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs**. *IEEE transactions on medical imaging*, 29(1):185–195.

[Nooralahzadeh et al., 2021] Nooralahzadeh, F., Gonzalez, N. P., Frauenfelder, T., Fujimoto, K., and Krauthammer, M. (2021). **Progressive transformer-based generation of radiology reports**. *arXiv preprint arXiv:2102.09777*.

[O'Keefe, 2022] O'Keefe, G. D. (2022). **Optical coherence tomography angiography**. https://eyewiki.aao.org/Optical_Coherence_Tomography_Angiography.

[Orlando et al., 2018] Orlando, J. I., Prokofyeva, E., del Fresno, M., and Blaschko, M. B. (2018). **An ensemble deep learning based approach for red lesion detection in fundus images**. *Computer methods and programs in biomedicine*, 153:115–127.

[O'Shea et al., 2015] O'Shea, K., and Nash, R. (2015). **An introduction to convolutional neural networks**. *arXiv preprint arXiv:1511.08458*.

[Pachade et al., 2023] Pachade, S., Porwal, P., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., and Meriaudeau, F. (2023). **RFMiD: Retinal Image Analysis for multi-Disease Detection Challenge**. Under Review.

[Pachade et al., 2021] Pachade, S., Porwal, P., Thulkar, D., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Giancardo, L., Quellec, G., and Mériaudeau, F. (2021). **Retinal fundus multi-disease image dataset (RFMiD): a dataset for multi-disease detection research**. *Data*, 6(2):14.

[Panwar et al., 2016] Panwar, N., Huang, P., Lee, J., Keane, P. A., Chuan, T. S., Richhariya, A., Teoh, S., Lim, T. H., and Agrawal, R. (2016). **Fundus photography in the 21st century—a review of recent technological advances and their implications for worldwide healthcare**. *Telemedicine and e-Health*, 22(3):198–208.

[Piette et al., 2006] Piette, S. D., and Sergott, R. C. (2006). **Pathological optic-disc cupping**. *Current opinion in ophthalmology*, 17(1):1–6.

[Pizer et al., 1987] Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J. B., and Zuiderveld, K. (1987). **Adaptive histogram equalization and its variations**. *Computer vision, graphics, and image processing*, 39(3):355–368.

[Porwal et al., 2018] Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., and Meriaudeau, F. (2018). **Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research**. *Data*, 3(3):25.

[Porwal et al., 2020] Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., Liu, L., Wang, J., Liu, X., Gao, L., and others (2020). **IDRiD: Diabetic retinopathy–segmentation and grading challenge**. *Medical image analysis*, 59:101561.

[Quinn, 1996] Quinn, C. (1996). **Cystoid macular edema.** *Optometry clinics: the official publication of the Prentice Society*, 5(1):111–130.

[Radford et al., 2018] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., and others (2018). **Improving language understanding by generative pre-training**.

[Rehak et al., 2008] Rehak, J., and Rehak, M. (2008). **Branch retinal vein occlusion: pathogenesis, visual prognosis, and treatment modalities**. *Current eye research*, 33(2):111–131.

[Reichel et al., 2015] Reichel, E., Duker, J., Goldman, D., Fein, J., and Vora, R. (2015). **Handbook of Retinal Disease: a Case-based Approach**. JP Medical Ltd.

[Ridnik et al., 2021] Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., and Zelnik-Manor, L. (2021). **Asymmetric loss for multi-label classification**. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91.

[Riordan-Eva et al., 2011] Riordan-Eva, P., and Cunningham, E. T. (2011). **Vaughan & Asbury's general ophthalmology**. McGraw Hill Professional.

[Rodriguez et al., 2022] Rodriguez, M., AlMarzouqi, H., and Liatsis, P. (2022). **Multi-label Retinal Disease Classification Using Transformers**. *IEEE Journal of Biomedical and Health Informatics*.

[Rosenbaum et al., 2016] Rosenbaum, J. T., Sibley, C. H., and Lin, P. (2016). **Retinal vasculitis**. *Current opinion in rheumatology*, 28(3):228.

[Rothova, 2007] Rothova, A. (2007). **Inflammatory cystoid macular edema**. *Current Opinion in Ophthalmology*, 18(6):487–492.

[Rotsos et al., 2008] Rotsos, T. G., and Moschos, M. M. (2008). **Cystoid macular edema**. *Clinical ophthalmology (Auckland, NZ)*, 2(4):919.

[Roychowdhury et al., 2014] Roychowdhury, S., Koozekanani, D. D., and Parhi, K. K. (2014). **Blood vessel segmentation of fundus images by major vessel extraction and subimage classification**. *IEEE journal of biomedical and health informatics*, 19(3):1118–1128.

[Ruia et al., 2021] Ruia, S., and Tripathy, K. (2021). **Fluorescein angiography**. In *StatPearls [Internet]*. StatPearls Publishing.

[Russakovsky et al., 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., and others (2015). **Imagenet large scale visual recognition challenge**. *International journal of computer vision*, 115(3):211–252.

[Saeedi et al., 2019] Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., Colagiuri, S., Guariguata, L., Motala, A. A., Ogurtsova, K., and others (2019). **Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas**. *Diabetes research and clinical practice*, 157:107843.

[Saine et al., 2002] Saine, P. J., and Tyler, M. E. (2002). **Ophthalmic photography: retinal photography, angiography, and electronic imaging**, volume 132. Butterworth-Heinemann Boston.

[Sandler et al., 2018] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). **Mobilenetv2: Inverted residuals and linear bottlenecks**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520.

[Santos et al., 2021] Santos, M., and Janigian, R. H. (2021). **Branch retinal artery occlusion**. https://eyewiki.aao.org/Branch_Retinal_Artery_Occlusion.

[Sathiyamurthy et al., 2007] Sathiyamurthy, K., Sophia, S. A. M., Ashalatha, A., and Sujitha, P. (2007). **Automated reasoning tool for the detection of race conditions in web services**. In *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)*, volume 2, pages 61–65. IEEE.

[Schmidt, 2008] Schmidt, D. (2008). **The mystery of cotton-wool spots-a review of recent and historical descriptions**. *European journal of medical research*, 13(6):231.

[Schuster et al., 2022] Schuster, T., Fisch, A., Gupta, J., Dehghani, M., Bahri, D., Tran, V. Q., Tay, Y., and Metzler, D. (2022). **Confident adaptive language modeling**. *arXiv preprint arXiv:2207.07061*.

[Schwartz et al., 1973] Schwartz, B., Reinstein, N. M., and Lieberman, D. M. (1973). **Pallor of the optic disc: quantitative photographic evaluation**. *Archives of Ophthalmology*, 89(4):278–286.

[Sengar et al., 2021] Sengar, N., Joshi, R. C., and Dutta, M. K. (2021). **An efficient artificial intelligence-based approach for diagnosis of media haze disease**. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE.

[Singh et al., 2021] Singh, D., and Tripathy, K. (2021). **Retinal macroaneurysm**. In *StatPearls [Internet]*. StatPearls Publishing.

[Son et al., 2017] Son, J., Park, S. J., and Jung, K.-H. (2017). **Retinal vessel segmentation in fundoscopic images with generative adversarial networks**. *arXiv preprint arXiv:1706.09318*.

[Sowka et al., 1999] Sowka, J., and Aoun, P. (1999). **Tilted disc syndrome.** *Optometry and vision science: official publication of the American Academy of Optometry*, 76(9):618–623.

[Sowka et al., 2014] Sowka, J. W., and Kabat, A. G. (2014). **Collateral damage**. https://www.reviewofoptometry.com/article/collateral-damage.

[Spaide et al., 2018] Spaide, R. F., Fujimoto, J. G., Waheed, N. K., Sadda, S. R., and Staurenghi, G. (2018). **Optical coherence tomography angiography**. *Progress in retinal and eye research*, 64:1–55.

[Spraul et al., 1997] Spraul, C. W., and Grossniklaus, H. E. (1997). **Vitreous hemorrhage**. *Survey of ophthalmology*, 42(1):3–39.

[Staal et al., 2004] Staal, J., Abràmoff, M. D., Niemeijer, M., Viergever, M. A., and Van Ginneken, B. (2004). **Ridge-based vessel segmentation in color images of the retina**. *IEEE transactions on medical imaging*, 23(4):501–509.

[Stitt et al., 2016] Stitt, A. W., Curtis, T. M., Chen, M., Medina, R. J., McKay, G. J., Jenkins, A., Gardiner, T. A., Lyons, T. J., Hammes, H.-P., Simo, R., and others (2016). **The progress in understanding and treatment of diabetic retinopathy**. *Progress in retinal and eye research*, 51:156–186.

[Straatsma et al., 1981] Straatsma, B. R., Foos, R. Y., Heckenlively, J. R., and Taylor, G. N. (1981). **Myelinated retinal nerve fibers.** *American journal of ophthalmology*, 91(1):25–38.

[Sun et al., 2022a] Sun, K., He, M., He, Z., Liu, H., and Pi, X. (2022a). **EfficientNet embedded with spatial attention for recognition of multi-label fundus disease from color fundus photographs**. *Biomedical Signal Processing and Control*, 77:103768.

[Sun et al., 2022b] Sun, K., He, M., Xu, Y., Wu, Q., He, Z., Li, W., Liu, H., and Pi, X. (2022b). **Multi-label classification of fundus images with graph convolutional network and LightGBM**. *Computers in Biology and Medicine*, 149:105909.

[Tan et al., 2019] Tan, M., and Le, Q. (2019). **Efficientnet: Rethinking model scaling for convolutional neural networks**. In *International conference on machine learning*, pages 6105–6114. PMLR.

[Tarabishy et al., 2007] Tarabishy, A. B., Alexandrou, T. J., and Traboulsi, E. I. (2007). **Syndrome of myelinated retinal nerve fibers, myopia, and amblyopia: a review**. *Survey of ophthalmology*, 52(6):588–596.

[Topcon,] Topcon. https://www.ophthalmicmart.com/product/topcon-trc-nw8f-plus/.

[Turbert, 2020] Turbert, D. (2020). **Fundus**. https://www.aao.org/eye-health/anatomy/fundus.

[Van et al., 2007] Van, S., and P, G. (2007). **Optic disc edema**. In *Seminars in neurology*, volume 27, pages 233–243. Thieme Medical Publishers.

[Varma et al., 2013] Varma, D., Cugati, S., Lee, A., and Chen, C. (2013). **A review of central retinal artery occlusion: clinical presentation and management**. *Eye*, 27(6):688–697.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). **Attention is all you need**. *Advances in neural information processing systems*, 30.

[Vaughan et al., 1977] Vaughan, D., and Asbury, T. (1977). **General ophthalmology**. In *General ophthalmology*, pages 379–379.

[Walker et al., 1990] Walker, H. K., Hall, W. D., and Hurst, J. W. (1990). **Clinical methods: the history, physical, and laboratory examinations**.

[Walter et al., 2007] Walter, T., Massin, P., Erginay, A., Ordonez, R., Jeulin, C., and Klein, J.-C. (2007). **Automatic detection of microaneurysms in color fundus images**. *Medical image analysis*, 11(6):555–566.

[Wang et al., 2017] Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X., and Wang, F.-Y. (2017). **Generative adversarial networks: introduction and outlook**. *IEEE/CAA Journal of Automatica Sinica*, 4(4):588–598.

[Wang et al., 2020] Wang, S., Zhou, W., and Jiang, C. (2020). **A survey of word embeddings based on deep learning**. *Computing*, 102(3):717–740.

[Wen et al., 2020] Wen, S., Liu, W., Yang, Y., Zhou, P., Guo, Z., Yan, Z., Chen, Y., and Huang, T. (2020). **Multilabel image classification via feature/label co-projection**. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(11):7250–7259.

[WHO, 2021] WHO (2021). **Diabetes**. https://www.who.int/news-room/fact-sheets/detail/diabetes.

[Wu et al., 2019] Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzębski, S., Févry, T., Katsnelson, J., Kim, E., and others (2019). **Deep neural networks improve radiologists' performance in breast cancer screening**. *IEEE transactions on medical imaging*, 39(4):1184–1194.

[Xia et al., 2021] Xia, H., Lan, Y., Song, S., and Li, H. (2021). **A multi-scale segmentation-to-classification network for tiny microaneurysm detection in fundus images**. *Knowledge-Based Systems*, 226:107140.

[Xiao et al., 2020] Xiao, B., Liao, Q., Li, Y., Weng, F., Jin, L., Wang, Y., Huang, W., Yi, J., Burton, M. J., and Yip, J. L. (2020). **Validation of handheld fundus camera with mydriasis for retinal imaging of diabetic retinopathy screening in china: a prospective comparison study**. *BMJ open*, 10(10):e040196.

[Xie et al., 2020] Xie, Y., Zhang, J., Lu, H., Shen, C., and Xia, Y. (2020). **SESV: Accurate medical image segmentation by predicting and correcting errors**. *IEEE Transactions on Medical Imaging*, 40(1):286–296.

[Xu et al., 2022] Xu, P., Zhu, X., and Clifton, D. A. (2022). **Multimodal learning with transformers: A survey**. *arXiv preprint arXiv:2206.06488*.

[Yamashita et al., 2018] Yamashita, R., Nishio, M., Do, R. K. G., and Togashi, K. (2018). **Convolutional neural networks: an overview and application in radiology**. *Insights into imaging*, 9(4):611–629.

[Yan et al., 2021] Yan, Y., and Liao, Y. J. (2021). **Updates on ophthalmic imaging features of optic disc drusen, papilledema, and optic disc edema**. *Current Opinion in Neurology*, 34(1):108.

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

[Yang et al., 2019] Yang, H. K., Oh, J. E., Han, S. B., Kim, K. G., and Hwang, J.-M. (2019). **Automatic computer-aided analysis of optic disc pallor in fundus photographs**. *Acta Ophthalmologica*, 97(4):e519–e525.

[Yao et al., 2022] Yao, X., Son, T., and Ma, J. (2022). **Developing portable widefield fundus camera for teleophthalmology: Technical challenges and potential solutions**. *Experimental Biology and Medicine*, 247(4):289–299.

[Yau et al., 2012] Yau, J. W., Rogers, S. L., Kawasaki, R., Lamoureux, E. L., Kowalski, J. W., Bek, T., Chen, S.-J., Dekker, J. M., Fletcher, A., Grauslund, J., and others (2012). **Global prevalence and major risk factors of diabetic retinopathy**. *Diabetes care*, 35(3):556–564.

[Yoshihara et al., 2014] Yoshihara, N., Yamashita, T., Ohno-Matsui, K., and Sakamoto, T. (2014). **Objective analyses of tessellated fundi and significant correlation between degree of tessellation and choroidal thickness in healthy eyes**. *PloS one*, 9(7):e103586.

[Zhang et al., 2011] Zhang, W., Liu, H., Rojas, M., Caldwell, R. W., and Caldwell, R. B. (2011). **Anti-inflammatory therapy for diabetic retinopathy**. *Immunotherapy*, 3(5):609–628.

[Zhang et al., 2021] Zhang, X., Wu, J., Meng, M., Sun, Y., and Sun, W. (2021). **Feature-transfer network and local background suppression for microaneurysm detection**. *Machine Vision and Applications*, 32(1):1–13.

[Zhang et al., 2020] Zhang, X., Xiao, Z., Zhang, F., Ogunbona, P. O., Xi, J., and Tong, J. (2020). **Shape-based filter for micro-aneurysm detection**. *Computers & Electrical Engineering*, 84:106620.

[Zhao et al., 2021a] Zhao, J., Yan, K., Zhao, Y., Guo, X., Huang, F., and Li, J. (2021a). **Transformer-based dual relation graph for multi-label image recognition**. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 163–172.

[Zhao et al., 2021b] Zhao, J., Zhao, Y., and Li, J. (2021b). **M3tr: Multi-modal multi-label recognition with transformer**. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 469–477.

[Zhou et al., 2021] Zhou, F., Huang, S., and Xing, Y. (2021). **Deep semantic dictionary learning for multi-label image classification**. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3572–3580.

[Zhou et al., 2022] Zhou, Q., Zou, H., and Wang, Z. (2022). **Long-Tailed Multi-label Retinal Diseases Recognition via Relational Learning and Knowledge Distillation**. In

International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 709–718. Springer.

[Zhu et al., 2017] Zhu, F., Li, H., Ouyang, W., Yu, N., and Wang, X. (2017). **Learning spatial regularization with image-level supervisions for multi-label image classification**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5513–5522.

Document réalisé avec LaTeX et :
le style LaTeX pour Thèse de Doctorat créé par S. Galland — http://www.multiagent.fr/ThesisStyle
la collection de paquets `tex-upmethodology` — http://www.arakhne.org/tex-upmethodology/

école doctorale **sciences pour l'ingénieur et microtechniques**

Title: Computer-Aided-Diagnosis for Ocular Abnormalities from A Single Color Fundus Photography with Deep Learning

Keywords: Deep learning, image processing, computer-aided-diagnosis, microaneurysms detection, multi-label detection, ocular abnormalities.

Abstract:

Any damage to the retina can lead to severe consequences like blindness. This visual impairment is preventable by early detection of ocular abnormalities. Computer-aided diagnosis (CAD) for ocular abnormalities is built by analyzing retinal imaging modalities, for instance, Color Fundus Photography (CFP). The main objectives of this thesis are to build two CAD models, one to detect the microaneurysms (MAs), the first visible symptom of diabetic retinopathy, and the other for multi-label detection of 28 ocular abnormalities consisting of frequent and rare abnormalities from a single CFP by using deep learning-based approaches. Two methods were proposed for MAs detection: ensemble-based and cascade-based methods. Ensemble-based MAs detection aims to find the best combination of input channels while the goal of cascade-based MAs detection is to reduce the false positive predictions with high sensitivity. The MAs detection with the cascade learning method achieves 0.792 sensitivity, the highest sensitivity on the E-Ophta dataset in 8 false positives per image. Two methods were also proposed for multi-label detection: Convolutional Neural Network (CNN)-based and Transformer-based methods. These proposed methods combine the visual features extracted from a color fundus image and the label co-occurrence dependencies extracted from linguistic features. The correlation between visual and linguistic features is learned by a semantic dictionary. CNN-based multi-label detection aims to adapt the model with out-of-vocabulary words. The results of this model show the positive impact of linguistic input interference in multi-label detection. Transformer-based multi-label detection enhances the linguistic input interference in multi-label detection.

Titre : Diagnostic Assisté par Ordinateur grâce à l'Apprentissage Profond des Anomalies Oculaires à partir d'Images du Fond d'œil.

Mots-clés : Apprentissage profond, traitement des images, diagnostic automatique, détection de microanévrismes, détection multi-étiquettes, anomalies oculaires.

Résumé :

Cette thèse présente des modèles de diagnostics médicaux assistés par ordinateurs (CAD) pour la détection d'anomalies oculaires à partir de photographies du fond d'œil. Le premier modèle CAD est mis en œuvre pour détecter les microanévrismes (MAs) avec une sensibilité élevée et un faible nombre de faux positifs. Le principal défi dans la détection des Mas basée sur des approches apprentissages profonds résulte de jeux de données fortement déséquilibrés. Nous avons proposé une détection des MAs qui se compose de trois processus principaux: le prétraitement, l'extraction des candidats MAs et la classification des MAs. Pour réduire le besoin d'un grand nombre de données (notamment de MAs), les candidats MAs sont extraits dans une approche non supervisée. Pour la classification finale, cette thèse propose deux approches, une méthode basée sur un ensemble de classifieurs et l'autre sur des classifieurs en cascade. Les expériences sont menées sur les jeux de données IDRiD et E-Ophta. Le classifieur basé sur un apprentissage en cascade atteint la sensibilité la plus élevée pour 8 FPI par rapport aux autres méthodes existantes dans la littérature. Le deuxième modèle proposé dans cette thèse est destiné à la la détection multi-étiquettes (labels) de 28 anomalies oculaires dont certaines très rares. Les anomalies oculaires rares sont généralement ignorées en raison d'un manque de caractéristiques visuelles. Pour enrichir notre modèle, nous proposons une approche associant les données linguistiques (description des pathologies) aux caractéristiques visuelles. Le modèle apprend la relation entre les caractéristiques spatiales et les caractéristiques linguistiques représentées sous forme de dictionnaire sémantique. Deux approches pour la détection multi-étiquettes avec apprentissage profond sont proposées dans cette thèse : l'apprentissage par dictionnaire sémantique basé CNN et basé Transformer. L'apprentissage du dictionnaire sémantique basé CNN se concentre sur l'apprentissage du dictionnaire sémantique avec la contrainte de représentation visuelle. Les résultats montrent que l'ajout de la modalité linguistique et la génération du dictionnaire sémantique peuvent augmenter les performances du modèle par rapport au modèle qui utilise uniquement la donnée image. La deuxième méthode proposée est construite sur la base de Transformers. Le dictionnaire sémantique agit comme la requête (query) tandis que les caractéristiques issues des images sont assignées à la clé (key) et la valeur (value). Les expérimentations sont menées sur le jeu de données RFMiD.

UBFC UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

Université Bourgogne Franche-Comté
32, avenue de l'Observatoire
25000 Besançon, France