DEVELOPMENT OF DEEP LEARNING FOR DIABETIC RETINOPATHY CLASSIFICATION SYSTEM BASED ON FUNDUS IMAGE

By
Mr. Rapeephat Yodsungnoen

A Thesis Submitted in Partial Fulfillment of the Requirements for the
Degree of Master of Science in Integrated Science and Innovation
Suranaree University of Technology
Academic Year 2024

M.Sc. Thesis

Mr. Rapeephat Yodsungnoen **ID. No. M6500580**

School of Integrated Science and Innovation **Institute of Science**

Thesis Commitee

A) Asst. Prof. Dr. Ittipon Fongkaew Thesis advisor

Student's Signature _____

(Mr. Rapeephat Yodsungnoen)

Date_____

Advisor's Signature _____

(Asst. Prof. Dr. Ittipon Fongkaew)

Date_____

RAPEEPHAT YODSUNGNOEN: DEVELOPMENT OF DEEP LEARNING FOR DIABETIC RETINOPATHY CLASSIFICATION SYSTEM BASED ON FUNDUS IMAGE
ADVISOR : ASST. PROF. DR. ITTIPON FONGKAEW, Ph.D. 73 PP.

DEEP LEARNING/DIABETIC RETINOPATHY/IMAGE PROCESSING/QUALITY ASSESSMENT

This thesis presents a comprehensive end-to-end diabetic retinopathy (DR) classification system based on retinal fundus images, consisting of two integrated modules: image screening and DR grading. The image screening module employs template-based correlation filtering to detect the optic disc and macula, followed by rule-based and machine learning algorithms to assess medical image suitability. Experimental evaluations demonstrate a high recall of 0.906 and a false discovery rate of 0.065, meeting clinical screening benchmarks. The DR grading module leverages the Swin Transformer backbone, augmented with synthetic oversampling (SMOTE) and fine-tuning strategies to address data imbalance. On the APTOS 2019 dataset, the proposed model achieves an F1 macro score of 0.693 and a quadratic weighted kappa (QWK) of 0.903, surpassing existing methods. Extensive ablation studies confirm the effectiveness of data augmentation, loss function selection, and sampling techniques. The proposed system offers both scientific contribution and practical applicability, targeting scalable and accessible AI-assisted DR screening in resource-limited settings.

School of Department of Interdisciplinary Science and Internationalization
Academic Year 2024

Student's Signature _____
Advisor's Signature _____

ACKNOWLEDGEMENTS

First of all, I would like to express my gratitude to Asst. Prof. Dr. Ittipon Fongkaew for his supervision and many possible opportunities he has given to me.

CONTENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

4.1 Background

The retina plays a crucial role among ocular structures, including blood vessels, the optic nerve, and photoreceptors. Hence, any disruption within these structures can lead to abnormalities in retinal function. Currently, when we observe the top 3 common diseases that lead to blindness in Thailand, namely cataracts, refractive lens error, and diabetic retinopathy (DR) (Isipradit, 2014). Although, cataracts are the most common cause of blindness but they are surgically treatable and often present clear early symptoms. On the other hand, diabetic retinopathy is irreversible and asymptomatic in its early stages. Consequently, DR emerges as the most common disease associated with irreversible vision loss and negligible early detection. As a result, we have to rapidly investigate and determine the solution for addressing this disease. Diabetic retinopathy (DR) is a condition stemming from elevated blood pressure and abnormal glucose levels associated with diabetes. DR can result in leakage and swelling of blood and fluid within the retina, ultimately leading to blindness and posing a significant threat to visual health. Additionally, studies have reported that approximately one-third of diabetes patients are at risk of developing DR (Wong, 2018) and once diagnosed with DR, these patients are 2.5 to 4 times more likely to develop sustained blindness (Wykoff, 2021). Moreover, statistics from 2010 reveal that 3.7 million individuals experienced visual impairment, and 0.8 million suffered from blindness due to DR (Wong, 2018). Interestingly, The prevalence of diabetic retinopathy (DR) is estimated to increase from 103 million in 2020 to 130 million in 2030 and further to 160 million in 2045 (Teo, 2021). However, A report from 2015 indicated that there were approximately 230,000 ophthalmologists across 194 countries, including Thailand, with an annual growth rate of 2.6 % (Resnikoff, 2019). As a result, there is a trend to face a shortage of ophthalmologists in the future, resulting in a higher chance of neglect in diabetic patients, who require regular screening, as well as among DR patients, who need periodic re-evaluation and accurate grading to prevent the onset of severe scenarios. Moreover, research in Thailand revealed a significant disparity in the distribution of ophthalmologists across different regions. While the average ratio stands at approximately one ophthalmologist per fifty thou-

sand individuals nationwide (Estopinal, 2013), this ratio is heavily skewed due to the concentration of ophthalmologists in urban areas. For instance, in the capital city, there are approximately 437 ophthalmologists (Royal College of Ophthalmologists of Thailand, 2024) per 5.5 million individuals (Department of Provincial Administration, 2024). Conversely, in larger provinces like Ubon Ratchathani, there are only 9 ophthalmologists (Royal College of Ophthalmologists of Thailand, 2024) serving a population of 1.8 million individuals (Department of Provincial Administration, 2024). This uneven distribution poses challenges in providing timely and accurate diagnoses, as nurses and physicians in hospitals and healthcare centers outside major cities may have to assume responsibilities held by ophthalmologists, increasing the risk of delayed or incorrect diagnoses.

Various object detection algorithms currently are being employed to automatically detect critical structures such as the optic disc and macula. These algorithms aim to replicate the screening methods used by expert ophthalmologists, to address the issue associated with collecting unusable images by nurses or physicians during patient consultations. In (Sinthanayothin, 1999), the method utilizes an inverted Gaussian template to detect the macula and employs the brightest pixel to locate the optic disc in the fundus image. Additionally, morphological techniques are deployed to address an exudates and blood vessels in the image, facilitating the locating of the optic disc and macula (Sekhar, 2008). In (Welfer, 2011), an approach for removing unwanted lesions and noise is proposed, along with optic disc detection using information from the vascular tree in the fundus image. In (Tariq, 2012), the method utilizes a Gaussian Mixture Model to locate the macula by aggregating five feature vectors related to the macula. In (Zheng, 2014), the approach employs a two-stage detection process, comprising coarse and fine stages, to determine the location of the optic disc. Subsequently, the macula is located through circular scanning around the optic disc. In (Deka, 2015), the method extracts blood vessels from the fundus image to utilize the vessel-free area for detecting the macula location. In (Kamble, 2017), intensity lines are sampled from the image to create an intensity profile. Then, signal processing techniques, specifically the Daubechies 4 wavelet, are then applied to determine the locations of the optic disc and macula.

Furthermore, to mitigate the risk of incorrect diagnosis, several research studies have employed deep learning for diabetic retinopathy (DR) severity grading. In (W.

Zhang, 2019), the method introduces ensemble backbone networks to enhance feature extraction capabilities and utilizes a customized fully connected neural network, termed SDNN, as a classifier. Additionally, Bayesian optimization for hyperparameter tuning is proposed to enhance the performance of the Inception-V4 network in (Shankar, 2020). In (A. He, 2020), the DR grading network is enhanced by introducing a novel attention block called the category attention block (CAB), specifically designed to tackle the imbalance issue present in various datasets. Moreover, in (R. Sun, 2021), the method introduces the bio-marker establishing from explainable attention map by leveraging the attention mechanism inherent in the customized Vision Transformer (ViT) network. In (Li, 2022), a pyramid network is introduced, capable of processing retinal images at various resolutions, thereby enhancing the network's ability to understand fine-to-coarse details present in the images. Additionally, the method proposes the utilization of attention maps as guidelines for training the networks. In their study, (Tusfiqur, 2022) introduced a comprehensive training approach aimed at developing a robust diabetic retinopathy (DR) grading network. This approach comprises three learning approaches: adversarial learning, supervised learning, and expert feedback learning. Adversarial learning is employed to train the lesion segmentation network, utilizing lesion map predictions to refine the grading network. The supervised learning approach trains the grading network using retinal images alongside lesion maps. Additionally, expert feedback learning involves leveraging expert ophthalmologist feedback to validate predictions and tune the grading network.

In clinical practice, the grading of diabetic retinopathy (DR) is conventionally conducted by trained ophthalmologists or retinal specialists, who evaluate retinal fundus images according to standardized clinical protocols, such as the International Clinical Diabetic Retinopathy (ICDR) scale. This grading process necessitates the identification of pathological features, including microaneurysms, hemorrhages, hard exudates, and neovascularization, which are indicative of disease progression. Each image is examined independently and categorized into one of five DR severity levels (0–4). In many hospitals, especially in Thailand, this process is manual, time-consuming, and subject to inter-observer variability, particularly in borderline cases. Due to limited specialist availability, it also causes delays in diagnosis and treatment in rural or resource-limited regions, thereby timely and accurate automated DR grad-

ing is crucial to address these issue. Nevertheless, building such systems requires large, diverse, and balanced datasets, criteria that are at odds with the characteristics of our current dataset, which is relatively small and highly imbalanced. While this limitation can be addressed by relying on ophthalmologists to generate new labeled data, the process is hindered by another practical challenge: a substantial portion of the image database consists of medically unsuitable images. Consequently, ophthalmologists must expend significant effort in screening out medically unsuitable images before proceeding with diagnostic labeling, which diminishes efficiency and slows the development of reliable DR grading models. Consequently, we are interested in establishing an end-to-end automated expert system that can perform the screening and grading processes on the retinal image. Additionally, to achieve this both tasks, the expert system must comprise two sub-systems. Firstly, in image screening system, we implement the template matching technique and anatomical knowledge of ocular structures to extract relevant features. Then, these features are utilized to locate the optic disc and macula, fulfilling the criteria of expert requirements. Lastly, we utilize the model or rule to classify image as medically unsuitable and medically suitable retinal image. The reason to use handcrafted feature extraction for ocular composition detection instead of deep learning is driven by the limited size of our available dataset, which comprises approximately 500 images. Secondly, for DR grading, we turn to deep learning to address this task. This decision is motivated by the availability of various public datasets and the outstanding performance of deep learning methods in this task.

4.2 Research objective

1. Implementing the correlation filtering technique to screen the retinal fundus image
2. Implementing the deep learning to classify the severity level of DR based on retinal fundus image

4.3 Scope and Limitations

This research project investigates the object detection algorithm for fundus image quality assessment that is utilized to screen the fundus image and establish our own dataset, as well as the deep learning for grading the DR severity level from

the fundus image. Moreover, the dataset in this work consists of IDRiD for ocular object detection, APTOS2019 used to train the DR grading network. In object detection, we use the template matching technique to detect the optic disc and macula. The reference template is created by averaging the N number of optic disc or macula images, and we will increase the macula detection by utilizing the ROI of macula. In deep learning, we trained five pretrained networks, including ResNet 50, VGG 19, Inception V3, DenseNet 161, Swin Transformer, or the other networks that can outperform the baseline score, based on DR grading. Then, we select the optimal network or ensemble for improving and tuning. Eventually, we aim to achieve an false discovery rate (FDR) of 0.05 and a recall score of 0.90 in image screening [(Coyner, 2018; Fleming, 2006)]. In the DR grading network, we expect to exceed human performance, which is typically 0.894 in accuracy, 0.714 in F1 macro score and 0.871 in QWK (Krause, 2018).

CHAPTER 2

LITERATURE REVIEW

5.1 Diabetic retinopathy

Diabetic Retinopathy (DR) is the ophthalmic disease that is triggered by diabetes, or it is a complication disease of diabetes. The dreadfulness of disease are the impact from abnormal blood pressure and blood glucose levels, causing the following symptoms: Vascular leakage in the retina causes blood and fluid to leak into the vitreous humor (gel-like fluid inside the eyeball). This result in fluid and blood congestion in the retina, which blurs vision. Additionally, new vascular growth, leading to blindness because these new blood vessels are fragile, thus they can leak massive amounts of blood, which can create a large dark spots and block the vision. The healthy retinas and DR patients' retinas are compared in Figure 1.
Moreover, the figure illustrates that the significant evidence used to indicate the presence of DR is the identified lesions. The following is a list and description of the lesions identified in DR:

- **Microaneurysm (MA)** is the earliest lesion which indicate evidence of DR existence. The lesion character is the microscopic dark red dot, appearing on the retina. The size is frequently less than 125 μm and the margins are sharp. this lesion is occurred by the bulges of the smallest intra-retinal blood vessels, called capillary. As shown in Figure 2.

- **Retinal Hemorrhage (HM)** is the next stage of MA because this lesion is indicative of capillary leakage, caused by severe hypertension, which allows the plasma constituents to leak into the retina. These hemorrhage's size is usually larger than 125 μm and various shapes such as dots, blots, and flame-shape with obscure margin. Clinically, this lesion will be indistinguishable from MA if the lesions are tiny and have the shape of a dot or blot. As shown in Figure 2.

- **Hard Exudate (HE)** is the cholesterol accumulation after the leaking of MA. HE is irregularly shaped, a variety of size, with a yellow or white color, and it often spreads surrounding the leaking microaneurysms in a circular formation. Furthermore, as a result of the association between HE and edema, the patients with this lesion have the risk of developing a complication disease named macula edema,

which occurs when edema appears adjacent to the macula and will cause blurry vision. As shown in Figure 2.

- **Soft Exudate (SE)**, referred to as a cotton wool spot, manifests as a deposition of deceased neuronal cells resulting from ischemia consequent to capillary closure. Moreover, this lesion will be indicative of a complication disease called macula ischemia if it appears in close proximity to the macula. Generally, this lesion appears as a fluffy, whitish, or cottony-like spot. As shown in Figure 2.

- **Intraretinal Microvascular Anomalies (IRMA)** manifests as anomalous branching or dilation of extant blood vessels, specifically capillaries, within the retinal. This phenomenon is instigated by hypoxic or ischemic conditions affecting the capillaries, thereby inducing a restructuring of pre-existing blood vessels or the formation of new blood vessels through endothelial cell proliferation. These newly developed blood vessels serve as conduits for the supply of essential resources to capillary non-perfusion regions. Typically, these blood vessels exhibit a characteristic pattern of crossing over each other, while avoiding intersections with major veins or arteries. As shown in Figure 2.

- **Venous Beading (VB)** delineates a critical manifestation within the retina wherein the elasticity and localized areas of major retinal vein walls is compromised. Resulting in the distortion of their inherent alignment and morphology, transitioning from a cylindrical string to a sausage-like string. Physically, IRMA and VB always occur in the late stages of non-proliferative disease. Thus, this occurrence serves as compelling evidence indicative of progression to proliferative disease. As shown in Figure 2.

- **Neovascularization (NV)** appears during the initial stages of proliferative disease. This lesion arises in response to localized hypoxia within the retina, prompting the secretion of vascular endothelial growth factor (VEGF). VEGF, a protein known to induce angiogenesis, stimulates the formation of new blood vessels within the retinal tissue. The apprehension surrounding this lesion pertains to its growth behavior and vascular characteristics, as the neovascular structures are inherently delicate and extend on top of the retinal surface, leading them susceptible to leakage and hemorrhage. Additionally, these vascular formations exhibit a

propensity to traverse one or multiple retinal veins and arteries, often presenting a florid, blossom-like appearance, similar to a flower bud. As shown in Figure 2.

- **Fibrous Proliferation (FP)** becomes evident subsequent to the emergence of NV, as a protective response within the oculus. This lesion is rooted in the imperative to strengthen the newly formed blood vessels. Accordingly, the eye initiates the construction of a supportive structure, wherein fibrous tissue interlaces proximal to these new blood vessels. Typically, the fibrous tissue exhibits a white color and displays a strong affinity for adherence to both the retinal tissue and the new blood vessels. Consequently, the fibrous adhesions pose a potential risk of accidentally tearing the neovascular structures, thereby leading to hemorrhages within the retinal space. Conversely, if these fibrous adhesions exert sufficient traction on the retina, they will have the propensity to induce retinal detachments. As shown in Figure 2.

- **Preretinal and Vitreous Hemorrhage (PRH, VH)** ensue when the fragile neovascular formations experience blood leakage or disruption due to the adhesion exerted by fibrous proliferations. The nomenclature of these two lesions distinguishes them based on their respective anatomical locations. In the event that blood permeates the potential space between the retinal tissue and the internal limiting membrane, which lines the surface of the retina, it is termed a preretinal hemorrhage. Conversely, if the blood permeate into the vitreous or the posterior chamber, it will be denoted as a vitreous hemorrhage. As shown in Figure 2.

Currently, diabetic retinopathy is classified into two stages: Non Proliferative Diabetic Retinopathy (NPDR) and Proliferative Diabetic Retinopathy (PDR) and into five categories including Normal, Mild, Moderate, Severe, Proliferation, following the International Clinical Diabetic Retinopathy (ICDR) severity scale [(Wilkinson, 2003; Gulshan, 2016)]. The details of these categories are explained below:

Class 1: No diabetic retinopathy

None of the above mentioned lesion appears in the patient, following examination guidelines published jointly by the International Council of Ophthalmology (ICO) and the American Diabetes Association (ADA) in 2018. Furthermore, a diabetic patients with no diabetic retinopathy have a less than 1 percent chance to become

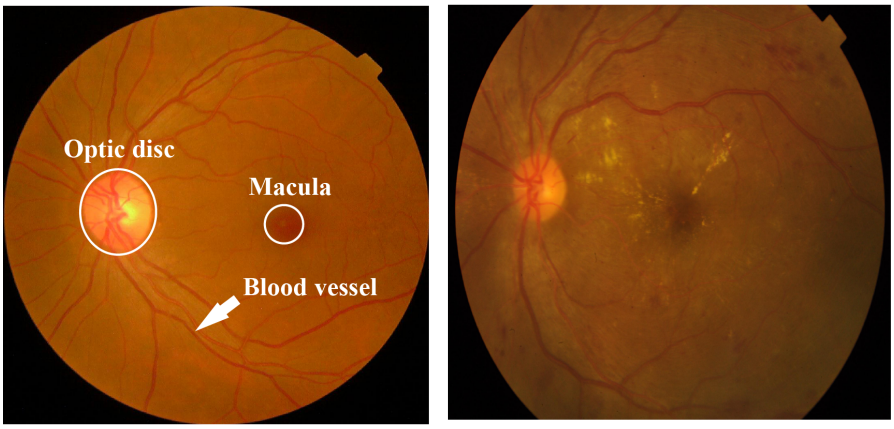

Figure 1 The retinal image comparison of normal and DR.

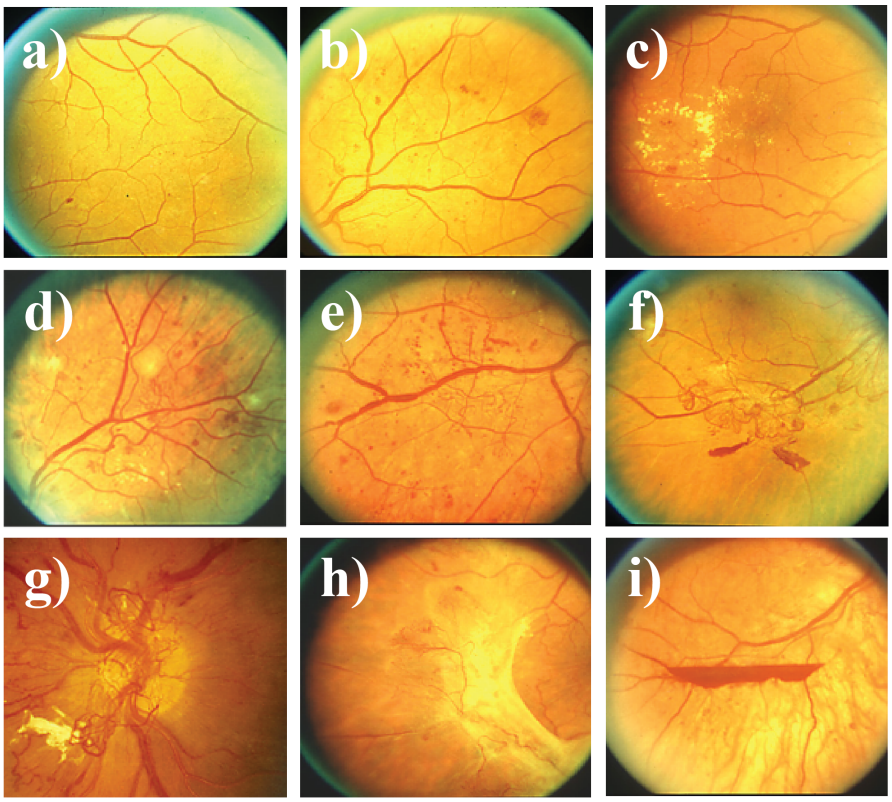

Figure 2 Diabetic retinopathy (DR) lesions are categorized based on the progression of lesion development; ranging from mild to severe stages. a) Microaneurysms (MA); b) Retinal Hemorrhage (HM); c) Hard Exudates (HE); d) Soft Exudates (SE); e) Intraretinal Microvascular Anomalies (IRMA) and Venous Beading (VB); f) and g) Neovascularization (NV); h) Fibrous Proliferation (FP); i) Preretinal and Vitreous Hemorrhage (PRH, VH) (Barbara Davis Center for Diabetes School of Medicine, 2024).

a PDR in the next four year and have to re-examination in 1-2 years based on American Academy of Ophthalmology (AAO) guidelines (Wong, 2018).

Class 2: Mild NPDR

In this class, the only discernible lesions have only microaneurysms in the diabetic patients. The recommended re-examination schedule is contingent upon the nation resource setting, occurring either every 6–12 months or 1–2 years. Additionally, over the subsequent four years, diabetic patients falling within this categories bear a chance of less than 5 percent for the development of PDR.

Class 3: Moderate NPDR

In the moderate NPDR class, the examination results consist of microaneurysms, dot and blot hemorrhages, hard exudates, soft exudates, or venous beading, but less than the 4:2:1 rule of severe NPDR, explained in the next topic. Indispensably, the patients in this category require a referral to an ophthalmologist and the recommended re-examination schedule, occurring either 3-6 months or 6–12 months, depending on the nation's resource setting.

Class 4: Severe NPDR

Severe NPDR pertains to a categories of diabetic retinopathy patients who follow to the 4-2-1 lesion rule:

- Each quadrant (Niemeijer, 2009) of the retina exhibits 20 or more intraretinal hemorrhages
- 2 or more quadrants exhibit the definite venous beading (VB)
- 1 or more quadrants exhibit the intraretinal microvascular abnormalities (IRMA)
- No signs of proliferative retinopathy

Patients in this category require a referral to an ophthalmologist and the recommended re-examination schedule, occurring less than 3 months. Individuals diagnosed with severe NPDR face a 17 percent chance of progressing to high-risk PDR within one year. Additionally, the chance increases to 40 percent for the development of high-risk PDR within three years.

Class 5: PDR

This is the most advanced category of the disease. In this category, hypoxic conditions stimulate the emergence of new, delicate, and anomalous blood vessels along the retinal wall. Consequently, the examination is imperative to detect one or more of the PDR lesions, namely neovascularization, fibrous proliferation, and vitreous or preretinal hemorrhage. Indispensably, the patients require a referral to an ophthalmologist, and the re-examination schedule occurs in less than 1 month because this category can cause irreversible damage to vision, leading to blindness. Figure illustrate the five categories Figure 3.

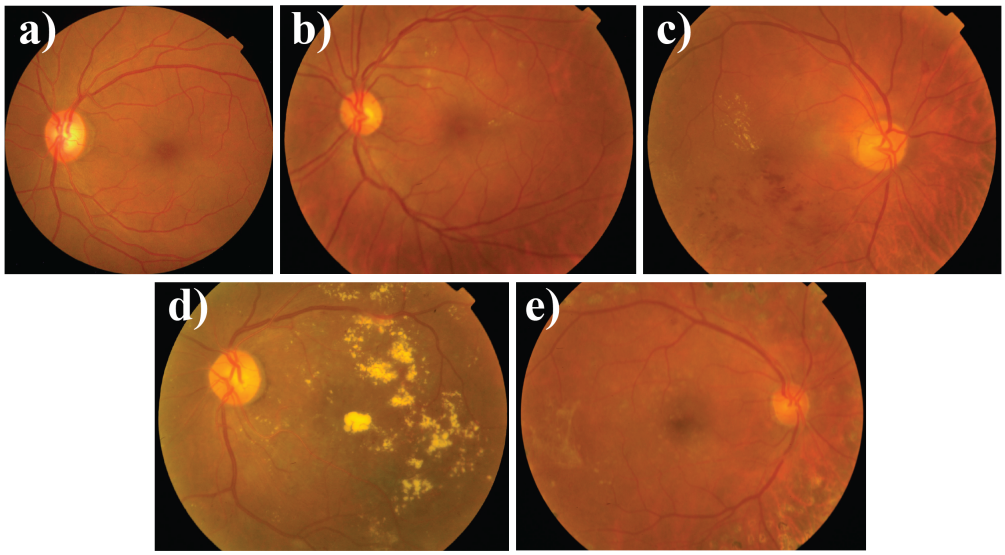

Figure 3 The figure illustrates the various severity levels of DR in the patient's retina: a) No DR; b) Mild NPDR; c) Moderate NPDR; d) Severe NPDR; and e) PDR (Karthik, 2019).

5.2 Macula and optic disc detection

The diagnosis of ocular disease by an ophthalmologist assesses not only the presence of lesions within the retina but also necessitates a comprehensive analysis of crucial anatomical ocular structures, namely the macula, optic disc, and blood vessels. This evaluation is imperative for the ophthalmologist to derive relevant information, enabling precise determinations regarding the severity of the ocular disease. The analysis of the macula and optic disc are particularly crucial, as the macula, housing a masses of cone cells, plays a crucial role in human color perception. Concurrently, the optic disc serves as the departure point for ocular structures, a collection of optic nerve fibers and blood vessels. Therefore, timely and precise detection of abnormalities in this region can significantly enhance the precision of disease diagnosis and abate the risk of developing severe stages or blindness. Currently, there are two categories of approaches that address the issue of locating the macula and optic disc: conventional approach, also known as the handcrafted feature approach, and novel approach, also known as deep learning approach.

5.2.1 Conventional detection approaches

In this approach, algorithm developers are obligated to manually analyze and extract the features of images for subsequent application in the detection of the macula and optic disc. Furthermore, these below anatomic information are leveraged to locate the macula (Sigut, 2023).

- the macula is circular and vessel-free area
- the macula is darker than surrounding area
- the distance between macula and the center of optic disc is approximately 2.5 times of optic disc diameters

Primarily, the methodology within this approach commences with the application of image processing techniques to identify the optic disc, a structure characterized by its large size and brightness. Subsequently, the acquired anatomical information pertaining to the macula is employed to declare the Region of Interest (ROI). This ROI is then utilized as a basis for detecting the macula within the retinal image. The optic disc detection process typically involves two stages: image pre-processing, and optic disc detection. Firstly, image pre-processing enhances the quality of the

images to facilitate subsequent stages. Various techniques are employed in this stage, including image enhancement to improve contrast and luminosity, thereby making the optic disc more distinguishable from the background [(Sinthanayothin, 1999; Deka, 2015; Medhi, 2016; Kamble, 2017; Palanisamy, 2023)]. Additionally, noise and background removal techniques are utilized to eliminate bright lesions and background elements that could interfere with the detection algorithm, potentially leading to false detection [(Sekhar, 2008; Usman Akram, 2010; Mvoulana, 2019)]. Furthermore, resizing the images is commonly performed to reduce computational resources and accelerate processing speed [(Sinthanayothin, 1999; Zheng, 2014; Palanisamy, 2023)]. Lastly, optic disc detection involves employing various feature extraction techniques to extract valuable information and locate the optic disc. (Sinthanayothin, 1999) propose the method by segmenting the image into 7 patch images, each sized 80 x 80 pixels, with the objective of identifying the optic disc based on adjacent pixels, showing the highest intensity variation. In [(Sekhar, 2008), (Usman Akram, 2010)], the methods select the location of the optic disc candidate as the pixel with the highest intensity on a gray fundus image, followed by the utilization of circular Hough transform for optic disc detection. Furthermore, vascular system is leveraged to detect the optic disc location, as the vascular is high density at the optic disc area, thereby various technique attempt to determine the vascular before proceeding with optic disc detection [(Welfer, 2011; Medhi, 2016; Chalakkal, 2018; Fu, 2022)]. Mostly, these techniques leverage the dense vascular structure as a reference point and subsequently employ circular Hough transform operations or region-based active contour models (Zheng, 2014) for precise optic disc detection. Additionally, the object detection technique known as template matching is employed to address this task due to the outstanding characteristics of the optic disc, which frequently exhibit a large size, brightness, and circular shape. In [(Chalakkal, 2018), (Mvoulana, 2019)], methods create three templates of the optic disc corresponding to the three channels of the RGB image. To ensure that these templates encapsulate rich information about the optic disc, they decide to create them by averaging N optic disc images. However, in (H. Yu, 2012), a method is proposed that utilizes template matching and a voting approach to enhance the accuracy of optic disc detection. This approach involves creating various templates resembling optic discs and subsequently matching these templates to the target image to detect the

optic disc location, which corresponds to the pixel containing the highest correlation value.

Similar to optic disc detection, the process of detecting the macula also involves two stages: ROI creation and macula detection. ROI creation is a essential initial stage because the macula typically appears as a small dark spot within the vessel-free area. Moreover, its shape may become unclear, particularly when the image is captured under inappropriate luminosity conditions or is affected by ocular diseases. As a result, leading to confusion between the macula and small dark lesions such as microaneurysms or retinal hemorrhages. Hence, the ROI creation stage plays a crucial role in assisting the macula detection algorithm by scoping the search space, leading in effective locating. The majority of proposed methods rely on anatomical knowledge regarding the relationship between the diameter and size of the OD and the location of the macula. Typically, these methods create the ROI by delineating a region that is located away from the OD's center, within a range of 1.5 to 3 times the diameter of the optic disc [(Sinthanayothin, 1999; Welfer, 2011; Chalakkal, 2018; Fu, 2022; Dinç, 2023; Palanisamy, 2023)]. The ROI, often created based on this knowledge, is typically in the shape of a rectangle. However, some methods deviate from the rectangular shape, which create the ROI in the form of a cone-like or half-circle shape by conducting circular scanning around the OD's center, ranging from -30 degrees to 30 degrees or -90 degrees to 90 degrees, respectively [(Sekhar, 2008; Zheng, 2014)]. Furthermore, the vascular structure serves as significant evidence to indicate the region of the macula. In [(Deka, 2015; Medhi, 2016)], these methods utilize the anatomical knowledge that the macula is always located in the vessel-free area. As a result, they employ techniques such as Discrete Wavelet Transform (DWT) and morphological operations to extract the blood vessel structure. Subsequently, these blood vessels are divided into three horizontal strips, and the ROI is selected from the strip that provides the least total number of blood vessels. Interestingly, in (Fu, 2022), the method utilizes the vascular structure to create a blood vessel vector model, which can roughly locate the macula. This model is then utilized for delineating the ROI. Following the ROI creation, various image processing techniques are employed to obtain the macula location. Ultimately, Following the ROI creation, various image processing techniques are employed to determine the location of the macula. Morphological operations, such as dilation,

erosion, top-hat, and bottom-hat, are applied to enhance the outstanding of the macula region against the background [(Sekhar, 2008; Welfer, 2011; Zheng, 2014; Deka, 2015; Chalakkal, 2018)]. Subsequently, thresholding algorithms, such as the brute-force or Otsu algorithm, are utilized to segment the macula [(Medhi, 2016; Chalakkal, 2018; Fu, 2022; Palanisamy, 2023)]. Additionally, in [(Sinthanayothin, 1999; Nayak, 2009)], template matching technique is introduced to determine the macula location by using the inverse gaussian function as a template and the normalized correlation coefficient function as a correlation function.

In summary, conventional approaches mostly utilize image enhancement techniques such as Contrast Limited Adaptive Histogram Equalization (CLAHE) to enhance image quality and leverage the analysis of vascular structures to roughly locate the optic disc (OD). However, relying solely on vascular structures for OD detection might not be practical in real-world scenarios, as this approach is sensitive to ocular diseases that can modify vascular structures, such as IRMA, VB, NV, etc., leading to inaccurate estimations of vascular density and OD position. Therefore, the application of template matching might offer a more practical solution, as the shape and luminosity of the OD are relatively consistent and only slightly affected by these diseases. Due to the sensitivity of vascular structures, creating the ROI based on anatomical knowledge of the macula appears to be a more sensible approach compared to relying on vascular structures such as the vessel-free area and the adaptive parabola model. Moreover, thresholding and morphological operations emerge as the most commonly utilized techniques for segmenting and localizing the macula, delivering appropriate results across various proposed methods. Ultimately, conventional approaches notably demonstrate the advantage of not necessitating extensive data for parametric tuning or creating the detection model.

5.2.2 Novel detection approaches

In this approach, algorithms are developed using deep learning techniques, which can captivate developers due to the versatility of deep learning in tasks namely feature extraction, enhancement of feature quality, and detection of ocular structures, including the optic disc and macula. As a result of this outstanding approach, various methods are emerging based on it. Initially, methods introduce convolutional neural network (CNN) architectures to tackle ocular detection.

For instance, (Tan, 2017) proposes a custom CNN with 125k learnable param-

eters, setting the stage for this trend. Subsequently, subsequent methods aim to enhance detection performance by dividing the task into two steps: a coarse step and a fine step. In (Sedai, 2017), the VGG-16 network is employed to roughly detect the macula location in the coarse step, followed by using a custom CNN to precisely locate the macula. Moreover, in (Al-Bander, 2018), a custom CNN is utilized for coarse detection of both the OD and macula, and then two custom networks are employed for fine detection. In (Y. Huang, 2020), the region proposal network (RPN) is introduced to extract features of the OD, subsequently leveraging these features to determine the OD's location using a fully connected neural network (FCNN). Finally, this method utilizes the detected OD to create the ROI of the macula, which is then passed to three additional custom networks for macula detection. In [5], a three-stage network is proposed for macula localization by using VGG-19 network as a backbone. Since the emergence of U-net in 2015 (Ronneberger, 2015), it has garnered significant attention from retinal deep learning researchers in leveraging the U-net, particularly localization and segmentation. For example,[(Hasan, 2021; Bhatkalkar, 2021)] utilize U-net as a foundational network for OD and macula detection, enhancing detection performance by modifying residual skip connections and employing gaussian heatmaps as labels for training instead of exact location coordinates. Moreover, attention-based networks like ViT, DeiT, and Swin have shown superior performance in encoder tasks, leading to their adoption in this domain. In [(Song, 2022; H. He, 2023)], attention networks are utilized to encode fundus images, with U-net in the up-sampling phase, serving as the decoder. Additionally, to improve the performance of the transformer-Unet (Transunet) architecture (Chen, 2021), the approach integrates a vascular segmentation network into the encoder in (Song, 2022). Furthermore, in (H. He, 2023), vessel-pretrained weights are utilized for the encoder, coupled with multitasking during training to further enhance performance. Indeed, the use of raw datasets for training in this approach poses a challenge, mainly due to the relatively small size of public datasets, typically containing less than 1200 images each. Consequently, the significance of augmentation becomes apparent, as it serves as a method to increase dataset diversity and quantity by leveraging existing data. Commonly, augmentation techniques include horizontal and vertical flips, random contrast enhancement, random color distortion, and the addition of Gaussian noise.

In summary, the proposed methods in this approaches commonly leverage CNN namely ResNet-50, VGG-16, VGG-19, and custom CNN as feature extractors. The incorporation of coarse and fine stages is a prevalent strategy, enhancing both localization and segmentation performance. Initially, there is a trend toward increasing these stages, anticipating improved network intelligence with increased complexity. However, with the introduction of the U-net architecture in later phases, developers shift their attention towards incorporating new features into theirs detection network, such as blood vessels segmentation and multi-tasking training, instead of increasing network complexity. The novel architectures demonstrate superior performance over traditional approaches in both localization and segmentation. Nevertheless, this outperformance frequently results in requiring for a massive amount of data for effective training.

5.3 Diabetic retinopathy classification

Diabetic retinopathy classification is a critical task that classifies patients into different stages according to a predefined protocol. The promptness and accuracy of this classification play a significant role in clinical detection and treatment procedures. A high-performance classification system can accelerate the treatment process, thereby abating the likelihood of progression to severe stages in patients. Typically, there are two classification types: binary classification, which classifies between normal and abnormal, and multi-classification, which classifies the severity level of the patient according to the standard protocol, 5.1. In this literature review, we demonstrate only multi-classification because it aligns with our research.

5.3.1 Multi-classification

The multi-classification of DR based solely on images presently remains a challenging task due to the complexity of distinguishing between severity levels, particularly within the NPDR group where distinctions are slight. However, emerging technologies at the frontier, denoted as deep learning, offer promising avenues to break through this challenge because deep learning architectures, exemplified by convolutional neural networks (CNNs) or vision transformers (ViTs), possess the capability to extract and understand image features effectively. Consequently, researchers in the DR field frequently turn to deep learning methodologies to address the multi-classification challenges. Similar to common deep learning methodologies, the pro-

cess for DR multi-classification entails pre-processing and DR classification stages. Pre-processing plays a crucial role in this task due to the imbalanced nature of the datasets used for DR classification, arising from the uneven distribution of severity levels among patients, with the majority falling into the normal and mild classes, Figure 9. Consequently, pre-processing in this context heavily emphasises image augmentation to generate new images from existing images. Techniques include random horizontal and vertical flipping, image rotation, random cropping, color jitters, the addition of Gaussian noise, and random brightness and contrast adjustments are commonly employed. Additionally, methods often incorporate Contrast Limited Adaptive Histogram Equalization (CLAHE) to enhance image contrast and resize each image to match the input size of the networks.

In DR classification, various methods have been proposed to enhance classification performance, with ensemble networks being one prominent approach. For instance, in (W. Zhang, 2019), a method employs three distinct CNN models—Inception-V3, Xception, and Inception-ResNet-V2—for feature extraction, supplemented by a customized deep learning block called SDNN for severity level classification. Additionally, in (Qummar, 2019), an ensemble of five different networks is utilized for this task. Notably, this method evaluates the ensemble network's performance across diverse types of datasets, including imbalanced, up-sampled, and down-sampled datasets, finding that up-sampled datasets can significantly improve classification performance compared to imbalanced datasets during training. In another approach detailed in (Shakibania, 2023), an ensemble network combines ResNet50 and EfficientNet-B0, with the feature vectors from these networks concatenated and passed to fully connected neural network layers for classification. Moreover, this method addresses the imbalance issue by filling minority classes with data from other datasets. Finally, in (Parsa, 2024), the method use the feature vector of VGG-16 network, ResNet 50, and Alexnet for classification. Additionally, the development of DR classification necessitates a massive amount of labelled data to enable the network to learn the complicated patterns of the task and overcome data imbalance issues. Nevertheless, procuring a large labelled dataset is a challenge, demanding considerable investments in human effort, costs, and time for labelling. Conversely, large, unlabeled datasets are more readily available. Hence, to leverage these numerous unlabeled datasets, a semi- and self-supervised learning

approach has been proposed in DR classification tasks to address these challenges. In (Islam, 2022), the approach introduces supervised contrastive learning to train the network, utilizing contrastive learning to extract and learn the mutual information of the DR dataset, while supervised learning is employed for classification. This method enables contrastive learning to capture features from both labelled and unlabelled data. Additionally, in (Tusfiqur, 2022), two networks, S-Net for lesion segmentation and G-Net for DR grading, are proposed. Notably, to enhance S-Net's segmentation capabilities and achieve accurate lesion segmentation, three discrimination networks are introduced to guide S-Net through adversarial learning. Subsequently, the segmentation results from S-Net, along with the fundus image, are utilized to train G-Net for classification. In (Ouyang, 2023), a self-supervised learning framework called SimCLR is introduced to train ResNet 50 using an unlabelled dataset instead of relying on transfer learning, which typically relies on labelled datasets. The trained network is then fine-tuned with labelled DR data. Furthermore, in (Parsa, 2024), ensemble networks comprising three distinct networks are employed, trained using a BYOL framework, and lastly fine-tuned with labelled data to enhance classification performance. Ultimately, the development of a DR classification network using complicated and interesting methods is good. Nevertheless, DR classification, being a subset of medical classification tasks, demands meticulousness and conciseness at every stage because the severity level predicted by these networks for each patient holds significant implications for their future health and well-being. Therefore, ensuring accuracy and reliability in DR classification is pivotal, as it directly influences clinical decisions and patient outcomes. As a result of these reasons, explainable AI (XAI) has also been developed in the field of DR classification, as XAI allows us to comprehend and trust the prediction results of the networks. Typically, the explanation approach in these methods leverages score maps generated by computing attention maps of neural layers to explain the decision-making process of these networks. These score maps frequently assign higher values to significant locations in the image, which commonly correspond to blood vessels, lesions, or critical ocular structures that can serve as indicators of the DR severity level present in the image. For instance, in (A. He, 2020), the category attention block (CAB) is introduced, designed to address the imbalance issue in various diabetic retinopathy datasets. The CAB is a plug-and-play module that

commonly connects to the last layer of the backbone In order to enhance class attention prior to feeding into the classifier. Interestingly, CAB's role is to enhance the attention of each category, resulting in being able to utilize the CAB's attention map for network explanation purposes. In (La Torre, 2020), a novel pixel-wise score propagation method is introduced, enabling the assignment of scores to individual pixels of the input image. These scores explain each pixel's contribution to the final classification results, thereby facilitating network debugging and providing diagnostic assistance for ophthalmologists. Furthermore, in (R. Sun, 2021), the method proposes a network architecture that integrates ResNet 50 as a backbone and incorporates a transformer network for classification purposes. A notable feature of this method is its utilization of the attention mechanism inherent to the transformer architecture. This mechanism facilitates the creation of a lesion-aware tensor, offering versatility for both classification and diagnostic assistance purposes. In (Quellec, 2021), an explainable classification network is proposed, wherein intermediate layers are trained using diverse lesion maps, including microaneurysms, exudates, and hemorrhages. This approach enables the attention maps of these layers to describe the lesions associated with the classification outcome. As a consequence of this training strategy, the attention map produced by this network is expected to be more reliable compared to those generated by networks trained solely on classification data. Additionally, in (Li, 2022), the proposed method is the lesion-attention pyramid network (LAPN), which consists of three sub-networks designed to process different input sizes. Interestingly, LAPN utilizes the lesion attention map generated by the first sub-network as a guiding factor for predicting the DR severity level, resulting in enabling the use of the attention map from the first sub-network for result explanation.

In conclusion, diabetic retinopathy classification often relies on ResNet and Inception networks for feature extraction, coupled with custom classifiers for classification tasks. High-performance methods frequently employ large or ensemble networks to achieve superior results. Moreover, to leverage unlabeled datasets, self-supervised frameworks like SimCLR and BYOL are employed to improve feature extraction. Additionally, efforts are made to explain network decisions by utilizing attention maps, which help understand the network's decision-making process or the input data's contribution to classification results. These attention mechanisms can

be categorized into two groups: module-based, designed as plug-and-play blocks to enhance classification performance and network interpretability, and activation-based, generating attention score maps from intermediate layer activation solely for explanatory purposes. Beyond proposals in classification development, addressing the imbalanced issue stands out as an intriguing aspect. Various methods have been proposed to address this challenge, including up-sampling minority classes, augmenting datasets by filling them with data from other datasets, utilizing attention blocks for class balancing, and employing weighted loss functions during training to address the imbalance in dataset distributions. These approaches aim to ensure that models are trained effectively and accurately across all classes, despite variations in class sizes within the dataset.

5.4 Evaluation matrix

The evaluation matrix, or metrics, serves as a crucial tool for assessing the performance of a model. It plays a crucial role in the model development process, providing insights into the model's strengths and weaknesses. In this literature, we identify significant metrics involving with our research.

5.4.1 Confusion matrix

The confusion matrix serves as a representation of the model's performance, providing four potential outcomes: true positive (TP), where the positive class is correctly identified as positive; false positive (FP), or type 1 error, indicating the negative class is erroneously classified as positive; true negative (TN), correctly identifying the negative class as negative; and false negative (FN), or type 2 error, where the positive class is incorrectly classified as negative. Nevertheless, relying on these outcomes from the matrix is insufficient to measure the extensive performance of the model. Therefore, it is essential to integrate these outcomes with other evaluation matrices for a more extensive measurement of the model's performance. The confusion matrix shown in Figure 4 can demonstrate the principle components of the confusion matrix.

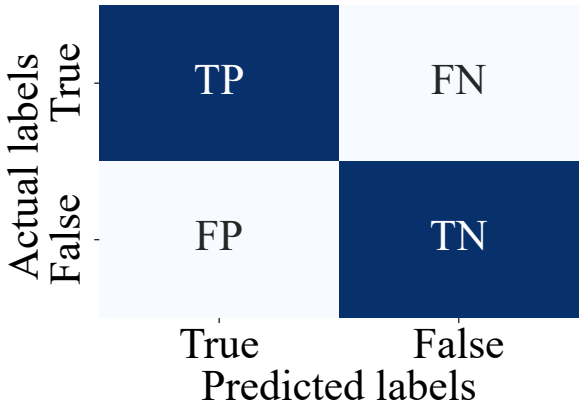

Figure 4 The confusion matrix.

5.4.2 Accuracy, Precision, and Recall

Accuracy measures the model performance by calculating the proportion of correct predictions over the entire prediction sample, as depicted in (Eq. 1). However, accuracy can be misleading when dealing with an imbalanced test set, as the majority class prone to dominate the results. Consequently, to address this issue, novel evaluation metrics are proposed, namely precision and recall. Precision measures the model's performance in accurately predicting positive samples, while recall measures the proportion of actual positives correctly identified. The equations for these metrics are expressed in (Eq. 2) and (Eq. 3), respectively.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

5.4.3 F1 score

The F1 score is introduced to address the precision-recall trade-off issue, which occurs because these metrics often cannot simultaneously improve. Consequently, when precision increases, recall tends to decrease, and vice versa. This trade-off creates a challenging decision point, as it is unknown where the optimal balance

lies between precision and recall. Therefore, the F1 score, representing the harmonic mean of precision and recall (Eq. 4), is employed to weigh these metrics and determine the optimal point. The F1 score values range between 0 and 1, with 1 indicating perfect harmony and 0 representing the worst harmony.

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \tag{4}$$

5.4.4 Receiver Operating Characteristic (ROC) curve

The ROC curve is a visualization of performance measurement for classification tasking across all possible thresholds. It is designed to evaluate classification performance without worrysome by thresholding. The curve consists of two essential components: the true positive rate (TPR), as expressed in (Eq. 5a), and the false positive rate (FPR), as expressed in (Eq. 5b). Additionally, Figure 5 provides an illustration of the curve's mechanism, where an increase in the threshold leads to a decrease in FPR followed by TPR. This is due to the reduction in positive class identification, resulting in an increase in false negatives (FN) and true negatives (TN), and conversely, a decrease in false positives (FP) and true positives (TP).

$$TPR = \frac{TP}{TP + FN} \tag{5a}$$

$$FPR = \frac{FP}{FP + TN} \tag{5b}$$

Furthermore, in Figure 6, we show the interpretation of the ROC curve in various patterns, which assist the reader in clearly understanding the ROC curve.

5.4.5 Area Under the Curve (AUC)

the interpreting performance through the visualization of the ROC curve can be challenging when faced with an unfamiliar curve, requiring professional expertise for accurate interpretation. To address this issue, the evaluation metric known as Area Under the Curve (AUC) is introduced. AUC provides a quantitative assessment that is simpler and more user-friendly than direct interpretation from the curve. The AUC represents the integral measurement of the area under the ROC curve. Ultimately, a higher AUC value corresponds to higher classification performance, Figure 7.

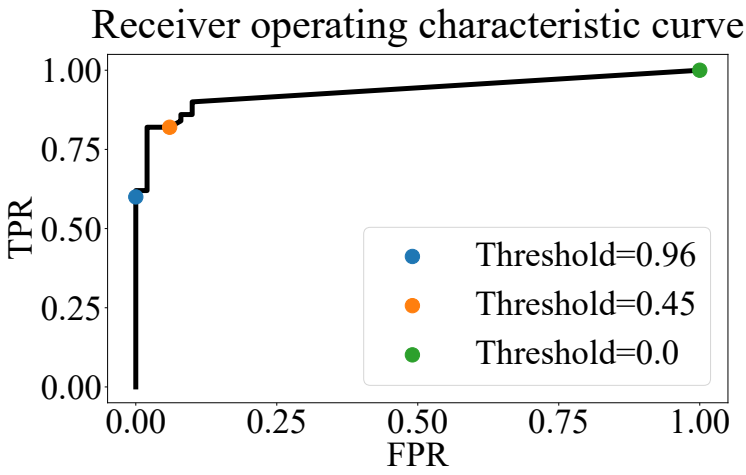

Figure 5 The ROC curve illustrate the trending of FPR and TPR as a function of threshold.

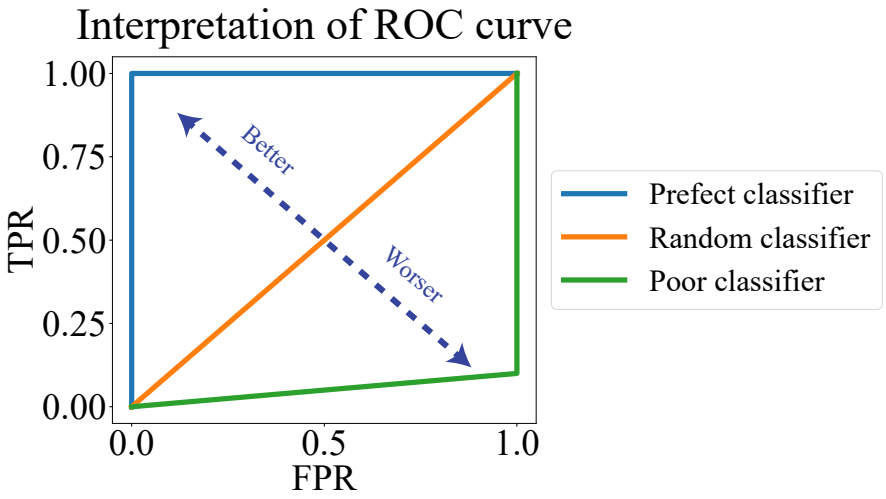

Figure 6 ROC curve interpretation by curve reading.

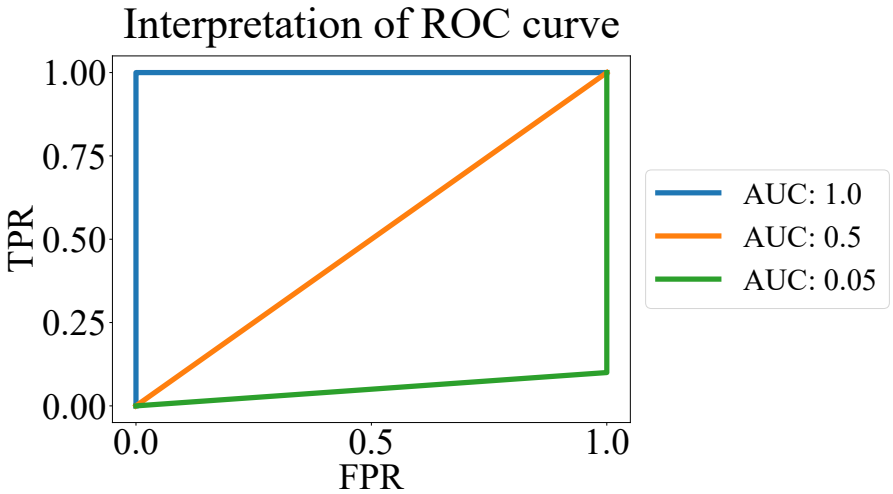

Figure 7 ROC curve interpretation by utilizing AUC.

5.4.6 Quadratic Weighted Kappa (QWK) coefficient

The quadratic weighted kappa coefficient is a measure of inter-rater reliability commonly used for ordinal categories. It generally assesses the degree of agreement between two raters. Notably, the quadratic weighted kappa coefficient is sensitive to large differences between the ratings given by the two raters. For instance, if the first rater assigns a value of 1 and the second rater assigns a value of 5, the resulting kappa coefficient will be worse compared to a scenario where the second rater assigns a value of 2. This sensitivity makes the quadratic weighted kappa a valuable metric for evaluating agreement between raters in ordinal category assessments. The kappa's equation expressed below:

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} o_{i,j}}{\sum_{i,j} w_{i,j} e_{i,j}} \tag{6}$$

$$w_{i,j} = \frac{(i-j)^2}{(C-1)^2} \tag{7a}$$

$$e_{i,j} = \frac{\sum_k o_{i,k} \sum_l o_{l,j}}{\sum_{i,j} o_{i,j}} \tag{7b}$$

Where $i, j, k, l \in \{0, 1, 2, \ldots, C-1\}$ and C is the number of classes. The κ is the kappa's coefficient, which is in the range of -1 and 1, where κ = -1, 0, and 1, denoted as complete disagreement, random agreement, and complete agreement, respectively. Furthermore, in the context of the quadratic weighted kappa coefficient, w, o, and e represents the weight matrix, defined as (Eq. 7a) in the quadratic case, observed rating matrix, which commonly is a confusion matrix in this research, and the expect rating matrix, defined as (Eq. 7b). These matrices have dimensions of $C \times C$, where C is the number of classes. In this research, the first rater corresponds to the predicted label, while the second rater corresponds to the ground truth label. Currently, there are the standard of agreement: $\kappa = 1$ representing the poor agreement, $0.01 \leq \kappa \leq 0.20$ slight, $0.21 \leq \kappa \leq 0.40$ fair, $0.41 \leq \kappa \leq 0.60$ moderate, $0.41 \leq \kappa \leq 0.60$ moderate and $0.61 \leq \kappa \leq 0.80$ almost perfect.

5.4.7 Euclidean Distance (ED)

The Euclidean distance is a metric that calculates the straight-line distance between two points in Euclidean space, utilizing their Cartesian coordinates (Eq. 8). In the context of a detection task, this distance represents the separation between the predicted location and the ground truth location in unit of pixels. A lower Euclidean distance indicates better model performance in accurately predicting the location.

$$d(X_1, X_2) = \|X_1 - X_2\| \tag{8}$$

Let $d(.,.)$ is the Euclidean distance and $X_1, X_2 \in \mathbb{R}^n$, n is a dimensions.

5.4.8 Intersection over Union (IoU)

The Intersection over Union (IoU) is a metric used to measure the similarity between two samples, typically bounding boxes in the context of object detection. The equation for IoU is as follows:

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{9}$$

Where A represents the ground-truth area and B represents the predicted area. The IoU value is a normalized metric ranging between 0 and 1. A high IoU value, close to 1, indicates a good prediction where the predicted area closely matches the ground-truth area in size and location. Conversely, an IoU value of 0 signifies an unacceptable prediction where there is no overlap between the ground-truth and predicted areas. Ultimately, the IoU indicates the degree of overlap between the predicted area and the ground-truth area.

5.4.9 Average Precision (AP) and Mean Average Precision (mAP)

Leveraging the IoU score alone to evaluate prediction performance in the context of object detection might not suffice, as classification is also a crucial aspect in real-world scenarios. Therefore, an evaluation metric that combines both object detection and classification has been developed, known as the precision-recall (PR) curve, illustrated in Figure 8. The PR curve, which plots recall against precision, is commonly utilized to visualize the prediction performance of object detection al-

gorithms. Additionally, the PR curve is threshold-independent, as each data point on the curve is generated by varying confidence thresholds, typically arranged from low to high confidence levels. Notably, the PR curve is particularly relevant in this context, where the true-negative (TN) sample represents the background, causing it to be difficult to define precisely because, in an image, everywhere except the ground-truth area is considered background. Therefore, utilizing precision and recall metrics that disregard the TN helps explain prediction performance, which is both sensible and acceptable. Nevertheless, adjusting the IoU threshold can alter the PR curve, rendering it challenging to compare prediction performance across diverse IoU thresholds. To address this issue, the average precision (AP) score, which integrates the area under the PR curve, is employed using (Eq. 10a). However, the AP score may differ slightly based on the quantity of data points in the PR curve, indicating that is a non-standard metric. To mitigate this, the 11-point interpolation approach is commonly used on the PR curve before computing the AP score, (Eq. 10b). Eventually, in multi-classification tasks, the mean average precision (mAP) score is calculated by averaging the AP scores across all classes. This provides a single value that represents the model's performance in both object detection and classification, as shown in (Eq. 10c).

$$AP = \int_{r=0}^{r=1} p(r)dr \tag{10a}$$

$$AP = \frac{1}{11} \sum_{r} p(r) \tag{10b}$$

$$mAP = \frac{1}{k} \sum_{i} AP_i \tag{10c}$$

Let r is a interpolated recall score, $r \in \{0, 0.1, 0.2, \ldots, 1\}$, $p(r)$ is a interpolated precision score as a function of interpolated recall score, k is a number of classes, $i \in \{1, 2, 3, \ldots, k\}$, and AP_i is the average precision score corresponding to i class.

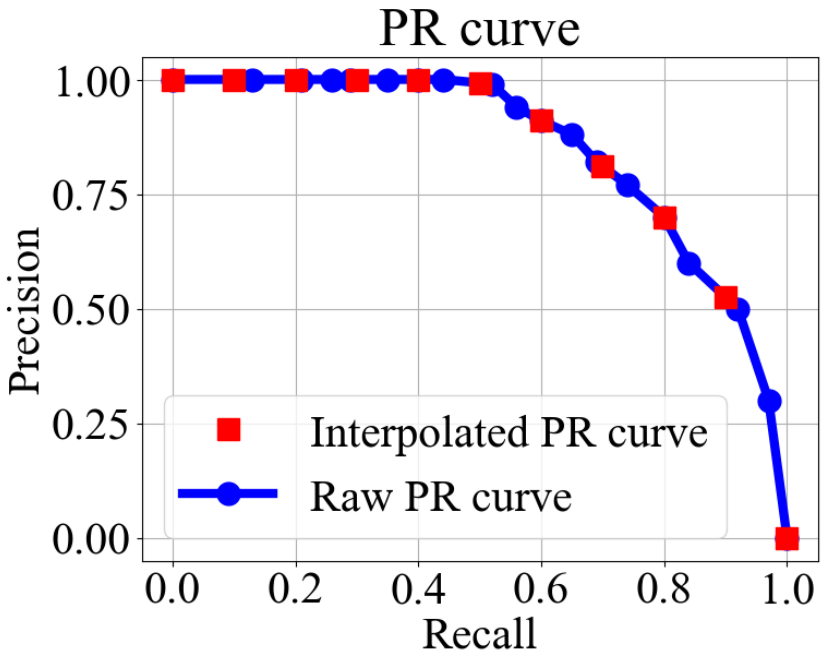

Figure 8 The figure illustrate the two PR curve including raw PR curve and interpolated PR curve, which is obtained by utilizing the 11-point interpolation approach.

CHAPTER 3

RESEARCH METHODOLOGY

In this chapter, we present the methodology for our research project, which is divided into two main parts: image screening and DR grading. The image screening process involves the detection of ocular structures, specifically the optic disc and macula, in retinal fundus images and the screening algorithm to identify the medically suitable retinal image. The detection is achieved through the application of a correlation filtering technique. The second part focuses on DR grading, where we utilize deep learning techniques to classify the severity level of diabetic retinopathy.

6.1 Dataset

- **IDRiD dataset (Porwal, 2018):** This dataset has been curated to support various challenges related to diabetic retinopathy (DR) such as lesion segmentation, DR and diabetic macula edema (DME) severity grading and localization of the optic disc and fovea center. The images in the dataset were acquired using a Kowa VX-10α digital camera with a 50° field of view (FOV) and a resolution of 4288x2848 pixels in jpg file format. Annotation of the dataset was implemented by individuals with expertise, including a master's student, a PhD student, and a medical expert, with validation conducted by a retinal specialist. The dataset contains a total of 516 images, which have been split into a training set (413 images) and a testing set (103 images) for both DR/DME severity grading and ocular structure localization. However, for the lesion segmentation task, there are only 81 images.

- **Messidor (Decencière, 2014):** The dataset was created between 2006 and 2008, collecting 1200 retinal fundus images from three ophthalmology departments in France. These images were captured using a 3CCD color video camera mounted on a Topcon TRC NW6 non-mydriatic retinography device with a 45-degree field of view (FOV). The images have resolutions of 1440x960, 2240x1488, and 2304x1536 pixels. The dataset serves multiple purposes, including diabetic retinopathy (DR) severity grading, where labels range from 0 to 4; lesion segmentation, which was manually segmented by expert ophthalmologists for 30 percent of the dataset; and macula localization, where ground-truth locations were labeled for 1136 images.

- **Private datasets :** The datasets were collected from the ophthalmology departments of two hospitals in Thailand: Maharaj Nakhon Ratchasima Hospital and Suranaree University of Technology Hospital (SUTH). Each dataset was annotated into two classes: positive, indicating medically suitable retinal images, and negative, indicating medically unsuitable retinal images. The first dataset comprises 428 images of varying sizes, with 328 positive and 100 negative cases. The second dataset consists of 610 images with a resolution of 2976×2976 pixels, including 337 positive and 273 negative cases. Due to the small size of dataset, we use the cross-validation testing to test the algorithm's performance .

- **APTOS 2019 Blindness Detection (Karthik, 2019):** This dataset has been established with the aim of developing a medical screening solution for Aravind Eye Hospital in India, specifically to address the needs detect and prevent DR disease among numerous rural patients. The most effective solution identified will be distributed with other ophthalmologists through the 4th Asia Pacific Tele-Ophthalmology Society (APTOS) Symposium. The dataset comprises a total of 3,662 images in the training set, as illustrated in Figure 9 and 1,928 images in the testing set. These images exhibit various resolutions and are stored in png format, distributed across five distinct categories. The dataset is publicly accessible on Kaggle at the following link: (https://www.kaggle.com/competitions/aptos2019-blindness-detection).

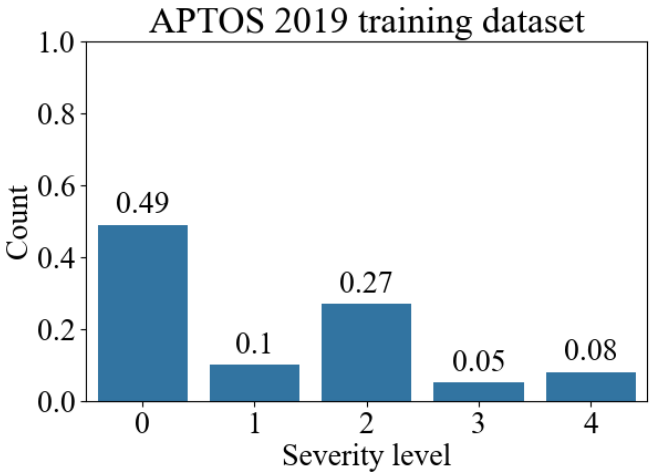

Figure 9 The class distribution of these datasets reveals that class 0 (no diabetic retinopathy) is the majority class, indicating an imbalance issue. The total number of images in the training sets of APTOS 2019 is 3,662.

6.2 Image screening

In this section, we implement 2 steps to achieve the screening of retinal images. The first step related to construct the templates for both the optic disc and macula. Then, we cooperate these templates and correlation filtering technique to detect the existence of optic disc and macula in retinal image. The second step is to screen the retinal image by using the screening algorithm to identify the medically suitable retinal image.

6.2.1 Optic disc and macula detection

Due to our screening algorithm relating to apply the correlation filtering technique for detect the existence of optic disc and macula in retinal image. Hence, we initially utilize the training set of the IDRiD dataset to generate reference templates (or mask) for both the optic disc and macula. Before generating the reference templates, we preprocess the fundus images by cropping the redundant dark areas surrounding the images. This preprocessing step is essential for enhancing image quality by eliminating background noise and improving overall consistency. Subsequently, all images are resized to a standardized resolution of 1280×1280 pixels, Figure 10.

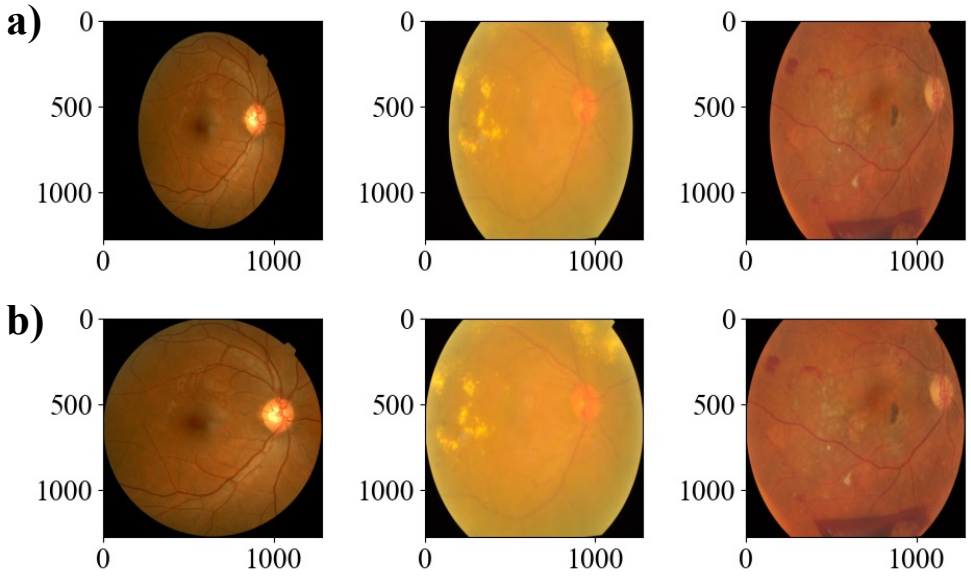

Figure 10 the cropped dark field image, a) without cropping, b) with cropping.

These preprocessing procedures ensure uniformity in both image dimensions and retinal region coverage, thereby supporting precise and consistent cropping during template generation. Then cropping the resized images around the optic disc and

macula. For optic disc, the cropping size is varied in the range of 150 to 400 with the step size of 50 in both symmetric and asymmetric shape. For macula, the cropping size is varied in the range of 100 to 350 with the step size of 50 in both symmetric and asymmetric shape. Currently, we use the cropping dimensions of 300x350 and 200x300 for the optic disc and macula. This cropping procedure was repeated N times, and the resulting cropped images were averaged to create reference templates for the optic disc and macula, respectively, define as,

$$\bar{T}(x,y) = \frac{1}{N}\sum_{i=1}^{N} T_i(x,y) \tag{11}$$

where $\bar{T}$ is a reference template and T_i is a cropped image number i^{th}. In the context of ocular detection, we deploy the correlation filtering method on the resized fundus image using an optic disc template to identify the optic disc.

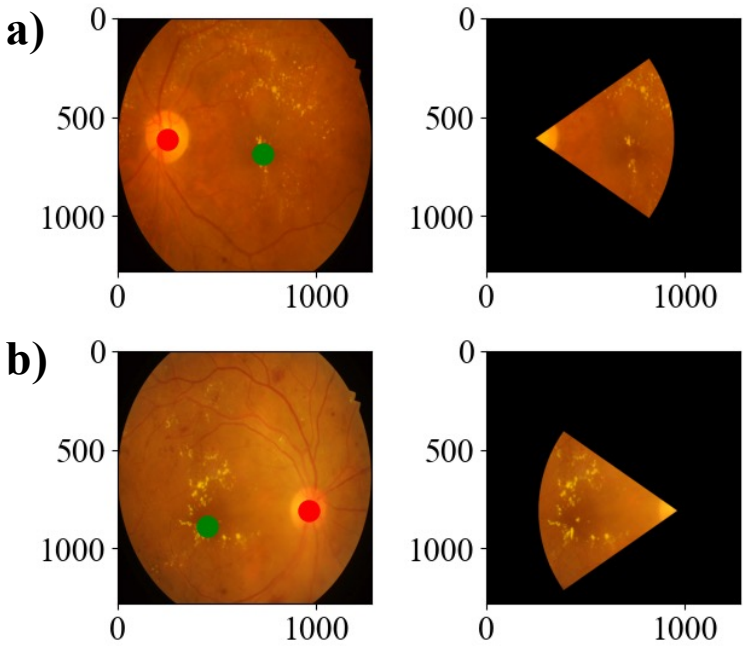

Figure 11 The figure shows the result of ROI cropping from both eyes, a) the left eye and b) the right eye. The macula and optic disc's ground-truth locations are represented by the red and green dots.

Following this, we delineate the region of interest (ROI) for macula detection by utilizing the determined location of the optic disc as a central reference point and selecting an area that locate to threefold the size of the optic disc within a range of -30 to 30 degrees (Sekhar, 2008), resulting in the ROI in a like-conical shape, illus-

trated in Figure 11. Subsequently, we employ the correlation filtering technique on the ROI image to locate the macula. Ultimately, thresholding the correspondence space is applied to determine the reliability of the maximum point, which serves as an indicator for locating the optic disc and macula. This process involves establishing a threshold value within the correspondence space, beyond which points are considered reliable. By applying thresholding, we can effectively identify the maximum point that accurately represents the location of these ocular structures. Furthermore, we currently use matching method, named normalized correlation coefficient (Eq. 12), to establish the correspondence space in the correlation filtering.

$$R(x,y) = \frac{\sum_{x',y'} \bar{T}'(x',y') \cdot I'(x+x',y+y')}{\sqrt{\sum_{x',y'} \bar{T}'(x',y')^2 \cdot \sum_{x',y'} I'(x+x',y+y')^2}} \tag{12}$$

where,

$$\bar{T}'(x',y') = \bar{T}(x',y') - \frac{1}{w \cdot h} \cdot \sum_{x'',y''} \bar{T}(x'',y'') \tag{13a}$$

$$I'(x+x',y+y') = I(x+x',y+y') - \frac{1}{w \cdot h} \cdot \sum_{x'',y''} I(x+x'',y+y'') \tag{13b}$$

Let I is a retinal fundus image.

6.2.2 Screening algorithm by rulebased

In this work, we use and develop a method proposed in (Şevik, 2014) to be suitable for our task in classifying the MURI and MSRI. We will call the retinal image MSRI if it can pass these criteria: first, the optic disc and macula must locate within the acceptance region, called R1; and second, the optic disc must locate outside the specific region, called R2. R2 is added to become a criterion because we don't need a nasal field of a retinal image. An example of these regions is illustrated in Figure 12. The R1 is defined by the lower and upper boundaries, following the below

equations:

$$L_{R1} = y_c - \epsilon \tag{14a}$$

$$U_{R1} = y_c + \epsilon \tag{14b}$$

where y_c is a vertical centerline of retinal image, ϵ is an adjusted parameter, L_{R1} is a lower boundary, and U_{R1} is an upper boundary. Moreover, the lower and upper boundary of R2 is computed as:

$$L_{R2} = y_c - \delta \tag{15a}$$

$$U_{R2} = y_c + \delta \tag{15b}$$

where x_c is a horizontal centerline of retinal image, δ is an adjusted parameter, L_{R2} is a lower boundary, and U_{R2} is an upper boundary. In this work, we use grid search on 50 images of private dataset to determine the optimal parameters, thereby ϵ and δ is 180 and 220 pixels, respectively.

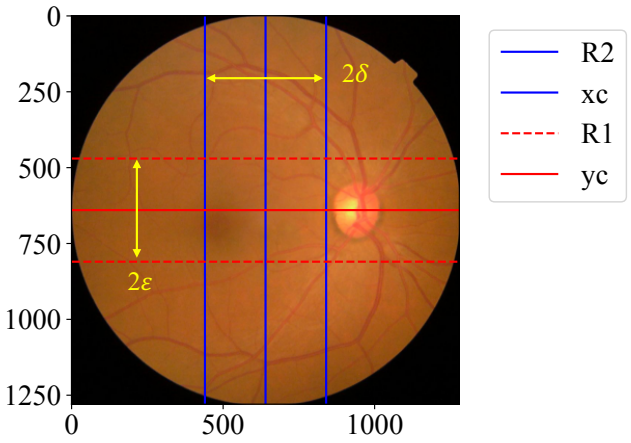

Figure 12 The acceptance region of R1 and R2.

6.2.3 Screening algorithm by Machine learning (ML)

We further investigate the screening algorithm by integrating machine learning (ML) techniques. Upon successful ocular detection, tabular data are produced, comprising features extracted from retinal images, including $OD_{confidence}$, $M_{confidence}$, $OD_{position}$, and $M_{position}$. The confidence scores, ranging from 0 to 1, indicating the certainty of each detection. $OD_{position}$ and $M_{position}$ represent the coordinates of the detected optic disc and macula, respectively, in the [x, y] format. Furthermore, we derive additional features, including $Dy_{intercept}$ and $Dr_{distance}$, based on these primary features. $Dy_{intercept}$ denotes the vertical deviation between the linear regression line of $OD_{position}$–$M_{position}$ and a horizontal reference line positioned at the midpoint of the y-axis, evaluated at x=0, Figure 13. The $Dr_{distance}$ quantifies the displacement between $M_{position}$ and the center of the retinal image, measured by Euclidean distance, as shown in Figure 13. A representative example of the tabular data is provided in Table 1.

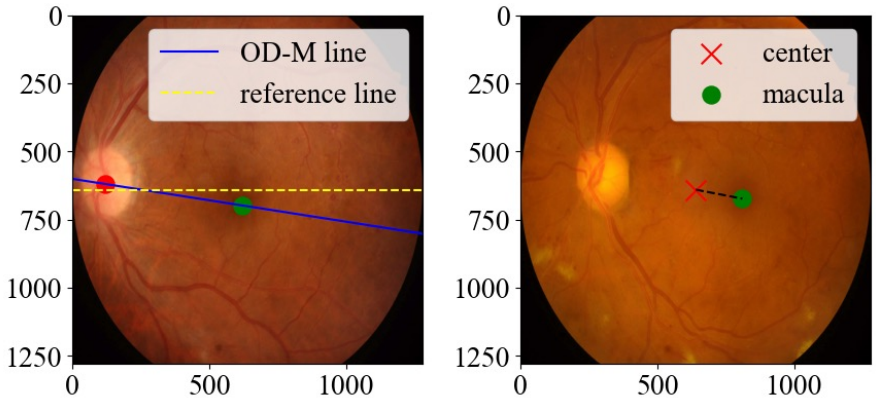

Figure 13 The figure illustrates components of the image that are used to calculate the $Dy_{intercept}$ and $Dr_{distance}$ features. (left) The $Dy_{intercept}$ is the deviation between OD-M line and reference line on the y-axis at x = 0. (right) The $Dr_{distance}$ is the distance from the center of the image to the center of the macula.

Table 1 The tabular data features for ML model.

$OD_{confidence}$	$M_{confidence}$	$OD_{position}$	$M_{position}$	$Dy_{intercept}$	$Dr_{distance}$
0.719	0.903	[934, 808]	[535, 821]	198.508	43577
0.835	0.774	[261, 512]	[753, 589]	168.848	15370
0.771	0.847	[229, 560]	[456, 811]	333.212	63097

We employ the tabular data extracted from the training set of our private

dataset to train a variety of machine learning (ML) algorithms, including logistic regression, decision tree, support vector machine (SVM), random forest, histogram gradient boosting, light gradient boosting, and XG boosting. To optimize model performance prior to comparison, we apply grid search methods to each ML algorithm. The evaluation of each hyperparameter configuration is conducted using stratified 5-fold cross-validation. ML models, pipelines, and cross-validation procedures are mainly implemented using the Scikit-learn package (Pedregosa, 2011).

6.3 Evaluation of image screening

In the context of ocular detection, we evaluate the performance of our algorithm using the testing set of IDRiD. Initially, we quantify the accuracy of our algorithm's predictions by measuring the error between the predicted location and the ground-truth location of ocular structures. This analysis is conducted using the Euclidean distance (ED), providing insights into the precision of our predictions at the pixel level. Additionally, we generate the negative sample, which is a background image, to measure the algorithm's performance using other evaluation metrics, namely precision, recall, and AP score. This background image resembles the fundus image but lacks the optic disc and macula Figure 14. This step is crucial as it allows us to accurately determine false positives (FP), which are difficult to define in fundus images containing all ocular structures. Therefore, in the quantitative measurement, we calculate the AP score to determine the algorithm's accuracy in bounding box prediction, higher AP score indicates better algorithm performance in object detection tasks. Moreover, we use a R-criterion score as mentioned in (Gegundez-Arias, 2013), to evaluate the performance of our macula detection on the Messidor dataset and due to the Messidor dataset comprising images of varying sizes, we apply different R values corresponding to each image size: R = 68 for images sized 1440 x 960, R = 103 for images sized 2240 x 1488, and R = 108 for images sized 2304 x 1536. In the context of screening, we employ various metrics, including the confusion matrix, false discovery rate (FDR), and recall of positive images, to assess the performance of our screening process on our private dataset. The false discovery rate is utilized to evaluate the proportion of negative samples that remain in the dataset after screening, while recall is used to measure the proportion of positive samples that are retained in the dataset post-screening.

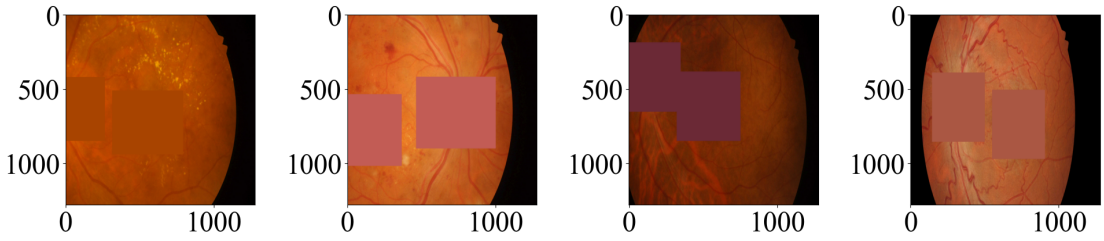

Figure 14 The example of background images which generate by covering the optic disc and macula by the mean value of fundus image.

6.4 DR grading

6.4.1 Data preparation

The APTOS 2019 dataset provide both training and testing set. Unfortunately, the testing set has no label, so we have to split the training set into training, validation and testing set with portion of 60%, 20% and 20%, respectively. Hence, the training set consists of 2,200 images, the validation set contains 731 images, and the testing set comprises 731 images. The APTOS 2019 dataset is highly imbalanced, with the distribution of severity levels as follows: 0 (No DR) - 1,083 images (49.2%), 1 (Mild DR) - 222 images (10.1%), 2 (Moderate DR) - 601 images (27.4%), 3 (Severe DR) - 117 images (5.3%), and 4 (Proliferative DR) - 177 images (8.0%).

6.4.2 Data augmentation and balancing

To address the imbalanced data issue, we implement oversampling technique, named Synthetic Minority Over-sampling Technique (SMOTE) (Chawla, 2002), to generate synthetic samples for the minority classes. However, our input data is not tabular data, but rather images. Therefore, we flatten the image into a vector with the size of HxWxC, where H is image height; W is image width; and C is number of image channel, and then, apply the SMOTE technique to generate new samples through interpolation between existing samples, as shown in Figure 15. Additionally, we also apply data augmentation techniques to enhance the diversity of our training set. These techniques include random horizontal flipping, random color jitter, random Gaussian blur, random adjust sharpness, random auto contrast, and random cropping. The augmentation process is performed on-the-fly during training to ensure that the model encounters a wide range of variations in the input data.

6.4.3 Architecture

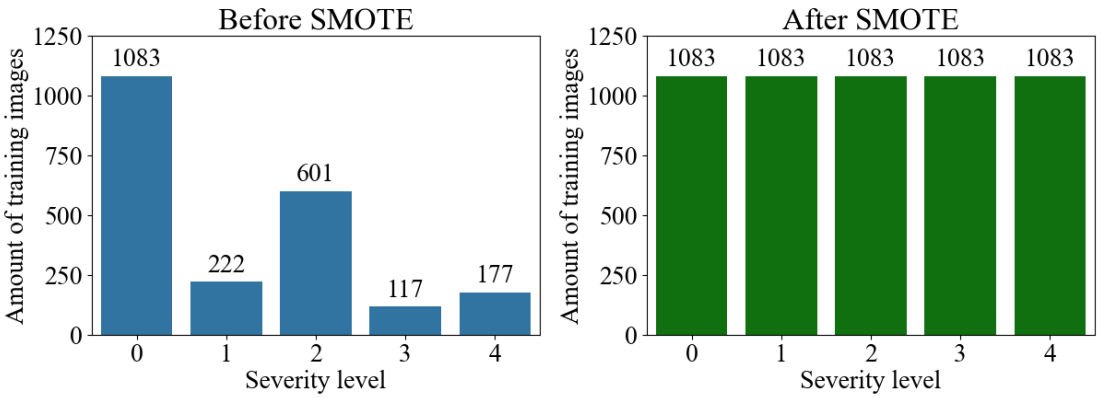

Figure 15 The distribution of the synthetic samples generated by SMOTE in the prior and posterior stage.

Our grading network comprises of two modules: a backbone model for feature extractor and a custom FCNN for classifier. The feature extractor is responsible for extracting high-level features from the input fundus images, and then, classifier leverages these features to predict the serverity level of diabetic retinopathy (DR). Our backbone model is Swin Transformer, which is a type of vision transformer that has shown promising results in various computer vision tasks. The classifier is FCNN, which contain 5 layers and number of node in each layer is 256, 128, 128, 64, and 5, respectively. Moreover, the essential detail of Swin Transformer is described in below.

6.4.3.1. Swin Transformer

The Swin Transformer (Liu, 2021) is a hierarchical vision transformer that is designed to address the challenge of processing high-resolution images efficiently in ViT-base model because Transformers is designed for language task, thereby adapting to vision task lead to the issue of computaional complexity, that is quadratic growing corresponding to image size. Therefore, the Swin Transformer introduces a Window-based self-attention mechanism, which partition the input image into non-overlapping windows and applies self-attention within each window, resulting in reformulating the complexity time to be linear, while the standard is quadratic complexity as a function of number of patch. As a result of this windowing scheme, Swin Transformer can understand the local context of image. Thus, to compromise the global context, Shifted Window-based self-attention is introduced, which shifts the windows between consecutive patch, allowing the model to capture global context in the image. This approach significantly reduces the computational cost while

maintaining the ability to capture long-range dependencies in the image. Moreover, the Swin Transformer architecture is characterized by its hierarchical structure, where the feature maps are progressively downsampled, allowing the model to learn multi-scale representations of the input image similar to CNN 16. Eventually, this design enables the Swin Transformer to balance efficiency and complexity resulting in state-of-the-art performance across a various fields of computer vision tasks, including image classification, object detection, and semantic segmentation.

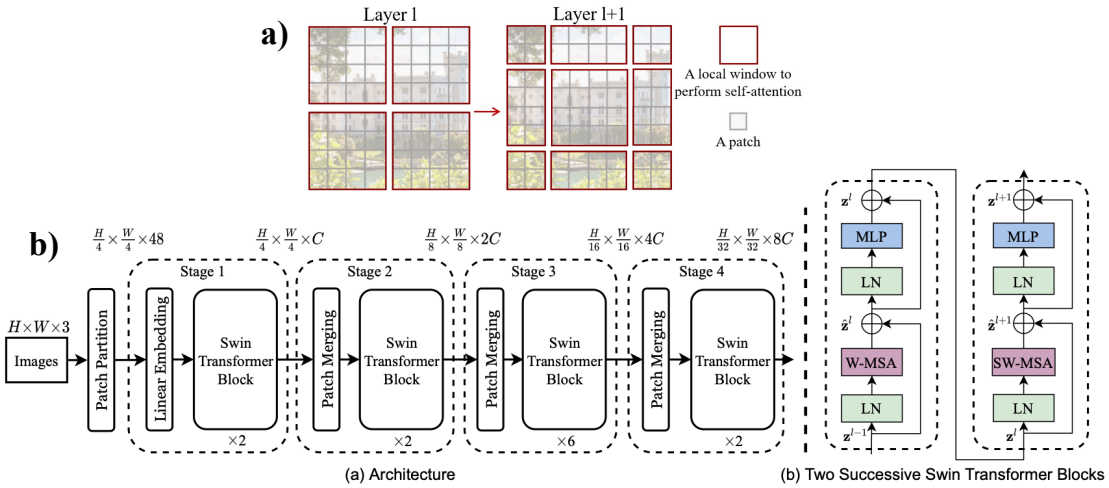

Figure 16 The architecture and attention mechanism of the Swin Transformer. a) The window shifting mechanism allow the model to capture global and nearest neihbour context. b) Swin Transformer architecture and blocks

6.4.4 Training setting and strategy

The training images used are RGB images with a resolution of 512x512 pixels and a batch size of 8. Model optimization is performed using the cross-entropy loss function and the AdamW optimizer with default parameters. To improve model robustness and generalization, various data augmentation techniques are employed during training. Model fine-tuning is conducted using a slow unfreezing strategy, in which the pretrained model's layers are gradually unfrozen and trained for a limited number of epochs to allow adaptation to the target task while preserving the learned representations. Initially, the classifier head is trained while the backbone remains frozen. Subsequently, the training proceeds with a gradual unfreezing of the backbone layers, starting from the last 20 layers of the Swin Transformer, unfrozen and trained for 50 epochs. The learning rate is set to 1e-3, and a step learning rate scheduler is employed with a step size of 10 and a decay factor (gamma) of 0.1. Model checkpoints are saved every 5 epochs, and the lowest validation loss check-

point is retained to mitigate overfitting. Early stopping is not applied, as the entire 50-epoch training progression is analyzed. Subsequently, the number of unfrozen layers is incrementally increased to 70, 160, and 250, respectively, with each stage undergoing an additional 50 epochs of training. This results in a total of 250 training epochs. The unfreezing schedule is designed to incrementally increase the number of trainable layers by the percentage of backbone parameters including 20, 40, 60, and 80 percents.

6.5 Evaluation of DR grading

Deep learning has frequently demonstrated remarkable performance across a multitude of tasks. However, leveraging deep learning without meticulous evaluation can be likened to deploying an unaccredited ophthalmologist, trained but lacking certification to assess performance, thereby compromising the reliability of diagnoses. Therefore, employing concise evaluation metrics is imperative to increase the reliability and confidence in the predictions generated by our network. Primarily, we measure the prediction performance by using the multi-class confusion matrix, which provides insights into the number of correct and missed predictions. As a result of this matrix, we can compute metrics such as precision, recall, and F1 score for each class, thereby offering a comprehensive evaluation of the network's classification capabilities. Notably, all evaluation results are reported using three significant figures, corresponding to the scale of the smallest amount of classes used in this study, which contains approximately 100 samples. This level of precision ensures consistency with the data resolution. Furthermore, we also demonstrate the network performance on the threshold-independent metric, including the ROC curve and AUC. Finally, we utilize the QWK metric to evaluate the grading performance of the network, providing a holistic assessment of its efficacy in classifying DR severity levels.

6.6 Computational resources

The experiments in this work were conducted on a personal computer equipped with an NVIDIA RTX 3070 Ti GPU with 8 GB of VRAM, an Intel Core i7-13700K CPU, and 32 GB of RAM. The operating system used is Ubuntu 20.04 LTS, and the deep learning framework employed is PyTorch version 2.6 (Paszke, 2019).

CHAPTER 4

RESULTS AND DISCUSSION

7.1 Results of image screening

7.1.1 Optical disc and macula detection

In Table 2, the performance of the proposed method reveals that the Euclidean distance (ED) error for the optic disc and macula is 19.9 and 20.3 pixels, respectively. When expressed as a percentage of the image width, these errors are approximately 1.55% and 1.59%, respectively. Additionally, we evaluate the proposed method using various average precision (AP) scores, including AP at an IOU threshold of 0.50 (AP_{50}), AP at an IOU threshold of 0.75 (AP_{75}), and the mean AP score between IOU thresholds of 0.50 and 0.95 with a step size of 0.05. The results indicate that macula detection is highly sensitive to the IOU threshold, as evidenced by the AP score dropping from 0.847 to 0.650. Conversely, optic disc detection shows less sensitivity when comparing AP_{50} with AP, indicating that the template matching method for optic disc detection is relatively stable. Moreover, we calculate the mAP to serve as the representative AP score for the proposed method. The mAP is widely used to compare the performance of object detection algorithms, providing a comprehensive measure of accuracy across different IOU thresholds. Generally, to compare the proposed method with other approaches, the Messidor dataset is used to measure the performance of macula detection algorithms. This measurement is conducted by comparing the predicted macula location with the ground truth location across various sizes of agreement areas and then computing the score, called the R-criterion score, where R represents the radius of the optic disc, being roughly 208 pixels each image. As shown in Figure 17, an increase in the denominator leads to a reduction in the size of the agreement area. Thus, the R-criterion effectively demonstrates the accuracy, precision, and stability of the algorithm in predicting the macula's location as the agreement area decreases.

Table 2 The general performance of proposed method on IDRiD dataset.

Ocular name	ED (pixels)	AP_{50}	AP_{75}	AP	mAP	runtime (s)
Optic disc	19.9	0.938	0.840	0.740	0.659	0.102
Macula	20.3	0.847	0.650	0.578		

Table 3 presents the outcomes of the proposed method across different agreement areas. It is notable that the values for R and R/2 are relatively close, indicating that over 90 percent of the predicted macula locations fall within the radius of R/2, which equals 104 pixels, around the ground-truth macula location. Similarly, the predicted macula is distributed around the ground-truth location within the radius of 52 pixels at 84.2 percent and within 26 pixels at 38.7 percent. This demonstrates the proposed method's capability to accurately predict the macula's location with varying levels of precision.

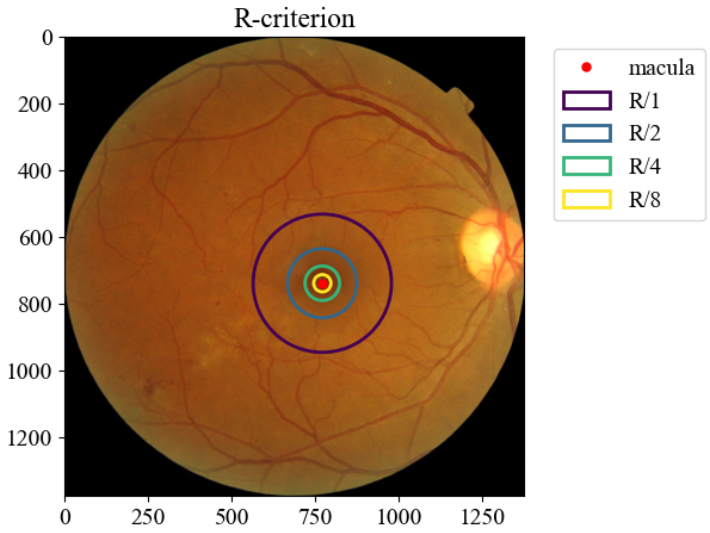

Figure 17 The agreement area of each R-criterion on Messidor dataset.

Table 3 The macula detection on Messidor dataset.

Ocular name	R/8	R/4	R/2	R
Macula	0.387	0.842	0.916	0.922

The qualitative results of the proposed method are illustrated in Figure 18 and Figure 19. Sub-figures (a-c) demonstrate high-quality predictions, whereas sub-figures (d-f) show poor-quality predictions. Notably, the IDRiD dataset results highlight frequent false detection of the optic disc, frequently caused by bright lesions and white fibers that obscure the optic disc's location. Additionally, false macula detection is influenced by factors such as medium-sized hemorrhages, dark spots, and uneven illumination in the retinal image because these dark-like regions can resemble the macula on grayscale images during the matching process. Furthermore, erroneous prediction in optic disc detection can adversely impact macula detection, as the

predicted location of the optic disc is used as a reference to delineate the ROI for macula detection. Furthermore, in the Messidor dataset, the proposed method frequently struggles to detect the macula due to the challenges in distinguishing blood vessels from the macula. Eventually, these qualitative results demonstrate the effectiveness of the method under optimal conditions and simultaneously expose its vulnerability to image noise and pathological obstructions.

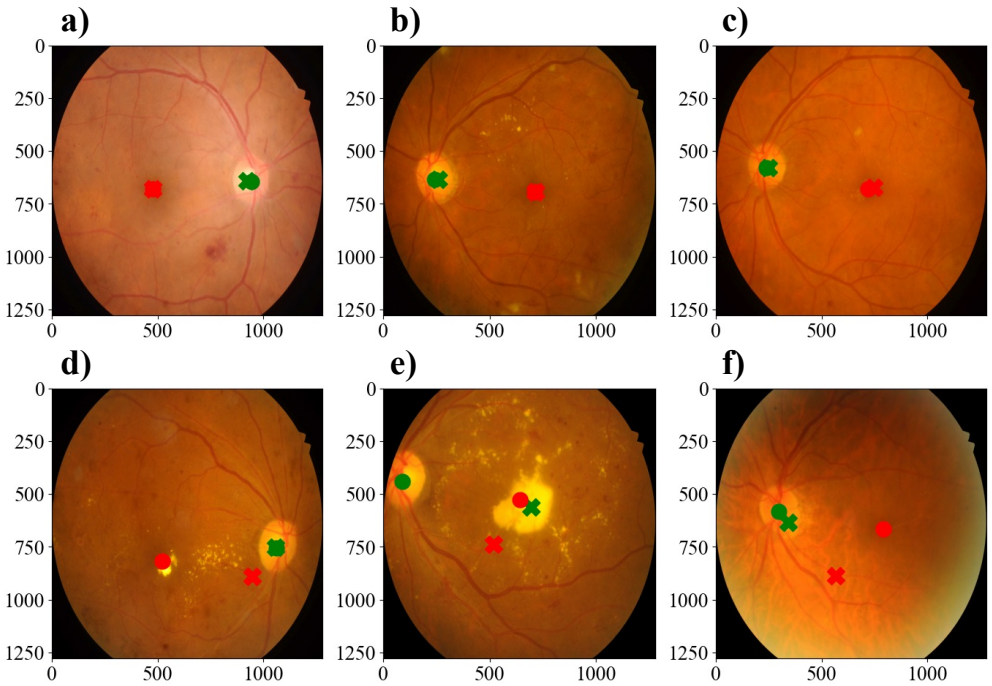

Figure 18 The detection results of the proposed method on the IDRiD dataset. The three images above (a–c) showcase good quality predictions, whereas the images below (d–f) exhibit poor quality predictions. Where, cross sign (✕) is a predicted location, the dot sign (●) is a ground-truth location, the green color represents the optic disc, and the red color represents the macula.

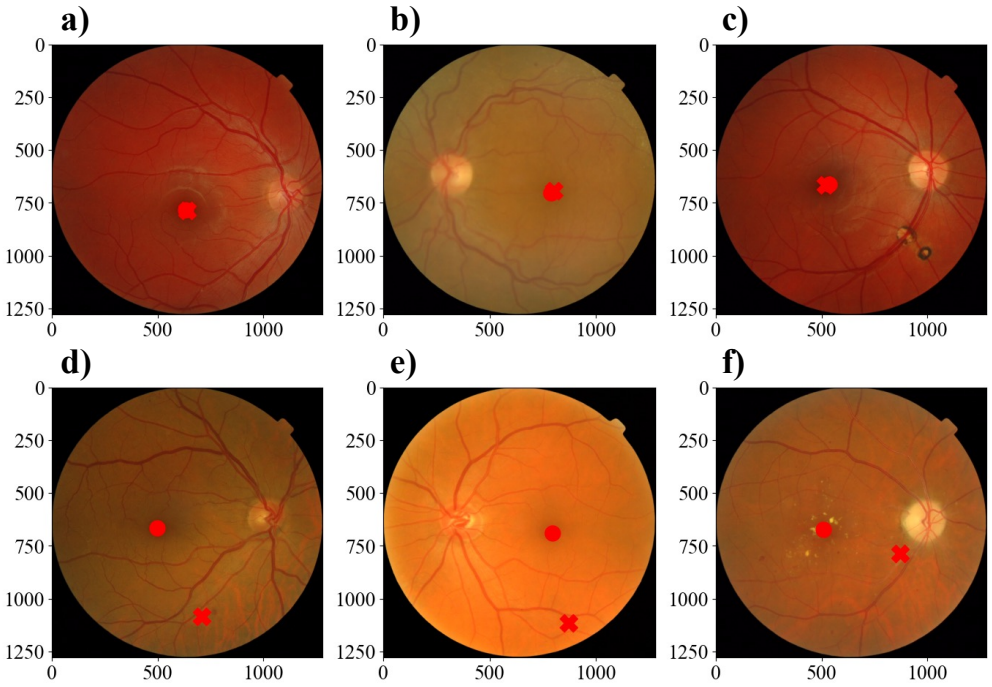

Figure 19 The detection results of the proposed method on the Messidor dataset. The three images above (a–c) showcase good quality predictions, whereas the images below (d–f) exhibit poor quality predictions. Where, cross sign (✕) is a predicted location, the dot sign (●) is a ground-truth location, and the red color represents the macula.

7.1.2 Screening algorithm

The performance of the proposed method is demonstrated through the confusion matrix presented in Figure 20. This matrix illustrates that the algorithm exhibits superior ability in correctly classifying positive images as opposed to negative images. This is evidenced by a lower incidence of false positives compared to false negatives. This discrepancy suggests that the features employed by the algorithm might not be sufficiently robust for accurately distinguishing negative images. Table 4 provides further insight, indicating the high precision score and reliable false discovery rate of the proposed method, which closely align with the goal score of 0.05. Moreover, the method achieves a high recall of 0.906, which can tackle the established goal. In practical terms, the proposed method successfully reduces the proportion of negative images in the dataset to roughly 7 percent while retaining 90 percent of the positive images after screening.

The qualitative results of the proposed method, as illustrated in Figure 21, reveal distinct characteristics associated with each type of prediction. Notably, false-positive cases frequently occur from false detections where lesions or haemorrhages

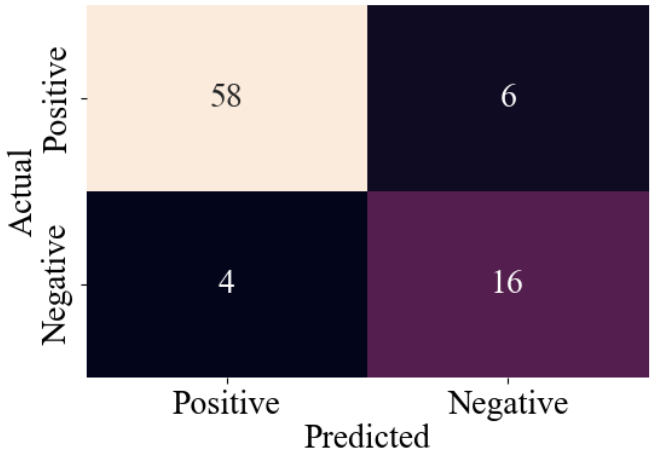

Figure 20 The confusion matrix of proposed image screening method.

Table 4 The performace of proposed image screening

	Accuracy	Precision	Recall	FDR
Goal (Coyner, 2018; Fleming, 2006)	-	0.950	0.900	0.050
Proposed method	0.881	0.935	0.906	0.065

are incorrectly identified as resembling the macula. Similarly, false negatives result from misdetections, as depicted in Figure 21.e. Conversely, the case shown in Figure 21.f indicates a scenario where image overcropping causes the macula to be located outside the acceptable region, leading to false negative predictions. These observations suggest that enhancing the stability of the retinal fundus field could potentially improve the accuracy of the screening process. In the true positive case, the proposed method demonstrates the ability to correctly classify an image as positive, even in the presence of abnormal retinal conditions. Additionally, images of the nasal field are accurately identified as true negatives, further validating the method's effectiveness.

In the further study of the ML model for the screening task, as presented in Table 5, most models demonstrate high performance, exhibiting reliable precision, recall, and false discovery rate (FDR) scores. While no single model consistently outperforms others across all metrics, certain models exhibit optimal performance for specific objectives. Based on this dataset, the Histogram Gradient Boosting (HGB) model is most effective for identifying negative images, achieving a precision of 0.908 and an FDR of 0.092, reflecting strong screening capability. In contrast, DecisionTree and LightGBM models demonstrate the highest recall scores of 0.938, making them

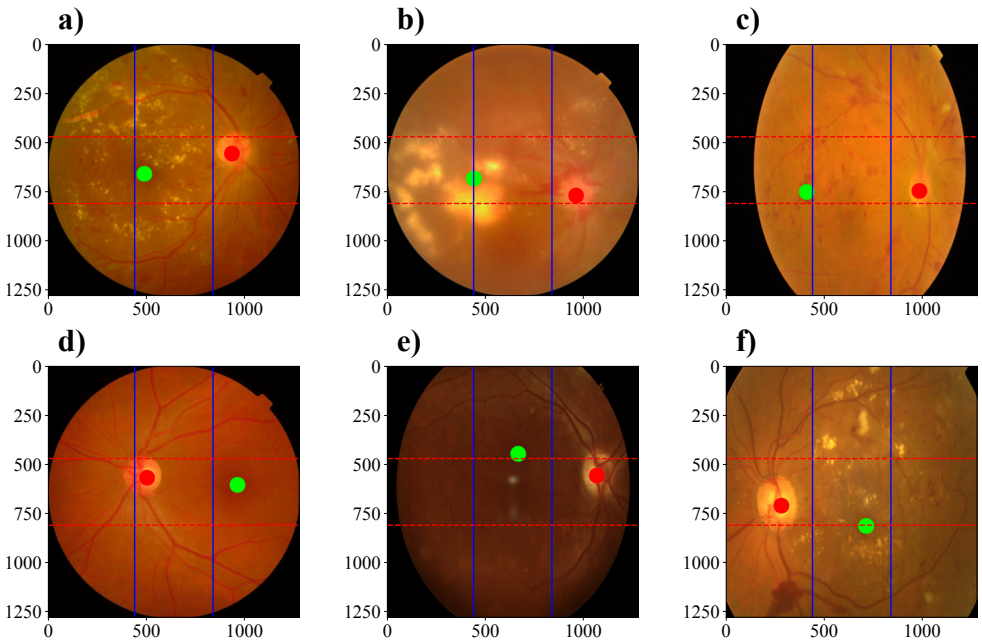

Figure 21 Example of image screening results: (a) True positive; (b-c) False positive; (d) True negative; and (e-f) False negative.

suitable for preserving positive images. Ultimately, if a single model were to be selected for real-world deployment, the HGB model would be preferred due to its superior precision and sufficiently high recall, which exceeds the target threshold of 0.900.

Table 5 The performace of machine learning in image screening. **Bold** represents the best score and underline represents the second best score.

Model name	Accuracy	Precision	Recall	FDR
Logistic Regression	0.845	0.905	0.891	0.095
Decision Tree	0.857	0.882	0.938	0.118
SVC	0.857	0.871	**0.953**	0.129
Random Forest	0.857	0.906	0.906	0.094
HistGradientBoosting	**0.869**	**0.908**	0.922	**0.092**
XGBoost	0.833	0.868	0.922	0.132
LightGBM	**0.869**	0.896	0.938	0.104

7.2 Ablation study of image screening

In this section, we empirically investigate the impact of the components in the algorithm. The metric considered for this entire study is an AP score at an IOU threshold of 0.50 due to ease and fairness.

7.2.1 Template size and sampling amount

First, we investigate the influence of optic disc template size and the number of sampled templates using a grid search strategy. Template sizes are varied from 200 to 400 pixels, incorporating both symmetric and asymmetric configurations, while the number of samples ranged from 5 to 50 in increments of 5. Higher performance is represented by brighter colors on the corresponding color bar. As depicted in Figure 22, the optimal template dimensions lie within the range of 250–350 pixels in width (x-axis) and 300–400 pixels in height (y-axis). Due to the variability in optimal sizes, we apply a 3x3 average kernel to smooth the score distribution and identify an optimally representative size which corresponds to a size range of (300, 350). Conversely, templates that are excessively small or disproportionate yield suboptimal results due to insufficient structural representation. Furthermore, to determine the appropriate number of samples, we utilize a line plot, revealing that performance improves with an increasing number of templates up to approximately 30, after which the gains plateau. Based on this observation, we select 35 as the optimal number of samples, as it resides within the performance plateau while avoiding unnecessary computational overhead.

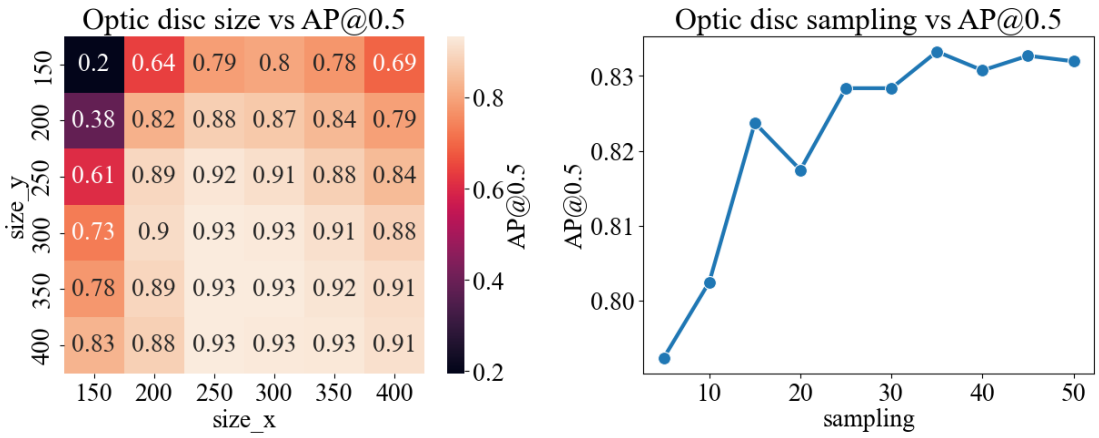

Figure 22 The template parmeter of optic disc. Heatmap depictes the sensitivity of template dimensions via the AP score. The line plot shows the impact of sampling.

Similarly to the optic disc, we conduct an investigation into the optimal parameters for macula template generation, focusing on both template size and the number of samples. Given the smaller anatomical size of the macula, the template size search space is restricted to 100–350 pixels, while the number of samples ranges from 5 to 50 in steps of 5. As illustrated in Figure 23, the heatmap indicates that template widths (x-axis) between 150–250 pixels and heights (y-axis) between 250–350 pixels yield the highest performance. To address the variability in optimal size—similar to the optic disc case—we apply an average kernel, resulting in an optimal template size of (200, 300). Moreover, further analysis demonstrates very small or excessively large templates degrade performance due to inadequate or overly diffuse anatomical representation. The corresponding line plot reveals a rapid increase in performance as the number of samples rises to 15, beyond which the performance stabilizes. Only minimal gains are observed beyond 35 samples. Ultimately, considering resource efficiency similar to the approach taken for the optic disc, we select 35 as the optimal number of samples, that balance efficiency and accuracy in macula detection.

Based on our investigation, we determined the optimal templates as depicted in Figure 24. For the optic disc, the optimal template size was determined to be (200, 300), and the sampling amount was set to 20 images. As same manner, for the macula, we selected a template size of (200, 200) with a sampling amount of 35 images.

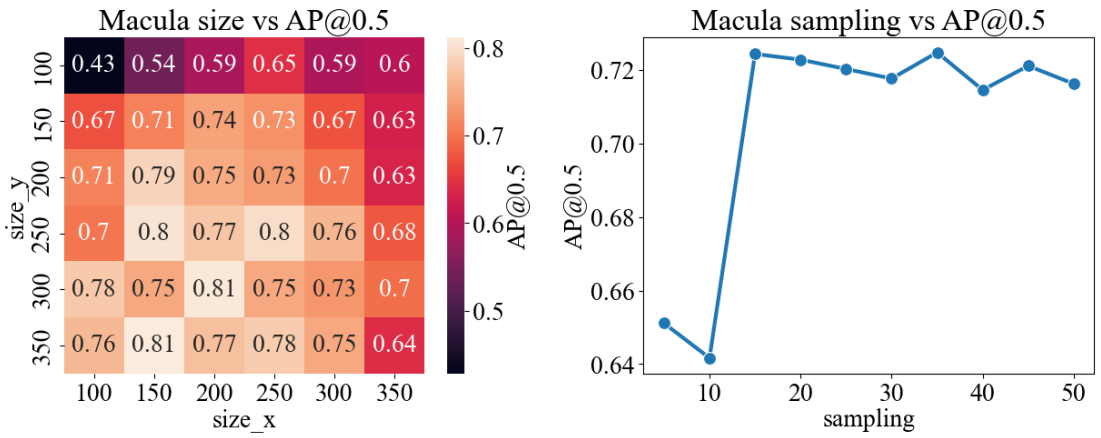

Figure 23 The template parmeter of macula. Heatmap depictes the sensitivity of template dimensions via the AP score. The line plot shows the impact of sampling.

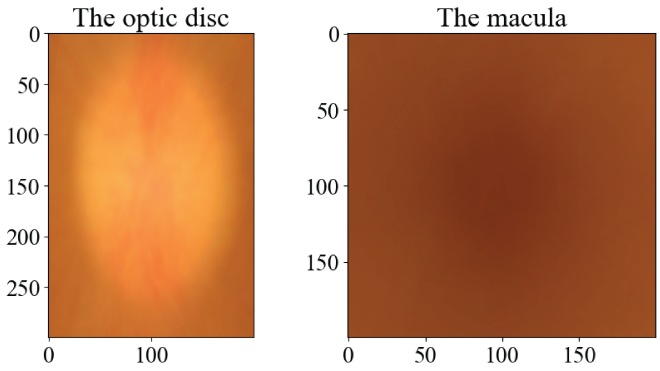

Figure 24 The optimal template for optic disc and macula.

7.2.2 Matching functions

Further study concerns the influence of the different matching functions on an AP50 score. These functions are shown below:

- **Sum of square differences (SQDIFF)** and **Normalized sum of square differences (SQDIFF_NORMED)**

  SQDIFF function:

$$R(x,y) = \sum_{x',y'} (T(x',y') - I(x+x',y+y'))^2 \tag{16}$$

  SQDIFF_NORMED function:

$$R(x,y) = \frac{\sum_{x',y'}(T(x',y') - I(x+x',y+y'))^2}{\sqrt{\sum_{x',y'}(T(x',y')^2 \cdot \sum_{x',y'} I(x+x',y+y')^2}} \tag{17}$$

- **Cross correlation (CCORR)** and **Normalized cross correlation (CCORR_NORMED)**

  CCORR function:

$$R(x,y) = \sum_{x',y'} (T(x',y') \cdot I(x+x',y+y')) \tag{18}$$

  CCORR_NORMED function:

$$R(x,y) = \frac{\sum_{x',y'}(T(x',y') \cdot I(x+x',y+y'))}{\sqrt{\sum_{x',y'}(T(x',y')^2 \cdot \sum_{x',y'} I(x+x',y+y')^2}} \tag{19}$$

- **Correlation coefficient (CCOEFF)** and **Normalized correlation coefficient (CCOEFF_NORMED)**

  CCOEFF function:

$$R(x,y) = \sum_{x',y'} (T'(x',y') \cdot I'(x+x', y+y')) \tag{20}$$

  CCOEFF_NORMED function:

$$R(x,y) = \frac{\sum_{x',y'} (T'(x',y') \cdot I'(x+x', y+y'))}{\sqrt{\sum_{x',y'} (T'(x',y')^2 \cdot \sum_{x',y'} I'(x+x', y+y')^2}} \tag{21}$$

  where,

$$T'(x',y') = T(x',y') - \frac{1}{w \cdot h} \cdot \sum_{x'',y''} T(x'',y'') \tag{22a}$$

$$I'(x+x', y+y') = I(x+x', y+y') - \frac{1}{w \cdot h} \cdot \sum_{x'',y''} I(x+x'', y+y'') \tag{22b}$$

In Table 6, we observe that the normalized function consistently yields superior scores compared to the non-normalized function. This discrepancy likely occurs due to variations in luminosity and contrast present in retinal fundus images while the template remains in fixed conditions. Consequently, these variations can lead to false detections when using non-normalized functions. Conversely, the normalized function is specifically designed to normalize both the target image and the template, thereby mitigating the influence of these variations and enhancing algorithm stability and accuracy. Ultimately, we choose CCOFF_NORMED to be the matching function for ocular structure detection, as it can outperform the other normalized functions.

7.2.3 Region of interest (ROI)

Further on, we examine the influence of the ROI technique on the overall algorithm performance, as detailed in Table 7. The AP_{50} score shows a improvement, from 0.811 to 0.847. However, when measured using the ED error, macula detection

Table 6 CCOEFF_NORMED achieves the highest AP_{50} score for both detection task.

	Optic disc	Macula
SQDIFF	0.386	0.087
SQDIFF_NORMED	0.420	0.090
CCORR	0.387	0.000
CCORR_NORMED	0.542	0.438
CCOEFF	0.381	0.000
CCOEFF_NORMED	0.948	0.830

with ROI yields a higher value compared to without ROI. This increase is occured to the ROI technique constraining the detection algorithm to identify the macula location within the specific area, which occasionally leads to the prediction of locations that merely resemble the macula. As a result of the quantitative results, the ROI technique appears to be slight improvement for detection performance. Addtionally, upon examining the qualitative results in Figure 25, reveals that the ROI technique significantly enhances detection accuracy by effectively guiding the macula detector to focus within the appropriate anatomical region.

Table 7 The quantitative result of ROI technique.

Method	ED (pixels)	AP_{50}
Macula w/o ROI	19.4	0.811
Macula w ROI	20.4	0.847

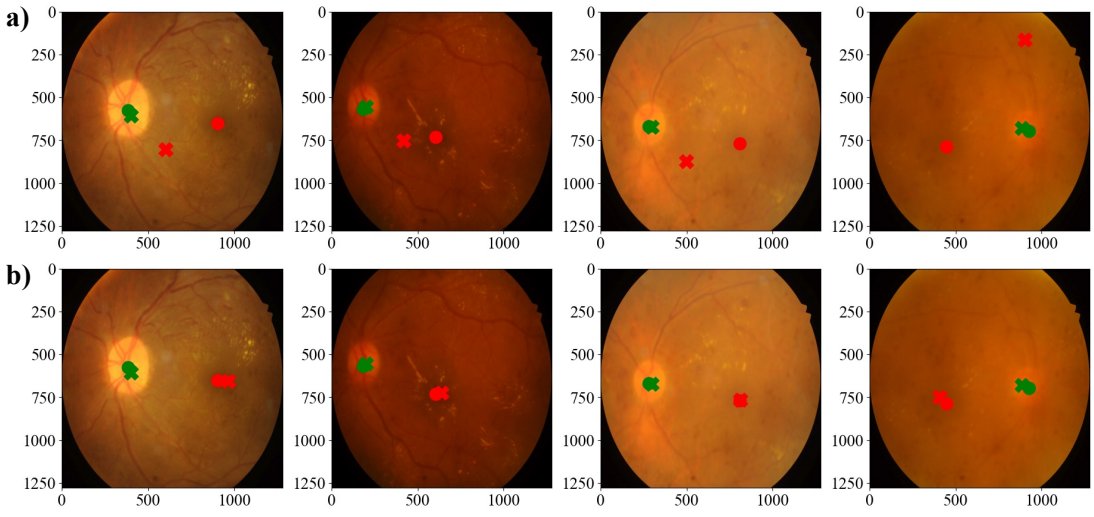

Figure 25 The qualitative result of macula detection, a) without ROI, b) with ROI. Where, cross sign (✕) is a predicted location, the dot sign (●) is a ground-truth location, the green color represents the optic disc, and the red color represents the macula.

7.2.4 The generalization of the proposed method

In this subsection, we investigate the generalizability of our approach using an Out-of-Distribution (OOD) testing framework. In this testing, models are trained on one dataset and evaluated on a different dataset, as summarized in Table 8. This form of testing presents a significant challenge, as it introduces distribution shift—an inherent issue in real-world applications that non-generalized models often struggle to address. The overall results suggest that our proposed method can effectively manage this challenge. Although it does not achieve the top performance in the second testing scenario, it still delivers competitive results. Within the machine learning context, inherently well-generalized models such as logistic regression and random forest exhibit stable performance across unfamiliar distributions. In particular, random forest, which employs a bagging algorithm, benefits from reduced variance and improved adaptability to diverse data distributions. Additionally, the observed asymmetry in generalization performance may occur from dataset characteristics: the SUTH dataset is balanced, whereas the Maharaj dataset is imbalanced. This discrepancy can significantly affect model transferability and the reliability of screening outcomes.

Table 8 The performace of machine learning in image screening. **Bold** represents the best score and underline represents the second best score.

Train	Test	Model name	Accuracy	Precision	Recall	FDR
Maharaj	SUTH	**LogisticRegression**	0.755	0.742	0.854	0.258
		DecisionTree	0.670	0.650	0.872	0.350
		SVC	0.685	0.651	**0.926**	0.349
		RandomForest	**0.767**	0.739	0.893	0.261
		HistGradientBoosting	0.737	0.703	0.908	0.297
		XGBoost	0.745	0.706	0.923	0.294
		LightGBM	0.714	0.679	0.914	0.321
		rulebase	0.757	**0.758**	0.821	**0.242**
SUTH	Maharaj	**LogisticRegression**	0.758	0.821	**0.875**	0.179
		DecisionTree	0.765	0.891	0.791	0.109
		SVC	**0.815**	0.906	0.847	0.094
		RandomForest	0.765	0.940	0.741	0.060
		HistGradientBoosting	0.787	0.939	0.772	0.061
		XGBoost	0.799	**0.947**	0.781	**0.053**
		LightGBM	0.772	0.934	0.756	0.066
		rulebase	0.748	0.932	0.725	0.068

7.3 Results of DR grading

Currently, the architecture of the proposed model includes Swin s as the backbone network for feature extraction, followed by three fully connected layers with 2208, 64, and 5 nodes, respectively, designed for the grading task. In terms of hyperparameters, the input data is an RGB retinal images with a resolution of 512x512 pixels. The model is trained with a learning rate of 0.0002, a batch size of 8, and using the AdamW optimizer with default parameters from the PyTorch library. During training, data augmentations such as random horizontal flips, random equalization, and random rotations are applied to enhance model robustness and generalization.

The performance of the proposed model is illustrated by the confusion matrix in Figure 26, which reveals a substantial disparity in predictive accuracy between majority and minority classes. Specifically, the model correctly classifies 356 out of 361 instances in class 0, while achieving only 15 correct predictions out of 38 for class 3. Misclassification patterns are particularly evident in minority classes, especially classes 3 and 4, underscoring the difficulty in distinguishing advanced stages of

diabetic retinopathy (DR), where lesion characteristics tend to overlap. This indicates that while the model demonstrates strong performance in detecting early-stage DR, further enhancement is needed for accurate classification of advanced stages. In addition to overall prediction trends, the confusion matrix also reveals instances of severe misclassification. For example, 22 images with a ground truth of class 4 were misclassified as classes 1 and 2. Therefore, analyzing these misclassifications could provide insights into the potential areas for model improvement, particularly in advanced DR detection.

Figure 26 The confusion matrix of proposed model.

In the same manner as the confusion matrix, the classification report in Table 9 indicates that the model has a better understanding of retinal images in the majority class compared to the minority class. The F1-scores for classes 1 and 2 are 0.993 and 0.785, respectively, while the average F1-score for the minority class is approximately 0.561. This significant gap underscores the model's performance discrepancy between the majority and minority classes. Additionally, the average F1-score across these classes shows a substantial difference between the F1 macro score of 0.693 and the micro score of 0.843. This disparity occurs from the model's poor performance on the minority classes, as evidenced by the precision and recall scores. Additionally, in the practica term, the precision macro score of 0.730 indicate that the model is confident to predict the severity of DR level as each class around 73 percent of the time, while the recall macro score of 0.693 indicate that the model can detect only 69 percent of all positive images. Consequently, to improve the

model's performance, it is imperative to address the imbalance issue, achieved by either balancing the amount of data in each class or by modifying the model's architecture or input data to better define the boundary conditions between classes, particularly the minority classes.

Table 9 The classification report of proposed model on the APTOS 2019 dataset.

class	Precision	Recall	F1-score	Support
0	1.000	0.986	0.993	361
1	0.615	0.793	0.694	74
2	0.757	0.814	0.785	199
3	0.441	0.395	0.412	38
4	0.839	0.441	0.578	59
Macro avg	0.730	0.687	0.693	731
Micro avg	0.853	0.845	0.843	731

Conversely, Figure 27 illustrates the ROC curve, which demonstrates satisfactory results due to the favorable shape of the curve and the high AUC across all classes. For instance, class 0 has an AUC of 1.00, and class 3 has an AUC of 0.92, which starkly contrasts with the findings from the confusion matrix and classification report. This conflict arises due to the imbalance issue, where a large amount of true negative data leads to consistently high AUC values on the ROC curve. Therefore, in the context of imbalanced datasets, the ROC curve may be non-informative and unable to accurately reflect the model's true performance. Consequently, to better visualize the model performance in this context, we introduce the Precision-Recall (PR) curve, which offers a more appropriate representation despite having similar logical foundations as the ROC curve as shown in Figure 28. However, in the table summarizing the model's general performance, we still have to report the AUC score of the ROC curve instead of the AUC of the PR curve, referred to as the average precision (AP) score, as the AUC score is commonly utilized for performance comparison in numerous research papers.

Additionally, we comprehensively investigate the model's performance on a balanced test dataset. In this experiment, 38 images are randomly selected from each class, matching the number of images in the minority class, to eliminate the influence of class imbalance. The results, summarized in Table 10 and visualized in Figure 29, surprisingly reveal the outstanding performance for class 0 (F1 = 0.993)

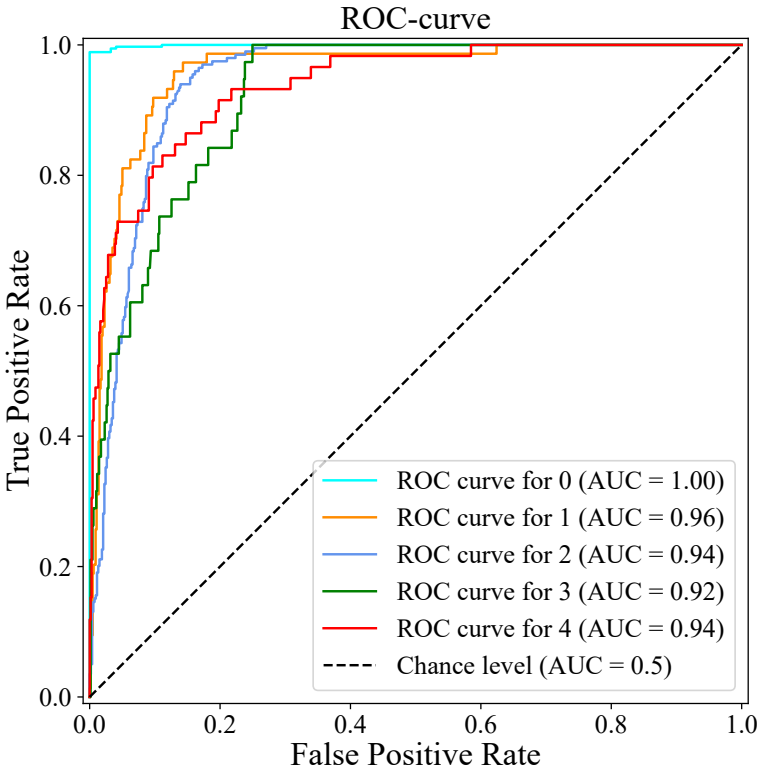

Figure 27 The figure illustrates the ROC curve and AUC for each class, indicating that the proposed model outperforms in classifying class 1 across a variety of threshold values. However, the other classes exhibit constraints related to the threshold value, impacting their lower classification performance.

and class 1 (F1 = 0.786). This contrasts with the performance observed on the imbalanced test dataset, where the model is expert on class 0 and class 2. Furthermore, analysis of the confusion matrix shows that the model continues to struggle with accurately classifying classes 3 and 4. These stages are frequently misclassified as class 2, resulting in a high false positive rate for class 2. Actually, this misclassification pattern also appears in the imbalanced test set, it is less apparent due to the disproportionately large number of class 2 samples, which masks the underlying issue. Further insights are provided by the ROC curves in Figure 30 and the PR curves in Figure 31. Class 0 achieves perfect classification, while classes 1, 3, and 4 also demonstrate acceptable performance. However, class 2 consistently underperforms across both evaluation metrics, indicating its inherent difficulty due to overlapping feature characteristics with adjacent severity levels. Eventually, these results confirm the model's robustness in detecting early DR stages and its potential for reliable performance in balanced conditions. However, the limited accuracy in classifying advanced stages (particularly classes 3 and 4) remains a challenge. To improve grading

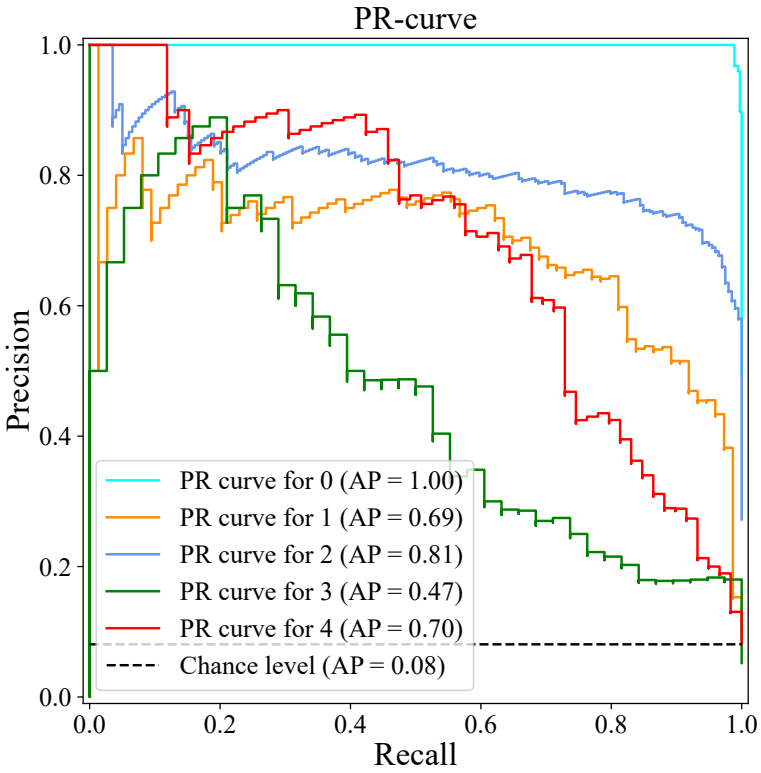

Figure 28 The PR curve and AUC for each class, which is more informative than the ROC curve in context of imbalance dataset.

performance, it is essential to enhance the model's capacity to distinguish advanced DR stages. This may be accomplished through balancing the training dataset or by modifying the model architecture to better define the boundary conditions between classes, especially for minority classes.

Table 10 The classification report of proposed model on the APTOS 2019 dataset.

class	Precision	Recall	F1-score	Support
0	1.000	0.921	0.959	38
1	0.717	0.868	0.786	38
2	0.441	0.790	0.566	38
3	0.714	0.395	0.509	38
4	0.850	0.447	0.586	38
Macro avg	0.745	0.684	0.681	190
Micro avg	0.745	0.684	0.681	190

Figure 29 The confusion matrix of proposed model.

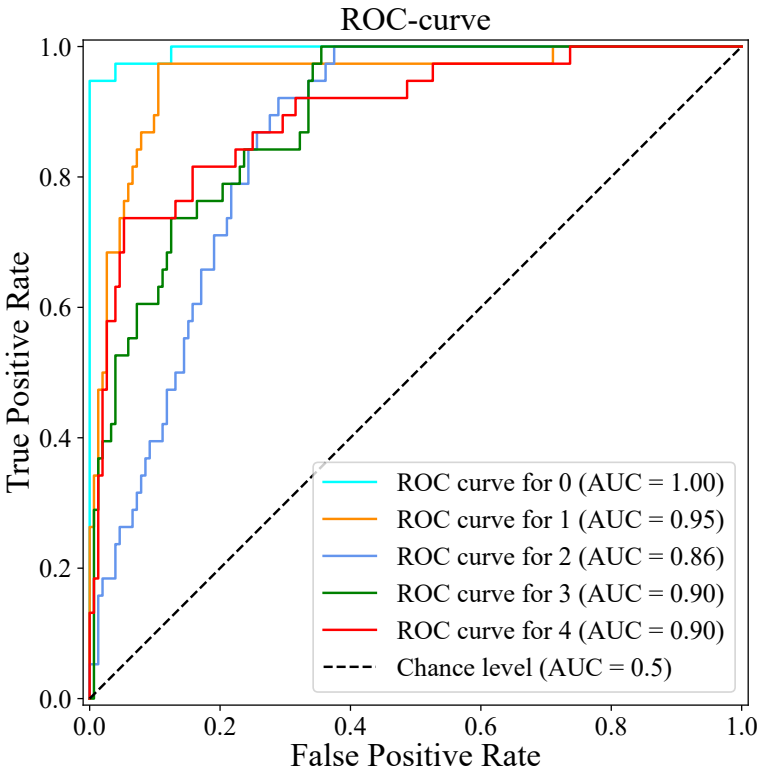

Figure 30 The figure illustrates the ROC curve and AUC for each class, indicating that the proposed model outperforms in classifying class 1 across a variety of threshold values. However, the other classes exhibit constraints related to the threshold value, impacting their lower classification performance.

Ultimately, Table 11 compares the performance between the proposed model and previosu related work, which is from difference vision architecture such as CNN and Vision-mamba, on the same dataset, indicating that the proposed model can

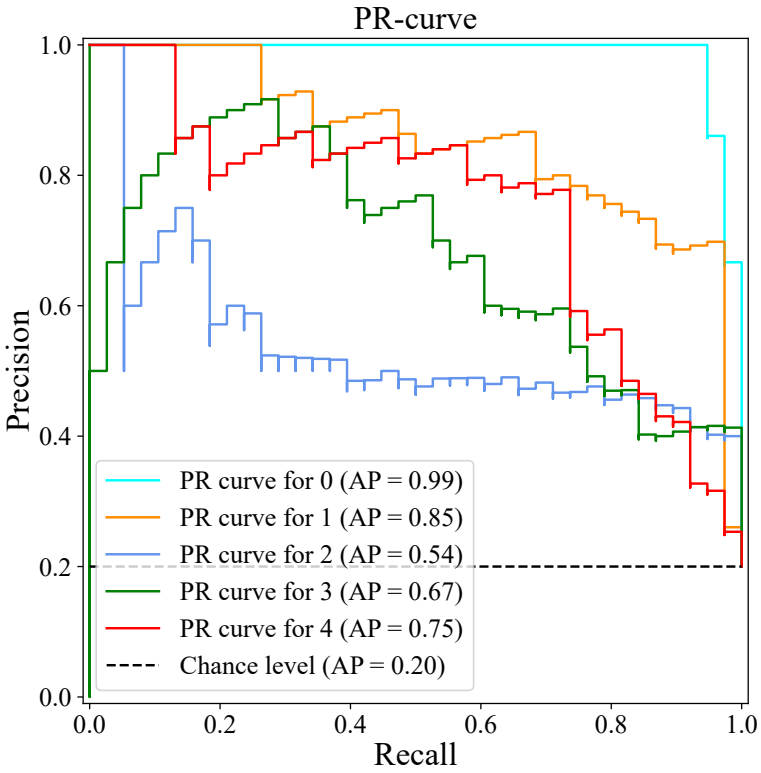

Figure 31 The PR curve and AUC for each class, which is more informative than the ROC curve in context of imbalance dataset.

outperforms the other models. This comparison reveals the backbone network based on Vision Transformer (ViT) architecture, can provide a significant improvement in the grading task. The QWK score of the proposed model is 0.903, which is higher than the previous work, suggesting that the proposed model can achieve a comparable grading performance as ophthalmologist. Nevertheless, the significant disparity between the F1 scores and accuracy reveals the proposed model's inferior performance in the classification task compared to the ophthalmologists. Consequently, to achieve our criteria, substantial improvements are necessary for the classification task.

Table 11 The general performance of proposed model compared the human performance. **Bold** represents the best score.

Dataset	Method	Acc.	Micro		Macro		QWK
			F1	AUC	F1	AUC	
EyEPACS-2	Ophthalmologists (Krause, 2018)	0.895	0.895	-	0.714	-	0.871
APTOS 2019	VGG16+Xception+CNN (Bodapati, 2021)	0.827	0.818	-	0.663	-	0.864
	DenseNet169+CBAM+INS (Farag, 2022)	0.822	0.825	-	0.685	-	0.888
	VMamba-m (Xue, 2024)	0.786	0.786	-	0.661	-	0.784
	Proposed model	**0.845**	**0.843**	**0.967**	**0.693**	**0.952**	**0.903**

7.4 Ablation study of DR grading

In this section, we examine the influence of various training improvement techniques on the validation set of the APTOS 2019 dataset. The DR severity grading task inherently faces the challenge of class imbalance. Therefore, appropriate metrics for this task are macro scores, including the F1 macro score and the AUC-ROC macro score. These metrics are suitable because they measure performance across all classes with equal weight, in contrast to micro scores, which balance each class according to its own size.

7.4.1 Backbone model selection

First, we investigate the backbone networks to determine the optimal feature extraction network for our task. Given the numerous networks currently proposed for DR grading, it is impractical to thoroughly evaluate them. Therefore, we emphasize five of the most commonly used networks: VGG 19 (Simonyan, 2014), Inception V3 (Szegedy, 2016), ResNet 50 (K. He, 2016), DenseNet 161 (G. Huang, 2017), and Swin Transformer (Liu, 2021). As shown in Table 12, DenseNet 161 achieves the highest scores for both macro-metrics, with an F1 score of 0.579 and an AUC of 0.933, and ranks second for the other metrics, which is impressive for a non-tuned network. However, this is a grading task, so the Quadratic Weighted Kappa (QWK) score must also be considered. On this metric, Swin Transformer significantly outperforms the other networks and also ranks first or second for the other scores. As demonstrated, DenseNet 161 and Swin Transformer exhibit impressive performance, presenting a dilemma in network selection. To address this challenge effectively, we consider employing both networks as backbone models.

7.4.2 Data sampler

Table 12 The general performance of each backbone model. **Bold** represents the best score and underline represents the second best score.

Model name	Acc.	Micro		Macro		QWK
		F1	AUC	F1	AUC	
VGG19	0.726	0.726	0.866	0.333	0.731	0.747
Inception V3	0.761	0.761	0.943	0.512	0.888	0.785
ResNet50	0.762	0.762	0.953	0.526	0.910	0.809
DenseNet161	0.792	0.792	0.963	**0.579**	**0.933**	0.822
Swin s	**0.793**	**0.793**	**0.964**	0.515	0.931	**0.843**

For further study, we investigated the data sampler strategy for effective network training. We compared conventional data samplers, typically sequential or random, with a sampler designed to address the imbalance issue, termed the imbalance data sampler. This sampler attempts to distribute data evenly across classes within each training batch, as illustrated in Figure 32. Consequently, the network gradually and equally comprehends each class, leading to a slight improvement in performance. Table 13 illustrates that the imbalanced data sampler notably enhanced the performance of DenseNet 161, particularly in the F1 macro score and QWK, which increased from 0.579 to 0.662 and 0.822 to 0.840, respectively. Similarly, the Swin Transformer also exhibited improvement in several metrics, particularly in the F1 macro score, which increased by approximately 0.14 points, and in the QWK, which improved from 0.843 to 0.869. In conclusion, the imbalanced data sampler demonstrates potential for improving network performance by optimizing the data sampling process.

Table 13 The influence of imbalance data sampler to DenseNet 161 and Swin Transformer. **Bold** indicates the best score.

Model name	Acc.	Micro		Macro		QWK
		F1	AUC	F1	AUC	
DenseNet161	0.792	0.792	**0.963**	0.579	**0.933**	0.822
+Imbalance sampler	**0.798**	**0.798**	0.956	**0.662**	0.928	**0.840**
Swin s	0.793	0.793	**0.964**	0.515	**0.931**	0.843
+Imbalance sampler	**0.810**	**0.810**	0.962	**0.657**	0.927	**0.869**

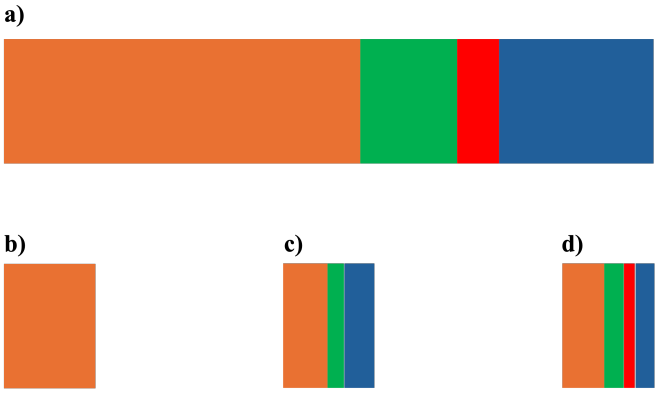

Figure 32 The figure illustrates the data sampler strategies: a) the entire training dataset; b) the training data in a batch with a sequential sampler strategy, which sequentially selects the data from beginning to end; c) the training data in a batch with a random sampler strategy, which randomly selects the data from the entire dataset; and d) the training data in a batch with an imbalance sampler strategy, which attempts to select the data from each class. Each color represents data from a different class.

7.4.3 Loss functions

In our third study, we investigated different loss functions employed in training neural networks, as selecting the appropriate loss function is crucial for guiding the network towards optimal performance.

- **Cross-entropy loss (CE):** Cross-entropy loss is used to measure model performance by quantifying the difference between the predicted distribution **Q** and the ground truth distribution **P**, which can be described in mathematical form as (Eq. 23).

$$CE(P,Q) = -\sum_{i=1}^{C} P(x_i)\log(Q(x_i)) \tag{23}$$

  where, C is the number of classes, x_i is the data point for i^{th} class.

- **Weighted cross-entropy loss (WCE):** Weighted cross-entropy loss is the extension of the standard cross-entropy loss by adding a weighted parameter for each class to mitigate the impact of imbalanced data during training, as shown in (Eq. 24).

$$WCE(P,Q) = -\sum_{i=1}^{C} w_i P(x_i)\log(Q(x_i)) \tag{24}$$

  where, C is the number of classes, x_i is the data point and w_i is the weight for i^{th} class, which typically computed as N/N_i. Here, N_i represents the number of data points in the i^{th} class, and N is the total number of data points.

- **Focal loss (FL) (Lin, 2017):** Focal loss is an extension of the cross-entropy loss, specifically designed to mitigate the imbalance between foreground and background classes in object detection tasks. Therefore, due to our imbalanced dataset, we decided to utilize this loss function. The focal loss is defined as follows:

$$FL(P,Q) = -\sum_{i=1}^{C} P(x_i)(1-Q(x_i))^{\gamma}\log(Q(x_i)) \tag{25}$$

  where, C is the number of classes, x_i is the data point for i^{th} class. γ is the focusing parameter $\gamma \geq 0$, influencing the model 's attention on the misclassified example or minority data point.

Table 14 illustrates that the CE loss yields the best performance across multiple metrics, particularly the F1 scores, which highlight the superior classification capabilities among the evaluated loss functions. In contrast, the FL loss exhibits better performance for the grading task, achieving roughly 0.853 for the QWK score. Interestingly, the WCE and FL losses do not perform as well as anticipated across various metrics. This issue might arise from the use of an imbalanced data sampler, which assists every batch to become a balanced batch, thereby causing the focusing parameter γ in FL loss to negatively affect the loss value during training. Additionally, the weight parameter in WCE loss is a fixed, global parameter that does not adapt to the class distribution within each batch, resulting in an ineffective weight for the loss value during training. Eventually, based on the results presented in the table, we can conclude that CE loss is the optimal loss function for training the model in conjunction with the imbalanced data sampler.

7.4.4 The impact of SMOTE

In this investigation, we explore the influence of the Synthetic Minority Oversampling Technique (SMOTE) through the model's performance and K-nearest neighbour performance. Table 15 illustrates the performance of Swin model with and without SMOTE. The result indicate that model is trained with synthetic data from SMOTE can achieve a higher score on both severity level and the overall performance, as shown that the F1 macros score increase from 0.637 to 0.659 in Swin s model. As a same manner, SMOTE also imrpove the performance of DenseNet 161 model, as illustrated in Table 16, where the F1 macro score increases from 0.637 to

Table 14 The influence of loss functions on DenseNet 161 and Swin Transformer. **Bold** represents the best score.

Model name	Loss function	Acc.	Micro		Macro		QWK
			F1	AUC	F1	AUC	
DenseNet161	**CE**	**0.798**	**0.798**	0.956	**0.662**	**0.928**	0.840
	WCE	0.778	0.778	0.950	0.633	0.921	0.842
	FL (γ = 1)	0.796	0.796	0.956	0.639	0.924	0.857
	(γ = 2)	0.795	0.795	0.956	0.656	0.925	0.851
	(γ = 3)	0.792	0.792	**0.958**	0.635	0.926	**0.856**
Swin s	**CE**	**0.810**	**0.810**	**0.962**	**0.657**	0.927	**0.869**
	WCE	0.791	0.791	0.957	0.630	0.925	0.851
	FL (γ = 1)	0.803	0.803	0.960	0.643	**0.930**	0.858
	(γ = 2)	0.808	0.808	0.961	0.637	**0.930**	0.846
	(γ = 3)	0.780	0.780	0.955	0.600	0.924	0.840

0.657. This improvement is particularly significant for the minority classes such as class 3 and 4, while compensating with performance to classify the class 1. However, the majority class (class 1) also experiance a slight increase in performance, which from roughly 0.7 to 0.72. This indicates that the model can learn more robust features from the synthetic data generated by SMOTE, leading to better generalization and classification performance across various classes.

Table 15 The classification report of Swin Transformer model with and without SMOTE.

class	Support	Witout SMOTE			With SMOTE		
		Precision	Recall	F1	Precision	Recall	F1
0	361	0.949	0.983	0.966	0.962	0.983	0.973
1	74	0.535	0.730	0.617	0.577	0.662	0.616
2	199	0.753	0.643	0.694	0.729	0.729	0.729
3	38	0.421	0.421	0.421	0.400	0.474	0.434
4	59	0.604	0.492	0.542	0.768	0.424	0.544
Macro avg	731	0.652	0.654	0.648	0.685	0.654	0.659
Micro avg	731	0.799	0.796	0.794	0.814	0.810	0.808

To further investigate the models' understanding, we analyze the quality of the feature vectors extracted from the backbone networks by applying the K-Nearest Neighbors (KNN) algorithm within the feature space. As shown in Table **??**, the KNN classifier trained with SMOTE achieves higher F1 macro scores on both the training

Table 16 The classification report of DenseNet161 model with and without SMOTE.

class	Support	Witout SMOTE			With SMOTE		
		Precision	Recall	F1-score	Precision	Recall	F1-score
0	361	0.973	0.986	0.979	0.952	0.989	0.970
1	74	0.505	0.743	0.601	0.612	0.554	0.582
2	199	0.739	0.668	0.702	0.708	0.779	0.742
3	38	0.360	0.474	0.409	0.500	0.421	0.457
4	59	0.808	0.356	0.494	0.684	0.441	0.536
Macro avg	731	0.677	0.646	0.637	0.691	0.637	0.657
Micro avg	731	0.817	0.798	0.797	0.806	0.814	0.807

and validation sets. Particularly, the feature vectors extracted from the Swin model show a performance gain of approximately 5.7% with the application of SMOTE. Therefore, the results ensure the effectiveness of SMOTE in improving the model's performance by addressing class imbalance.

Table 17 The classification performance of KNN model with different pretrained models. Notably, we use only **macro** score for comparison

Model name	Without SMOTE		SMOTE	
	F1 (Train)	F1 (Val)	F1 (Train)	F1 (Val)
DenseNet161	0.858	0.535	0.950	0.552
Swin s	0.752	0.489	0.895	0.546

An additional investigation into the impact of SMOTE on model performance using a balanced test dataset is presented in Table 18. The results indicate that the F1 macro score of the Swin model increases from 0.648 to 0.680 when trained with SMOTE, demonstrating a positive effect. In contrast, DenseNet161 exhibits a decline in performance, with its F1 macro score decreasing from 0.627 to 0.601 under the same conditions. This decline is arised to reduced classification accuracy in classes 1 and 4, as well as an increased false positive rate in class 2 when trained with SMOTE 19. As a result, we face with a dilemma because SMOTE generally enhances model performance, except for the DenseNet161 on the balanced dataset. To further evaluate the robustness of SMOTE, we apply a KNN classifier on features extracted from both models and test it on a balanced dataset, as shown in Table 20. The KNN results demostrate consistent improvements, with performance gains of approximately 0.6 for DenseNet161 and 0.9 for Swin, confirming the effectiveness of

SMOTE in enhancing classification ability. In conclusion, the SMOTE technique proves beneficial in improving model performance on imbalanced datasets, particularly for the Swin architecture.

Table 18 The classification report of Swin Transformer model with and without SMOTE.

class	Support	Witout SMOTE			With SMOTE		
		Precision	Recall	F1	Precision	Recall	F1
0	38	0.900	0.947	0.923	0.878	0.947	0.911
1	38	0.674	0.763	0.716	0.750	0.711	0.730
2	38	0.433	0.684	0.531	0.446	0.763	0.563
3	38	0.643	0.474	0.546	0.750	0.474	0.581
4	38	0.790	0.395	0.526	0.792	0.500	0.613
Macro avg	190	0.688	0.653	0.648	0.723	0.679	0.680
Micro avg	190	0.688	0.653	0.648	0.723	0.679	0.680

Table 19 The classification report of DenseNet161 model with and without SMOTE.

class	Support	Witout SMOTE			With SMOTE		
		Precision	Recall	F1-score	Precision	Recall	F1-score
0	38	0.857	0.947	0.900	0.783	0.947	0.857
1	38	0.700	0.737	0.718	0.714	0.526	0.606
2	38	0.426	0.605	0.500	0.389	0.737	0.509
3	38	0.593	0.421	0.492	0.700	0.421	0.525
4	38	0.630	0.447	0.523	0.714	0.395	0.509
Macro avg	190	0.641	0.632	0.627	0.659	0.605	0.601
Micro avg	190	0.641	0.632	0.627	0.659	0.605	0.601

Table 20 The classification performance of KNN model with different pretrained models. Notably, we use only **macro** score for comparison

Model name	Without SMOTE		SMOTE	
	F1 (Train)	F1 (Val)	F1 (Train)	F1 (Val)
DenseNet161	0.858	0.499	0.950	0.561
Swin s	0.752	0.455	0.895	0.564

7.4.5 The impact of fine-tuning

In this section, we examine the impact of fine-tuning on the performance of backbone models. Fine-tuning serves as a critical step in adapting pre-trained models to task-specific domains by leveraging representations learned from large-scale datasets. We evaluate the performance of both pre-trained and fine-tuned models on the APTOS 2019 dataset, as summarized in Table 21. Both DenseNet161 and Swin s show impressive improvements following fine-tuning, achieving F1 macro scores of 0.680 and 0.693, respectively, compared to 0.657 and 0.659 in their pre-trained model. As a result, this fine-tuning provide the substantial benefit, especially in this challenging task, which involves imbalanced class distributions and overlapping lesion features in advanced stages (classes 2, 3, and 4). In such a context, even a 5% gain in F1 macro score is a significant enhancement in model performance.

Table 21 The classification performance of DenseNet161 and Swin s models before and after fine-tuning. Pretrained models are trained on ImageNet-1K dataset. Fine-tuned models are pretrained model and then, tuned on APTOS 2019 dataset. Notably, we use only **macro** scores for comparison.

Model name	Pretrained			Fine-tuned		
	Precision	Recall	F1	Precision	Recall	F1
DenseNet161	0.691	0.637	0.657	0.710	0.662	0.680
Swin s	0.685	0.654	0.659	0.730	0.687	0.693

To assess the model's ability to understand and represent data, we extract feature vectors from both pre-trained and fine-tuned backbone models. We check these representations in two ways: by using a K-Nearest Neighbors (K-NN) classifier for quantitative evaluation and t-distributed stochastic neighbor embedding (t-SNE) for qualitative evaluation. The K-NN results reveal that fine-tuning substantially enhances classification performance for both RGB and grayscale images, with average improvements of approximately 9.2% for DenseNet 161 and 16.1% for Swin s, as shown in Table 22. Notably, the fine-tuned Swin Transformer using RGB images achieves the highest validation score of 0.716, demonstrating its strong representational capacity. This result underscores the model's suitability for downstream classification tasks, as it effectively captures meaningful and discriminative features from the input data. The t-SNE visualizations in Figures 33 and 34 visualize the feature spaces of Swin and DenseNet161, demonstrating the influence of fine-tuning and input modality. Fine-tuning significantly improves class separability in both models, with a clear improvement in classifying abnormal classes, corresponding to diabetic retinopathy (DR) severity levels greater than 1. In the fine-tuned models, abnormal

classes form well-defined clusters, whereas in the pre-trained models, these classes are more intermingled and thus more challenging to classify. In conclusion, these studies confirm that fine-tuning enhances the quality of learned representations, enabling better identification of diabetic retinopathy stages.

Table 22 The classification performance of KNN model with different image types and pretrained models. Notably, we use only **macro** score for comparison

Model name	Image type	Pretrained		Fine-tuned	
		F1 (Train)	F1 (Val)	F1 (Train)	F1 (Val)
DenseNet161	Gray	0.968	0.593	0.978	0.647
	RGB	0.950	0.552	0.983	0.682
Swin s	Gray	0.935	0.532	0.977	0.684
	RGB	0.895	0.546	0.984	0.716

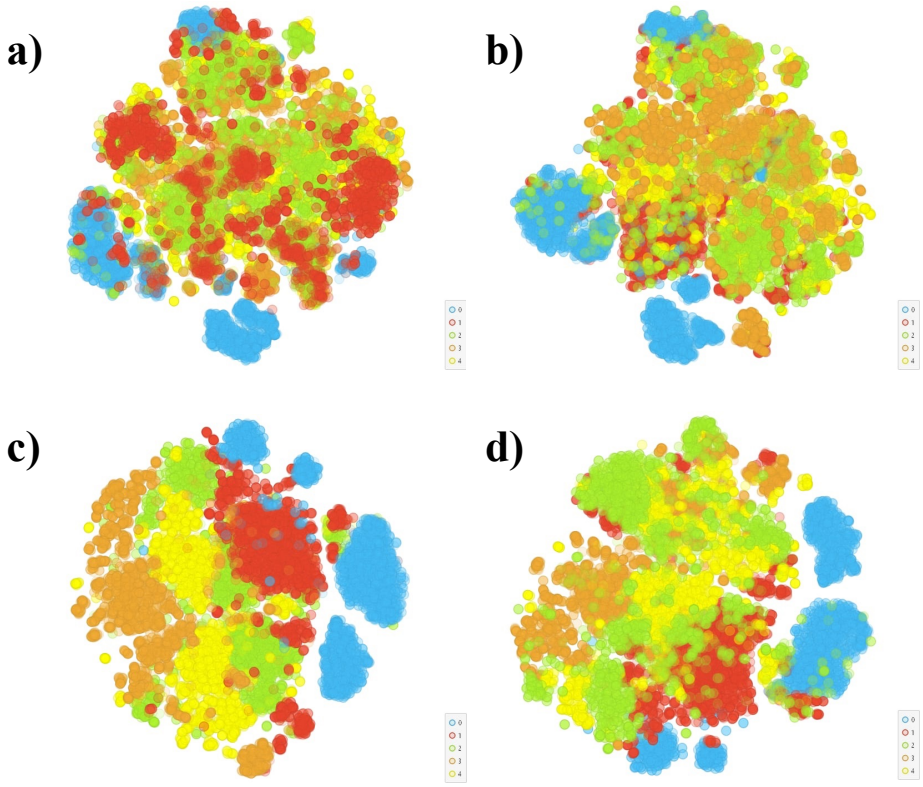

Figure 33 The t-SNE visualization of the DenseNet 161 model's feature space. The colors represent different classes; cyan is class 0; red is class 1; green is class 2; orange is class 3; and yellow is class 4, respectively. a) and b) illustrate the feature space of model without tuning while c) and d) illustrate the feature space of model with tuning. Moreover, a) and c) are the feature space of RGB images while b) and d) are the feature space of grayscale images.

To further examine the impact of fine-tuning on model performance under balanced test conditions, the results are summarized in Table 23. The fine-tuned DenseNet161 model achieves an improved F1 macro score of 0.657, compared to

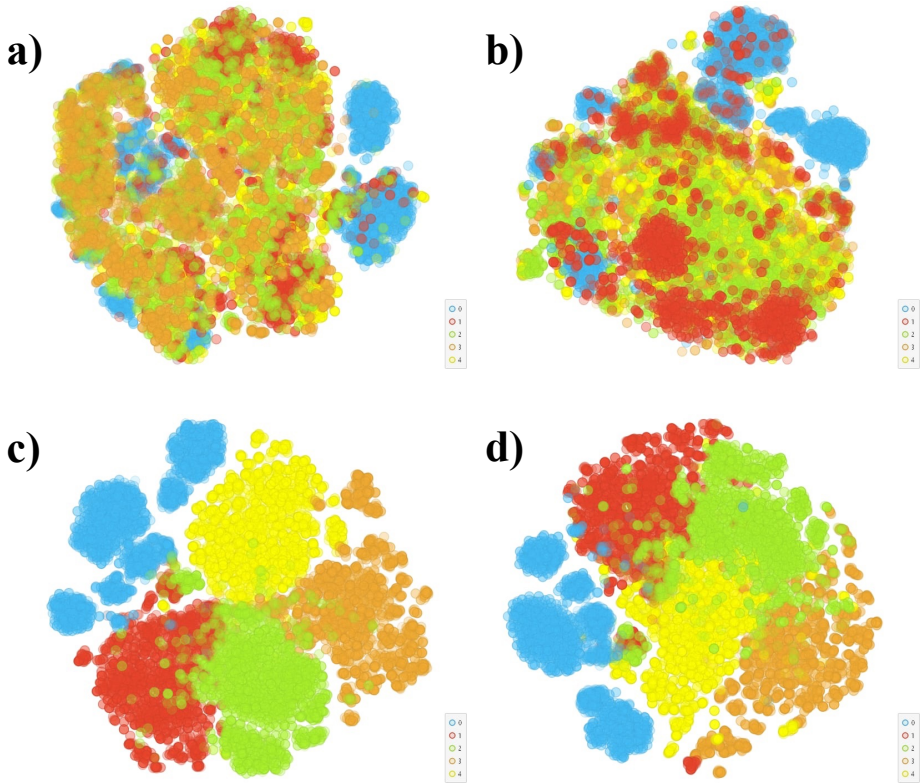

Figure 34 The t-SNE visualization of the Swin s model's feature space. The colors represent different classes; cyan is class 0; red is class 1; green is class 2; orange is class 3; and yellow is class 4, respectively. a) and b) illustrate the feature space of model without tuning while c) and d) illustrate the feature space of model with tuning. Moreover, a) and c) are the feature space of RGB images while b) and d) are the feature space of grayscale images.

its pre-trained model. In contrast, the Swin Transformer exhibits only a marginal improvement of 0.01 in F1 macro score after fine-tuning. These results show lower performance than those model obtained on the imbalanced test set, suggest that the models still face challenges in accurately classifying DR severity levels, although after fine-tuning. Furthermore, the KNN result highlights that RGB input images yield greater performance gains than grayscale images, especially in the fine-tuned setting 24. The highest validation F1 score of 0.701 is achieved by the KNN classifier using features extracted from the fine-tuned Swin Transformer trained on RGB inputs. These results confirm the importance of fine-tuning and color information in boosting the discriminative capacity of vision models for DR grading.

Table 23 The classification performance of DenseNet161 and Swin s models before and after fine-tuning. Pretrained models are trained on ImageNet-1K dataset. Fine-tuned models are pretrained model and then, tuned on APTOS 2019 dataset. Notably, we use only **macro** scores for comparison.

Model name	Pretrained			Fine-tuned		
	Precision	Recall	F1	Precision	Recall	F1
DenseNet161	0.659	0.605	0.601	0.733	0.658	0.657
Swin s	0.723	0.679	0.680	0.745	0.684	0.681

Table 24 The classification performance of KNN model with different image types and pretrained models. Notably, we use only **macro** score for comparison

Model name	Image type	Pretrained		Fine-tuned	
		F1 (Train)	F1 (Val)	F1 (Train)	F1 (Val)
DenseNet161	**Gray**	0.968	0.635	0.978	0.622
	RGB	0.950	0.562	0.983	0.692
Swin s	**Gray**	0.935	0.576	0.977	0.684
	RGB	0.895	0.564	0.984	0.701

CHAPTER 5

CONCLUSION

This thesis introduces a comprehensive end-to-end diabetic retinopathy (DR) classification framework based on retinal fundus imaging, comprising two primary modules: an image screening module and a DR grading module. The screening module employs a template-based correlation filtering approach to detect crucial ocular structures, namely the optic disc and macula, followed by either a rule-based or machine learning (ML)-based decision mechanism to classify images as medically suitable or unsuitable. Extensive evaluation on private datasets confirms the effectiveness of the screening method, which achieves a high recall of 0.906, ensuring the retention of diagnostically valuable images, and a low false discovery rate (FDR = 0.065), indicating a high chance of excluding clinically relevant images. The performance is further enhanced by the use of anatomically tailored template designs and optimal sampling strategies, with the normalized correlation coefficient (CCORR_NORMED) demonstrating robust matching performance under varying illumination conditions. Additionally, the integration of a region-of-interest (ROI) framework further improves the reliability of macula detection. Ultimately, out-of-distribution testing confirms that the proposed method maintains competitive performance when generalized to unseen datasets. In the ML context, generalized models such as Random Forest and Logistic Regression demonstrate promising robustness to distributional shifts, highlighting their potential utility in real-world screening applications. As a result, they reveal the opportunity to improve this work with advanced ML in the future.

Our diabetic retinopathy (DR) grading module employs the Swin Transformer (S) network as the backbone, coupled with a fully connected neural network for classification. In the context of imbalanced data, we address it by using SMOTE to synthesize data along with various data augmentation strategies. The model is further enhanced through fine-tuning on the APTOS 2019 dataset. Moreover, comprehensive evaluation demonstrates that the proposed model achieves an F1 macro score of 0.693 and a quadratic weighted kappa (QWK) score of 0.903, surpassing previously reported methods in both metrics. Moreover, the classification report indicates strong performance in detecting early DR stages, suggesting that the model

is well-suited for early-stage DR screening applications. Nevertheless, detailed analysis via the confusion matrix and classification report reveals underperformance in minority classes, especially severe and proliferative DR, due to overlapping lesion characteristics and limited training data. Additionally, a performance gap remains between the model and expert ophthalmologists, with an F1 macro score of 0.714. To further improve the performance of deep learning-based algorithms, three critical cores must be considered: data quality and quantity, model capacity (i.e., number of parameters), and computational resources. Given the current limitations, our primary focus is on improving the dataset, as the existing data are both imbalanced and limited in size. This constraint has led to suboptimal performance, particularly in minority classes such as severe NPDR and PDR. To address this, we plan to expand the dataset by collecting more diverse samples and by leveraging generative models, such as generative adversarial networks (GANs), variational autoencoders (VAEs), or diffusion models, to synthesize images of advanced DR stages. Once the data-related issues are mitigated, we will focus on increasing model capacity by incorporating architectures with a larger number of parameters, followed by scaling computational resources to support training on larger models and datasets.

In the context of innovation, our work focuses on the development of a comprehensive end-to-end system that integrates both DR grading and image screening modules. The DR grading module is designed to assess the severity level of diabetic retinopathy solely based on a patient's fundus image, thereby offering a user-friendly and accessible diagnostic tool. In parallel, the image screening module evaluates the quality of the fundus images, which is a critical component. By automatically filtering out medically unsuitable images prior to manual labeling, the system reduces the burden on ophthalmologists, enabling them to concentrate on their primary task of annotating images. As a result of this module, we obtain a more informative and higher-quality dataset, which in turn enhances the performance of the DR grading model. Ultimately, by integrating these modules with existing resources, including ophthalmologists, a data lake, and a diagnostic web application, we establish a functional end-to-end system. This platform supports automated DR severity grading through telemedicine while enabling continuous model improvement through feedback and data accumulation. From a market perspective, the global artificial intelligence market for retinal image analysis is valued at approximately USD 147.8

million in 2024 (Transpire, 2024), with a projected compound annual growth rate (CAGR) of 13.5%. Of this, clinical diagnosis and early detection screening account for an estimated 45.6%, or USD 67.4 million (market.us, 2024). Major players in the field, such as Topcon Corporation, RetinAI Medical AG, and Eyenuk Inc., leverage advantages like 24/7 ophthalmologist-backed systems and advanced AI technologies. However, these solutions are frequently associated with high costs and limited scalability to local or rural areas. Our approach distinguishes itself by targeting affordability and accessibility. We aim to develop both hardware and software solutions that are affordable cost for deployment in local areas, including local healthcare centers, medium-sized hospitals, and small hospitals, estimated at approximately 890 facilities nationwide. Notably, there are currently no domestic competitors in Thailand offering a fully integrated solution comprising both diagnostic instruments and software. The present situation presents a great opportunity for market leadership within the country, with the potential to scale further into the Southeast Asian region.

In the context of practice, The system is currently at the proof-of-concept stage but has been developed with practical deployment. The image screening module has been evaluated on both public and private datasets, including those collected in Thailand, indicating its readiness for real-world implementation. In contrast, while the DR grading module demonstrates reliable performance, surpassing related works and closing the level of ophthalmologists, it has only been validated on a single dataset. As a result, further evaluation across multiple datasets from local hospitals is necessary to assess generalizability and strengthen clinical reliability. At present, the project comprises a fully integrated pipeline, including a data lake, labeller, diagnostic web application, and an image screening system. Thus, the DR grading component remains the final element requiring refinement before full deployment as an automated clinical decision-making system. Nevertheless, if the goal is to establish a decision support system rather than full automation, the current DR grading model is sufficiently reliable, particularly for early-stage detection, and can already serve as a valuable tool for assisting ophthalmologists and physicians. This is especially beneficial in local or rural healthcare settings where access to retinal specialists is limited. Finally, the system can act as a front-line screening tool to flag potentially abnormal cases and support clinicians in making timely

and accurate decisions, ultimately improving early detection and accelerating referrals for further examination. In the long-term strategic plan, the proposed system is designed for continuous improvement. Upon integration of new datasets from additional hospitals, the image screening module will automatically filter medically suitable images and then store them into a cleaned database, and the diagnostic outcomes provided by ophthalmologists will be used to continuously retrain the DR grading model. Once the model achieves a reliable and clinically acceptable level of performance, we plan to deploy the system in partner hospitals such as Suranaree University of Technology Hospital (SUT) and Maharat Nakhon Ratchasima Hospital to collect real-world feedback and further optimize the system. In parallel, we will make efforts directed toward expanding collaborations with other hospitals to increase data diversity and model robustness. Simultaneously, we aim to secure small-scale research funding to support the development of intellectual property, including the pursuit of patents. These patents can serve as collateral for acquiring larger-scale funding to further scale development and deployment. Moreover, as there is not a current domestic competitor offering both integrated software and hardware solutions, this positions us with a unique opportunity to lead the national market. If the implementation proceeds successfully, we plan to expand regionally, targeting the Southeast Asian healthcare market, where similar needs and infrastructure gaps exist.

In summary, this work introduces an end-to-end system for diabetic retinopathy classification, integrating image screening and DR grading modules. The screening component demonstrates readiness for deployment with high recall and low FDR, while the grading model, based on a fine-tuned Swin Transformer, achieves competitive performance with a macro F1-score of 0.693 and QWK of 0.903 but performance on minority classes remains a challenge. Therfore, future work will emphasize on dataset expansion in both diversity and size, and model scaling. As a result of existing infrastructure, including a data lake, web application, and screening system, this work is positioned for real-world decision support application, particularly in local and rural healthcare settings. Fianlly, the system holds strong potential for clinical integration, continuous improvement, and national deployment as a AI-driven diagnostic solution.

REFERENCES

Al-Bander, B., Al-Nuaimy, W., Williams, B. M., and Zheng, Y. (2018). Multiscale sequential convolutional neural networks for simultaneous detection of fovea and optic disc. *Biomedical Signal Processing and Control 40*, 91–101.

Barbara Davis Center for Diabetes School of Medicine (2024). *Discussion and Pictures of Diabetic Retinopathy Lesions*. Retrieved from https://medschool.cuanschutz.edu/barbara-davis-center-for-diabetes/patient-care/ophthalmology/discussion-and-pictures-of-diabetic-retinopathy-lesions. Accessed: 2024-02-14.

Bhatkalkar, B. J., Nayak, S. V., Shenoy, S. V., and Arjunan, R. V. (2021). FundusPosNet: A deep learning driven heatmap regression model for the joint localization of optic disc and fovea centers in color fundus images. *IEEE Access 9*, 159071–159080.

Bodapati, J. D., Shaik, N. S., and Naralasetti, V. (2021). Composite deep neural network with gated-attention mechanism for diabetic retinopathy severity classification. *Journal of Ambient Intelligence and Humanized Computing 12*(10), 9825–9839.

Chalakkal, R. J., Abdulla, W. H., and Thulaseedharan, S. S. (2018). Automatic detection and segmentation of optic disc and fovea in retinal images. *IET Image Processing 12*(11), 2100–2110.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research 16*, 321–357.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., and Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

Coyner, A. S., Swan, R., Brown, J. M., Kalpathy-Cramer, J., Kim, S. J., Campbell, J. P., Jonas, K., Chan, R. P., Ostmo, S., and Chiang, M. F. (2018). Deep learning for image quality assessment of fundus images in retinopathy of prematurity. *Investigative Ophthalmology & Visual Science 59*(9), 2762–2762.

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

Deka, D., Medhi, J. P., and Nirmala, S. (2015). Detection of macula and fovea for disease analysis in color fundus images. *2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)*. IEEE, 231–236.

Department of Provincial Administration (2024). *Number of people throughout the kingdom According to evidence of civil registration*. Retrieved from https://www.dopa.go.th/news/cate1/view9265. Accessed: 2024-02-14, Created: 2024-02-09.

Dinç, B. and Kaya, Y. (2023). A novel hybrid optic disc detection and fovea localization method integrating region-based convnet and mathematical approach. *Wireless Personal Communications 129*(4), 2727–2748.

Estopinal, C. B., Ausayakhun, S., Ausayakhun, S., Jirawison, C., Joy Bhosai, S., Margolis, T. P., and Keenan, J. D. (2013). Access to ophthalmologic care in Thailand: a regional analysis. *Ophthalmic epidemiology 20*(5), 267–273.

Farag, M. M., Fouad, M., and Abdel-Hamid, A. T. (2022). Automatic severity classification of diabetic retinopathy based on densenet and convolutional block attention module. *IEEE Access 10*, 38299–38308.

Fleming, A. D., Philip, S., Goatman, K. A., Olson, J. A., and Sharp, P. F. (2006). Automated assessment of diabetic retinal image quality based on clarity and field definition. *Investigative ophthalmology & visual science 47*(3), 1120–1125.

Fu, Y., Zhang, G., Li, J., Pan, D., Wang, Y., and Zhang, D. (2022). Fovea localization by blood vessel vector in abnormal fundus images. *Pattern Recognition 129*, 108711.

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.

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

Hasan, M. K., Alam, M. A., Elahi, M. T. E., Roy, S., and Martí, R. (2021). DRNet: Segmentation and localization of optic disc and Fovea from diabetic retinopathy image. *Artificial Intelligence in Medicine 111*, 102001.

He, A., Li, T., Li, N., Wang, K., and Fu, H. (2020). CABNet: Category attention block for imbalanced diabetic retinopathy grading. *IEEE Transactions on Medical Imaging 40*(1), 143–153.

He, H., Lin, L., Cai, Z., Cheng, P., and Tang, X. (2023). JOINEDTrans: Prior Guided Multi-task Transformer for Joint Optic Disc/Cup Segmentation and Fovea Detection. *arXiv preprint arXiv:2305.11504*.

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

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Huang, Y., Zhong, Z., Yuan, J., and Tang, X. (2020). Efficient and robust optic disc detection and fovea localization using region proposal network and cascaded network. *Biomedical Signal Processing and Control 60*, 101939.

Isipradit, S., Sirimaharaj, M., Charukamnoetkanok, P., Thonginnetra, O., Wongsawad, W., Sathornsumetee, B., Somboonthanakij, S., Soomsawasdi, P., Jitawatanarat, U., Taweebanjongsin, W., (2014). The first rapid assessment of avoidable blindness (RAAB) in Thailand. *PloS one 9*(12), e114245.

Islam, M. R., Abdulrazak, L. F., Nahiduzzaman, M., Goni, M. O. F., Anower, M. S., Ahsan, M., Haider, J., and Kowalski, M. (2022). Applying supervised contrastive learning for the detection of diabetic retinopathy and its severity levels from fundus images. *Computers in Biology and Medicine 146*, 105602.

Kamble, R., Kokare, M., Deshmukh, G., Hussin, F. A., and Mériaudeau, F. (2017). Localization of optic disc and fovea in retinal images using intensity based line scanning analysis. *Computers in biology and medicine 87*, 382–396.

Karthik Maggie, S. D. (2019). *APTOS 2019 Blindness Detection*. Retrieved from https://kaggle.com/competitions/aptos2019-blindness-detection.

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.

La Torre, J. de, Valls, A., and Puig, D. (2020). A deep learning interpretable classifier for diabetic retinopathy disease grading. *Neurocomputing 396*, 465–476.

Li, X., Jiang, Y., Zhang, J., Li, M., Luo, H., and Yin, S. (2022). Lesion-attention pyramid network for diabetic retinopathy grading. *Artificial Intelligence in Medicine 126*, 102259.

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

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

market.us (2024). *AI Powered Retina Image Analysis Market*. "Retrieved from https://market.us/report/ai-powered-retina-image-analysis-market/.

Medhi, J. P. and Dandapat, S. (2016). An effective fovea detection and automatic assessment of diabetic maculopathy in color fundus images. *Computers in biology and medicine 74*, 30–44.

Mvoulana, A., Kachouri, R., and Akil, M. (2019). Fully automated method for glaucoma screening using robust optic nerve head detection and unsupervised segmentation based cup-to-disc ratio computation in retinal fundus images. *Computerized Medical Imaging and Graphics 77*, 101643.

Nayak, J., Bhat, P. S., and Acharya, U. (2009). Automatic identification of diabetic maculopathy stages using fundus images. *Journal of medical engineering & technology 33*(2), 119–129.

Niemeijer, M., Abràmoff, M. D., and Van Ginneken, B. (2009). Fast detection of the optic disc and fovea in color fundus photographs. *Medical image analysis 13*(6), 859–870.

Ouyang, J., Mao, D., Guo, Z., Liu, S., Xu, D., and Wang, W. (2023). Contrastive self-supervised learning for diabetic retinopathy early detection. *Medical & Biological Engineering & Computing 61*(9), 2441–2452.

Palanisamy, G., Ponnusamy, P., and Gopi, V. P. (2023). An adaptive enhancement and fovea detection technique for color fundus image analysis. *Signal, Image and Video Processing 17*(3), 831–838.

Parsa, S. and Khatibi, T. (2024). Grading the severity of diabetic retinopathy using an ensemble of self-supervised pre-trained convolutional neural networks: ESSP-CNNs. *Multimedia Tools and Applications*, 1–34.

Paszke, A. (2019). Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research 12*, 2825–2830.

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).

Quellec, G., Al Hajj, H., Lamard, M., Conze, P.-H., Massin, P., and Cochener, B. (2021). ExplAIn: Explanatory artificial intelligence for diabetic retinopathy diagnosis. *Medical Image Analysis 72*, 102118.

Qummar, S., Khan, F. G., Shah, S., Khan, A., Shamshirband, S., Rehman, Z. U., Khan, I. A., and Jadoon, W. (2019). A deep learning ensemble approach for diabetic retinopathy detection. *Ieee Access 7*, 150530–150539.

Resnikoff, S., Lansingh, V. C., Washburn, L., Felch, W., Gauthier, T.-M., Taylor, H. R., Eckert, K., Parke, D., and Wiedemann, P. (2019). Estimated number of ophthalmologists worldwide (International Council of Ophthalmology update): will we meet the needs? *British Journal of Ophthalmology*.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 234–241.

Royal College of Ophthalmologists of Thailand (2024). *Ophthalmologists in Thailand*. Retrieved from http://rcopt.org. Accessed: 2024-02-14.

Sedai, S., Tennakoon, R., Roy, P., Cao, K., and Garnavi, R. (2017). Multi-stage segmentation of the fovea in retinal fundus images using fully convolutional neural networks. *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, 1083–1086.

Sekhar, S., Al-Nuaimy, W., and Nandi, A. K. (2008). Automated localisation of optic disk and fovea in retinal fundus images. *2008 16th European Signal Processing Conference*. IEEE, 1–5.

Şevik, U., Köse, C., Berber, T., and Erdöl, H. (2014). Identification of suitable fundus images using automated quality assessment methods. *Journal of biomedical optics 19*(4), 046006–046006.

Shakibania, H., Raoufi, S., Pourafkham, B., Khotanlou, H., and Mansoorizadeh, M. (2023). Dual Branch Deep Learning Network for Detection and Stage Grading of Diabetic Retinopathy. *arXiv preprint arXiv:2308.09945*.

Shankar, K., Zhang, Y., Liu, Y., Wu, L., and Chen, C.-H. (2020). Hyperparameter tuning deep learning for diabetic retinopathy fundus image classification. *IEEE Access 8*, 118164–118173.

Sigut, J., Nuñez, O., Fumero, F., Alayon, S., and Diaz-Aleman, T. (2023). Fovea localization in retinal images using spatial color histograms. *Multimedia Tools and Applications*, 1–19.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sinthanayothin, C., Boyce, J. F., Cook, H. L., and Williamson, T. H. (1999). Automated localisation of the optic disc, fovea, and retinal blood vessels from digital colour fundus images. *British journal of ophthalmology 83*(8), 902–910.

Song, S., Dang, K., Yu, Q., Wang, Z., Coenen, F., Su, J., and Ding, X. (2022). Bilateral-vit for robust fovea localization. *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 1–5.

Sun, R., Li, Y., Zhang, T., Mao, Z., Wu, F., and Zhang, Y. (2021). Lesion-aware transformers for diabetic retinopathy grading. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10938–10947.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Tan, J. H., Acharya, U. R., Bhandary, S. V., Chua, K. C., and Sivaprasad, S. (2017). Segmentation of optic disc, fovea and retinal vasculature using a single convolutional neural network. *Journal of Computational Science 20*, 70–79.

Tariq, A., Shaukat, A., and Khan, S. A. (2012). A Gaussian mixture model based system for detection of macula in fundus images. *Neural Information Processing: 19th International Conference, ICONIP 2012, Doha, Qatar, November 12-15, 2012, Proceedings, Part II 19*. Springer, 33–40.

Teo, Z. L., Tham, Y.-C., Yu, M., Chee, M. L., Rim, T. H., Cheung, N., Bikbov, M. M., Wang, Y. X., Tang, Y., Lu, Y., (2021). Global prevalence of diabetic retinopathy and projec-

tion of burden through 2045: systematic review and meta-analysis. *Ophthalmology 128*(11), 1580–1591.

Transpire (2024). *AI in Diabetic Retinopathy Market, Forecast to 2032*. "Retrived from https://www.transpireinsight.com/report/ai-in-diabetic-retinopathy-market.

Tusfiqur, H. M., Nguyen, D. M., Truong, M. T., Nguyen, T. A., Nguyen, B. T., Barz, M., Profitlich, H.-J., Than, N. T., Le, N., Xie, P., (2022). DRG-Net: Interactive Joint Learning of Multi-lesion Segmentation and Classification for Diabetic Retinopathy Grading. *arXiv preprint arXiv:2212.14615*.

Usman Akram, M., Khan, A., Iqbal, K., and Butt, W. H. (2010). Retinal images: optic disk localization and detection. *Image Analysis and Recognition: 7th International Conference, ICIAR 2010, Póvoa de Varzin, Portugal, June 21-23, 2010, Proceedings, Part II 7*. Springer, 40–49.

Welfer, D., Scharcanski, J., and Marinho, D. R. (2011). Fovea center detection based on the retina anatomy and mathematical morphology. *Computer methods and programs in biomedicine 104*(3), 397–409.

Wilkinson, C. P., Ferris III, F. L., Klein, R. E., Lee, P. P., Agardh, C. D., Davis, M., Dills, D., Kampik, A., Pararajasegaram, R., Verdaguer, J. T., (2003). Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology 110*(9), 1677–1682.

Wong, T. Y., Sun, J., Kawasaki, R., Ruamviboonsuk, P., Gupta, N., Lansingh, V. C., Maia, M., Mathenge, W., Moreker, S., Muqit, M. M., (2018). Guidelines on diabetic eye care: the international council of ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings. *Ophthalmology 125*(10), 1608–1622.

Wykoff, C. C., Khurana, R. N., Nguyen, Q. D., Kelly, S. P., Lum, F., Hall, R., Abbass, I. M., Abolian, A. M., Stoilov, I., To, T. M., (2021). Risk of blindness among patients with diabetes and newly diagnosed diabetic retinopathy. *Diabetes care 44*(3), 748–756.

Xue, J., Wu, J., Bian, Y., Zhang, S., and Du, Q. (2024). Classification of Diabetic Retinopathy Based on Efficient Computational Modeling. *Applied Sciences 14*(23), 11327.

Yu, H., Barriga, E. S., Agurto, C., Echegaray, S., Pattichis, M. S., Bauman, W., and Soliz, P. (2012). Fast localization and segmentation of optic disk in retinal images using directional matched filtering and level sets. *IEEE Transactions on information technology in biomedicine 16*(4), 644–657.

Zhang, W., Zhong, J., Yang, S., Gao, Z., Hu, J., Chen, Y., and Yi, Z. (2019). Automated identification and grading system of diabetic retinopathy using deep neural networks. *Knowledge-Based Systems 175*, 12–25.

Zheng, S., Pan, L., Chen, J., and Yu, L. (2014). Automatic and efficient detection of the fovea center in retinal images. *2014 7th International Conference on Biomedical Engineering and Informatics*. IEEE, 145–150.

APPENDIX

PUBLICATIONS

CODE AVAILABILITY

The code for this project is available on GitHub at the following link: https://github.com/men31/Development-of-Deep-Learning-for-Diabetic-Retinopathy-Classification-System-Based-on-Fundus-Image. The code is open-source and can be freely accessed, modified, and distributed under the term of MIT lincese.

ETHICAL APPROVAL

This thesis is a part of project, named "Prototype of a Three-dimensional Retina Scanner", which is approved by the Ethics Committee of Suranaree University of Technology (SUT). The project is conducted under the supervision of Assoc. Prof. Dr. Panomsak Meemon. Moreover, the student who is the author of this thesis has passed the "Human Subject Protection Course" from SUT and "Ethics of AI" course from University of Helsinki.

สถาบันวิจัยและพัฒนา
มหาวิทยาลัยเทคโนโลยีสุรนารี

ขอมอบเกียรติบัตรฉบับนี้ไว้เพื่อแสดงว่า

รพีพัฒน์ ยศสูงเนิน

ได้ผ่านการอบรมจริยธรรมการวิจัยในมนุษย์

หลักสูตร "Human Subject Protection Course"
Basic Good Clinical Practice (GCP) Training Course

ในระหว่างวันที่ 24-25 สิงหาคม 2566 ผ่าน Application Zoom Cloud Meetings ณ มหาวิทยาลัยเทคโนโลยีสุรนารี

ให้ไว้ ณ วันที่ 25 สิงหาคม 2566

วันหมดอายุ 25 สิงหาคม 2569

รองศาสตราจารย์ ดร.ระพี อูทเคอ
ผู้อำนวยการสถาบันวิจัยและพัฒนา
มหาวิทยาลัยเทคโนโลยีสุรนารี

ผู้ช่วยศาสตราจารย์ ดร.เบ็ญจมาศ จิตรสมบูรณ์
ประธานคณะกรรมการจริยธรรมการวิจัยในมนุษย์
มหาวิทยาลัยเทคโนโลยีสุรนารี

HSP-EC-SUT-135/2023

UNIVERSITY OF HELSINKI

Ethics of AI

Certificate of Completion

Rapeephat Yodsungnoen

has successfully completed the 2 ECTS online course

Jukka K. Nurminen
Professor, University of Helsinki

November 4, 2023

https://certificates.mooc.fi/validate/me50evy8ue

FCAI HiDATA Helsinki City of Amsterdam SUPPORTED BY MAYOR OF LONDON

CURRICULUM VITAE

Name	Rapeephat Yodsungnoen
Date of Birth	14 May 2000
Place of Birth	Nakhon Ratchasima, Thailand
Education	2022–2024 M.Sc. Integrated Science and Innovation (ISI), Module of Applied Machine Learning and Scientific Data Analysis, Institute of Science, Suranaree University of Technology 2018–2022 B.Sc. Physics, Institute of Science, Suranaree University of Technology