CHAPTER II
LITERATURE REVIEW

2.1 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 2.1. 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 2.2.

- **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 2.2.

- **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 2.2.

- **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 2.2.

- **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 2.2.

- **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 2.2.

- **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 2.2.

- **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 2.2.

- **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 2.2.

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 (Gulshan et al., 2016; Wilkinson et al., 2003). 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 a

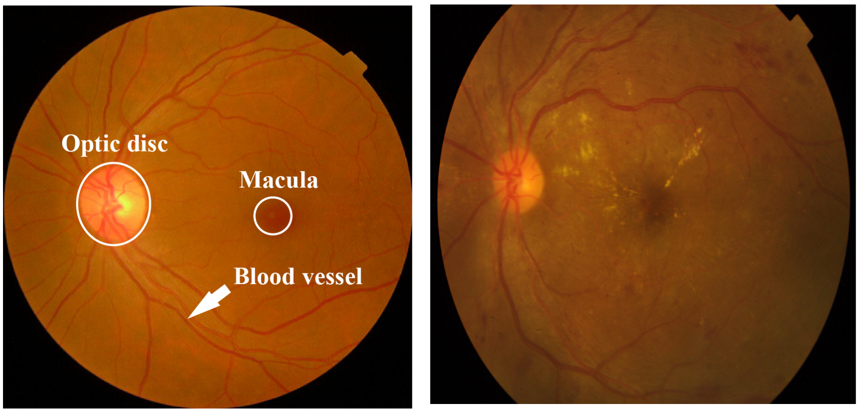

Figure 2.1 The retinal image comparison of normal and DR.

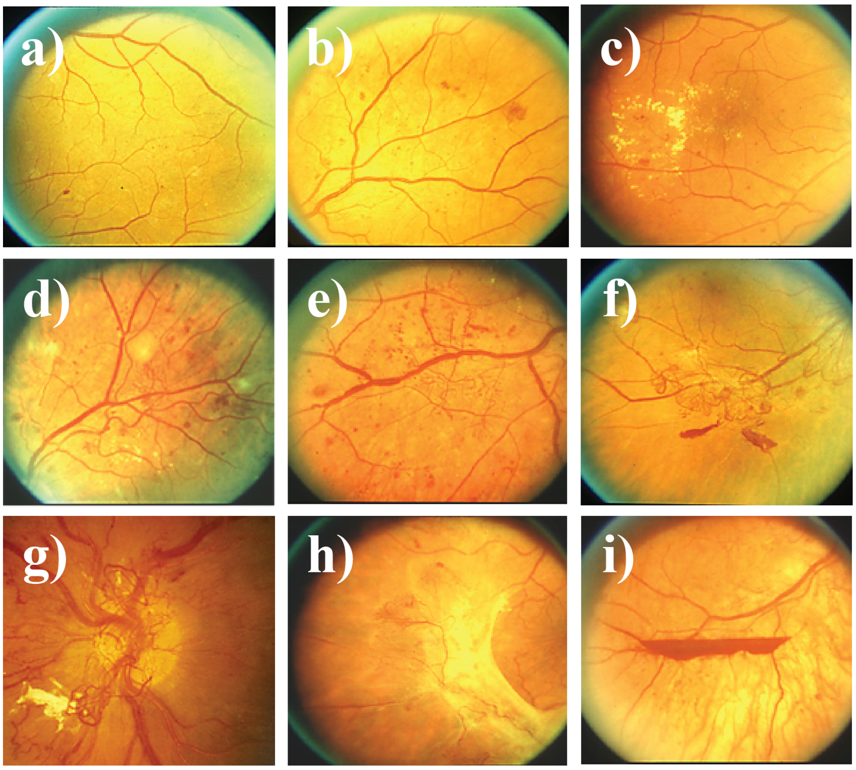

Figure 2.2 Diabetic retinopathy (DR) lesions are categorized based on the progression of lesion development; ranging from mild to severe stages. a) Microaneurysms (MA); b) Retinal Hemorrhage (HM); c) Hard Exudates (HE); d) Soft Exudates (SE); e) Intraretinal Microvascular Anomalies (IRMA) and Venous 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.

PDR in the next four year and have to re-examination in 1-2 years based on American Academy of Ophthalmology (AAO) guidelines (Wong et al., 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 et al., 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 2.3.

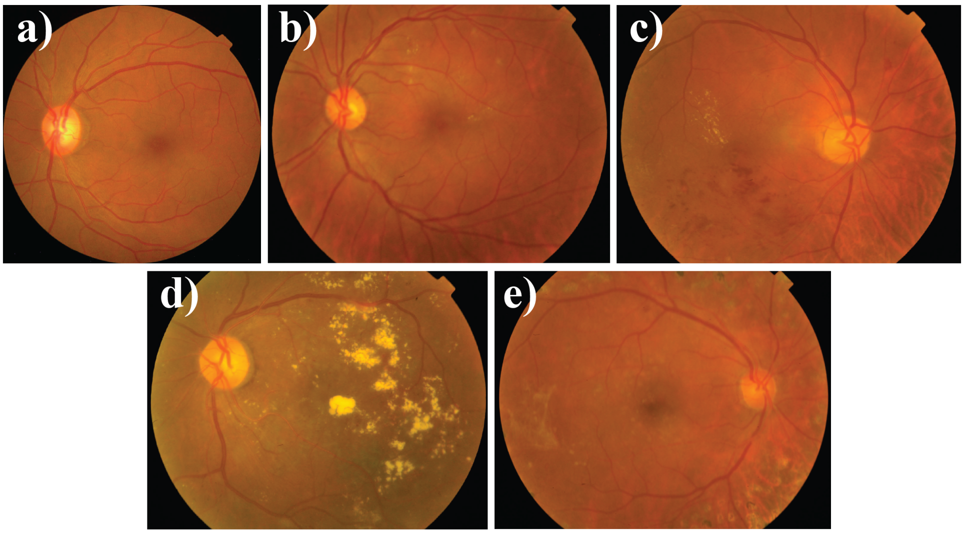

Figure 2.3 The figure illustrates the various severity levels of DR in the patient’s retina: a) No DR; b) Mild NPDR; c) Moderate NPDR; d) Severe NPDR; and e) PDR Karthik, 2019.

2.2 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.

2.2.1 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 et al., 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 (Deka et al., 2015; Kamble et al., 2017; Medhi and Dandapat, 2016; Palanisamy et al., 2023; Sinthanayothin et al., 1999). 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 (Mvoulana et al., 2019; Sekhar et al., 2008; Usman Akram et al., 2010). Furthermore, resizing the images is commonly performed to reduce computational resources and accelerate processing speed (Palanisamy et al., 2023; Sinthanayothin et al., 1999; Zheng et al., 2014). Lastly, optic disc detection involves employing various feature extraction techniques to extract valuable information and locate the optic disc. (Sinthanayothin et al., 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 et al., 2008), (Usman Akram et al., 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 (Chalakkal et al., 2018; Fu et al., 2022; Medhi and Dandapat, 2016; Welfer et al., 2011). 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 et al., 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 et al., 2018; Mvoulana et al., 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 (Yu et al., 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 (Chalakkal et al., 2018; Dinç and Kaya, 2023; Fu et al., 2022; Palanisamy et al., 2023; Sinthanayothin et al., 1999; Welfer et al., 2011). 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 et al., 2008; Zheng et al., 2014). Furthermore, the vascular structure serves as significant evidence to indicate the region of the macula. In (Deka et al., 2015; Medhi and Dandapat, 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 et al., 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 (Chalakkal et al., 2018; Deka et al., 2015; Sekhar et al., 2008; Welfer et al., 2011; Zheng et al., 2014). Subsequently, thresholding algorithms, such as the brute-force or Otsu algorithm, are utilized to segment the macula (Chalakkal et al., 2018; Fu et al., 2022; Medhi and Dandapat, 2016; Palanisamy et al., 2023). Additionally, in (Nayak et al., 2009; Sinthanayothin et al., 1999), 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.

2.2.2 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 et al., 2017) proposes a custom CNN with 125k learnable parameters, 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 et al., 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 et al., 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 et al., 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 (Xie et al., 2020), 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 et al., 2015), it has garnered significant attention from retinal deep learning researchers in leveraging the U-net, particularly localization and segmentation. For example,(Bhatkalkar et al., 2021; Hasan et al., 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 (H. He et al., 2023; Song et al., 2022), 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 et al., 2021), the approach integrates a vascular segmentation network into the encoder in (Song et al., 2022). Furthermore, in (H. He et al., 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.

2.3 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, 2.1. In this literature review, we demonstrate only multi-classification because it aligns with our research.

2.3.1 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 process 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 3.1. 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 (Zhang et al., 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 et al., 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 et al., 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 and Khatibi, 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 et al., 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 et al., 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 et al., 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 and Khatibi, 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 et al., 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 (de La Torre et al., 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 (Sun et al., 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 et al., 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 et al., 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.

2.4 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.

2.4.1 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 2.4 can demonstrate the principle components of the confusion matrix.

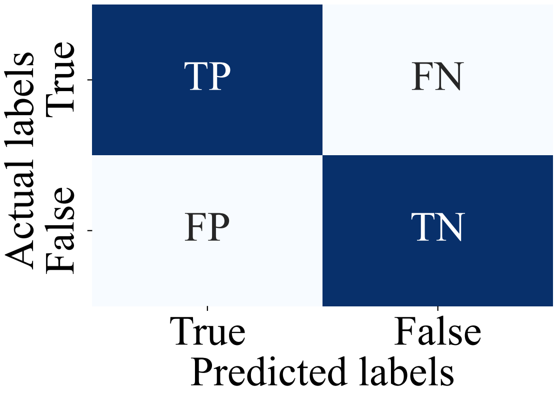

Figure 2.4 The confusion matrix.

2.4.2 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. 2.1). 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. 2.2) and (Eq. 2.3), respectively.

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{2.1}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2.2}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2.3}$$

2.4.3 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. 2.4), 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.

$$\mathrm{F1} = \frac{2}{\frac{1}{\mathrm{Precision}} + \frac{1}{\mathrm{Recall}}} \tag{2.4}$$

2.4.4 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. 2.5a), and the false positive rate (FPR), as expressed in (Eq. 2.5b). Additionally, Figure 2.5 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).

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2.5a}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{2.5b}$$

Furthermore, in Figure 2.6, we show the interpretation of the ROC curve in various patterns, which assist the reader in clearly understanding the ROC curve.

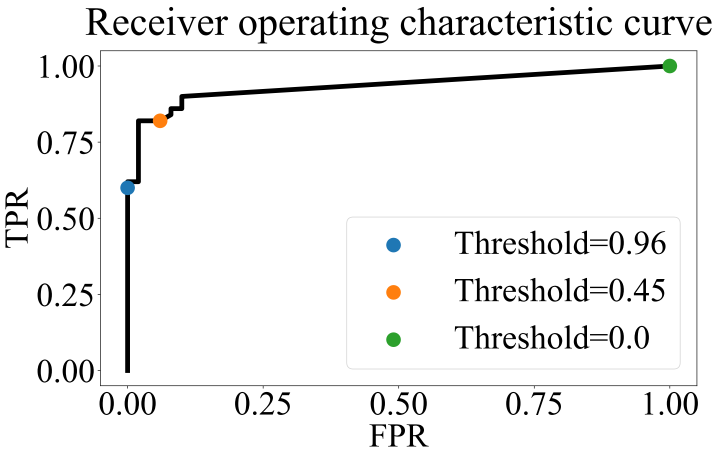

Figure 2.5 The ROC curve illustrate the trending of FPR and TPR as a function of threshold.

2.4.5 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

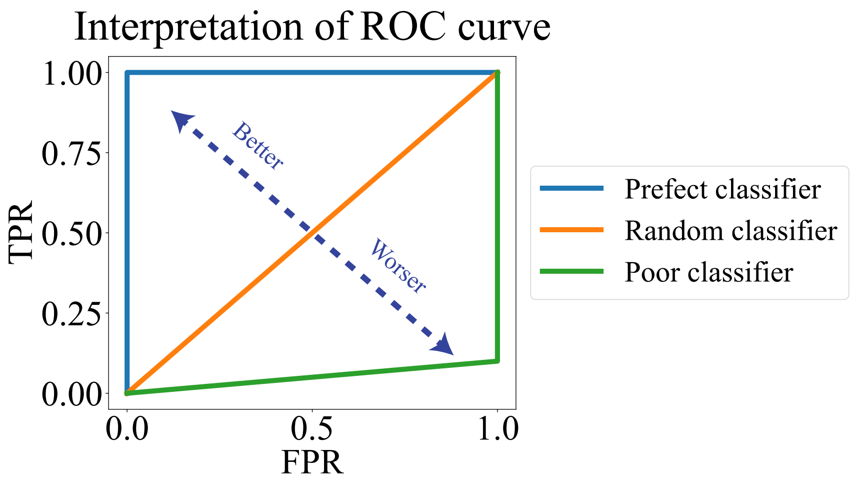

Figure 2.6 ROC curve interpretation by curve reading.

represents the integral measurement of the area under the ROC curve. Ultimately, a higher AUC value corresponds to higher classification performance, Figure 2.7.

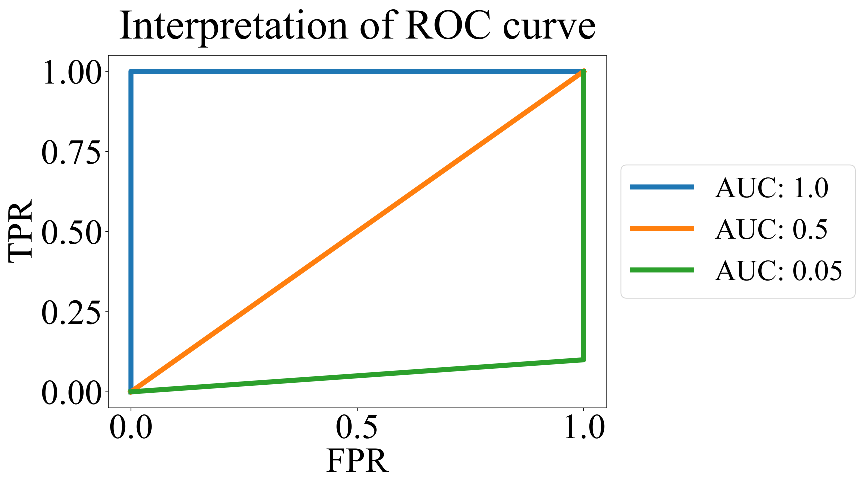

Figure 2.7 ROC curve interpretation by utilizing AUC.

2.4.6 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:

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} o_{i,j}}{\sum_{i,j} w_{i,j} e_{i,j}} \tag{2.6}$$

$$w_{i,j} = \frac{(i-j)^2}{(C-1)^2} \tag{2.7a}$$

$$e_{i,j} = \frac{\sum_k o_{i,k} \sum_l o_{l,j}}{\sum_{i,j} o_{i,j}} \tag{2.7b}$$

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. 2.7a) in the quadratic case, observed rating matrix, which commonly is a confusion matrix in this research, and the expect rating matrix, defined as (Eq. 2.7b). 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.

2.4.7 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. 2.8). 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.

$$d\left(X_1, X_2\right) = X_1 - X_2 \tag{2.8}$$

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

2.4.8 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:

$$IoU = \frac{\left|A \cap B\right|}{\left|A \cup B\right|} \tag{2.9}$$

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.

2.4.9 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 2.8. The PR curve, which plots recall against precision, is commonly utilized to visualize the prediction performance of object detection algorithms. 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. 2.10a). 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. 2.10b). 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. 2.10c).

$$\mathrm{AP} = \int_{r=0}^{r=1} p\left(r\right) dr \tag{2.10a}$$

$$\mathrm{AP} = \frac{1}{11} \sum_{r} p\left(r\right) \tag{2.10b}$$

$$\mathrm{mAP} = \frac{1}{k} \sum_{i} \mathrm{AP}_i \tag{2.10c}$$

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

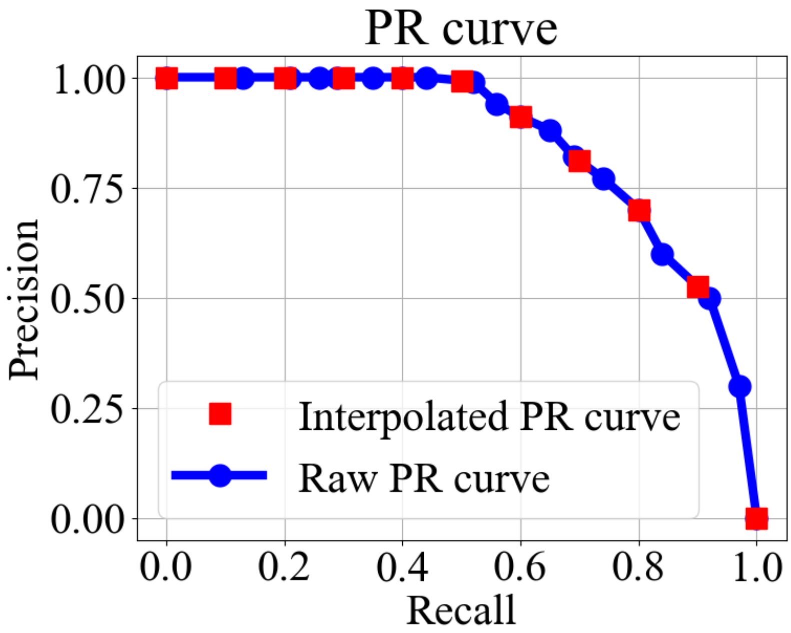

Figure 2.8 The figure illustrate the two PR curve including raw PR curve and interpolated PR curve, which is obtained by utilizing the 11-point interpolation approach.