

**MULTI-SCALE CONTRASTIVE
SELF-SUPERVISED LEARNING
FRAMEWORK FOR EARLY
DETECTION OF DIABETIC
RETINOPATHY SEVERITY
GRADING**

WADHA MUTAIB ALMATTAR

COMPUTER SCIENCE

NOVEMBER 2024

MULTI-SCALE CONTRASTIVE SELF-SUPERVISED LEARNING FRAMEWORK FOR EARLY DETECTION OF DIABETIC RETINOPATHY SEVERITY GRADING

BY

WADHA MUTAIB ALMATTAR

A Dissertation Presented to the
COLLEGE OF GRADUATE AND INTERDISCIPLINARY STUDIES
KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

In

COMPUTER SCIENCE

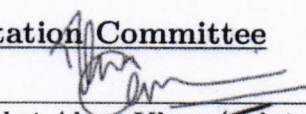
NOVEMBER 2024

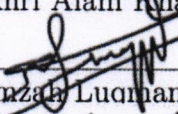
KING FAHD UNIVERSITY OF PETROLEUM & MINERALS
DHAHRAN 31261, SAUDI ARABIA

DEANSHIP OF GRADUATE STUDIES

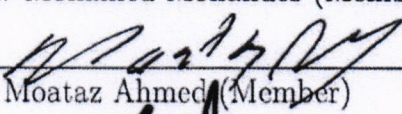
This dissertation, written by **WADHA MUTAIB ALMATTAR** under the direction of her dissertation adviser and approved by her dissertation committee, has been presented to and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **PHD THESIS IN COMPUTER SCIENCE**.

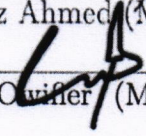
Dissertation Committee


Dr. Fakhri Alam Khan (Adviser)

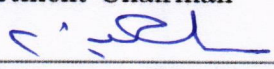

Dr. Hamzah Luchman (Co-Adviser)

m.mohandes
Prof. Mohamed Mohandes (Member)


Dr. Moataz Ahmed (Member)


Dr. Adi AlOwifer (Member)


Dr. Motaz Alfarraj
Department Chairman


Dr. Suliman Al-Homidan
Dean of Graduate Studies

Date



©Wadha Mutaib Almattar
2024

*To the soul of my father, Mutaib, who used to call me "Doctor,"
my mother, Haya,
my small mother, Haya,
my brothers and sisters
my husband, Ibrahim,
and my beating hearts, Wid, Humod, Faisal, and Saud.*

ACKNOWLEDGMENTS

In the name of Allah, the Most Gracious, the Most Merciful, I humbly extend my deepest gratitude to Allah for His endless blessings and unwavering support throughout this journey.

I would like to extend my deepest gratitude to my supervisor, Dr. Fakhri Alam Khan, and co-supervisor, Dr. Hamzah Luqman, for their continuous support, insightful guidance, and encouragement during my PhD journey. Their expertise and advice were invaluable in shaping my work. I am also incredibly thankful to Dr. Saeed Anwar and Dr. Saddam Al Azani for their constant support through fruitful research discussions and their invaluable contributions, which helped reshape my work into a much stronger version. My sincere thanks go to the committee members: Prof. Mohamed Mohandes, Dr. Moataz Ahmed, and Dr. Adi AlOwifier, for their time, insightful comments, and critical evaluation of my work.

I am deeply thankful to Imam Abdulrahman Bin Faisal University (IAU) for generously awarding me a scholarship to pursue my PhD. My sincere gratitude goes to the Center for Intelligent Secure Systems (IRC-ISS), whose funding in 2023 was instrumental in initiating my research journey. I would also like to extend my heartfelt thanks to the Joint Research Center for Artificial Intelligence (JRCAI) at King Fahd

University of Petroleum and Minerals (KFUPM) for their generous support in 2024. JRCAI not only provided financial backing but also offered a dedicated workspace, allowing me to maintain a consistent workflow and engage professionally with fellow researchers. I am truly fortunate to have collaborated with an exceptional group of researchers and friends at JRCAI, whose support and collaboration were invaluable throughout my journey.

To my beloved big family, thank you for always being my pillar of strength. My heartfelt gratitude to my mother, my stepmother, my brothers: *Mohammad, Saad, Naif, Bandar, AbdulAziz, AbdulAllah* and my sisters: *Azizah, Ebtisam, Nora, Njood, Mazna*, as well as my nephews and nieces for their unconditional love and support.

To my small family, my dear husband *Ibrahim*, who took on the responsibility entrusted to him by my beloved father to support me in completing my PhD. My father had entrusted him with this, and Ibrahim carried this responsibility with honor and dedication, providing me with unwavering support throughout this journey, especially during the times when I felt broken and overwhelmed. From the bottom of my heart, thank you for your constant support. To my wonderful children, *Wid, Humod, Faisal*, and ***my little boy Saud***, who began this journey with me when he was just 3 months old—we went through a tough journey together, but your support was endless. In many moments, you took care of me more than I took care of you. Thank you for your constant love, prayers, and understanding throughout this challenging time. I want you to know that the world is waiting for your achievements, and I believe in each one of you. I also extend my heartfelt thanks to our nanny *Rosana* for her continuous help in taking care of my children, which allowed me to focus on my PhD studies.

Finally, I would like to extend my heartfelt thanks to my dear friends: Sumayh Aljameel, Mashael Alkadi, Hussah Albinali, Razan Alshaikh, Khansa Al Mofeez, my PhD journey friends: Atheer Algherairy, Nora Alturayeif, Norah AlMubairik, whose encouragement and support have been a source of strength throughout this journey.

Contents

ACKNOWLEDGMENTS	vi
LIST OF TABLES	xv
LIST OF FIGURES	xix
LIST OF ABBREVIATIONS	xxiii
ABSTRACT (ENGLISH)	xxv
ABSTRACT (ARABIC)	xxvii
CHAPTER 1 INTRODUCTION	1
1.1 Problem Statement and Motivation	5
1.2 Research Objectives	8
1.3 Research Questions	9
1.4 Thesis Contributions	14
1.5 Publications and Funds	16
1.6 Thesis Structure	17
CHAPTER 2 DIABETIC RETINOPATHY DIAGNOSIS: MEDICAL BACKGROUND AND CHALLENGES	19
2.1 Introduction	19
2.2 Fundus images capturing cameras	20
2.3 Diabetic Retinopathy	22
2.3.1 DR Severity Levels	25
2.4 DR Diagnosis Challenges	27

CHAPTER 3 DIABETIC RETINOPATHY GRADING USING	
ARTIFICIAL INTELLIGENCE: LITERATURE REVIEW	29
3.1 Introduction	29
3.2 Search strategy	31
3.3 Related Surveys Comparison	33
3.4 Benchmark DR Grading Datasets	35
3.4.1 Pixel-Level Annotated Datasets	36
3.4.2 Image-Level Annotated Datasets	38
3.5 Supervised Learning Based Techniques	42
3.5.1 Classic DR Severity Grading Techniques	42
3.5.2 Deep Learning DR Severity Grading Techniques	46
3.6 Semi-supervised Learning	73
3.7 Discussion and Observations	75
3.7.1 DR Grading Datasets	75
3.7.2 DR Classification Tasks	79
3.7.3 Preprocessing Techniques	81
3.7.4 Learning Techniques	82
3.7.5 Performance Measures	85
3.8 Research Gaps and Future Directions	88
3.8.1 Data-centric Gaps	88
3.8.2 Model-centric Gaps	90
3.9 Chapter Summary	95
CHAPTER 4 DATASET CONSTRUCTION OF DUAL-VIEW	
FUNDUS IMAGES	98
4.1 Introduction	98
4.2 Single and Dual View Fundus Images	99
4.2.1 Diabetic Retinopathy Two-Field Image Dataset	101
4.2.2 Diabetic Retinopathy Grading and Image Quality Estimation	
Challenge	101

4.3	Self-supervised Learning Pre-training Dataset	102
4.3.1	Optic-disc and Macula-centered Dataset	102
4.3.2	ODM Variants	104
4.4	Evaluation Datasets	105
4.4.1	APTOS	106
4.4.2	Messidor-2	107
4.4.3	DDR	108
4.5	Remarks	109

CHAPTER 5 DIABETIC RETINOPATHY GRADING USING SUPERVISED TRANSFER LEARNING

111

5.1	Introduction	111
5.2	Convolutional Neural Networks and Vision Transformer	112
5.3	Supervised Transfer Learning in Medical Imaging Diagnosis	117
5.4	Methodology	118
5.4.1	Models	119
5.4.2	Datasets	121
5.4.3	Experimental Settings	122
5.4.4	Evaluation Metrics	122
5.5	Results and Discussion	124
5.5.1	Transfer Learning Performance of CNN Models	125
5.5.2	Transfer Learning Performance of Vision Transformer Models	132
5.5.3	CNN Models Class-Wise Performance	137
5.5.4	Vision Transformer Models Class-Wise Performance	144
5.5.5	Comparison Results with Previous Studies	145
5.6	Conclusion and Remarks	146

CHAPTER 6 DIABETIC RETINOPATHY GRADING USING CONTRASTIVE SELF-SUPERVISED TRANSFER LEARNING WITH LIMITED DATA

149

6.1	Introduction	149
------------	------------------------	-----

6.2	Background and Motivation	151
6.3	Related Work	154
6.4	Dual-view Fundus Images Datasets	158
6.5	Preliminaries and Notations	159
6.5.1	Contrastive Self-supervised Learning	159
6.5.2	Saliency Detection	161
6.6	Methodology	162
6.6.1	Channel-guided Self-supervised Learning Model (CGSSL)	
Framework Construction		163
6.6.2	Channel-Based Knowledge Utilization	164
6.6.3	Channel-Based Saliency Integration for Enhanced Fundus Anal-	
ysis		167
6.7	Experimental setup	168
6.7.1	Pre-training Dataset	168
6.7.2	Evaluation Datasets	169
6.7.3	Implementation details	169
6.7.4	Self-supervised Evaluation Protocols	170
6.8	Analysis and Discussion	171
6.8.1	Transferability and Generalisability	171
6.8.2	Adjacent DR Grades	173
6.8.3	Vision Transformer Supervised Learning	177
6.9	Ablation Studies	178
6.9.1	M_1 : Single View vs. Dual View Fundus Images	178
6.9.2	M_2 : The Sole Impact of Green-channel	180
6.9.3	M_3 : The Effect of Residual Connection	180
6.9.4	M_4 : The Impact of Channel-based Knowledge from Dual-view	
Fundus Images		181
6.9.5	M_5 : SSL baseline	182
6.10	Remarks	183

CHAPTER 7 MULTI-SCALE CONTRASTIVE SELF-SUPERVISED LEARNING FOR DIABETIC RETINOPATHY GRADING	185
7.1 Introduction	185
7.2 Background and Motivation	187
7.3 Self-supervised Learning (SSL) Transfer Learning in Medical Imaging .	189
7.4 Multi-scale Feature Fusion	191
7.5 MsSSL overall architecture	192
7.5.1 Multi-scale Features Extraction Encoder	192
7.5.2 Deep Learner	194
7.6 Contrastive loss objectives	197
7.7 Experiment Settings	199
7.7.1 Encoders.	199
7.7.2 Pretraining Datasets.	200
7.7.3 Downstream Datasets.	201
7.7.4 Preprocessing	202
7.7.5 Pretraining setup	203
7.7.6 Evaluation Protocols	203
7.8 Results and Analysis	203
7.8.1 Representation Transferability Assessment	204
7.8.2 Representation Quality Assessment	205
7.8.3 Comparative Analysis of Self-Supervised vs. Supervised Learning Approaches	207
7.8.4 Comparative Analysis of the Best ViT Backbone in Supervised and Self-Supervised Learning Settings	216
7.8.5 Qualitative Evaluation	221
7.8.6 Error Analysis of Challenging Cases	224
7.9 Discussion	229
7.10 Ablation Studies	231
7.10.1 FPN-backbone	233

7.10.2 ViT Variants as Backbone	234
7.10.3 Feature Fusing Approaches	236
7.10.4 Salient detection	237
7.10.5 The Impact of Deep Learner Module	238
7.10.6 Preprocessing Ablation Studies	241
7.10.7 Evaluation of Data Splitting Strategies	243
7.11 Conclusion and Remarks	247
CHAPTER 8 CONCLUSION	249
8.1 Contributions and Findings	249
8.2 Thesis Meets the Research Objectives and Questions	253
8.3 Limitations and Future Work	258
REFERENCES	263
VITAE	293

List of Tables

Table 2.1:	The DR severity levels depending on lesions classification according to ICDR [1]	24
Table 3.1:	Related Review Comparison Framework. (SSL: Semi-supervised, USL: unsupervised, G: analysis of gaps in public datasets, No: number of datasets analyzed, NA: not available)	35
Table 3.2:	Publicly available datasets for DR severity grading and detection	41
Table 3.3:	Classic DR techniques for DR classification	45
Table 3.4:	A summary of the end-to-end CNN models for DR severity grading and classification. G: DR grades, ...	51
Table 3.5:	A Summary of Transfer Learning Techniques	60
Table 3.6:	Summary of Hybrid techniques for DR severity grading. These techniques employ neural networks as feature extractors and ML algorithms for classification	68
Table 3.7:	A summary of neural network architecture used for DR grading.	71
Table 3.8:	Metrics for DR grading evaluation. True positives (TP), true negatives (TN), false positives (FP), false negatives (FN), P_0 is the observed agreement and P_e is the expected agreement	86

Table 4.1:	Optic Disc Macula dataset (ODM) dataset provides different variants. ODM-G and ODM-S datasets contain dual-view fundus images. OD and MA datasets contain single-view fundus images.	102
Table 5.1:	Summary of CNN architectures used in the study	119
Table 5.2:	Summary of Vision Transformer (ViT) architectures used in the study	120
Table 5.3:	Performance comparison of CNN and ViT architectures on the ODM dataset across multiple evaluation metrics.	125
Table 5.4:	Performance comparison of CNN and ViT architectures on the APTOS dataset across multiple evaluation metrics.	125
Table 5.5:	Performance comparison of CNN and ViT architectures on the Messidor dataset across multiple evaluation metrics.	126
Table 5.6:	Performance comparison of CNN and ViT Architectures on the DDR dataset across multiple evaluation metrics.	127
Table 5.7:	Original samples of severe NPDR (Grade 3) and corresponding Grad-CAM visualizations generated by various CNN models across three datasets. The visualizations highlight the areas of importance used by each model for prediction.	130
Table 5.8:	Comparison results of supervised learning studies on APTOS dataset.	145
Table 6.1:	Comparison results with state-of-the-art self-supervised models on three evaluation benchmark datasets: APTOS, Messidor, and DDR. Results are in Kappa score. FT stands for Fine-tuning.	171
Table 6.2:	Supervised Learning: performance of Vision Transformer variants (ViT) architectures trained on APTOS dataset with random weights.	176

Table 6.3:	Ablation Study: the efficiency of CGSSL components standalone and combined. M_1 to M_6 represent different models with/without a particular component. Results are reported in the Kappa score.	177
Table 6.4:	The Fine-tuning performance evaluation illustrates the impact of dual view over single view fundus images on the The Asia Pacific Tele-Ophthalmology Society (APTOS) dataset.	178
Table 6.5:	The fine-tuning evaluation performance of the SSL baseline on Convolutional Neural Network (CNN). M_4 model focuses on green-channel with the influence of dual-view fundus images. Results are reported in the Kappa score.	180
Table 7.1:	Comparison results of fine-tuning evaluation on APTOS dataset.	204
Table 7.2:	Comparison results of linear evaluation on APTOS dataset.	205
Table 7.3:	Performance Comparison of best CNN and best ViT in SL settings based on kappa score, to our proposed models CGSSL and MsSSL on the APTOS dataset across multiple evaluation metrics.	206
Table 7.4:	Performance comparison of best CNN and best ViT in SL settings, to our proposed models CGSSL and MsSSL on the Messidor dataset across multiple evaluation metrics. .	213
Table 7.5:	Performance Comparison of the ViT-S-16 Backbone in Supervised and Self-Supervised Learning Settings across Various Datasets.	216
Table 7.6:	Prediction results for samples of grade 4 fundus images using SL and MsSSL models.	224
Table 7.7:	Ophthalmologist Assessment.	225
Table 7.8:	MsSSL Best Settings.	231
Table 7.9:	MsSSL model with different FPN backbones.	233
Table 7.10:	MsSSL model with different ViT models backbones.	234

Table 7.11:	MsSSL model with different fusion methods.	235
Table 7.12:	MsSSL model with different saliency detection methods . .	237
Table 7.13:	MsSSL model with and without the Deep Learner (DLe) module.	238
Table 7.14:	Deep Learner (DLe) module with different layers design. . .	239
Table 7.15:	Ablation study on the effect of preprocessing techniques on the model’s performance.	241
Table 7.16:	Comparison of Model Performance Across Different Data Splitting Strategies on Messidor Dataset.	243
Table 7.17:	Comparison of Model Performance Across Different Data Splitting Strategies on APTOS Dataset.	244

List of Figures

Figure 1.1:	Prediction of the number of population with diabetic in different continents (2015-2040) [2]	2
Figure 2.1:	Fundus capturing cameras. (a) Desktop [3], (b) Portable [4]	21
Figure 2.2:	Normal fundus image of the right eye (left image) and a retina affected by Diabetic Retinopathy (DR) showing some lesions (right image) [5]	23
Figure 2.3:	DR severity levels. (a) No DR, (b) Mild, (c) Moderate, (d) Severe, (e) Proliferative DR [6].	24
Figure 2.4:	PDR levels. (a) NVD (b) NVE [7].	27
Figure 3.1:	Publicly Available DR Datasets Distribution.	75
Figure 3.2:	The distribution of DR Classification studies based on the task: binary classification, multi-classification, and lesions segmentation.	79
Figure 3.3:	The distribution of the supervised learning techniques proposed for DR grading.	82
Figure 4.1:	Examples of dual-views of fundus images from ODM dataset. (a) MA-centered DR fundus images for grades 2,3 and 4. (b) OD-centered DR fundus images for grades 2, 3 and 4.	102
Figure 4.2:	The distribution of ODM dataset for self-supervised learning, and the three evaluation datasets.	105

Figure 5.1:	ResNet50 architecture, using skip connections to enable deeper networks and improve performance in tasks like diabetic retinopathy grading. [8]	113
Figure 5.2:	The Vision Transformer architecture. ViT divides images into patches and applies self-attention to capture global context by enabling more comprehensive feature extraction. . .	115
Figure 5.3:	Confusion matrices for different CNN models on the APTOS dataset.	135
Figure 5.4:	Confusion matrices for different CNN models on the Messidor dataset.	137
Figure 5.5:	Confusion matrices for different CNN models on the DDR dataset.	139
Figure 5.6:	Confusion matrices for different Vision Transformers models on the APTOS dataset.	140
Figure 5.7:	Confusion matrices for different Vision Transformers models on the Messidor dataset.	141
Figure 5.8:	Confusion matrices for different Vision Transformers models on the DDR dataset.	142
Figure 6.1:	State-of-the-art Self-supervised network architectures. SimCLR, MoCo and BYOL all shared the same projection head design.	159

Figure 6.2:	The proposed Channel-guided self-supervised learning (CG-SSL) framework. The Green-channel extraction component extracts green-channel knowledge from RGB dual-view fundus image. Two augmented images are derived from the green dual-view image and fed to two encoders. The query encoder is the base encoder where the SSL learning is achieved. The key encoder is a momentum encoder that supports the query encoder by focusing on green-salient guidance. The projection head design consists of three layers with residual connection, batch normalization and ReLU.	161
Figure 6.3:	Channel-Based Extraction: Selected dual-view fundus images from ODM dataset. (a) Dual-view fundus images in RGB color space. (b) Corresponding green-channel fundus images.	164
Figure 6.4:	Signal-to-noise ratio (SNR) in different fundus image channels (Red, Green, and Blue) of 100 images randomly chosen from the ODM dataset. The green channel has the highest signal-to-noise ratio compared to other blue and red channels.	165
Figure 6.5:	The visualization of the salient maps of a fundus image. a) original fundus image in RGB color space. b) The salient map is detected from the RGB fundus image. c) The salient map is detected from the green-channel fundus image.	167
Figure 6.6:	Fine-tuning the three evaluation datasets for the downstream task using: a,b and c: ViT-S16 model with weights transferred from ImageNet, and d, e and f ViT-S16 with weights transferred from our proposed CG-SSL.	173
Figure 7.1:	Multi-scale Feature Extraction Encoder architecture. GFF: global feature extractor, FFE: fine-grained feature extractor.	192
Figure 7.2:	The architecture of Feature Pyramid Network (FPN)	194

Figure 7.3:	Deep Learner (DLe) architecture refines features through ConvTranspose2d layers with ReLU activations, ending with CBAM for enhanced feature attention and resolution. .	195
Figure 7.4:	The structure of Convolutional Block Attention Module (CBAM) [9].	196
Figure 7.5:	Confusion matrices for the best CNN, ViT models in supervised learning, compared to our proposed SSL models CGSSL and Multi-scale Self-supervise Learning Model (MsSSL) on the APTOS dataset.	207
Figure 7.6:	Confusion matrices for the best CNN, ViT models in supervised learning, compared to our proposed SSL models CGSSL and MsSSL on the Messidor dataset.	214
Figure 7.7:	Self-attention map visualization illustrating model decisions. a) ViT-S16 with weights pre-trained on ImageNet. b) ViT-S16 with weights from MsSSL. c) ViT-S16 with weights from CGSSL.	222
Figure 7.8:	The mean and the STD Distribution of different Splittings Strategies on Messidor Dataset.	243
Figure 7.9:	The mean and the STD Distribution of different Splittings Strategies on APTOS Dataset.	245

LIST OF ABBREVIATIONS

APTOS	The Asia Pacific Tele-Ophthalmology Society
CBAM	Convolutional Block Attention Module
CGSSL	Channel-guided Self-supervised Learning Model
CNN	Convolutional Neural Network
DDR	The Dataset of Diabetic Retinopathy
DLe	Deep Learner module
DR	Diabetic Retinopathy
FEE	Fine-Grained Feature Extractor
FPN	Feature Pyramid Network
GFE	Global Feature Extractor
Messidor-2	Methods to Evaluate Segmentation and Indexing Techniques in the Field of Retinal Ophthalmology
MFEE	Multi-scale Feature Extraction Encoder
MFF	Multi-scale Feature Fusion
MsSSL	Multi-scale Self-supervise Learning Model
NPDR	Non-Proliferative Diabetic Retinopathy
ODM	Optic Disc Macula dataset

PDR	Proliferative Diabetic Retinopath
SSL	Self-supervised Learning
ViT	Vision Transformer

DISSERTATION ABSTRACT

NAME: Wadha Mutaib ALmattar

TITLE OF STUDY: Multi-scale Contrastive Self-supervised Learning Framework for Early Detection of Diabetic Retinopathy Severity Grading

MAJOR FIELD: Computer Science

DATE OF DEGREE: November 2024

Diabetic Retinopathy (DR) is a leading cause of blindness worldwide, and its accurate diagnosis and severity grading are critical for timely intervention. However, grading DR severity is particularly challenging due to the subtle differences between adjacent stages and the difficulty in identifying late-stage DR cases, especially in imbalanced datasets. This thesis proposes a novel self-supervised learning (SSL) framework tailored specifically for DR grading. By leveraging rich fundus image data, the framework is designed to extract and transfer robust representations without relying on labeled data during the pretraining phase. Our approach integrates a multi-scale feature extraction mechanism, combining the strengths of Vision Transformers (ViTs) and convolutional neural networks (CNNs) through a Feature Pyramid Network (FPN), with attention mechanisms to enhance feature learning. We evaluate our SSL models—CGSSL and

MsSSL—on benchmark DR datasets, including APTOS and Messidor, and compare their performance against state-of-the-art supervised models. The results demonstrate that our SSL-based models significantly improve classification accuracy, particularly in distinguishing between adjacent DR grades and identifying severe DR cases. Additionally, the study explores the impact of dataset size and class imbalance on model performance, highlighting the limitations of conventional transfer learning from natural image datasets like ImageNet. The findings suggest that SSL provides a more effective solution for DR severity grading, offering superior generalization capabilities for real-world clinical applications. This research contributes to the growing body of work in medical image analysis by presenting an advanced, scalable approach for addressing the challenges of DR diagnosis.

ملخص الرسالة

الاسم:

وضحه بنت متعب بن محمد المطر

عنوان الدراسة:

إطار عمل التعلم الذاتي التوجيهي التبايني متعدد المقاييس للكشف المبكر عن تصنيف شدة اعتلال الشبكية السكري

التخصص:

علوم الحاسب

تاريخ الدرجة العلمية: نوفمبر ٢٠٢٤

يُعتبر اعتلال الشبكية السكري (Diabetic Retinopathy) أحد الأسباب الرئيسية للعمى على مستوى العالم، وتشخيصه بدقة وتحديد درجة شدته أمران حاسمان للتدخل في الوقت المناسب. ومع ذلك، فإن تصنيف شدة اعتلال الشبكية السكري يمثل تحديًا كبيرًا بسبب الفروقات الدقيقة بين المراحل المتجاورة والصعوبة في تحديد الحالات المتأخرة من اعتلال الشبكية، خاصةً في مجموعات البيانات غير المتوازنة. يقترح هذا البحث نموذجًا جديدًا يعتمد على التعلم الذاتي التوجيهي (SSL) مخصص لتصنيف درجات شدة اعتلال الشبكية السكري. من خلال الاستفادة من بيانات صور قاع العين الغنية بالمعلومات التشخيصية، يهدف النموذج إلى استخراج ونقل تمثيلات قوية دون الاعتماد على البيانات المصنفة خلال مرحلة ما قبل التدريب. يجمع نهجنا بين آلية استخراج الميزات متعددة المقاييس، التي تدمج بين نقاط القوة في المحولات البصرية (ViTs) والشبكات العصبية الالتفافية (CNNs) عبر شبكة الهرم المميز (FPN)، مع آليات الانتباه لتعزيز تعلم الميزات. تم تقييم نماذج التعلم الذاتي التوجيهي الخاصة بنا — CGSSL و MsSSL — على مجموعات بيانات معيارية لتصنيف اعتلال الشبكية السكري مثل APTOS و Messidor، ومقارنتها مع النماذج القائمة على التعلم الخاضع للإشراف. أظهرت النتائج أن نماذجنا المبنية على التعلم الذاتي التوجيهي تحسن بشكل ملحوظ دقة التصنيف، خاصةً في التفريق بين درجات اعتلال الشبكية المتجاورة وتحديد الحالات الشديدة. علاوةً على ذلك، استكشفت الدراسة تأثير حجم البيانات وتوازن الفئات على أداء النموذج، مسلطة الضوء على محدوديات التعلم المنقول من مجموعات بيانات الصور الطبيعية مثل ImageNet. تشير النتائج إلى أن التعلم الذاتي التوجيهي يقدم حلاً أكثر فعالية لتصنيف شدة اعتلال الشبكية السكري، ويوفر قدرات تعميم فائقة للتطبيقات السريرية الواقعية. يساهم هذا البحث في تطوير مجال تحليل الصور الطبية من خلال تقديم نهج متقدم وقابل للتوسع لمواجهة تحديات تشخيص اعتلال الشبكية السكري.

CHAPTER 1

INTRODUCTION

Diabetes is a chronic disease that occurs when the human body cannot effectively use and produce enough insulin in the human body [10]. Diabetes can harm different body organs such as the heart, eyes, blood vessels, kidneys, and nerves. It is estimated globally that the number of people with diabetes will be 366 million by 2030 which forms 4.4% of the world population [11]. Another study by the International Diabetes Federation (IDF)[2] estimated the number of people with diabetes to rise from 35.4 million in 2015 to 642 million in 2040. Figure 1.1 illustrates overall increasing of diabetic patients from 2015 to 2040 per region for people with age between 20 and 79 [2]. As shown in Figure 1.1, the number of diabetic patients in the Middle East was 35.4 million in 2015, and it is predicted to dramatically increase to 72.1 million by 2040.

diabetic retinopathy (DR) is one of the major consequences of having diabetic disease for a long period of time. DR harms and damages the blood vessels in the retina. According to [11], the diabetic patient has a probability of 30% of developing DR. DR is one of the major causes of blindness all over the world [11]. This silent

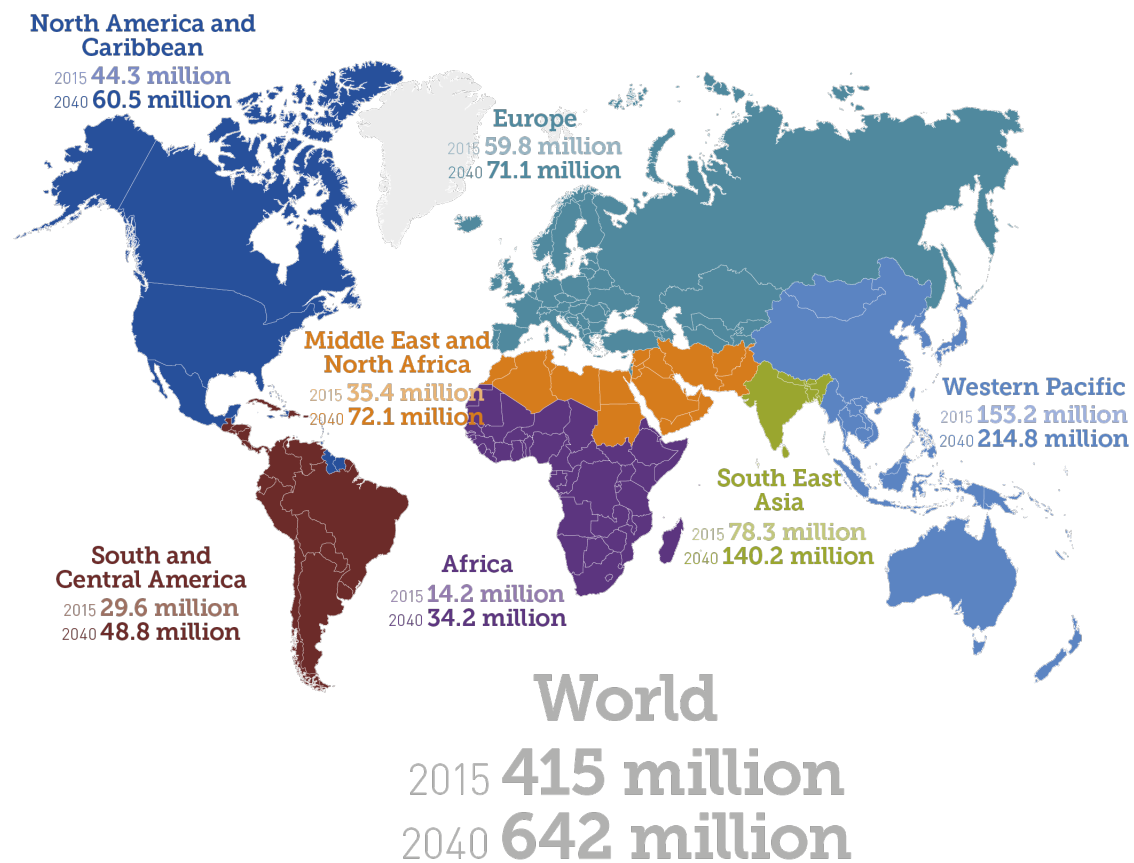


Figure 1.1: *Prediction of the number of population with diabetic in different continents (2015-2040) [2]*

illness that may not be discovered until retinal abnormalities have progressed to the point that treatment is ineffective or impossible.

DR has several levels related to its severity. The levels can be diagnosed using a combination of structural features (lesions) in fundus images, such as the presence of microaneurysms, exudates, hemorrhages, and neovascularization (more details about these lesions can be found in Section 2.1. These lesions appear as small objects in the retina fundus images and they are important features for DR grading in medical imaging. However, detecting these lesions is challenges and can be mistaken with other objects and lesions [6] [12].

DR severity is categorized into two main categories: Non-Proliferative Diabetic Retinopathy (NPDR), and Proliferative Diabetic Retinopath (PDR). NPDR represents the early stages of DR disease and it includes four severity grades: 0 indicates no DR, while grades from 1 to 3 represent DR cases with different severity levels. The PDR represents the later stages of DR and it includes two levels: neovascularization on optic disc (NVA), and neovascularization elsewhere (NVE). Late discovery of NPDR and PDR stages can result in floater, blurred vision, and eventually blindness.

An accurate diagnosis of the severity degree of DR helps ophthalmologists to apply appropriate treatment procedures to prevent visual degradation. Deploying a proper treatment procedure by ophthalmologists, is coupled with the effectiveness of diagnosis the DR to its right stage. The existing DR treatments are confined to halting or delaying vision loss. However, manual diagnosis of the DR lesions through fundus images is challenging. It is time-consuming and it demands specialized personnel and equipment, which are often lacking in remote area or low-income countries [13]. In

addition, low quality fundus images can result in lesion misclassification. Lesions size also contribute in DR grades misclassification.

The study of automatic DR detection from retinal-based images is currently generating a lot of attention in the ophthalmology and modern computer vision (CV) fields. Several attempts have been presented in the literature for automate DR detection. The majority of the early efforts used classic computer vision techniques to extract low and high scaled features from DR fundus images and used Machine Learning (ML) algorithms for DR classification. The main issue of these techniques is that are sensitive toward the quality of the extracted features [14]. Recent studies on DR detection paid more attention to deep learning (DL) techniques due to their exceptional performance for several problems.

The main difference that makes DL take over the traditional methods is that it extracts image features in end-to-end manner by learning the underlying representations [14]. DL showed a great important in DR detection. However, DL requires a large annotated data which is not always possible in medical domain. In addition, DL sometimes produces low accuracies with low-quality fundus images which prevent DL from properly capturing low-scaled features.

Thus, the research topic of automatic DR identification from retinal-based images is of tremendous interest in the fields of ophthalmology and contemporary computer vision. The current DR detection techniques has several limitations due to the DR lesions size at each DR stage. Adjacent DR grades share some lesions in terms of form, color, or physical appearance. This makes current techniques fail on distinguishing low scaled features that represent lesions at each stage. Moreover, Lesions may be

overlapped with retina structure that lead to misclassification. More challenges and limitations are discussed in Section 2.1. Most of the recent studies in the literature address the DR in its early stages (NPDR) only. Few studies tried to investigate the later stages (PDR).

This research primarily focuses on exploring the field of DR with the goal of automatically predicting the disease's stage. This chapter offers an overview of the problem from both medical and automation perspectives, as well as the motivations behind the study. It will detail the research objectives, specifying the study's goals, and outline the thesis structure to guide the reader through the following sections. Additionally, the chapter will emphasize the key contributions of this research.

1.1 Problem Statement and Motivation

DR is one of many consequences of diabetic disease. The progression of DR is gradually increased over time leading to sever levels. DR is considered as one of the major causes of blindness. Thus, early detection of DR is required to delay the disease's progression at NPDR and PDR stages. However, diagnosing the severity of DR is a time-consuming procedure in medical practice due to scarcity of well-trained ophthalmologists and the nature of DR lesions as will be discussed in Section 2.1, Section 2.4.

A fully automated screening system is required to increase the care quality and support ophthalmologists decision. It is also needed in remote sites that lack of ophthalmologists. Each DR stage has its lesions representing the DR severity level. Adjacent DR grades share some lesions characteristics such as color and appearance. As

reported in the literature, misclassifications between adjacent DR grades occur due to the similarity between lesions in terms of appearance and size. In addition, the overlapping between lesions and retina structure is another challenge for DR detection systems. To automate the DR severity grading procedure, the detection of the small DR lesions requires an efficient features extraction technique and image enhancement which are capable to leverage fine-grained features.

Several classical machine learning techniques have been proposed for detecting the presence of different types of lesions (red and white) and to distinguishing between DR severity grades. On the other hand, deep learning (DL) models showed a great improvement in DR detection compared with classic machine learning (ML) technique. The majority of the proposed studies in the literature used CNN models pre-trained on ImageNet. However, the properties of fundus images and DR lesions are different than the classes of ImageNet that explain the low scores of these models. In addition, the deep learning models require large dataset which is not available with medical problems including DR.

The challenge with DR datasets is two-folds: collection and annotation. In terms of geographic differences, deep learning algorithms derived from Western populations failed to detect a number of serious lesions in Saudi population [15]. Due to the fact that dark skin has more melanocytes in the retina's outermost layer compared to lighter skin. Unlike lighter skin, dark skin hides vascular changes. Regarding the DR detection tasks, most of the surveyed studies tackle DR in its early stages NPDR, while few studies have tackled late stages PDR where new vessels are growing round the optic disc or else-where and harming the vision. From the literature, PDR

grade is misclassified with grade 3 due to the similarity between adjacent grades as will reported in Chapter 3 at Section 2.4. In addition, in PDR late stages new vessels are growing (neovascularization) and these vessels are not consistence and the pattern is changeable. As results, the DL network learn the general pattern of neovascularization then a new pattern is appeared that lead to misclassification. Moreover, the complexity of the small vessels characteristics that appears differently at each pixel patch is considered a challenge in this task that leads to misclassification. [16].

Our research is focused on addressing the critical challenge of accurately grading diabetic retinopathy (DR) severity, particularly the common misclassification of adjacent DR grades and the difficulty in identifying late-stage DR cases. To tackle this, we aim to develop a self-supervised learning (SSL) framework tailored specifically for DR severity classification, leveraging the rich diagnostic information embedded in fundus images. By harnessing SSL, we seek to transfer robust, learned representations that can improve the model’s ability to distinguish between subtle variations in adjacent DR grades and better handle the underrepresented, severe cases. Our goal is to create a model that not only achieves state-of-the-art performance in DR grading but also addresses these specific challenges, making it comparable to or better than existing supervised and SSL-based methods in the field.

1.2 Research Objectives

The main aim of this thesis is to conduct both foundational and practical research in the area of computer vision for diabetic retinopathy grading. The key objectives of this thesis include:

1. Conduct a comprehensive systematic literature review (SLR) to assess the effectiveness, limitations, and emerging trends in DR severity grading methodologies.
2. Evaluate existing DR datasets, examining their availability, data quality, and how well they meet the specific needs of our research problem.
3. Investigate the scalability and effectiveness of Vision Transformer (ViT)s in DR grading, compared to CNNs, and to analyze the impact of transfer learning from ImageNet in medical imaging contexts.
4. Explore the potential of SSL in improving the quality of learned representations for DR grading, particularly when using out-domain and in-domain image data.
5. Develop strategies for enhancing SSL frameworks with a focus on improving domain adaptation and feature extraction for diabetic retinopathy (DR) grading. The proposed framework will aim to effectively distinguish between adjacent DR severity grades, particularly the more challenging late-stage grades, ensuring more accurate and reliable classification across all stages of the disease.
6. Share research results by publishing papers, enhancing the knowledge base of the research community and ensuring accessibility to the findings.

1.3 Research Questions

This section outlines the research questions designed to address five key gaps identified in our literature review (Chapter 3).

The **Gap-3** involves the over-reliance on convolutional neural networks (CNNs) for DR classification. This thesis explores the potential of Vision Transformers (ViTs) alongside CNNs to create a hybrid model that leverages the strengths of both architectures for more accurate grading.

The **Gap-4** addresses the use of transfer learning from ImageNet, which may not be ideal for medical images like fundus images. We will investigate the effectiveness of self-supervised learning (SSL) and in-domain transfer learning using medical datasets.

The **Gap-5** focuses on improving feature extraction, particularly for detecting tiny DR lesions often missed by CNNs. This research will develop a hybrid CNN-ViT model to better capture these small features, integrating the strengths of both architectures.

The **Gap-6** addresses the challenge of limited annotated data in medical imaging. This research will leverage self-supervised learning (SSL) to train models using large amounts of unlabeled data, reducing reliance on annotated datasets.

Lastly, the **Gap-8** highlights one of the most significant challenges in DR grading, the overlap of common lesions between adjacent classes. This overlap often leads to misclassification in deep learning (DL) models, resulting in a high rate of false positives.

By answering the following research questions and developing a novel approach for

an automation DR grading in this thesis, we aim to make a significant contribution to the field of computer vision in the context of medical diagnosis, especially DR, by fulfill the listed gaps earlier. Following, we introduce our research questions, each aligned with the corresponding research objectives mentioned previously.

1. RQ1: What are the current advancements and challenges in Diabetic Retinopathy (DR) severity grading techniques?

Answering this research question would involve an investigation of the literature to identify the characteristics of currently available DR current techniques in used to automate the diagnosis and publicly available datasets utilized in this area. This exploration aims to shed light on their strengths and limitations, helping to pinpoint gaps and challenges within existing datasets and automated diagnosis approached using machine and deep learning methods, as reported in **Chapter 3**. This aligns with **Gap-3 and Gap4**, which underscores the need to understand the current state of the problem in hand.

2. RQ2: Which publicly available datasets for Diabetic Retinopathy are most suitable for our research, and how do they align with our requirements in terms of availability, relevance, and applicability?

Answering this research question involves conducting a comprehensive literature review to identify the publicly available datasets commonly used for diabetic retinopathy (DR) grading. Through this process, we compiled a list of suitable datasets, evaluating them in terms of their availability, relevance to DR grading, and applicability to our research needs (**Chapter 3**).

Next, we constructed a dataset specifically for self-supervised learning (SSL), drawing from these publicly available resources. A key focus was on datasets that provide dual-view fundus images, which are valuable for SSL due to their potential for learning robust image representations. This aligns with **Gap-6**, where deep learning (DL) models traditionally rely heavily on large datasets for effective training. However, our SSL model is designed to learn representative knowledge from relatively smaller datasets compared to what SSL models typically require, addressing the challenge of data scarcity.

Additionally, this approach addresses **Gap-4**, where SSL is highlighted as a method capable of effective learning in scenarios with limited data. The constructed dataset for SSL is introduced in **Chapter 4**, along with the selected datasets for fine-tuning and evaluation, ensuring they meet our criteria for quality and compatibility with our research objectives.

3. RQ3: How does Vision Transformer perform in DR severity grading in supervised transfer learning, compared to CNN models? Additionally, how does transfer learning from ImageNet influence performance in DR severity grading?

In this research question, we explored the ViT models, a recently introduced architecture that has demonstrated strong performance in computer vision tasks. We compared the performance of CNN models and ViT in a supervised setting. By examining the effectiveness of these two prominent architectures, our goal is to address the gap in diabetic retinopathy (DR) grading by proposing an

innovative technique that leverages both models as encoders within an SSL framework to enhance model transferability and quality. This approach aligns with **Gap-3 and Gap-4**, which highlight the strengths of these architectures in computer vision tasks, particularly in DR grading. In **Chapter 5**, we offer insights into the use of pre-trained models on ImageNet.

4. RQ4: How effectively does self-supervised learning (SSL) enhance representation learning when applied to both out-domain and in-domain images for DR grading?

Given the limited research on using SSL for in-domain transfer learning, as noted in our literature review (**Gap-3, Gap-4, and Gap-6**), we propose this research question. This question specifically targets the critical need for effective pretraining in medical diagnosis to significantly enhance model performance. To explore this, we experimentally develop our SSL baseline, as introduced in **Chapter 6**. Through this experiment, we provide valuable insights into DR grading using SSL through the following contributions:

- Introduced the CG-SSL model, a novel framework pre-trained with a minimal amount of data yet enriched with comprehensive knowledge representations. This model is specifically guided by saliency detection in the green channel and leverages dual-view fundus images to optimize the learning process. The innovative projection head design, featuring residual connections, plays a crucial role in effectively utilizing detailed information and enhancing model performance.

- Highlighted the critical role of dual-view in medical field and how it is beneficial in knowledge representation in self-supervised learning, demonstrating its impact on the efficacy of feature extraction and overall model performance.

5. RQ5: How can the self-supervised learning (SSL) framework be enhanced for domain-adaptive learning to extract high-quality, fine-grained features to distinguish adjacent DR grades?

In this study, we proposed Multi-scaled Self-supervised Learning (MsSSL) framework that incorporating multi-scale feature extraction from various deep learning architectures will improve the diabetic retinopathy (DR) grading model by capturing both high-level and fine-grained features. This hypothesis is closely related to **Gap-3, Gap-4, Gap-5, Gap-6, and Gap-8**, highlighting the importance of investigating diverse feature extraction techniques. By exploring this research question in **Chapter 7**, our goal is to address these gaps and demonstrate how combining features at multiple scales from different encoders can enhance the accuracy and comprehensiveness of DR grading systems. Furthermore, we propose that our model is particularly effective in distinguishing late stages of DR, outperforming transfer learning approaches from ImageNet, which may lack domain-specific granularity for such fine distinctions.

These research questions are designed to address the gaps identified in **Chapter 3 (Section 3.8)** and focus on key topics such as dataset construction, the integration of advanced architectures like CNNs and Vision Transformers (ViTs) as feature extrac-

tors. Specifically, the questions aim to explore the potential of SSL for medical image diagnosis, and compare the performance of different deep learning models. These research questions will serve as the foundation for our research in the subsequent chapters, where we will further investigate these areas to advance the field of diabetic retinopathy diagnosis and improve deep learning methodologies in medical imaging.

1.4 Thesis Contributions

This thesis presents several key contributions aimed at advancing the field of diabetic retinopathy (DR) diagnosis through the development of innovative deep learning techniques. By addressing critical gaps in data scarcity, model architecture, the proposed methods enhance the efficiency and effectiveness of automated DR grading systems. The contributions of this thesis are as follows:

1. **Comprehensive literature review on diabetic retinopathy (DR) grading:** We conducted an extensive review of the existing literature, addressing the lack of comprehensive reviews in this area, particularly focusing on datasets, deep learning models techniques used for DR diagnosis.
2. **Investigating supervised transfer learning for diabetic retinopathy (DR) grading:** We performed a series of experiments using both CNN and Vision Transformer (ViT) models to assess the effectiveness of transfer learning with pre-trained models from ImageNet. Our findings highlight the limitations of transferring knowledge from ImageNet, a general-purpose dataset, to

the domain of medical imaging, specifically fundus images in the context of DR grading, revealing important gaps in model performance and feature adaptation.

3. **Development of a self-supervised learning (SSL) framework for DR grading:** We designed an SSL approach utilizing dual-view fundus images to enhance feature representation, specifically addressing the challenge of data scarcity in medical imaging.
4. **Development of a hybrid CNN-ViT model with feature pyramid networks (FPN) to enhance SSL learning:** We integrated convolutional neural networks (CNNs) and Vision Transformers (ViTs) within the proposed SSL framework, using feature pyramid networks (FPN) to fuse multi-scale features from both architectures. This approach enabled more representative feature learning, improving the detection of small lesions and enhancing the transferability of features for accurate DR diagnosis. Specifically, the framework is designed to effectively distinguish between adjacent DR severity grades, particularly the more challenging late-stage grades, ensuring more accurate and reliable classification across all stages of the disease. This capacity to differentiate closely related grades is critical for early and precise medical intervention.
5. **Addressing gaps in dataset size:** Through the use of self-supervised learning (SSL) and dual-view fundus images, which are rich in information, we mitigated the reliance on large datasets, an issue commonly faced in deep learning models. This approach enables effective learning from smaller datasets while maintaining high performance in diabetic retinopathy grading tasks.

1.5 Publications and Funds

- Almattar, W., Luqman, H., & Khan, F. A. (2023). “Diabetic retinopathy grading review: Current techniques and future directions. Image and Vision Computing”, 104821.
- W. Almattar, S. Anwar, & F. A. Khan, “Limitations of Conventional Transfer Learning: A Comparative Study of CNN and ViT Models for Diabetic Retinopathy Grading”.
- W. Almattar, S. Anwar, & F. A. Khan, “Efficient knowledge transfer using channel-guided contrastive self-supervised learning for diabetic retinopathy severity grading with limited data”.
- W. Almattar, S. Anwar, S. , Al Azani, & F. A. Khan, “Multi Scale Self Supervised Learning for Deep Knowledge Transfer in Diabetic Retinopathy Grading”.

This thesis was supported by two funding sources, which were essential for conducting experiments, acquiring resources, and advancing the research. The financial backing enabled the development and evaluation of the proposed models and helped in sharing the findings with the academic community. This support significantly expanded the scope and impact of the work in medical image analysis. The two funding sources are:

- Interdisciplinary Research Center of Intelligent Secure Systems (IRC-ISS), Grant no. INSS2205, 2023.

- SDAIA-KFUPM Joint Research Center for Artificial Intelligence, Grant no. JRCAI-RG-14, 2024.

1.6 Thesis Structure

This thesis is organized to offer a clear and thorough examination of DR grading, starting from the initial problem statement and leading to the presentation of new contributions. The structure of the thesis is as follows:

Chapter 1, this introductory chapter lays the foundation for the thesis, introducing the research problem, outlining the motivations behind the study, and establishing the research objectives. It also provides a brief summary of the contributions made throughout this work, setting the stage for the detailed exploration that follows.

Chapter 2, this chapter provides a thorough examination of diabetic retinopathy (DR), discussing the disease’s progression, associated complications, and its significant impact on patient health if not properly managed. It also reviews the challenges present in current clinical practices, emphasizing the need for automated diagnostic solutions to improve accuracy, efficiency, and accessibility in detecting and managing DR.

Chapter 3, a comprehensive literature review is conducted in this chapter, examining existing research in DR grading. It identifies current trends, gaps, and the limitations in the field, providing a critical assessment of how these factors inform the current research.

Chapter 4, this chapter details the construction of the ODM dataset, a criti-

cal component of the research, particularly in the context of self-supervised learning (SSL). It discusses the dataset’s significance and the role it plays in advancing the study.

Chapter 5, here, the thesis explores the application of supervised transfer learning for DR grading, providing detailed insights into its methodologies and performance. It also critically evaluates the limitations of this approach, discussing why it may be insufficient for fully addressing the complexities of DR diagnosis.

Chapter 6, this chapter shifts focus to in-domain transfer learning through self-supervised learning. It presents the research findings, accompanied by a comprehensive analysis of the outcomes, and discusses how this approach addresses some of the limitations identified in the previous chapter.

Chapter 7, Building on previous discussions, this chapter introduces an enhanced SSL framework specifically designed for DR grading. It presents in detail the methodology of extracting high and fine-grained features, and discusses the findings, highlighting how this approach advances the field.

Chapter 8, The final chapter synthesizes the contributions and key findings of the thesis. It reflects on the limitations of the study and provides suggestions for future research directions, offering a roadmap for continuing work in this area.

CHAPTER 2

DIABETIC RETINOPATHY

DIAGNOSIS: MEDICAL

BACKGROUND AND

CHALLENGES

2.1 Introduction

This chapter offers a comprehensive introduction to DR and its varying levels of severity. It begins by detailing the different types of fundus images and the devices commonly used to capture them for DR diagnosis. Following this, the chapter delves into an in-depth discussion of diabetic retinopathy, including its progression, potential complications, and the significant impact it has on patient health if left untreated. Additionally, the challenges faced in clinical practice are examined, highlighting the

complexities and limitations of current diagnostic methods. This section underscores the pressing need for automation in DR diagnosis, aiming to enhance accuracy, efficiency, and accessibility in medical settings. Through this exploration, the chapter sets the stage for understanding why developing advanced automated diagnostic tools is essential for improving patient outcomes in the fight against diabetic retinopathy.

This chapter provides a general overview of disease and its severity levels. First, we present the various fundus image types and the devices used to acquire them for DR diagnosis. Then, we provide an overview of DR disease and its complications and the challenges in the practical clinical practice and the need of the automation of the DR diagnosis.

2.2 Fundus images capturing cameras

Fundus imaging involves capturing digital color images of the eye fundus to diagnose the condition of the retinal tissues. Abramoff and Kay [17] defined fundus imaging as *The process whereby reflected light is used to obtain a two-dimensional (2D) representation of the 3D, semitransparent, retinal tissues projected onto the imaging plane. Thus, any process that results in a 2D image where the image intensities represent the amount of a reflected quantity of light is fundus imaging.*

Fundus images and optical coherence tomography (OCT) are the two types of retinal imaging color. The main difference between the two types is that OCT images are two- and three-dimensional images that provide significant information regarding retinal structure and thickness using low-coherence light, whereas fundus images are

two-dimensional images that use reflected light to capture retinal structures in two-dimensions [18].

Several types of fundus cameras are used in the medical sector. Some types of fundus cameras require pupil dilation using mydriatic, a special eye drop, and a bright light flash. Non-mydriatic fundus cameras are advanced fundus cameras that do not require pupil dilation and use a low-intensity flashlight. These cameras are available in two forms: desktop and portable cameras (Figure 2.1). Desktop fundus cameras are the most widely used owing to their outstanding image quality, but they are expensive and cumbersome to transport. Portable fundus cameras are cheaper and more convenient for retinal imaging in remote areas or home use.

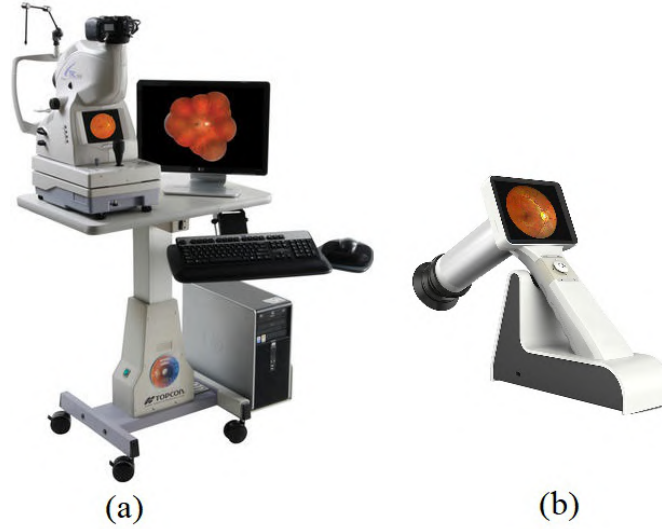


Figure 2.1: *Fundus capturing cameras. (a) Desktop [3], (b) Portable [4]*

Determining the severity of several eye diseases, such as DR, hypertension, and cardiovascular, depends on fundus image quality. The quality of the fundus image depends on different factors, such as camera type (for example, non-mydriatic and

mydriatic), retina dilation, examination settings, illumination, artifacts, and noise [19]. Therefore, capturing high-quality fundus images is crucial for the accurate diagnosis of the severity level of DR.

2.3 Diabetic Retinopathy

DR is an eye disease caused by diabetes, characterized by the presence of lesions in the retina. These lesions represent pathological features of the human eye's retina. The increasing occurrence of DR can be linked to the notable rise in the diabetic population over an extended duration. In fact, DR is considered the leading cause of legal blindness among the working-age population in developed countries [2].

DR lesions frequently arise due to elevated blood sugar levels in the retina, causing damage to the blood vessels [2], which results in bleeding or swelling that obstructs blood flow. As a result, the retina develops abnormal blood vessels that may impair vision.

The pathological consequences of DR can cause the formation of a diverse array of retinal lesions. These lesions include microaneurysms, hemorrhages, hard exudates, and cotton wool spots (also known as soft exudates). In particular, retinal lesions can be classified into two primary categories: red and white lesions. The former is characterized by the presence of bloody lesions and includes microaneurysms and hemorrhages. Various factors contribute to the development of white lesions in the retina. The presence of hard exudates, discernible as yellow deposits, resulting from the accumulation of lipoproteins within the retinal tissue is among the factors. Additionally,

white lesions can develop from nerve fiber-related yellowing or lymph leakage, both characterized by soft exudates, commonly known as cotton wool spots. Figure 2.2 shows a healthy retina and one with different DR lesions.

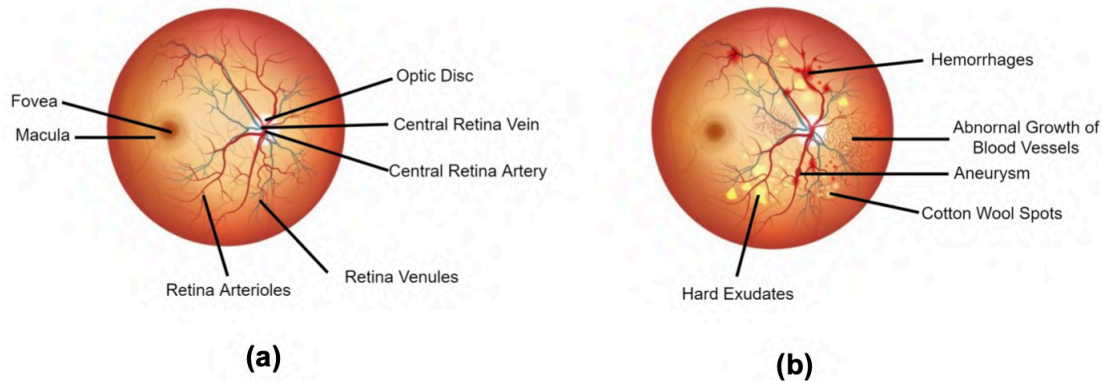


Figure 2.2: *Normal fundus image of the right eye (left image) and a retina affected by DR showing some lesions (right image) [5]*

The diagnosis of DR and the stages of the disease are closely related. In medical practice, it is typically performed by identifying abnormalities in the eye fundus image (retina). To standardize retinopathy screening, the American Diabetes Association (ADA) recommends that diabetic patients must begin DR screening the year of being diagnosed with type 2 diabetes, whereas type 1 diabetic patients should have the first DR screening after three years [20]. Subsequent examination should be done every 6-12 months, depending on the severity of the DR. Ophthalmologists follow a standardized protocol for fundus examination, as shown in Table 2.1. This protocol, which was proposed by the International Clinical DR Scale (ICDR) [1], associates each severity level with observable findings for DR diagnosis severity. Figure 2.3 presents examples of these levels.

Table 2.1: *The DR severity levels depending on lesions classification according to ICDR [1]*

Level	Severity	Lesions
0	No DR	No lesions
1	Mild NPDR	Microaneurysms
2	Moderate NPDR	More than just Microaneurysms but less than severe NPDR
3	Severe NPDR	Any of the following: - More than 20 intraretinal hemorrhages in each of 4 quadrants - Definite venous beading in 2+ quadrants - Prominent intraretinal microvascular abnormalities in 1+ quadrant - No signs of proliferative retinopathy
4	PDR	One or more of the following: - Neovascularization - Vitreous/ pre-retinal haemorrhage

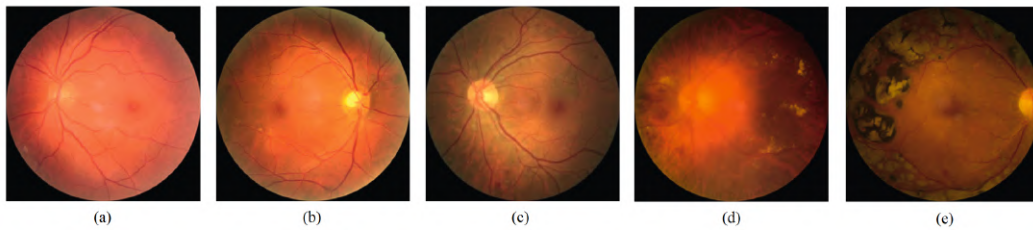


Figure 2.3: *DR severity levels. (a) No DR, (b) Mild, (c) Moderate, (d) Severe, (e) Proliferative DR [6].*

2.3.1 DR Severity Levels

According to the ICDR [1], DR severity can be classified into two classes: NPDR and PDR. Each class has its own severity levels representing the stage of the disease complication. Specific lesions in the retina can be used to describe each level.

Non-proliferative Diabetic Retinopathy

As shown in Figure 2.3, NPDR is divided into four stages: 0 representing non-DR, 1 representing mild, 2 representing moderate, and 3 representing severe. The non-DR is the one in which a diabetic patient's retina is healthy and has not yet developed any DR symptoms. Mild NPDR is the stage where microaneurysms appear, which are tiny red blood dots protruding from an artery or vein. Moderate NPDR is the stage characterized by the hardening of exudates, retinal hemorrhages, and microaneurysms from the previous stage. Hard exudates are lipid and proteinaceous materials that accumulate in the outer retinal layers, whereas hemorrhages are red spots that are bigger than aneurysms caused by abnormal bleeding inside the retinal blood vessels.

Severe NPDR is the third stage of DR severity, characterized by a group of lesions in any quarter of the four quadrants of the fundus image. The fundus image is graded as severe NPDR if one of the following statuses is in any quadrant. First, each of the four quadrants has more than 20 intraretinal hemorrhages. Second, definite venous beading in 2+ quadrants is detected when the vessels' shape resembles a sausage. Third, the presence of prominent intraretinal microvascular abnormalities (IRMA) in the 1+ quadrant with no signs of neovascularization. Abnormal blood

vessel branching, IRMA, is frequently misclassified as neovascularization. However, IRMA and neovascularization can be distinguished in several ways because IRMA develops deeper in the retina, away from the optic disc (OD), whereas neovascularization primarily develops around the OD. Additionally, IRMA leaks can be diagnosed after a shorter period of uncontrolled diabetes, whereas neovascularization takes longer to appear in the patient's retina. Therefore, the absence of neovascularization is the fourth indication of severe NPDR. Figure 2.3 (b-d) shows fundus images at mild, moderate, and severe stages of DR.

Proliferative Diabetic Retinopathy

As the disease progresses, PDR presents a more extreme level of DR where lesions can impair human vision. During this stage, new abnormal vessels called neovascularization develop after a prolonged period of diabetes. These vessels develop on the retina's surface and can bleed and form scar tissue that severely impairs vision.

The PDR is classified into two stages: neovascularization on the OD (NVD) and neovascularization elsewhere (NVE) [7]. NVD is a level where abnormal vessels have grown to one disc diameter (DD), whereas NVE is a level where new abnormal vessels appear further than one DD. The presence of NVD, NVE, or both indicates the presence of DR severity at the PDR level. However, this level of DR is easily distinguishable because the new abnormal vessels resemble normal retinal vessels. Figure 2.4 shows two fundus images with NVD and NVE levels.

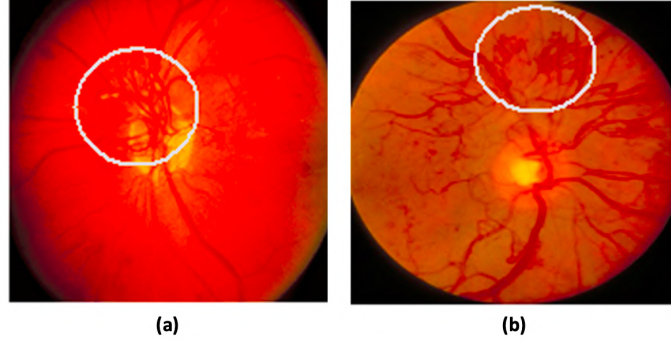


Figure 2.4: *PDR levels. (a) NVD (b) NVE [7].*

2.4 DR Diagnosis Challenges

The diagnosis of DR has several challenges, ranging from manual screening to automated diagnosis, with the difficulty of accurately classifying lesions being one of them. Lesions are tiny and difficult to detect using manual diagnosis. In clinical practice, the difference between two consecutive DR levels is difficult to distinguish, particularly during the early stages. The DR grade may be misclassified owing to these similarities [6]. For example, hemorrhages are sometimes misclassified as microaneurysms owing to their similar color. Lesions may also overlap with retina structure (optic disk and vessels), which makes DR screening more challenging [12]. This overlap causes the lesions to disappear, making it difficult to accurately grade the severity of DR. For example, microaneurysms are very close to retina vessels, and the contrast between them and the background is very low.

The quality of the fundus image significantly impacts the accuracy of the DR grading techniques [19]. Problems during the acquisition process can result in low-quality images with several challenges, such as noise, reflections, illuminations, and

low contrast [21], [19]. The local contrast of fundus images is occasionally poor owing to diverse sites when acquiring fundus images with different fundus cameras under different conditions and illumination [19].

It is essential to develop DL models to automate DR grading and overcome the previous challenges. However, some challenges are inherited from clinical practice to AI models. For example, the misclassification of adjacent grades with similar lesions is a common problem in DR grading models. Another problem of using DL models for DR screening is that they require a significant amount of annotated data to perform well. Collecting medical data for DR diagnosis is challenging owing to privacy concerns in the medical industry. Additionally, annotating fundus images is costly and time-consuming and demands expert input. These challenges make it more difficult to use DL models for DR grading.

Diverse data is important for developing robust automated DL grading systems. However, DL algorithms created for Western populations fail to detect some significant lesions in Eastern populations. According to [15], people with darker skin tones have more melanocytes in the retina’s outermost layer, the retinal pigmented epithelium (RPE). Compared with the lighter-colored retina, the darker-colored retina conceals vascular changes. This issue was noticed when [15], used the DL framework developed using a dataset of individuals with light-colored skin and retina. For an algorithm to be more broadly applicable, it must be adopted in patients with more diverse racial backgrounds and skin tones.

CHAPTER 3

DIABETIC RETINOPATHY GRADING USING ARTIFICIAL INTELLIGENCE: LITERATURE REVIEW

3.1 Introduction

DR is widely recognized as a leading cause of blindness among individuals with diabetes worldwide. Therefore, early diagnosis of DR plays a crucial role in preserving patients' vision and halting the progression of the disease to advanced stages. However, manual diagnosis of DR in clinical practice is time-consuming and susceptible to human error, especially during the early stages when the lesions associated with DR are often minute and challenging to identify. Furthermore, with the projected surge

in the number of diabetic patients and a concurrent shortage of ophthalmologists, there will be insufficient healthcare professionals available to examine all individuals at risk. The application of machine- and deep learning-based techniques has proven effective in diagnosing medical images, including those related to DR. In this chapter, we surveyed and analyzed 55 DR grading studies published between 2018 and 2022 extracted from four academic digital libraries: Scopus, Web of Science, Google Scholar, and Science Direct. The review provides a comprehensive discussion and analysis of these selected studies, considering various aspects such as benchmark DR datasets, classification tasks, preprocessing techniques, learning approaches, and performance evaluation measures. Within the literature on DR grading, supervised-based learning techniques have been found to be more prevalent than semi-supervised learning techniques. Furthermore, researchers predominantly addressed this problem as an image-level classification task, overlooking the distinctive characteristics of lesions within each grade. Numerous proposed techniques primarily concentrate on detecting the early stages of DR, while a limited number of studies address the disease’s advanced stages. The primary findings of our analysis reveal a potential direction for future research that emphasizes data- and model-centric approaches.

The objective of this review in this chapter is to provide a comprehensive analysis of the current literature concerning DR grading techniques. In addition, we aim to identify the open research gaps that should be investigated by researchers to develop more reliable DR grading methods. We categorize the DR grading techniques based on their learning approach (supervised, unsupervised, and semi-supervised) and conduct an in-depth analysis of each approach. The main contributions of this literature review

are outlined as follows:

- Proposing a comprehensive taxonomy to systematically classify existing research in the field of diabetic retinopathy (DR) severity grading.
- Critically analyzing publicly available datasets, highlighting their limitations, and offering recommendations for enhancing their utility in future studies.
- Evaluating the role of various image processing techniques in improving the performance of DR classification models, with a focus on their impact on feature extraction and lesion detection.
- Summarizing and comparing different learning techniques, including supervised, semi-supervised, and self-supervised methods, used to develop DR classification models, and providing a comparative analysis of their performance across various configurations.
- Identifying open research gaps and proposing areas for future investigation to advance the development of more accurate and reliable DR grading systems.

3.2 Search strategy

This review presents a comprehensive survey of automatic DR grading techniques, considering the abundance of studies published in the previous decade. Therefore, this review focused on studies published between 2018 and 2022 to analyze and discuss more recent studies in the field of DR grading. This selected time frame has witnessed

notable advancements in deep learning (DL) models and techniques, leading to a significant milestone where DL algorithms surpassed traditional machine learning (ML) approaches in accuracy and performance, especially in computer vision tasks. Furthermore, this period holds significance for DR research due to the availability of numerous newly released datasets. These datasets have played a crucial role in facilitating the development of more accurate and efficient DL models by allowing researchers to train their models on larger and more diverse image collections, consequently improving the models' performance in detecting DR. During the period (2018-2022), there was an increase in research efforts toward DR severity grading using DL models. Multiple studies were conducted during this period, focusing on various aspects of the modeling process, including data preprocessing, feature extraction, model architecture, and optimization. Reviewing studies conducted during this period provides a comprehensive understanding of the state-of-the-art methods employed in DR severity grading utilizing DL models.

In order to encompass all studies pertinent to automatic DR grading, we targeted the four main academic repositories: Scopus, Web of Science, Google Scholar, and Science Direct. We searched for all relevant studies using the title, abstract, and keywords. The keywords used in the search query were "diabetic retinopathy classification," "diabetic retinopathy detection," "diabetic retinopathy grading," "diabetic retinopathy recognition," and "fundus image". These keywords returned 239 papers.

To identify and retain the most relevant papers, we established a set of inclusion and exclusion criteria. These criteria included the following: the study should be written in English and published in peer-reviewed journals or conferences. Additionally,

we only included empirical studies that used ML and DL techniques for DR grading. The selected studies provided sufficient details about their models and results. Applying these criteria resulted in 55 studies being included in this review.

3.3 Related Surveys Comparison

Several recent reviews have focused on summarizing the techniques proposed for grading DR. Table 3.1, shows selected reviews on DR and compares them with this review. These reviews have primarily focused on DL techniques, with only one review discussing classic techniques [22]. While many reviews have focused on supervised learning techniques [23], [24], [25], only a few surveys have discussed unsupervised and semi-supervised learning techniques for DR grading [22], [26]. To the best of our knowledge, there has been no review surveying semi-supervised techniques for DR severity grading. Therefore, many of the supervised learning techniques discussed in these reviews are based on CNNs, and only a limited number of reviews discussed models utilizing other types of NNs. The number of reviewed studies in the related literature ranged from 18 to 74. However, some of these reviews [23], [24], [26] neglected to mention the period of the extracted studies or the digital libraries from which the studies were extracted [22], [23], [24], [26]. These reviews discussed various datasets, from image-level labeling datasets aimed at DR classification to pixel-level segmentation for lesion localization. However, none of them have analyzed the drawbacks of these datasets or examined how public datasets were utilized, such as whether they combined samples from multiple datasets or used a single dataset exclusively.

The future work discussed in these reviews has mostly concentrated on investigating the model itself and has not given enough attention to the importance of focusing on data usage.

A comprehensive review that discusses recent DR grading approaches is required. To fill this gap, this survey provides a comprehensive review of supervised, unsupervised, and semi-supervised approaches that have been proposed for DR grading between 2018 and 2022. Contrary to most recent reviews that focused only on CNN models for DR grading, our review will explore all NNs, including CNN-based ones, that have been used for DR grading. We will also analyze and discuss different pre-processing techniques, classification tasks, and performance measures employed by the identified studies. Our review aims to offer research communities a better understanding of the constraints present in current public DR datasets. Additionally, we extensively explore how various datasets have been employed in the studies we reviewed.

Table 3.1: *Related Review Comparison Framework. (SSL: Semi-supervised, USL: unsupervised, G: analysis of gaps in public datasets, No: number of datasets analyzed, NA: not available)*

Ref.	Year	Period	Supervised			Datasets				
			ML	DL		SSL	USL	G	Studies	
				CNN	non-CNN				No.	
[26]	2019	NA	×	✓	✓	×	✓	×	18	NA
[22]	2020	2013-2018	✓	✓	×	×	✓	×	15	74
[24]	2020	NA	×	✓	×	×	×	×	13	33
[25]	2020	2015-2019	×	✓	×	×	×	×	5	18
[23]	2021	NA	×	✓	×	×	×	×	10	NA
Ours	2022	2018-2022	✓	✓	✓	✓	✓	✓	18	55

3.4 Benchmark DR Grading Datasets

Several datasets have been proposed in the literature regarding DR severity grading and lesion segmentation tasks by providing image and pixel-level annotations. This section reviews these datasets and compares the specifications of the available DR grading and segmentation datasets. The datasets are compared based on the number of fundus images, DR grades, field-of-view (FOV), capturing device, and the availability of lesions annotation, as shown in Table 3.2. In this section, we categorize

the DR datasets into pixel and image-level annotated datasets based on the level of annotation.

3.4.1 Pixel-Level Annotated Datasets

Several DR datasets have been annotated at the pixel level. The annotation includes the lesion’s location, type, and blood vessels. Some datasets may also include the DR grade level of each image in a dataset. These datasets can be used for lesion segmentation and classification but have size limitations. This can be attributed to the cost and time of annotating fundus images at the lesion and vessel levels.

DRIVE [27] is the first dataset that has been proposed for DR. The dataset consists of 40 retinal images collected from 40 patients. The dataset was divided into training and testing, each with 20 images. Seven images represent mild NPDR, whereas 33 are free of DR lesions. The image has a resolution of 68×584 pixels, a FOV of 45° , and is compressed into a JPEG format. The images were captured using a Canon CR5 non-mydratic 3CCD camera. The dataset, which contains pixel-level binary vessel masks, is primarily used for evaluating the blood vessel segmentation of fundus images. The retinal vessel image set for estimation of width database (Review-DB) was created to facilitate research on vessel segmentation [28]. It has 16 images containing 193 vessel segments.

The CHASEDB1 dataset has 28 fundus images with a resolution of 1280×960 [29]. The dataset was collected from 14 children by capturing both eyes. A pair of images is used to represent one case. The dataset contains ground truth vessel annotations. The high-resolution fundus (HRF) contains 30 fundus images [30], which are divided

into two categories: 15 healthy images and 15 images that have DR signs. Experts manually segmented each DR-containing image. Each image ground truth in this dataset, representing the entire vasculature, contains complete thick and thin vessels.

The fine-grained annotated diabetic retinopathy (FGADR) dataset contains two sets. The first one is used for the segmentation task and contains 1,842 pixel-level annotated images. The other is used for DR grading and contains 1,000 images [31]. The FGADR dataset provides four DR lesion annotations: microaneurysms, hemorrhages, exudates, and soft exudates. Additionally, it provides image-level labeling to ease the identification of grades three and four by displaying laser marks and proliferate membranes, both of which are known to be useful signs in these grades. The Indian DR image segmentation dataset (IDRiD-S) [32] is a publicly available dataset that contains 81 retinal images that have been manually annotated for DR segmentation. The Dataset for Diabetic Retinopathy (DDR) dataset contains 13,673 fundus images of various sizes [33]. The dataset consists of 6266 images with DR severity levels of 0 and other images with samples from other levels. The distribution of the dataset is imbalanced, with more normal images than DR images. The dataset contains 757 images annotated with a bounding box to locate all types of DR lesions.

The E-ophtha dataset has 82 and 381 fundus images for EXs and MAs, respectively [34]. The EX-set contains 47 images with EX-signs, while the rest are normal. In the MA set, 148 fundus images have an MA sign, while the remaining images are healthy. The resolution of the images in the dataset ranges from 1440×960 to 2544×1696 pixels.

3.4.2 Image-Level Annotated Datasets

Many datasets that have been proposed for DR grading have image-level annotations. The annotation at this level indicates the DR grade of the image sample without providing any information regarding the spatial location of the DR lesions on the image. Although collecting and labeling these types of images is time-consuming, pixel-level annotated datasets require more resources to locate the DR lesions.

The retina image bank (RiB) platform, launched in August 2012, covers all aspects of retinal illness, anatomy, education, and treatment [35]. Each month, new images are added to this dynamic resource, and experts are invited to contribute their best images, case studies, and diagrams. It currently has 23,650 fundus images.

EyePACS is among the earliest datasets to be proposed for DR grading [36]. The dataset was created in 2015 as part of the DR detection competition sponsored by the California Healthcare Foundation - EyePACS and published in Kaggle. The dataset contains 88,704 fundus images with two resolutions (5184×3456 , 3888×2592). Ophthalmologists classified the lesions in these images into five DR stages from 0 to 4. Furthermore, the retinal images in the EyePACS dataset were collected using various fundus cameras under various lighting settings with a 45° FOV, making DR classification more challenging. The images were captured using different types of digital fundus cameras (Centervue DRS, Opovue iCam, Canon CR1/DGi/CR2, and Topcon NW).

The Asia Pacific Tele-Ophthalmology Society (APTOS) dataset contains 3662 retina images of different image sizes annotated into five DR stages scaled from 0

to 4 citekaggle2019. The dataset is publicly available on Kaggle, providing only the training set with labels, whereas the testing set has 702 samples that do not have any corresponding labels and 702 private leaderboard samples. The distribution of the dataset is imbalanced.

The Messidor dataset has two published versions: Messidor-1 and Messidor-2 [37]. The Messidor-1 dataset contains 1200 fundus images, 400 of which were captured without pupil dilation and 800 of which were captured with pupil dilation. The images were annotated into four DR grades, and the risk of macular edema was also provided. The fundus images were captured using a color video 3CCD camera mounted on a Topcon TRC NW6 non-mydratic retinography with a 45° FOV. The images in the dataset have different resolutions: 1440×960, 2240×1488, and 2304×1536 pixels. The Messidor-2 dataset contains two macula-centered ocular fundus images (one per eye). The dataset contains 1748 fundus images captured from 874 examinations across four grades and does not contain ground truth. However, other studies have annotated the dataset, although they are not part of the official Messidor-2 database. The images were collected using the same Messidor-1 settings. The deep DR (DeepDR) dataset contains 1200 fundus images graded into five DR severity scales [38].

Other datasets with few samples have been proposed for DR grading. The IDRiD-D [32] dataset, which is another version of IDRiD, was proposed for DR classification tasks. This dataset, which has 413 images with a resolution of 4288 × 2848, is publicly available. The dataset was manually classified into five DR severity levels. According to the risk of macular edema (ME), the dataset contains ME for three classes, ranging from 0 to 2. Kowa VX-10 alpha digital fundus camera was used for image capturing

with 50° FOV centered close to the macula. The STARE dataset has 400 fundus images annotated with 14 retina-related diseases [39]. The dataset was annotated for all DR grades. The images were captured using TopCon TRV-50 fundus camera with 35° FOV, which is approximately 650×550 pixels. The DIARETD0 database has 130 fundus images, 20 of which are healthy, and 110 have diabetic symptoms [40]. Fundus images were captured with 50° FOV. The DIARETDB1 dataset was proposed with a few samples using the same settings [41], and the dataset has 89 fundus images, 84 mild NPDR images, and 5 healthy images.

Table 3.2: *Publicly available datasets for DR severity grading and detection*

Dataset	Annotation level	Grades	Size	Annotation
DRIVE[27]	Pixel	-	40	Vessels
Review[28]		-	16	Vessels
CHASEDB1[29]		-	28	Vessels
E-Ophtha[34]		-	463	lesions
IDRiD-S[32]		-	81	Vessels
HRF[30]	Pixel and image	2	30	Vessels
FGADR[31]		5	1000	Lesions
DDR[33]		5	13673	Lesions
RiB[35]	Image	All retina diseases	-	
EyePACS[36]		5	88,704	
APTOS[42]		5	3662	
Messidor-1[37]		4	1200	
Messidor-2[37]		4	1748	
DeepDR[38]		5	1200	
IDRiD-D[32]		5	413	
STARE[39]		5	400	
DIARETD0[40]		2	130	
DIARETD1[41]		5	89	

3.5 Supervised Learning Based Techniques

Several techniques have been proposed to diagnose DR in its early stages. Many DR grading techniques apply supervised learning-based techniques. These techniques require a substantial amount of data for model training, which is typically unavailable for DR disease. Supervised learning-based techniques can be classified into classic and DL techniques.

3.5.1 Classic DR Severity Grading Techniques

Traditional DR classification methods rely on computer vision techniques to extract a set of features that are subsequently employed for training machine learning models or classifiers. The effectiveness of these approaches is heavily influenced by the quality of the features that are extracted. To enhance the performance of the models, various feature extraction methods, such as Gabor, local binary patterns (LBP), and gray level run length matrix (GLRLM), were used to extract more pertinent features. Additionally, various ML algorithms have been used in the DR grading literature, such as support vector machine (SVM), multiple support vector machine (SVM), k-nearest neighbor (kNN), NN, and extreme learning machine (ELM).

Mahiba et al. [43] used the Haralick texture feature (HTF) to extract features from fundus images. These features were used as inputs to three architectures: SVM, CNN, and Multi-CNN. This approach was evaluated on 550 retinal images collected from local hospitals. According to the reported results, the proposed approach outperformed other methods in moderate and severe NPDR grades. The proposed HTF+MCNN

outperformed the existing HTF with SVM and HTF with CNN methods in terms of overall accuracy, achieving 98.41% versus 97.84% and 96.65%, respectively. To extract discriminative features from low-resolution and noisy fundus images, Nazir et al. [44] proposed a novel technique for DR severity grading using the tetragonal local octal pattern (T-LOP) feature extraction technique. The extracted features were then fed into ELM for DR classification. The proposed approach was evaluated using the EyePACS, DRIVE, STARE, and Review-DB datasets. This approach obtained an accuracy of 99.6% on EyePACS, 99.96% on DRIVE, 99.95% on STARE, and 98.94% on Review-DB.

Jadhav et al. [45] used intensive image preprocessing steps to enhance DR grading and lesion detection. The preprocessing stages included converting the RGB fundus image into a green channel, enhancing the image using contrast limited adaptive histogram equalization (CLAHE), removing the OD through open-close watershed transformation (O-CWT), and segmenting the blood vessels using gray level thresholding (G-LTH). Top-hat transformation (T-HT) and Gabor filtering were used to segment the abnormalities, whereas LBP, TEM, Shannon's, and Kapur's entropy were used to extract the features. The modified gear steering-based rider optimization algorithm (MGS-ROA), which uses the least amount of correlation, was used for feature selection. A deep belief network (DBN) classification algorithm was trained for the four-class grading. Contrary to NN, KNN, and SVM, DBN outperformed other techniques with an accuracy, sensitivity, specificity, and precision of 93.182%, 86.364%, 95.455%, and 86.364% on the DIARETDB1 dataset, respectively.

Abdelmaksoud et al. [46] proposed an ML-CAD system that could extract texture

features from different angles of a fundus image using a gray-level run length matrix (GLRLM). Their study focused on two- and four-class tasks and lesion detection. To enhance the image quality, the median filter was used to remove noise, and HEBPDS was used to enhance the image contrast. To distinguish normal from DR cases, by using GLRLM, features from the green channel were extracted using GLRLM and passed to the SVM classifier for binary classification. Then, a customized U-Net segmentation was used for segmenting hemorrhages, exudates, blood vessels, and microaneurysm lesions. Finally, the four-class classification task was performed using a multi-label SVM (MLSVM). This study used the following eight publicly benchmark datasets: DRIVE [27], CHASEDB1 [29], STARE [39], HRF [30], IDRiD [32], DIARETDB1 [41], Messidor-1 [37]. The last dataset, E-ophtha [34], was used for training U-Net. These datasets were trained in various settings, but only the binary classification results are shown in Table 4. The proposed model had the worst accuracy on the Messidor dataset (75%), while it had a binary classification accuracy of 96.66% on the HRF dataset. In this study, the proposed system was compared with other frameworks in terms of lesion segmentation, and the proposed approach achieved the best accuracy of 96.58% on the DRIVE dataset.

Yaqoob et al. [47] used a histogram of oriented gradients (HOG) to extract statistical and geometric features from fundus images. The proposed approach was evaluated on Messidor-2 and EyePA datasets. The proposed approach had an accuracy of 97.01% on the Messidor-2 dataset in 26 seconds and 97.88% on the EyePACS in 50 seconds. Table 3.3 summarizes the classic techniques for DR grading.

Table 3.3: *Classic DR techniques for DR classification*

Ref.	Task	Pre-processing	Features extraction	Classifier	Grading	Dataset	Acc.
[44]	Multi	—	T-LOP	ELM	5	EyePACS	99.6%
						DRIVE	99.96%
						STARE	99.95%
						Review-DB	98.94%
[43]	Multi	Background removal	F(V1)	MCNNs	5	Private	98.41%
		Resizing	HTF	SVM			97.84%
		HE, adaptive HE	image-F(V3)	CNNs			96.65%
[47]	Multi	Canny edge detector	HOG	ResNet	5	Messidor-2	97.1%
						EyePACS	97.88%
[45]	Multi	Green channel	LBP	MGS-ROA	4	DIARETDB	193.182%
		CLAHE					
		O-CWT					
		G-TH					
		T-HT					
		Gabor filtering					
[46]	Binary	Median filter	GLRLM	SVM	2	MESSIDOR	75%
		HEBPDS		MLSVM		HRF	96.66%
		Green channel extraction				DRIVE	96.58%

3.5.2 Deep Learning DR Severity Grading Techniques

Most proposed DR grading techniques depend on DL for feature extraction and classification, and this section surveys these DL techniques. We categorize these techniques into two groups: CNN- and non-CNN-based models, which employ alternative network architectures to CNN. Several proposed DR grading DL models employed CNN to extract and classify features. CNN efficiently extracts spatial information from input images for several computer vision problems. This review categorizes the surveyed CNN-based DR models into three configurations.

The first configuration is end-to-end CNN, which requires training CNN from scratch on DR images. Another configuration depends on transfer learning (TL) and optimizes pre-trained models for DR classification tasks. The third configuration is the hybrid approach, which uses CNN for feature extraction and classical ML algorithms for classification.

End-to-End CNN-Based Model

End-to-end CNN-based models learn and classify DR grades using a single CNN model. This model is fully trained on fundus images. Rahim et al. [48] conducted three classification experiments using the proposed CNN model. Two of these experiments are binary classification (DR or Non-DR and maculopathy or non-maculopathy) and multiclassification by merging the DR stages with and without maculopathy. The proposed model was evaluated on a private dataset that contains 600 fundus images collected from the Eye Clinic at Melaka Hospital, Malaysia. The reported results

demonstrated that in some experiments, data augmentation aids in improving classification performance. The accuracy obtained from the training CNN model in the DR binary classification experiment was 63.33% with data augmentation, whereas without data augmentation it dropped to 61.11%. The maculopathy detection experiment yielded an accuracy of 77.22% without data augmentation and 22% with no improvement. In multi-class DR and maculopathy, the accuracy dropped with augmentation from 44.20% to 42.54%.

Sanaullah et al. [49] proposed a customized CNN model to classify the DR in fundus images into five classes corresponding to the DR grades. During the preprocessing stage, they used the Coye filter for contrast enhancement. Several experiments have been conducted to evaluate the proposed model on the EyePACS dataset. According to the reported results, the model fed with images preprocessed with the Coye filter outperformed RGB and green channel images with an accuracy of 89.07%.

Chakravarthy et al. [50] proposed the DR-Net model for DR classification. The model was trained on four public datasets: EyePACS, DRIVE [27], Messidor-2 [37], and private samples provided by ophthalmologists. During the preprocessing stage, CLAHE and Gaussian blurring were applied on the green channel of the fundus images. The proposed approach achieved an accuracy of 81% for the binary classification of the DR fundus images. Arora et al. [51] also proposed a customized CNN model for the binary classification of DR. This study was evaluated using a dataset consisting of 1000 fundus images from the EyePACS study. The images were augmented during the training and were cropped to a size of 128 x 128. The CNN model achieved 74% validation accuracy on the EyePACS dataset.

Abdelmaksoud et al. [52] proposed a hybrid model for DR grading. To diagnose the five grades of DR, the proposed model used EyeNet and DenseNet models to create the E-DenseNet model. The model was trained on the EyePACS and APTOS datasets, achieving an average accuracy of 91.6% and a Kappa score of 0.883 on the EyePACS dataset. Hemanth et al. [53] investigated the effect of histogram equalization (HE) and CLAHE techniques for DR classification. The image processing techniques were applied separately to each image channel (R, G, and B) and recombined for subsequent stages. The model was trained using 400 fundus images extracted from the Messidor-2 dataset. According to the results, CLAHE is more effective with a CNN model compared with HE, achieving an accuracy, specificity, and sensitivity of 97%, 98%, and 94%, respectively. According to Hemanth et al. [53], the proposed model is currently being used in two hospitals in Turkey. Qian et al. [54] also used CLAHE for DR grading. This study proposes an AD2NET model that uses DenseNet and Res2Net models. Furthermore, the attention mechanism was used to improve the model performance by focusing on the important regions in the fundus image. The proposed model was evaluated on the EyePACS dataset and achieved an accuracy of 83.2% and a Kappa of 0.8.

DeepPCANet was proposed [55] to overcome the limitations of DR annotation shortage and a large number of CNN parameters. This study used three image-level DR grading scenarios in the field: a multi-class, two-class (normal vs. DR from grade 1 to 4), and two-class (non-referral (grade 0 and 1) vs. referral (grades 2-4)). Three versions of DeepPCANet were proposed, each trained on a different dataset: DeepPCANet-4, DeepPCANet-5, and DeepPCANet-16. DeepPCANet-5,

DeepPCANet-4, and DeepPCANet-16 were trained on a King Saud University Medical City, Saudi Arabia, APTOS2019, and EyePACS datasets, respectively. The model is considered lightweight because its architecture was derived from fundus images. To automatically identify the depth and width of the DeepPCANet model, they used K-medoid clustering, principal component analysis (PCA), and inter- and intra-class variations. The findings demonstrate competitive performance using a few layers and parameters. In the five-class scenario, DeepPCANet-4 and DeepPCANet-16 achieved an accuracy of 98.21% and 94.22% on the APTOS and EyePACS datasets, respectively. In the two-class scenario, DeepPCANet-4 had the highest accuracy of 99.7% on the APTOS dataset. The highest accuracy that was achieved in binary and multi-classification experiments is reported in Table.3.4

Some studies have proposed methods to localize lesions and classification tasks [56], [57]. Liu et al. [56] proposed a weighted path CNN (WP-CNN) model with a weighted path to capture and reconstruct the features with different receptive fields. A private dataset containing 60,000 images was used for model training, while the STARE dataset was used for testing. This model achieved an accuracy, sensitivity, and specificity of 94.23%, 90.94%, and 95.74%, respectively. Additionally, WP-CNN outperformed DenseNet-121, ResNet-101, and SeNet-101 models in lesion localization. Alyoubi et al. [57] proposed two models: one to classify the fundus image into five stages using CNN512, and the other to localize DR lesions using YOLOv3. Both models were fused to perform DR multiclassification and lesion localization. Models were trained and evaluated on DDR [28] and APTOS public datasets. An accuracy of 89% sensitivity and 97.3% specificity were reported on the DDR dataset. Table 3.4

summarizes the end-to-end CNN techniques proposed for DR grading.

Transfer Learning

Many studies investigating supervised learning have employed transfer learning (TL) via pre-trained models. Transfer learning facilitates the adjustment of a model that was initially trained for a specific task, enabling its utilization for a different task. Several pre-trained models were trained on ImageNet [58]. The ImageNet project is a large visual database intended for use in research on visual object recognition tasks.

Wan et al. [59] investigated four pre-trained CNN models for DR classification: AlexNet, VGG, GoogLeNet, and ResNet. The results demonstrated that the VGG model had the highest accuracy of 95.68% on the EyePACS dataset. Lian et al. [60] investigated the impact of preprocessing, CNN architectures, class imbalance, and fine-tuning on DR classification. The EyePACS dataset was used to train the models for five-class classification. The results demonstrated that the fine-tuned VGG pre-trained model with proper image preprocessing and imbalance samples outperformed other architectures such as ResNet-50 and AlexNet.

Zhao et al. [61] used the attention technique to address small lesion size limitations. Their study proposed BiRA-Net, a new DL architecture that combines two models: an attention model for feature extraction and a bilinear model for fine-grained classification. Furthermore, their study proposed a new loss function to improve the model’s overall convergence. The model was evaluated on the EyePACS dataset for five-class grading. According to the results, the model could accurately classify all stages except class one, which was primarily predicted as class 0. The results demon-

Table 3.4: A summary of the end-to-end CNN models for DR severity grading and classification. G: DR grades,

Ref.Task	Pre-processing	Model	G.Dataset	Accuracy
[50] Binary	CLAHE Green channel extraction Gaussian Blurring	DR-Net	2 Mixed:(Private EyePACS DRIVE Messidor-2)	81%
[51] Binary	Augmentation Cropping	CNN	2 EyePACS	Sens.:88.4% Speci.:83.5%
[49] Multi	Coye filter Green channel	CNN	5 EyePACS	89%
[52] Multi	—	E-DenseNet	5 EyePACS APTOS	96.8% 86.5%
[53] Multi	HE CLAHE	CNN	5 Messidor-2	97%
[54] Multi	CLAHE	AD2Net	5 EyePACS	83.2%
[56] Multi	—	WP-CNN	5 Training: private Testing: STARE	94.23%
[57] Multi	Augmentation Cropping	CNN	5 APTOS DDR	84.1% 88.6%
[55] Multi	—	DeepPCANet-55	APTOS EyePACS	98.21% 94.22%
Binary		DeepPCANet-42	APTOS	99.7%
[48] Multi Binary	Augmentation	DCNN	5 2 Private	44.20% 63.33%

strated that BiRA-Net, Marco F1, and Micro F1 achieved an accuracy of 54.31%, 57.25%, and 54.36%, respectively.

Lands et al. [62] evaluated four pre-trained models for five-class DR grading. These models include ResNet-50, DenseNet121, DenseNet169, and DenseNet256. The fundus images were preprocessed using Gaussian blur subtraction to increase the image's details before data augmentation. DenseNet121 obtained the best kappa score of 0.78 on the APTOS dataset.

Dai et al. [63] developed an automated DR grading system that contains three sub-networks for image quality assessment, lesion-aware, and DR grading. ResNet41 and Mask-RCNN57 were used to develop each sub-net. First, ResNet was trained on the fundus images to create the DR base network, which was used for sharing its weights in the three sub-networks. The features extracted from the lesion-aware sub-network and the original fundus image were fed to the grading sub-network. The model was trained on 666,383 labeled fundus images collected from a local diabetic center in Shanghai. The grading results demonstrated an AUC of 94.3%, 95.5%, 96%, and 97.2% for mild, moderate, severe, and proliferative lesions, respectively. Whereas, the lesion detection results demonstrated an AUC of 90.1%, 94.1%, 95.4%, and 96.7% for microaneurysms, cotton-wool spots, hard exudates, and hemorrhages, respectively. The results also demonstrated how image quality can improve DR mild grading results.

A multi-input system was proposed [64], where the network was fed with two different input modalities: RGB fundus images from the MISSADOR dataset and OCTA images from a local dataset KHUMC. The proposed architecture comprises a backbone convolutional network, ResNet, linked to a two-fold feature augmentation

network (TFA-Net). Multiple convolution blocks of ResNet extract representational characteristics at different scales. The TFA-Net is built using weight-sharing convolution kernels before deploying a reverse cross-attention (RCA) stream. This approach was originally proposed to solve the issue of high demand for enormous, labeled data samples. The authors concluded that multiple inputs to CNN architecture combined with strongly regularized feature levels can be a viable option for reducing the reliance on large amounts of labeled data. The results showed that ResNet101-RCA provided the best accuracy of 94.8% on the EyePACS dataset. Suedumrong et al. [13] investigated the performance of two pre-trained models, GoogLeNet and VGG-16, for DR classification. GoogLeNet performed better on binary DR classification, while VGG-16 provided better accuracy on DR severity grading. In the two-class classification task, an accuracy of 85.86% and 79.89% was reported using GoogLeNet and VGG-16 models on the EyePACS dataset, respectively. In the five-class classification task, VGG-16 achieved an accuracy of 71.65% compared with 61.58% for GoogLeNet.

One of the main challenges of DR classification is the distinction between adjacent grades, such as 0 and 1, which is primarily attributed to the small size of grade one lesions. Metan et al. [65] examined how residual connections with CNN-based models can degrade the performance of CNN models in detecting tiny lesions. The authors evaluated non-residual CNN, plain ResNet, ResNet with shallow skip connections, and ResNet with deep skip connections. These models were evaluated on the EyePACS dataset, and the performance of ResNet with deep skip connections outperformed plain ResNet on grades 0 and 2 predictions. However, all the architectures demonstrated that the accuracy of predicting class one was still low compared with other grades.

Several studies have attempted to use an ensemble technique to improve the performance of DR grading systems [66], [67], [68]. An ensemble-based framework was proposed in [66] to classify fundus images into five severity classes. Image augmentation was performed to balance the samples for each class, and up- and down-sampling were investigated. The proposed framework consists of five deep CNN models: ResNet-50, Inceptionv3, Xception, Dense-121, and Dense-169. The proposed models achieved an accuracy of 80.8% using the EyePACS dataset. Zhang et al. [67] proposed the DeepDR model for two- and five-class DR severity grading. A high-quality dataset containing 13,767 fundus images was collected from a local hospital. An adaptive HE was used to reveal hidden visual effects in the fundus images. Several experiments have been conducted to investigate different combinations of six pre-trained models for DR grading. The models were divided into two groups: weak and strong. The group of weak models includes InceptionV3, InceptionResNetV2, and Xception. The group of strong models includes ResNet-50, DenseNet169, DenseNet201, and DenseNet121. First, several ensembles were investigated in each group. Then, the best combination from each group was further combined to construct the final ensemble grading system framework. According to the results, ensemble models from the strong group produced the best accuracy of 96.50% in the five-class task and 97.67% in the two-class task. Furthermore, the results demonstrated that the optimal number of classifiers to ensemble might be less or equal to the number of classes of interest. Some local hospitals in China are currently using the DeepDR model. Similarly, five pre-trained CNN models (ResNet-50, Xception, Inceptionv3, DenseNet121, and DenseNet169) were ensemble [68] and used for DR grading. Additionally, the authors investigated the

effect of each RGB channel on the model performance. Therefore, each image channel (Red, Green, and Blue) was separately fed to each model, and the effect of each channel on the outcome was evaluated. The ensemble model that used the fundus image's green channel outperformed the other channels. The accuracy obtained using an ensemble with a green channel was 81.85% compared with red and blue channels that had accuracies of 74.98% and 73.86%, respectively, using the EyePACS dataset.

Other studies [52] [6] proposed hybrid models for DR classification. An E-DenseNet model was proposed in [52] by combining two pre-trained models, EyeNet and DenseNet, to classify the five DR grades from two DR datasets, EyePACS and APTOS. The proposed model achieved an accuracy of 86.5% and 96.8% on the APTOS and EyePACS datasets, respectively. Similarly, three hybrid model structures were proposed for DR classification by ensembling five CNN models (EfficientNetB4, EfficientNetB5, NASNetLarge, Xception, and InceptionResNetV2) [6]. Additionally, the study proposed an enhanced loss function that outperformed the cross-entropy loss function in terms of performance and training speeds. The EyePACS, APTOS, and DeepDR datasets were combined to train the proposed hybrid models. According to the study, adding more complexity to the hybrid models improves performance. To this end, the proposed algorithm achieved five-class accuracy, sensitivity, specificity, precision, and F1 score of 86.34%, 98.77%, 74.76%, 91.37%, and 93.9%, respectively.

Islam et al. [69] used a Gaussian filter for noise removal, and the filtered images were input into the VGG-16 pre-trained model. The APTOS dataset was used to evaluate the proposed model, and an accuracy of 91.32% was achieved. Fundus images were divided into four segments and fed into a multi-channel architecture for training

four Inception-V3 networks in parallel [70]. Then, the last layers from each channel were concatenated to train fully connected layers for five-class DR grading. The proposed approach, which uses the learned image downscaling (LID) method, achieved an accuracy and sensitivity of 85.2% and 83.4% on the APTOS dataset, respectively.

A lightweight model was proposed to classify DR fundus images into four grades [71]. The study focused on the preprocessing phase to achieve acceptable DR classification accuracy with a small and imbalanced dataset. The preprocessing phase includes contrast enhancement, eliminating the black background, CLAHE to add more contrast to the small lesions in the fundus image, clipping limit with 2.0, and weighted Gaussian blur to reduce noise. The pre-trained DenseNet model was used, achieving a kappa score of 0.88 and 0.98 on the validation and training data, respectively. Sugeno et al. [12] proposed a new approach for DR lesion detection and classification using two datasets. The APTOS dataset [42] was used to train the EfficientNet model [72] for four-class NPDR classification (no-DR, mild, moderate, and severe), while the DIARETDB1 database [41] was used for the lesion detector model. The studies conducted intensive preprocessing steps to exclude blurred images. The lesion detector aims to distinguish between red and white lesions. The proposed approach failed to distinguish red lesions from normal vessels. Red lesions, such as microaneurysms and hemorrhages, are very close to vessels, which makes it difficult to detect and distinguish normal vessels from red lesions. Furthermore, OD and contour regions were detected as white lesions, such as hard and soft exudates. The application of OD separation did not solve vessel separation because segmenting vessels can also remove white lesions. The studies also provided a visual explanation

for interpreting the model decision using three methods: grad CAM++, faster score CAM, and smooth grad. The results demonstrated an average accuracy of 94.6% for grading DR.

A hybrid approach that combined preprocessing steps, DL models, and an optimization method was proposed to achieve an end-to-end model [73]. They used the following three datasets: APTOS, EyePACS, and Messidor-2. Morphological gradient (MG) and segmentation approaches were applied during the preprocessing stage for image enhancement. MG includes dilation and erosion, whereas the segmentation process (SP) includes thresholding, closing, contour extraction, and masking. The two generated images from each preprocessing step—morphological and segmented—were fed into two models, ResNet-18 and VGG-16, to generate feature maps that were fed to the dominant activation selection process (DASP). DASP contains atom search optimization, dominant class selection, and a softmax layer for the classification process. The results demonstrated that using MG and SP with ResNet-18 and VGG-16 improved accuracy. ResNet-18 demonstrated an accuracy of 99.29% on the EyePACS and 98.70% on the mixed dataset (APTOS & Messidor-2). VGG-16 achieved an accuracy of 99.28% on the EyePACS and 98.52% on the mixed dataset (APTOS & Messidor-2).

To simplify and improve the efficiency of the models trained for two-class DR grading, Cinarer et al. [74] investigated appropriate preprocessing methods that could achieve considerable results. Images were resized to $224 \times 224 \times 3$ and converted from RGB into grayscale modality. Additionally, a Gaussian filter was used to remove noise. Data augmentation was applied to the training set but not to the validation set. The

performance of the AlexNET and VGG-16 deep TL algorithms was evaluated on the APTOS dataset. Only 3662 fundus images from the APTOS dataset were used. The proposed model achieved a high accuracy of 95.26% and 98.18% using AlexNet and VGG-16, respectively.

Vives et al. [75] proposed a weighting strategy in backpropagation to improve image processing by introducing synaptic metaplasticity. This technique uses four pre-trained models, VGG-16, InceptionV3, ResNet- 50, and DenseNet121. The results demonstrated that even with a small dataset, using synaptic metaplasticity in CNNs aids in detecting the most significant features with a noticeable improvement in the learning rate, convergence, and accuracy. InceptionV3 architecture with synaptic metaplasticity outperformed the other models with accuracy, F1-score, precision, and recall of 94.46%, 94.24%, 98.9%, and 90% on the APTOS dataset, respectively.

In contrast to earlier studies that aimed to improve the DR classification accuracy through extensive preprocessing, Murcia et al. [76] conducted several experiments using different CNN pre-trained models to achieve high performance without enhancing the image quality. Their study evaluated the following pre-trained models: Alexnet, VGG-16, VGG-19, Squeezenet, ZhangNet, ResNet-18, and ResNet-50. The DR classification task used in this study involves three-class grading, where grades 0 and 1 are considered as one level, followed by grades 2 and 3 for moderate and severe. ResNet-50-based architecture achieved the best results with an accuracy of 93% for grades 0 and 1, 81% for grade 2, and 92% for grade 3, overall accuracy of 92.3%. Binary classification (grades 0 vs. 3) yields higher than 97% accuracy. Saeed et al. [77] proposed an intelligent DR diagnosis system without applying any preprocessing. A

two-stage model was introduced to perform the DR classification task. The first stage is a method for fine-tuning pre-trained model adaptation, where the model was trained on the structure of the lesion (lesion ROIs) on the E-optha dataset. The second stage involves adapting high-level layers to extract the unique structures using PCA layers, which might reduce the model complexity. Three different combinations of DR classification tasks were applied: multi-class grading, two-class grading (no DR vs. the rest of grades), and two-class (non-referable (Normal + grade 1) vs. referable (grades 2, 3, and 4)). Notably, downsizing the image would have damaged the structure of the lesion, which is important for DR classification. Therefore, the DR model received fundus images with a size of 1000 x 1000. The proposed model, ResNetGB (ResNet + gradient boosting classifier), achieved 99.73% and 98.88% accuracy on the EyePACS and Messidor datasets for the multi-classification problem, respectively. The reported results showed high sensitivity and specificity with two-class grading 99.97% and 99.99% in the EyePACS dataset, and 99.54%; 98.90% in the Messidor dataset. The results reported in Table 3.5 include only the multi-class grading.

Some reviewed studies proposed a multi-task model for several eye diseases, including DR grading. Jhe et al. [78] developed a multi-task model by combining DR classification with other eye diseases. The study employed three CNN models based on AlexNet to classify DR severity grades and macular edema risk. The features extracted from each model were concatenated and fed into the classification layer. The model was trained using the IDRiD dataset, which has five- and three-class grading for DR and macular edema risk, respectively. The results demonstrated that the model better classified DR grades 0 and 2 than other grades. The low classifica-

tion results of grades 1 and 4 can be attributed to the small number of samples and the use of data augmentation techniques. The fusion network had an accuracy of 65.38%. A new architecture was proposed to maintain fundus image resolution and extract as many significant features as possible [79]. Their study proposed a multi-cell (M2CNN) architecture based on Inception-ResNet-v2 to facilitate the training of CNN with large-resolution images, where both the image resolution and network depth gradually increase. The MsCNN architecture selects the NN path based on the resolution of the input image. The proposed model achieved a Kappa of 0.841 on the EyePACS dataset. Table 3.5 summarizes the pre-trained CNN techniques proposed in the reviewed DR grading studies.

Table 3.5: *A Summary of Transfer Learning Techniques*

Ref.Task	Pre-processing	Model	GradingDataset		Accuracy
	Gray-scale				
[74] Binary	Gaussian filter	VggNet-16	2	APTOS	98.17%
	Augmentation				
[59] Multi	Augmentation	CNN	5	EyePACS	95.68%
	Normalization				
Tab. 3.5 Continues...					

Table 3.5 – continued from previous page

	Augmentation				
[60] Multi	Normlization	VGG-16	5	EyePACS	79.04%
	Color enhancement				
[61] Multi	Augmentation	BiRA-Net	5	EyePACS	54.%31
	HE				
	Augmentation				
[62] Multi	Gaussian Blur	DenseNet121	5	EyePACS	78%
	Subtraction			APTOS	-
[63] Multi	quality assessment	ResNet41	5	Private	-
		Mask-RCNN57			
	Augmentation				
[64] Multi	LMIS	ResNet	5	Private	94.8%
		TFA-Net		MISSADOR	
	intensity subtract				
[65] Multi	Cropping	Plain ResNet	5	EyePACS	81%
[13] Multi	Light adjustments	GoogLeNet	5	EyePACS	85.86%
		VGG-16			71.65%
[66] Multi	Augmentation	Ensemble of CNNs	5	EyePACS	80.8%
Tab. 3.5 Continues...					

Table 3.5 – continued from previous page

[68]	Multi	Augmentation RGB Separation	Ensemble of CNNs	5	EyePACS	81.85%
[78]	Multi	Augmentation	AlexNet	5	IDRiD	65.38%
[79]	Multi	-	M2CNN	5	EyePACS	0.841 (Kappa)
[52]	Multi	-	E-DenseNet	5	EyePACS APTOS	96.8% 86.5%
[6]	Multi	Augmentation Binary processing	Hybrid model	5	Mixed: (EyePACS APTOS DeepDR)	86.34%
[69]	Multi	Cropping Sharpening Gaussian filter	VGG16	5	APTOS	91.32%
[70]	Multi	Cropping Padding Gaussian filter downscaling algorithms	Parallel Inception-V35		Mixed: (IDRID EyePACS)	85.2%
Tab. 3.5 Continues...						

Table 3.5 – continued from previous page

[71] Multi	CLAHE	DenseNet	4	EyePACS	0.8836 (Kappa)
	Clipping				
	Gaussian blur				
[12] Multi	Laplacian filter	EfficientNet-B3	4	APTOS	94.6%
	Standard deviation				
	Optic disc separation				
[73] Multi	Dilation Erosion	ResNet-18	5	EyePACS	99.29%
				Mixed:	
				(APTOS Messidor-2)	98.70%
		VGG-16		EyePACS	99.28%
				Mixed:	
				(APTOS Messidor-2)	98.52%
[75] Multi	Resizing	Inception-3	5	APTOS	94.46%
Tab. 3.5 Continues...					

Table 3.5 – continued from previous page

[76]	Multi -	ResNet50	3	Missador-2	92.3%
	Binary		2		97%
[77]	Multi -	ResNetGB	5	EyePACS	99.73%
				Messidor-2	98.88%
[67]	Multi HE	DeepDR-Ens(456)	5	Private	96.50%
	BinaryAdaptive HE		2		97.67%
End of Tab. 3.5					

Hybrid CNN-ML based Approaches

Some studies [80], [81], [82], [83], [84], [85], [86], [87] used the CNN model for feature extraction and the conventional ML algorithm for classification. A CNN with ML classifier combination was proposed for two- and five-class DR grading [85]. The study intended to minimize the learnable parameters in the CNN model to allow real-time deployment and maintain classification accuracy [80]. The CNN model was used as a feature extractor, and its output was fed into different ML classifiers, SVM, AdaBoost, Naive Bayes, Random Forest, and J48. Each classifier was evaluated on three datasets: IDRiD, Messidor-2, and EyePACS. The results demonstrated that the J48 classifier outperformed other classifiers in terms of binary and multi-class classification. The Messidor-2, IDRiD, and EyePACS datasets were trained and evaluated separately. The accuracy for binary classification was 99.92%, 99.76%, and 99.99% on the Messidor-2, IDRiD, and EyePACS datasets, respectively. On the other hand, for

the multi-class tasks, the accuracy was 99.75%, 99.03%, and 99.99% on the Messidor-2, IDRiD, and EyePACS datasets, respectively.

Chun-Yan et al. [80] proposed a novel architecture using bilinear pooling CNN to extract second-order statistics of low features and tree-based approaches for high features. The compact bilinear pooling (CBP) CNN was trained on VGG-16, and the extracted features were fed into a gradient-boosted decision tree (GBDT) for classification. The proposed model was evaluated on the EyePACS dataset, which demonstrated superior performance for five-class DR classification. The results had a Kappa score, average F1-score, and micro-average AUC of 0.73, 0.79, and 0.95, respectively. Bhatti et al. [81] proposed an automated DR classification system that divides DR into three classes: mild, moderate, and severe DR, using a CNN-based model for feature extraction and multi-SVM (MSVM)s. The images were resized into uniform samples before being preprocessed using Gaussian filtering to remove noise. The model achieved an accuracy of 94.6% and an estimation capacity value of 93.8% on the DIARETDB0 dataset.

A hybrid DL model consisting of MobileNetV2 and SVM was proposed in [82]. The MobileNetV2-SVM architecture was described as a small-scale architecture with 4.2 million trainable parameters. The proposed model was trained and evaluated using the APTOS dataset. Compared with other studies using similar datasets, MobileNetV2-SVM achieved a kappa score, accuracy, and AUROC of 0.925, 0.9225, 85%, respectively. This study demonstrates that using the appropriate optimization method, a relatively modest and generic CNN architecture with minimal image input size can produce promising DR classification performance with a limited amount of data. De-

spite these excellent results, the study highlighted a significant misclassification at mild to moderate grades, moderate to both mild and severe grades, and proliferative to severe grades. These findings are consistent with a prior study that discovered substantial overlap between the classification of the mild, moderate, and severe grades, as well as the severe-3 and proliferative-4 grades, in the APTOS dataset, even after ensemble feature extraction.

Another study [83] proposed a CNN model for feature extraction and an SVM classifier for DR classification. Two approaches (prominent feature TL approach (PFTL) and image feature TL approach (IFTL)), were presented and evaluated. The PFTL approach used CNN models for feature extraction to produce image feature vector (IFV), which was subjected to a statistical feature analysis method to reduce the IFV complexity and remove irrelevant features. Consequently, PFV was produced, which was then fed into the SVM classifier. The IFTL approach followed the same steps as PFTL, but excluded the statistical step. Both approaches were trained and evaluated on the Messidor-2 and IDRiD datasets. As a preprocessing stage, images were augmented, and the background and OD were eliminated. Then, a morphological closing operation (MCO) was used to remove the noise represented as holes in the image. The PFTL approach with statistical feature reduction outperformed the IFTL approach with an accuracy of 90.51% using the IncpetionV3 model on the Messidor-2 dataset, and an accuracy of 90.01% on the IDRiD dataset. In terms of cost analysis and computational time, the PFTL approach achieved the best results.

A multipath CNN (M-CNN) was used to extract the local and global features [84]. Three ML classifiers were evaluated (SVM, random forest, and J48) using three

datasets (IDRiD, EyePACS, and Messidor-2) for four-class severity grading. The J48 classifier performed the best with an accuracy of 99.03%, 99.9%, and 99.75% on the IDRiD, EyePACS, and Messidor-2 datasets.

Deepa et al. [86] proposed a framework comprising seven functional units: preprocessing, patch extraction, CNN models, probability vector concatenation, ML classification based on an ANN classifier, ensemble technique, and DR classification. InceptionV3 and Xception models were re-trained on a dataset consisting of 2290 high-resolution fundus images collected from a local hospital, which was annotated into four-class severity (normal, mild DR, moderate DR, and PDR). More samples were added from the following public datasets: DIARETDB, STARE, retinopathy online challenge (ROC) [88], and EyePACS. The ensemble approach with SVM achieved the best performance with an accuracy of 96.2% compared with ANN, Adaboost, and random forest classifiers. The authors proposed a further classification into NVD, NVE, and neovascularization at the iris (NVI) to detect the high-risk characteristics of PDR.

Yaqoob et al. [87] proposed a DR classification approach using ResNet-50 as a feature extractor and random forest for classification. The proposed approach was trained for two- and five-class tasks on the Messidor-2 and EyePACS datasets, respectively. The proposed approach was compared with six architectures, ResNet-50, VGG-19, Inception-v3, MobileNet, Xception, and VGG-16. The results demonstrated that ResNet-50 outperformed other pre-trained models with 96% and 75.09% accuracy on the Messidor2 and EyePACS datasets, respectively. Table 3.6 summarizes the techniques for DR grading that employed Hybrid methods CNN with ML techniques.

Table 3.6: Summary of Hybrid techniques for DR severity grading. These techniques employ neural networks as feature extractors and ML algorithms for classification

Ref.Task	Pre-processing	Features extraction	Classifier	Grading	Dataset	Accuracy
[80] Multi	Augmentation	VGG-16	GBDT	5	EyePACS	95%
	Color enhancement	CBP				
[81] Multi	Gaussian filter	CNN	MSVM	3	DIARETDB09	94.6%
[82] Multi	Augmentation	MobileNetV2	SVM	5	APTOS	85%
[83] Multi	Augmentation	PFTL	SVM	3	IDRiD	90.01%
	Background elimination					
	OD					
	MCO					
[84] Multi	—	M-CNN	J48	4	IDRiD	99.03%
					EyePACS	99.9%
					Messidor-2	99.75%
[86] Multi	—	Ensemble approach (MPDCNN)	SVM	4	Private	96.2%
					Mixed:	
					(DIARETDB	
					STARE	
					ROC	
[85] Multi	—	CNN	J48	5	Messidor-2	99.75%
					IDRiD	99.03%
					EyePACS	99.99%

Other Neural Networks

Several studies have used different NN architectures other than CNN, either through new NNs or by combining CNN with other NN architectures such as Siamese networks, capsule networks (CapsNets), network-in-network, and transformers.

Zeng et al. [89] proposed a Siamese-like CNN model that accepts two input images representing the left and right eyes to reflect the clinical practice of DR diagnosis. The proposed model was fine-tuned on Inception-V3 for five DR severity gradings. A subtraction operation was applied to each pixel value of the fundus image using the weighted mean of its surrounding pixel values, and 50% grayscale was added. This operation made the blood vessels more explicit. The model determined the correlation between the two eyes and fused them into a fully connected layer for binary classification. The binary classification results on the EyePACS dataset showed an AUC, sensitivity, and specificity of 0.951, 82.2%, and 70.7%, respectively. Unpaired input fundus images were used to evaluate the effectiveness of the binocular model. The results demonstrated that the binocular model may be useful for DR grading.

Capsule Networks (CapsNets) are networks that can retrieve spatial information while preventing data loss that may occur during pooling operations in CNN models. A study [90] emphasized that the pooling layer in CNN loses some relevant information, even though the pooling layer's primary purpose is to reduce computation without sacrificing the data's essence. The model was trained on the Messidor-2 dataset for a five-class task. A Gaussian filter was used during the preprocessing stage to remove noise and CLAHE was used to improve the contrast. The green channel

was extracted owing to its high contrast. The proposed model attained an accuracy of 97.98% on the Messidor-2 dataset.

Another study [91] proposed the DRISTI model, which uses a combination of CapsNets and VGG-16 to classify DR into two- and five-class DR severity. APTOS, IDRiD, DIARETDB1, DIARETDB0, STARE, and DRIVE were used for the two-class task, while the Messidor-2 dataset was used for the five-class task. The two-class task had a testing accuracy of 95.50% on the mixed datasets setting, whereas, for the multi-class classification testing accuracy, the proposed model achieved 75.81% on the Messidor-2 dataset.

To develop a highly nonlinear scale-invariant deep model, Khan et al. [92] proposed the VGG-NiN model by stacking network-in-network (NiN) over the VGG-16 model. Because the NiN mode adds more nonlinearity to the model, it was added to improve the classification. The main goal of this study was to achieve DR classification with the lowest possible learnable parameters to speed up model training and convergence. According to the experimental results, the proposed model achieved the same micro-AUC of 0.95 with only 52% of the learnable parameters.

Matsoukas et al. [93] compared the performance of CNN and ViTs (ViTs) in the medical imaging domain. They used three common medical imaging datasets that cover a wide range of targeted domains, including fundus images, skin lesions, and masses in mammograms. Fundus images were taken from the APTOS dataset. According to the results, ViTs localize lesions such as hemorrhages and exudates that CNNs find difficult to localize. According to the study, ViTs and fine-tuned CNNs have similar performance. Table 3.7 summarizes the non-CNN techniques proposed

for DR grading.

Table 3.7: *A summary of neural network architecture used for DR grading.*

Ref.Task	Pre-processing	Model	GradingDataset	Accuracy
[89] Multi	Augmentation Subtraction operation	Siamese NN5	EyePACS	95.1%
[90] Multi	CLAHE Green channel	CapsNet	4 Messidor-2	97.98%
[92] Multi	Cropping Augmentation	VGG, NiN	5 EyePACS	95%(AUC)
[93] Multi	-	ViT	5 APTOS	0.896 (Kappa)
[91] Multi	Augmentation	DRISTI	5 Messidor-2	75.81%
			Mixed:IDRiD	
			DIARETDB1	
Binary			2 DIARETDB095.50%	
			STARE	
			DRIVE	

3.6 Semi-supervised Learning

Large-scale labeled data, which takes time to process and is prone to human error, is a key component of supervised learning performance for DL models. Medical images differ from other types of images in that more than one medical expert is required to label them to reach a certain level of agreement and confidence. Most reviewed studies are limited to supervised learning for grading DR, where model performance is bounded by the quality of images and the quantity of labeled images. Limited studies have used semi-supervised learning for DR grading using few labeled samples in the form of self-supervised learning which is usually trained on large unlabeled data. Self-supervised learning is considered a sub-type of unsupervised learning. Its primary distinction from unsupervised learning is that it does not cluster unlabeled data. Instead, self-supervised learning transforms an unsupervised problem (with no labeled data) into a supervised problem using a subset of labeled data samples to recognize the remaining samples.

Li et al. [94] used semi-supervised learning to mitigate the need for an accurate fundus image annotation labeled by an expert. They proposed a semi-supervised auto-encoder graph network (SAGN) using semi-supervised learning. The SAGN model consists of three modules: auto-encoder feature learning, neighbor correlation mining, and graph representation. The auto-encoder model learns to extract features from fundus images and reconstruct them into fundus images similar to the original inputs. The radial basis function was calculated to determine the similarities between labeled and unlabeled data using the neighbor correlation. Finally, they developed a graph

convolutional NN (GCN) to assign severity levels to the fundus. The proposed model achieved an accuracy of 94.4% using 30K labeled samples.

Zhang et al. [95] proposed a hybrid graph convolutional network (HGCN) for DR grading using a limited number of labeled samples. This semi-supervised framework includes three essential modules: feature extraction module, modularity-based graph learning module, and hybrid graph convolutional module. CNN was used to extract features that were fed into the modularity-based graph learning module to apply node-to-node learning. As the final representation of fundus images, an HGCN module was created to learn the synthetic features by mixing independent CNN and mutual GCN features. The MESSIDOR dataset was used for training and testing. The results showed that with 20% annotated samples, the HGCN model achieved 89.3%, while with 100% annotation, the results outperformed the state-of-the-art baselines on the MESSIDOR dataset, with 97.9% accuracy, 95.8% sensitivity, and 98.9% specificity.

Luo et al. [96] proposed a self-supervised fuzzy clustering network (SFCN) for classifying normal and abnormal fundus images. The proposed model consists of feature learning from unlabeled fundus images, self-supervision using a fuzzy clustering module, and the application of a fundus image classification task. The first basic component of the SFCN is built using multiple CNN deconvolutional layers (ResNet) to ensure that a sufficient feature representation with appropriate information is learned to construct itself. Then, a self-supervision model was used to exploit the underlying correlation in unlabeled images and transform the network output into fuzzy probabilities for each cluster. Model evaluation was conducted using three datasets, including Messdior, DRIVE, and DIARETDB. The proposed SFCN achieved an ac-

curacy of 87.6%, 81.7%, and 84.7% on the MESSIDOR, DRIVE, and DIARETDB1, respectively.

3.7 Discussion and Observations

In this section, we provide extensive analysis and discussion of the papers surveyed in this review. The DR literature will be analyzed based on the datasets used, proposed techniques, and obtained results. We also further discuss each factor and draw comparisons and conclusions from the selected studies.

3.7.1 DR Grading Datasets

Several datasets have been used in the selected DR grading and classification studies. Most of the datasets used in the literature are publicly available, while few are private. These datasets are used mainly for the proposed models' training and evaluation, for performance comparisons with other models for the same problem of interest.

The datasets differ in the annotation level (pixel or image level), size, and DR severity scale. Most existing DR datasets include only image-level grading labels with minimal pixel-level lesion-based annotations. All the dataset fundus images were captured using high-quality desktop cameras with a 45°–50° FOV. Additionally, many datasets are unbalanced and skewed, with the normal DR class dominating.

Publicly available datasets include a varying number of samples. The largest dataset, EyePACS [36], contains approximately 88,704 DR severity grading samples representing the five DR stages, whereas DIARETDB1 [41] contains the fewest samples

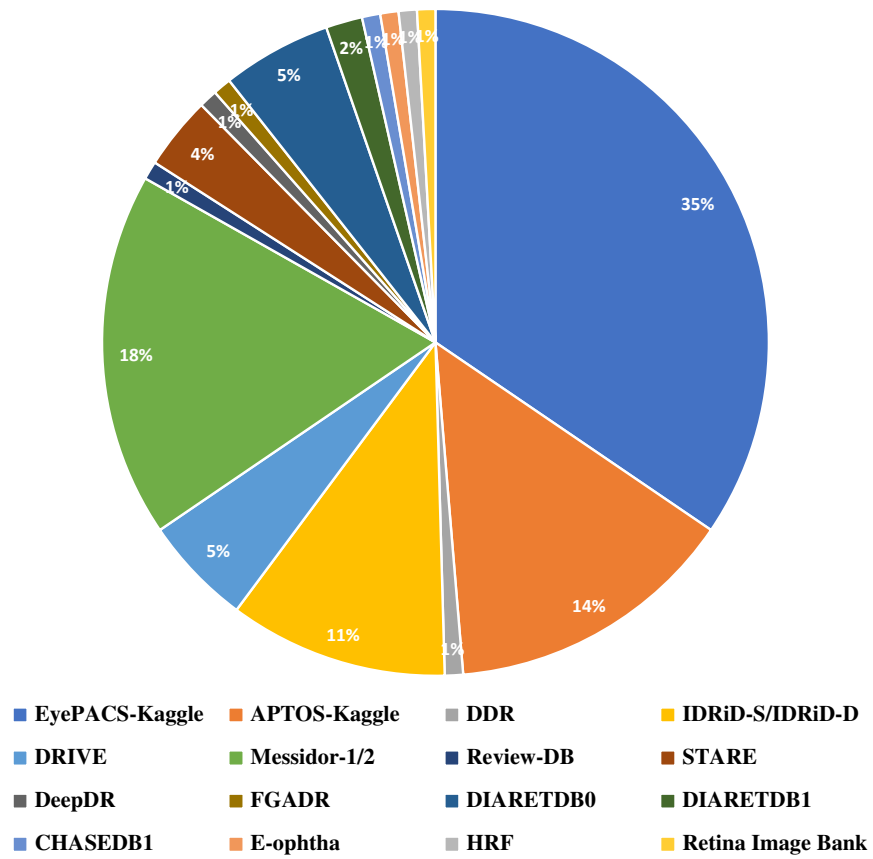


Figure 3.1: *Publicly Available DR Datasets Distribution.*

with five grades. Most DR datasets used in the literature are publicly available; however, few are private and restricted. Three datasets were extensively used in the literature, with 35% of studies using the EyePACS dataset, 15% using the APTOS dataset, and 17% using the Messidor dataset, as shown in Figure 3.1. Other DR datasets have rarely been used in the past five years. Several researchers used private datasets collected from local hospitals for developing models [48], [67], [43], [56], [63], [64], [86]. Some of these private datasets are limited in size [48], [43], [64], [86], while others are large [67], [56], [63]. The largest private dataset used in the literature consists of 666,383 fundus images collected from several patients [63].

Some studies combined different samples from several DR datasets to create cohort training samples. In [50], the target was high-quality images from three datasets, EyePACS, Drive, and Messidor, which were selected by physicians. Liu et al. [6] merged three datasets (EyePACS, APTOS, and DeepDR) to create a dataset of 39,988 fundus images. Some studies used a subset of samples from a dataset. For example, in [97] and [51], only 1000 fundus images from EyePACS were used. Other researchers [78] used a dataset to train their models and different datasets for validation. Studies that used privately owned datasets for training also used publicly available datasets for testing and validation, such as [56], which used the STARE dataset for testing.

Moreover, most of the public datasets were collected using high-quality desktop fundus cameras. As far as we know, there is not a publicly available dataset for DR that includes fundus images captured using a portable fundus camera. In medical practice, non-mydratic fundus cameras are preferred over portable fundus cameras because they have better resolution. Conversely, patients in remote places might not

have access to cutting-edge diagnostic tools such as fundus cameras. Therefore, there is a growing demand for data collected using a portable and low-cost camera [98]. This kind of dataset would be beneficial for creating DR grading systems designed for smartphones equipped with portable fundus cameras.

Most publicly available datasets have been collected from Western populations. According to [15], most DL algorithms trained on these datasets fail to highlight some major lesions when tested on fundus images collected from the eastern population. They observed that people from the Middle East or with a dark complexion have more melanocytes in the retina’s outermost layer. A Darker retina hides some vascular alterations than a lighter retina. Limited studies have attempted to collect cohort diversity samples, including light- and dark-colored retinas, to address this issue. Consequently, more focus must be placed on creating a dataset of cohort fundus images.

Most majority the available datasets only contain the early stages of NPDR. As far as we are aware, there is currently no accessible dataset publicly available that contains annotated DR late-stage PDR images along with severity grades and lesion annotations. We found that most public datasets provide image-level labeling, whereas few datasets provide pixel-level annotation. The pixel-level annotation varied between vessels and lesions during the early stages. Several studies have found that fundus image annotation is a time-consuming task requiring industry experts to agree on the annotation to ensure a higher level of confidence.

Patient medical data are another important factor that might affect the diagnosis decision. This information includes genetic factors, diabetes period, type of diabetes

(1 or 2), hemoglobin A1C score, and hypertension. These clinical factors are rarely used in DR classification because this information is unavailable in public datasets. To improve the overall accuracy of DR grading, researchers must understand what happens when DR medical data are combined with fundus images. They also must investigate what happens when different data types are fed into the models.

3.7.2 DR Classification Tasks

Three tasks are usually investigated in DR diagnosis studies, binary DR grading classification, multi-classification, and lesion segmentation. Early studies on DR grading focused on binary classification setup by determining whether the fundus image has DR signs. This task has received more attention, and high classification accuracies have been reported. Conversely, recent studies have focused more on multi-classification and lesion segmentation. Multi-classification aims to classify DR severity into five grades. This task is usually associated with the segmentation of lesions to increase confidence in the models' decisions. As shown in Figure 3.2, 60% of the reviewed studies mainly focused on the multi-classification tasks, while 26% performed both essential tasks, such as binary and multi-classification. Most studies reviewed categorized DR severity into five stages: no-DR, mild, moderate, severe, and PDR. Few studies focused on the four initial stages while excluding PDR [71], [45], [84], [12], [86], [46]. Furthermore, some studies focused on the first three stages [81], [83], [84].

Most studies used binary classification and classified the images as DR or no-DR. However, other studies attempted to use different binary classification setups. These setups not only examine the non-DR versus DR cases as an abnormality; but

also another approved binary setup from the clinical practice such as non-referable and referable. Whereas the non-referable includes grades 0 and 1, while referable includes all grades 2,3 and 4. From the literature, two studies conducted by [55], [77] proposed two binary setups: normal stage 0 vs. the rest of the DR grades from 1 to 4 representing DR. The second setup is non-referable, including normal grade and grade 1, vs. referable, including grade 2 to the highest grade. However, it should be noted that the studies using binary classification produced higher accuracy than those using models trained for multi-classification. This demonstrates the complexity of the problem when it is promoted to a multi-classification level.

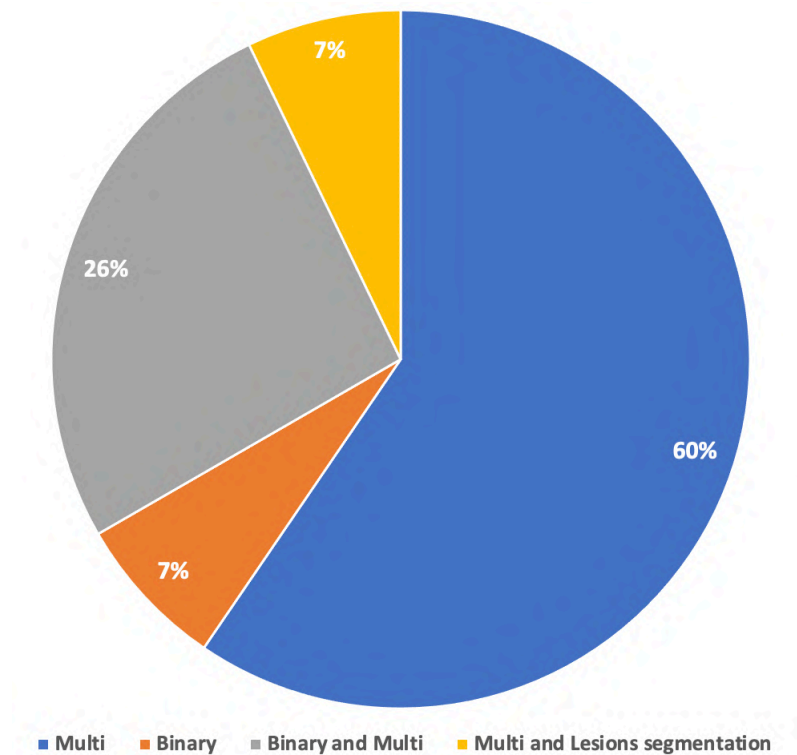


Figure 3.2: *The distribution of DR Classification studies based on the task: binary classification, multi-classification, and lesions segmentation.*

3.7.3 Preprocessing Techniques

Several preprocessing approaches were used to prepare and improve the quality of fundus images prior to feeding them to the proposed models. The preprocessing techniques improve the appearance of lesions and vessels in fundus images, which assists in segmenting and classifying lesions. Researchers have used different DR grading preprocessing techniques, such as color enhancement, median filter, top-hat transformation, Gabor filtering, subtraction operation, and morphological closing operation. Additionally, the surveyed studies have used different data augmentation methods to deal with dataset imbalance and limitation issues [59], [80], [60], [61], [78], [51], [66], [48], [89], [68], [82], [62], [6], [91], [83], [64], [57], [92]. These methods include image flipping, random image rotation with different degrees, shifting, mirroring, shearing, and random zooming.

Several DR grading approaches [68], [45], [46], [49], [50], [90] have used the green channel of the fundus image as an input for their model. The green channel stores more valuable information than the other channels. It has the most light and contrast, allowing for better performance in detecting retinal lesions, such as vessels, microaneurysms, and neovascularization [12]. Harihanth et al. [68] investigated the effect of each RGB channel on the model's performance. The ensemble model trained on the green channel outperformed the ensemble models trained on the red and blue channels. The reported accuracy with an ensemble for the green channel was 81.85% compared with 74.98% and 73.86% for the red and blue channels, respectively, with the EyePACS dataset.

To prepare fundus images for the following stages, some studies applied noise removal and contrast enhancement. Several approaches used the Gaussian filter to remove noise from fundus images [50], [62], [69], [70], [71], [74], [81]. To adjust the brightness of the fundus images, several DR grading techniques enhanced the contrast in the preprocessing stage [90], [50], [53], [54], [71], [83], [67]. Contrast enhancement improves the image’s appearance and reveals some hidden features that the model can use later. HE, adaptive HE, and CLAHE are the most used contrast stretching techniques in the surveyed studies. CLAHE is a computer image processing technique used to enhance the contrast of an image. It is particularly useful for improving the visibility of details in images especially in medical images and various other applications where local contrast enhancement is required without introducing unwanted side effects. In contrast, HE enhances overall contrast, adaptive HE improves local contrast, and CLAHE maintains local contrast while preventing excessive enhancement

3.7.4 Learning Techniques

The diagnosis of DR was largely carried out using DL techniques. According to the review, supervised learning-based techniques were used in 95% of DR diagnosis studies, and semi-supervised or self-supervised learning-based techniques were used in 5% of the studies. The survey clearly shows that recent DL techniques that require small, annotated datasets, such as self-supervised and active learning techniques, have not been explored deeply in the literature.

Figure 3.3 shows the distribution of supervised learning-based techniques used for DR grading. Most proposed techniques used pre-trained models, whereas approxi-

mately 19% of the surveyed techniques developed end-to-end CNN models for DR grading. These models were fully trained on DR datasets without using any pre-trained models. The remaining studies either used models other than CNN for DR classification, such as Transformers, or they used pre-trained models for extracting low-level features from the fundus images, which were then fed into the ML method for classification. This distribution demonstrates that non-CNN models have not been extensively investigated for DR grading.

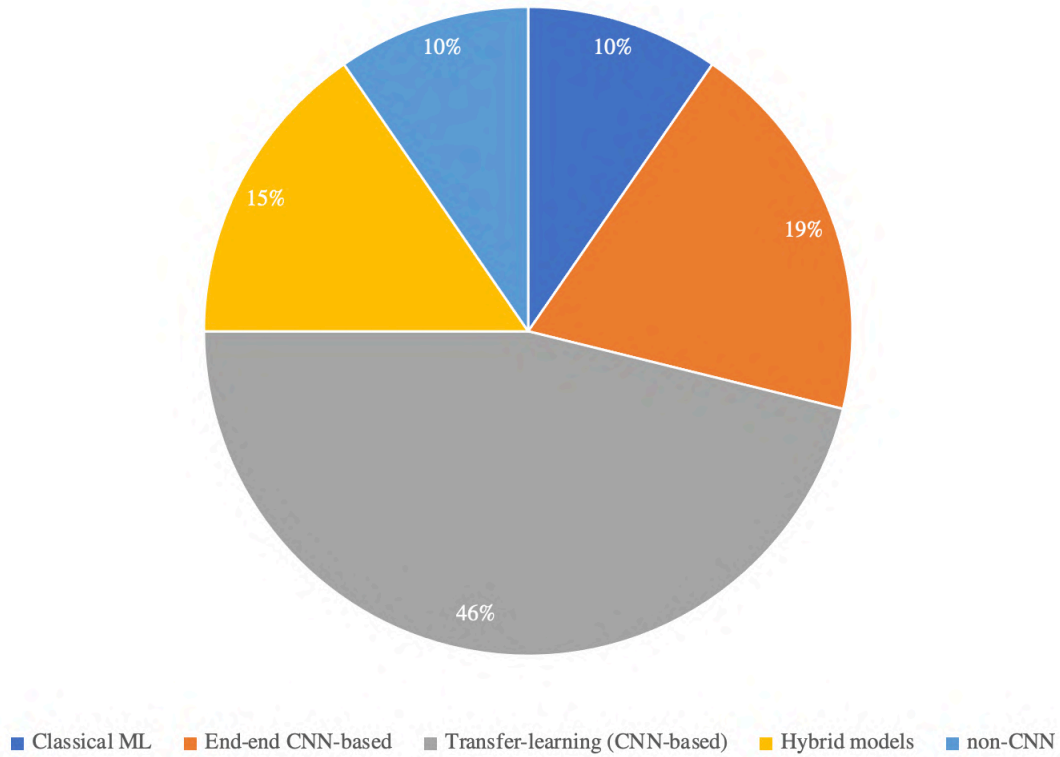


Figure 3.3: *The distribution of the supervised learning techniques proposed for DR grading.*

The use of TL has been identified as a promising solution for the lack of annotated DR fundus images. Thus, when comparing the performance of end-to-end CNN models trained entirely from scratch on fundus images with the same architecture but with

pre-trained models, the latter outperforms better [48]. Note that the pre-trained models used in the existing DR literature were primarily trained on the ImageNet dataset. This can be attributed to the scarcity of models trained specifically on medical images and the discrepancy in the quantity of available medical images versus those within the ImageNet dataset.

Few studies [61], [64] combined the strength of the pre-trained model as the backbone architecture and other model architectures to achieve more accurate results. In [61], the ResNet model was used as a backbone that works with the attention network, and the classification was performed using the Bilinear network. Similarly, [64] used ResNet as the backbone of their architecture and combined it with a twofold feature augmentation mechanism to develop a model that used two image modalities.

Several studies have used CNN models as feature extractors and ML classifiers for classifying fundus images. The overall results, as shown in Table 3.6, indicate that performance has improved with different datasets. The highest accuracy was obtained when the CNN and J48 classifiers were combined [85], [84]. SVM also performed well using three DR gradings [83] compared with five DR grades [82]. However, the J48 classifier outperformed the SVM classifier when combined with CNN using the same dataset. In general, the performance with these approaches was degraded when evaluated on grading DR with five stages compared with three stages.

Studies that used DL networks other than CNN achieved almost the same performance as CNN models. Among these architectures, CNN achieved the highest accuracy at 97.98% [90] compared with DRISTI on the same dataset (Messidor-2) [91]. However, transformer architecture with an attention mechanism can also en-

hance DR classification performance. The performance of ResNet-50 combined with self-supervised achieved a kappa score of 0.894, which was consistent with the previous results [93]. Similarly, ViT with self-supervised achieved a kappa score of 0.896. This indicates that CNN could replace ViT and benefit from ViT’s ability to localize lesions better than CNN.

Existing DL retinal image diagnosis techniques require large, labeled data to oversee network learning. An extensive team of knowledgeable professionals is required to manually annotate many fundus images to label the DR lesions. The lack of such expert annotators can cause inconsistent annotation, particularly with some tiny DR lesions. Therefore, to address the lack of annotated datasets, several studies have used semi-supervised learning techniques for DR grading [94], [95], [96], [99]. Despite using fewer examples, the performance of semi-supervised learning-based techniques is comparable to that of supervised learning. The results indicate it is possible to reduce the manpower and time required for labeling fundus images while maintaining the same level of performance as supervised learning for DR grading. Moreover, more labeled samples may result in better performance [95].

3.7.5 Performance Measures

The surveyed studies have used a range of performance metrics (accuracy, specificity, sensitivity, precision, recall, and the quadratic kappa score) to evaluate the performance of the proposed DR grading approaches 3.8. Although several studies reported their results using one metric, most reviewed studies used more than one metric. These metrics have been used to evaluate the performance of different ML DR grad-

ing models. Accuracy is a measure of the proportion of correct predictions made by the model, whereas specificity measures the proportion of true negatives correctly identified. Sensitivity, also known as recall, measures the proportion of true positives correctly identified. Precision is a measure of the proportion of true positives among all positive predictions made by the model. The recall is similar to sensitivity, but it can also be used to refer to the proportion of relevant items retrieved by the model. Finally, the quadratic weighted kappa is a metric that gauges the agreement between predicted and actual labels, taking into consideration the potential for chance agreement.

These metrics have been used to evaluate the performance of different machine-learning models in DR grading. The accuracy is a measure of the proportion of correct predictions made by the model, while the specificity measures the proportion of true negatives that were correctly identified. Sensitivity, also known as recall, measures the proportion of true positives that were correctly identified. Precision is a measure of the proportion of true positives among all positive predictions made by the model. The recall is the same as sensitivity, but it can also be used to refer to the proportion of relevant items that were retrieved by the model. Finally, the quadratic weighted kappa is a measure of agreement between the predicted and true labels, which takes into account the possibility of agreement occurring by chance.

Table 3.8: *Metrics for DR grading evaluation. True positives (TP), true negatives (TN), false positives (FP), false negatives (FN), P_0 is the observed agreement and P_e is the expected agreement*

Metric	Formula
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Specificity	$\frac{TN}{TN+FP}$
Sensitivity	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1-score	$\frac{2*Precision*Recall}{Precision+Recall}$
Kappa-score	$\frac{P_0+P_e}{1-P_e}$

Accuracy is usually used with balanced data, whereas other metrics are used with unbalanced datasets. Several studies in the literature have addressed unbalanced data using some image augmentation techniques to handle the unbalanced samples per DR grade, [79], [60], [66], [48], [89], [61], [65], [78], [51], [68], [82], [100], [101], [102], [103], [62], [53], [104], [6], [83], [97], [105], [92], [54]. As observed, 24 studies explicitly reported the issue, with 92% of them using image augmentation to resolve it, while 8% used up-down sampling techniques. Contrary to most studies that used data augmentation to address the imbalance issue, [65] addressed imbalance data by allocating gradient weights for each DR class during backpropagation. The study

discovered that this method handled the class imbalance issue much better than the other weighted methods.

3.8 Research Gaps and Future Directions

In this section, we aim to list number of potential gaps by analysing the results of this conducted review. The surveyed studies demonstrate that the medical industry is adopting technological advances to improve DR diagnosis services in healthcare. However, the proposed solutions must be improved in several aspects, such as preprocessing, DR grading, lesion detection, and data availability. Several attempts have been made to attain acceptable performance of healthcare systems that can deliver DR diagnosis services more efficiently and precisely that meet medical standards.

Several research gaps were identified in the DR grading studies, which emphasized the flaws in the currently available techniques. The reported gaps are divided into two categories: data-centric and model-centric. The data-centric category discusses potential research gaps related to the data wrapped in six gaps, whereas the model-centric category discusses gaps related to the techniques proposed for DR severity grading and wrapped in nine gaps. Additionally, we recommend certain areas where greater contributions can be made in DR grading.

3.8.1 Data-centric Gaps

1. **Portable fundus Camera:** Most datasets used in the surveyed studies were gathered using desktop fundus cameras. We are unaware of any dataset that

contains fundus images captured using a portable fundus camera. The desktop fundus camera is expensive and cumbersome to transport. Conversely, portable cameras are cheaper and more convenient for retinal imaging in remote areas or home use. Additionally, portable fundus cameras provide fundus images with resolution and quality different from desktop cameras. This should be a focus area for future research.

2. **Fundus images views:** Most publicly available fundus images are captured as single-view images, which restricts the amount of information that deep learning models can extract. To address this limitation, efforts have been made to gather dual-view fundus images from the same patients. Expanding the availability of dual-view fundus images would significantly enhance research in diabetic retinopathy (DR) grading.
3. **Quality assessments:** Most publicly available datasets do not apply image quality assessments to the collected data. Because model performance is associated with image quality, some studies automated image filtering to select only high-quality images, whereas others used experts to evaluate the images and select the ones with high quality. Hence, creating a DR dataset with high-quality samples is required.
4. **Datasets Diversity:** An additional gap to consider is the availability of public datasets collected from one country or race. Models that were trained on lighter-colored retinas failed when tested on darker-colored retinas [15]. To address this issue, few studies have attempted to collect cohort diversity samples, including

light- and dark-colored retinas. Therefore, there is no DR dataset collected from patients with different colored skin and races.

5. **DR late stages:** All publicly available datasets contain only the early stages of NPDR with its lesions. However, there is no publicly available dataset annotated for late PDR with its severity grades and lesions.
6. **Medical records:** To diagnose the severity grade of DR, ophthalmologists typically combine the patient’s medical history with fundus images in medical practice. This information can contain important clinical data, such as genetic factors, diabetes period, diabetes type, and hemoglobin A1C score. However, all publicly available datasets do not contain patient medical records. This information can be used to improve the performance of DL grading models.

3.8.2 Model-centric Gaps

1. **Deep learning Explainability:** The acceptance of AI models in clinical practice is still low owing to the black-box phenomenon of these models. Thus, the DL models lack interpretability, causing the DR grading systems to be less trusted in hospitals. Visualizing how an NN learns to identify different features in images provides a deeper understanding of how the model performs. Additionally, it is important to understand why the model may be incorrectly classifying some images, which might improve the accuracy and precision of the model. Thus, further research is required to reveal the black-box phenomenon and explain how the model works to identify different features present in images

that assist the model in decision-making.

2. **Modality:** Another area that deserves further exploration is the integration of patients' medical records with fundus images for DR detection. Current research predominantly focuses on single-modality approaches, relying solely on fundus images, but there is untapped potential in a multi-modal approach that combines these images with medical records, such as the duration of diabetes. This integration could help prevent misclassification between closely related stages of DR by incorporating the patient's clinical history. Additionally, the use of generative AI has not yet been explored in this context, representing a further opportunity for innovation in this domain.
3. **Deep Learning Architectures:** A promising research area lies in exploring and developing novel deep learning architectures beyond the commonly used CNN. The majority of current research has concentrated on CNN-based models for medical imaging tasks, including The Dataset of Diabetic Retinopathy (DDR) classification. However, there is growing interest in alternative architectures, particularly attention-based models and transformers, which have shown potential in various domains. These newer approaches could revolutionize medical imaging by providing more accurate and robust DR classification, potentially surpassing the performance of traditional CNNs [93]. Further investigation into the application of transformers in medical imaging could lead to significant advancements in the field.
4. **Transfer Learning in Medical Diagnosis:** A significant research gap iden-

tified in the literature review is the prevalent use of transfer learning with pre-trained models on ImageNet for medical diagnosis tasks, particularly for classification fundus images. The inherent characteristics of fundus images differ substantially from the public images in ImageNet, raising concerns about the effectiveness and appropriateness of this approach. In the LR, only one study [106] is identified, which proposed medical domain pretrained model on million publicly available medical images, covering histology, ophthalmology, and other fields, showing the potential of medical pretrained models. This suggests a need for further investigation into the efficacy of transfer learning from ImageNet compared to in-domain transfer learning, where models are pre-trained on medical images more closely related to the target domain. Additionally, few studies have explored the potential of self-supervised learning in this context, indicating another area ripe for further research and development.

5. **Feature Extraction:** The grading of the DR problem depends on the presence of lesions that represent each grade. Thus, the primary issue with retinal images is that the DR lesions are tiny and difficult to detect. According to [90], the architecture of CNN has some issues in DR grading, such as losing some information about the images with pooling layers. To address this issue and to retain as many features as possible, some studies [53], [63], [64], [45] have attempted to maintain the image quality before feeding it to the model. Other studies [61], [54] proposed a mechanism for highlighting the relevant features during model training. Therefore, models that can recognize these tiny lesions

effectively and extract the features during the training must be developed. This could be handled by combining computer vision techniques and DL for feature extraction.

6. **Data Sizes:** DL models require large data to represent each DR grade. However, obtaining super-labeled data can be challenging. In medical imaging, this challenge is driven by patient privacy issues and finding experts for the annotation process. This limitation can be addressed with semi-supervised learning techniques using a few labeled samples to label unlabeled samples. Another promising paradigm that has significant attention in computer vision is self-supervised learning, where the model is trained on large unlabeled data to leverage the representation. Then, transfer the knowledge to the target task in supervised learning with limited labeled data. Another promising paradigm is using the few-shot learning technique, an ML paradigm that can be generalized to a new category not encountered in training by providing only a few examples of a new category not seen in training. The discussion and analysis of this review show that manual labeling of fundus images is a time-consuming and expensive task. Weakly supervised learning is considered another direction to work with small data. This method alleviates the burden of manually labeling samples, which is expensive in medical practice. Moreover, achieving high performance over smaller training samples indicates the strength of a classifier and speed of convergence [44].

7. **PDR grading:** According to the conducted review, grading NPDR is the most

explored task compared with PDR. Few studies have focused on DR detection in its late severity levels (grades 3 and 4). This can be attributed to difficulties associated with detecting lesions at these levels and the lack of large, annotated datasets with PDR levels [86]. Most surveyed studies approached DR grading as a pure computer vision classification problem. Researchers used models proposed in the literature for other computer vision problems to grade DR at the image level. Few researchers have used pixel-level detection to support image-level classification. Fundus images can contain multiple lesion types, and some lesions may be shared between DR grades. Therefore, DR severity-level detection using lesion types rather than image-level classification is more accurate.

8. **DR adjacent grades:** Similar to DR grades 2 and 3, the differences between grades 0 and 1 are small. Thus, grades 0 and 1 features extracted using some models look similar [6], which causes misclassification. This is the same for grades 3 and 4. Therefore, a model that can distinguish grades 0 and 1 and grades 3 and 4 by extracting useful features representing different lesions must be developed. One possible research area could be the use of novel approaches such as most reviewed methods have focused on CNN models, and few studies have used other networks, such as attention-based and transformers. Transformers may pave the way for a new era in medical imaging, which may outperform CNN in DR classification [93].
9. **Lesions colors:** Red and white lesions might overlap with the structure of the retina (OD and vessels) [12], causing false-positive results of the OD and blood

vessels. The red lesions are located close to the blood vessels, which makes the distinction between them complex. This issue has not been extensively investigated.

3.9 Chapter Summary

This chapter presents a DR review that comprehensively investigates the prior contributions made to the automated diagnosis of DR by other studies. The utilization of artificial intelligence technologies and ML algorithms in the medical field has improved diagnostic applications for DR. Through this review, we presented an in-depth analysis and highlighted gaps and challenges of several ML and DL techniques proposed for DR grading. The proposed models were evaluated using benchmark datasets, classification tasks, preprocessing methods, learning techniques, and performance measures. We comprehensively analyzed 18 publicly available DR datasets, discussing research gaps and possible future works. Furthermore, we thoroughly analyzed different learning techniques and identified several gaps the research community could fill. Additionally, we examined the potential and difficulties of diagnosing DR to identify areas where researchers might enhance their methods. The suggested research possibilities can be pursued using either data- or model-centric approaches.

This thesis addresses five critical gaps identified in the literature related to model-centric challenges in diabetic retinopathy (DR) grading:

Gap-3: Deep Learning Architectures: While most current research focuses on convolutional neural networks (CNNs), this thesis investigates the potential of alter-

native architectures, particularly Vision Transformers (ViTs), which may offer more accurate and robust DR classification. This exploration into newer, attention-based models could lead to significant advancements in medical imaging.

Gap-4: Transfer Learning in Medical Diagnosis: The effectiveness of transfer learning using models pre-trained on ImageNet for medical diagnosis, particularly for DR, is questionable due to the differences between ImageNet images and medical images. This thesis aims to assess the efficacy of in-domain transfer learning, particularly through self-supervised learning (SSL), where models are pre-trained on medical images more relevant to the target domain.

Gap-5: Feature Extraction: DR grading relies heavily on the detection of small lesions in retinal images, which are challenging to detect. CNN architectures, which often lose important details due to pooling layers, may not be fully adequate. This thesis will explore methods to retain critical image features and enhance the detection of these tiny lesions, combining computer vision techniques with deep learning for more effective feature extraction.

Gap-6: Data Sizes: The training of DL models for DR grading requires large amounts of annotated data, which is difficult to obtain due to privacy concerns and the scarcity of expert annotations. This thesis will leverage self-supervised learning (SSL) to train models on limited amounts of unlabeled data and then fine-tune them on small labeled data. This involves constructing representative datasets with adequate samples and beneficial for SSL learning. Another direction of research is to leverage the potential of semi-supervised and few-shot learning techniques could be explored to mitigate the challenges of data scarcity in medical imaging.

Gap-8: DR adjacent grades: Due to the subtle differences between adjacent diabetic retinopathy (DR) grades (e.g., grades 0 and 1, and grades 3 and 4), models often struggle with misclassification, as features extracted from these grades appear similar. To address this, a model capable of distinguishing these grades by capturing distinct lesion features is needed. Most existing methods rely on CNNs, but exploring novel approaches, such as attention-based networks and transformers, may offer superior performance and usher in advancements in DR classification.

These gaps will be systematically addressed through the research questions and objectives outlined in Chapter 1, specifically in Sections 1.2 and 1.3. The research questions are designed to directly target the existing limitations in diabetic retinopathy (DR) severity grading, particularly with respect to feature extraction, model generalization, and domain adaptation. By formulating objectives that focus on improving the granularity of feature representation and optimizing self-supervised learning frameworks for DR classification, this study aims to close these gaps. Each research question is mapped to a specific challenge in current DR grading systems, ensuring that the proposed solutions are not only theoretically sound but also practically implementable. Through this structured approach, we aim to contribute new insights and methodologies that advance the field.

CHAPTER 4

DATASET CONSTRUCTION OF DUAL-VIEW FUNDUS IMAGES

4.1 Introduction

In this Chapter, we answer the research question of the thesis: “**RQ2**: Which publicly available datasets for Diabetic Retinopathy are most suitable for our research, and how do they align with our requirements in terms of availability, relevance, and applicability?”. We discuss the limitations of existing DR datasets in Chapter 4, and in this chapter we highlighting the need for a specialized dataset tailored to the self-supervised learning to extract more representative representation that can be transferred to the downstream task. Thus, we explore the publicly available datasets to be used for pretraining and evaluation in SSL framework. For the pretraining, we constructed ODM, a dual-view dataset designed for SSL learning. This chapter help to bridge the **Gap-6** listed in Chapter 3(Section 3.8),

Our contributions in creating this dataset are as follows:

1. We have constructed and publicly released ODM. This dataset comprises 5,089 dual-view fundus images with different variants serves the research community.
2. The constructed dataset is a multi-view dataset, with each variant has its own purpose. This dataset variants provides more opportunities to study the different between single and dual views in RGB color space and green-channel color space only.

4.2 Single and Dual View Fundus Images

Within the current research community, the presence of dual-view DR image datasets is notably scarce. While numerous benchmark datasets are publicly available, it is imperative to highlight that the majority of these resources predominantly provide single-view fundus images. The lack of dual-view datasets hinders the comprehensive analysis and evaluation of DR detection and grading algorithms relying on the combined information from macula-centric and optic disc-centric images. Consequently, researchers encounter limitations in accurately capturing the entirety of the retinal structures and identifying subtle abnormalities that may be obscured in single-view images.

Single-view fundus images primarily focus on either the macula or the optic disc, which limits the scope of the analysis. The macula-centric view is crucial for detecting conditions that affect central vision, while the optic disc-centric view is vital for assessing the optic nerve and detecting conditions such as glaucoma. Without dual-view datasets, it becomes challenging to develop and validate algorithms that need compre-

hensive retinal assessments. Subtle pathologies, which might be missed when only one view is available, could potentially be detected when both macula and optic disc views are analyzed together. This comprehensive view is essential for an accurate diagnosis and staging of DR, as it allows for the detection of microaneurysms, hemorrhages, exudates, and other signs that are critical in different regions of the retina.

There is a pressing need to address this gap by developing and releasing benchmark datasets encompassing dual-view DR images. Such datasets will enable the advancement of robust algorithms and systems for improved DR diagnosis and management. Dual-view datasets can significantly enhance the performance of deep learning models by providing richer contextual information and enabling the models to learn more comprehensive features. This, in turn, can lead to more accurate detection, grading, and monitoring of DR, ultimately improving patient outcomes.

In response to this need, we present two novel publicly available datasets for dual-view: DRTiD [107] and DeepDRiD [108]. These datasets provide paired macula-centric and optic disc-centric images, facilitating the development and evaluation of advanced DR detection algorithms. Additionally, we have constructed a dataset for self-supervised learning, the ODM, derived from the aforementioned datasets. The ODM dataset serves as a valuable resource for training models that can leverage the complementary information from both views, promoting advancements in the field of DR research and enhancing the capabilities of diagnostic systems.

4.2.1 Diabetic Retinopathy Two-Field Image Dataset

The Diabetic Retinopathy Two-Field Image Dataset (DRTiD) [107] is a recent benchmark consisting of 3,100 two-field fundus images captured from various practical screening scenarios, offering a diverse range of high-quality images. DRTiD [107] comprises a pair of macula-centric and optic disc-centric images, capturing the essence of two-field imaging. This dataset provides annotations for DR severity grades and macula and optic disc localization, adding significant value to DR detection research.

4.2.2 Diabetic Retinopathy Grading and Image Quality Estimation Challenge

Diabetic Retinopathy—Grading and Image Quality Estimation Challenge (DeepDRiD) dataset [108] provides regular fundus images for left and right eyes from each patient in dual views. Fundus images are acquired with two centered macula-centered and optic-disc-centered, as shown in Figure 4.1. The total number of collected samples is 2000 fundus images. DeepDRiD provides five grades, ranging from 0 to 4. In addition, ultra-widefield (UWF) images are available. Both dual-views and UWF fundus images are labeled for quality assessment.

4.3 Self-supervised Learning Pre-training Dataset

4.3.1 Optic-disc and Macula-centered Dataset

We created a novel dataset by combining the datasets mentioned above. Our dataset, ODM, which stands for ODM, is designed explicitly for self-supervised learning and differs from the original dataset used for supervised learning. In contrast, combining datasets from different sources acquired from various sites can enrich diverse data structures and representations, which is the basis of learning in SSL. This fusion of diverse data can provide more comprehensive knowledge for the SSL model. Furthermore, it is crucial to provide SSL models with varying difficulty levels of samples, where easy examples can guide the model’s initial learning, while harder ones can push its boundaries and enhance representation learning.

Our constructed ODM dataset comprises 5089 pairs of two-field fundus images, each representing a specific grade. The ODM dataset exhibits imbalances, similar to the original dataset and many benchmark datasets in supervised learning, Figure 4.2. The development of this extensive dataset and its variations addresses a significant gap in the study community and drives future progress in dual-view DR diagnosis research using the self-supervised learning paradigm. The fundus images in Figure 4.1 display several DR samples from the ODM dataset.

Table 4.1: *ODM dataset provides different variants. ODM-G and ODM-S datasets contain dual-view fundus images. OD and MA datasets contain single-view fundus images.*

Dataset	Samples	Description
ODM	5089	Dual-view fundus images in RGB color space
ODM-G	5089	Green-channel of dual-view fundus images
ODM-S	2545	Dual view of fundus images, the number of images as same as OD and MA
OD	2545	Single-view of fundus images where the optic disk is centered
MA	2544	Single-view of fundus images, where the macula is the centered

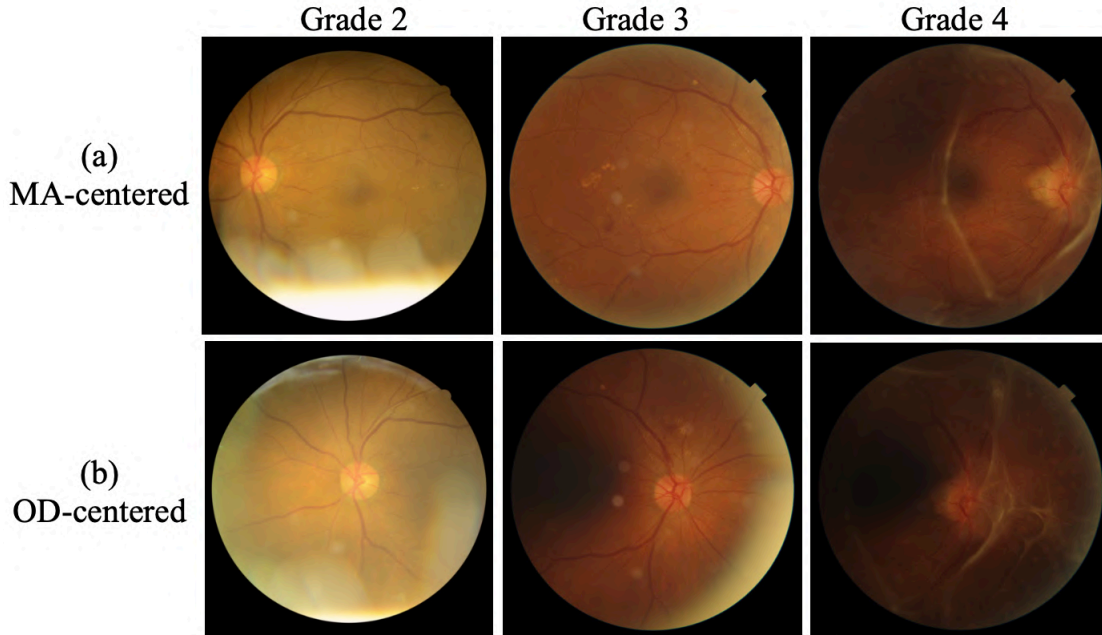


Figure 4.1: *Examples of dual-views of fundus images from ODM dataset. (a) MA-centered DR fundus images for grades 2,3 and 4. (b) OD-centered DR fundus images for grades 2, 3 and 4.*

4.3.2 ODM Variants

We create four versions of the Optic-disc Macula dataset (ODM): Optic-disc (OD), Macula (MA), Optic-disc and Macula Small (ODM-S), and Optic-disc and Macula Green (ODM-G). Table 4.1 describes briefly each variant.

The OD dataset comprises 2545 single-view fundus images obtained with the optic disc (OD) positioned at the center. These images provide critical information about the optic nerve head, which is essential for diagnosing conditions such as glaucoma and optic neuropathy.

A new perspective was created using the MA dataset, which consists of 2544 fundus images collected with a focus on the macula. The macula-centric view is crucial for detecting conditions that affect central vision, such as age-related macular degeneration and diabetic maculopathy. This dataset allows for a detailed examination of the macular region, enhancing the detection of fine retinal details and pathologies localized in the central retina.

ODM-S is a subset of the original ODM dual-view dataset, with the same number of samples as the OD and MA datasets. This subset is specifically curated to emphasize examples of dual views, providing paired images where both the optic disc and macula are visible. This dual-view approach facilitates comprehensive retinal analysis, enabling the development of more robust and accurate DR detection algorithms by leveraging the combined information from both critical regions of the retina.

The last variant, ODM-G, includes dual-view fundus images that solely contain the green channel while maintaining the same number of samples as ODM. The green

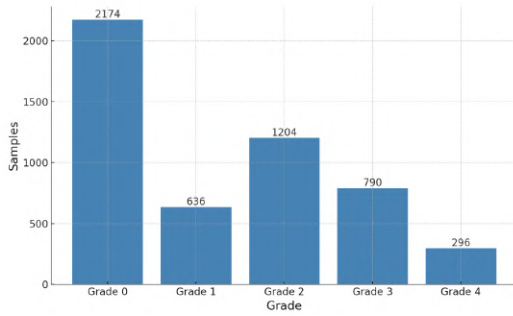
channel is particularly valuable in retinal imaging as it provides higher contrast and better visualization of the retinal vasculature and subtle pathological changes. By focusing on the green channel, ODM-G enhances the visibility of microaneurysms, hemorrhages, and other fine details, thereby improving the sensitivity and specificity of DR detection algorithms.

Together, these four ODM variants offer a comprehensive set of resources for developing and evaluating advanced algorithms for DR diagnosis and management. Each variant caters to different aspects of retinal imaging, providing researchers with the flexibility to explore various approaches and enhance the accuracy of their diagnostic models.

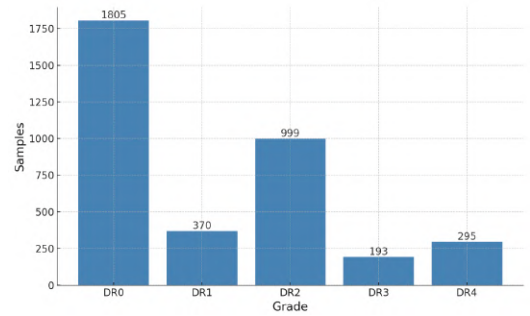
4.4 Evaluation Datasets

This study leverages three well-established evaluation datasets: APTOS2019, Messidor-2, and DDR, to rigorously assess the performance of the proposed models throughout the thesis. These datasets have been selected not only for their widespread use in the field of DR grading and their comprehensive annotations, which are crucial for both training and evaluating deep learning models in medical imaging, but also to provide a range of dataset sizes. This variation in dataset size is important for assessing how the proposed models perform under different data regimes, enabling a thorough evaluation of their robustness and adaptability. By analyzing model performance across datasets of varying sizes, we can better understand how these models handle different amounts of training data and how well they generalize to real-world

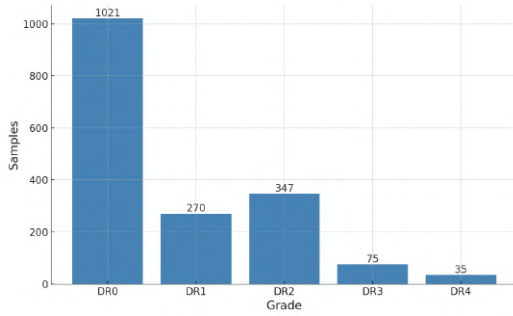
scenarios. Additionally, these datasets are used to explore and assess the knowledge transfer from the pretraining task using unlabeled samples to the downstream task using labeled samples. While the datasets consist of single-view data, dual-view fundus images serve as the source of transferred knowledge learned from ODM. It is worth noting that all the datasets are highly imbalance with the grade zero (No DR) is the prominent class, Figure 4.2.



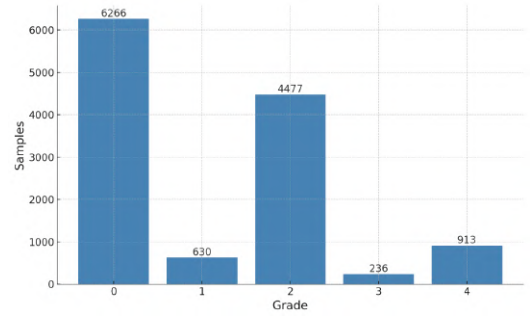
(a) *ODM for self-supervised learning*



(b) *APTOS*



(c) *Messidor*



(d) *DDR*

Figure 4.2: *The distribution of ODM dataset for self-supervised learning, and the three evaluation datasets.*

4.4.1 APTOS

The APTOS dataset is part of the Asia Pacific Tele-Ophthalmology Society challenge, designed to advance the detection and grading of DR [109]. This dataset contains a

large collection of fundus images labeled according to the severity of DR, ranging from no DR to proliferative DR. It provides a diverse set of images, making it an ideal resource for evaluating the generalizability of the models developed in this thesis. The dataset is particularly valuable for testing the ability of the proposed models to handle varying levels of image quality and pathological features.

This dataset provides a total of 5,590 single-view fundus images. The annotations are publicly available only for 3,662 single-view fundus images. Consequently, we randomly split the available annotated fundus images into 70%, 15%, and 15% for training, validation, and testing, respectively.

4.4.2 Messidor-2

Methods to Evaluate Segmentation and Indexing Techniques in the Field of Retinal Ophthalmology (Messidor-2) is a well-known dataset in the ophthalmology and medical imaging communities, containing high-resolution fundus images with annotations for DR severity [37]. It is widely used for benchmarking DR grading algorithms. In this thesis, Messidor-2 plays a critical role in evaluating the final models' performance and comparing them against existing state-of-the-art approaches. The dataset's established reputation in the field ensures that the evaluation results are both reliable and relevant to current research trends. It consists of 1748 single-view fundus images. The fundus images were captured with the macula centered. We randomly split the available annotated fundus images into 70%, 15%, and 15% for training, validation, and testing, respectively.

4.4.3 DDR

The Diabetic Retinopathy Detection (DDR) dataset is a large-scale dataset that includes high-resolution fundus images [33]. It is annotated with detailed DR severity levels, providing a rich resource for evaluating the efficacy of knowledge transfer techniques. The DDR dataset is crucial for assessing the robustness of the models, especially when transferring knowledge from the pretraining task on unlabeled data to the downstream task of DR grading using labeled data. This dataset’s extensive size and diversity help ensure that the models can perform well across different populations and imaging conditions.

DDR contains 13,673 single-view fundus images collected from 147 hospitals in China. All images were classified according to DR severity grading into five classes, with the image quality assessment (ungradable category) as the sixth class. Poor-quality fundus images without clearly displayed lesions are considered ungradable. As we are not considering the image quality, an ungradable class is eliminated, ending with 6,320, 2,503, and 3,759 single-view fundus images for training, validation, and testing. In this study DDR dataset was used for evaluation with supervised learning in Chapter 5, and the SSL baseline proposed in Chapter 6. Then this datasets was included in the pre-training to the enhancement SSL models and our final proposed model in Chapter 7, due to its images diversity to enrich the learning representation along side with ODM dataset.

4.5 Remarks

The ODM dataset, specifically constructed for self-supervised learning (SSL), serves as the primary source of transferred knowledge in this thesis. By combining diverse fundus images from multiple sources, the ODM dataset allows the model to learn rich, generalizable feature representations in an unsupervised manner. The inclusion of dual-view fundus images (optic disc and macula) enhances the model’s ability to capture spatial relationships, providing a deeper understanding of diabetic retinopathy (DR) severity.

Following SSL pretraining on the ODM dataset, the models are evaluated on three well-established datasets: APTOS2019, Messidor-2, and DDR. These datasets, widely used in DR grading research, vary in size and complexity, offering a robust assessment of the models’ performance across different data regimes. This variation helps to gauge the models’ adaptability and generalization capabilities, particularly in real-world clinical settings where data availability is often limited.

These datasets also serve to evaluate the effectiveness of knowledge transfer from the SSL phase (using the unlabeled ODM) to the supervised task of DR grading with labeled data. While the evaluation datasets consist of single-view images, the dual-view knowledge learned from ODM is expected to enhance model performance. Additionally, these datasets exhibit significant class imbalances, particularly with an overrepresentation of grade zero (no DR), as shown in Figure 4.2, providing a realistic challenge for assessing model robustness in handling skewed data distributions.

In the upcoming chapters, we will utilize the ODM dataset to pretrain our proposed

models and subsequently evaluate the transferred knowledge using the APTOS2019, Messidor-2, and DDR datasets. This process will allow us to rigorously assess the effectiveness of the pretraining strategy and its impact on downstream DR diagnosis tasks.

CHAPTER 5

DIABETIC RETINOPATHY GRADING USING SUPERVISED TRANSFER LEARNING

5.1 Introduction

DR is a severe complication of diabetes that can lead to vision impairment and blindness if not accurately diagnosed and treated. The early detection and grading of DR are crucial for effective intervention, making it a significant area of study in medical image analysis. In recent years, deep learning techniques, particularly CNN and ViT, have demonstrated remarkable performance in various image classification tasks, including medical imaging.

This chapter focuses on the application of supervised transfer learning for the grading of Diabetic Retinopathy using CNN-based and ViT-based models. Transfer learn-

ing, which involves leveraging pre-trained models on large datasets like ImageNet[58], has become a powerful method to enhance performance on specific tasks, especially when labeled data is limited [110]. This aligned with our **RQ3**, and fulfill **Gap-3** and **Gap-4**.

This chapter critically examines the limitations of using ImageNet pre-trained models for medical diagnosis tasks, such as DR grading. While these models have achieved success in various computer vision problems, the knowledge they transfer comes from images that differ significantly from the nature of medical images. As a result, the effectiveness of these models in medical contexts is questionable. Through a comparative analysis of transfer learning with ImageNet pre-trained weights. This chapter aims to highlight the challenges of applying these conventional models to DR grading. The findings underscore the need for more specialized approaches, setting the stage for the introduction of a SSL framework with in-domain transfer learning in subsequent chapters, where weights and knowledge are transferred directly from fundus images to enhance diagnostic accuracy.

5.2 Convolutional Neural Networks and Vision Transformer

Convolutional Neural Networks (CNNs) have evolved over the years, with each new architecture addressing limitations in its predecessors and introducing innovations to improve performance and efficiency. The progression from simpler to more advanced architectures reflects an ongoing effort to optimize deep learning models for increas-

ingly complex tasks.

VGG-16, introduced in 2014, was one of the earliest deep CNN architectures that achieved significant breakthroughs in image classification [111]. It simplified the design by stacking deep layers of 3x3 convolutional filters, resulting in a model that extracts hierarchical features as the depth increases. However, VGG-16's main limitation is its computational cost due to the large number of parameters and absence of specialized mechanisms to improve efficiency. The simplicity of its architecture, while effective, posed challenges in training deeper networks, particularly due to vanishing gradients and heavy memory usage.

To address these issues, ResNet (Residual Networks) was introduced, with ResNet-18 being the more shallow variant and ResNet-50 a deeper counterpart [8], [112]. ResNet introduced a key innovation: residual connections, which allow information to bypass layers via identity mappings Figure 5.1. This architecture effectively mitigates the vanishing gradient problem, enabling the training of much deeper networks. The "shortcut connections" in ResNet models allow gradients to propagate more easily during backpropagation, resulting in better convergence in very deep networks. ResNet-50, with 50 layers, significantly increases the model's capacity to learn complex features compared to ResNet-18's 18 layers, but both offer improvements over earlier architectures like VGG in terms of training deep networks and computational efficiency.

Building upon the idea of efficient architecture scaling, Inception-V3 [113] introduced the concept of "Inception modules," which allow for multi-scale feature extraction within a single layer. This architecture reduces computational cost by factorizing

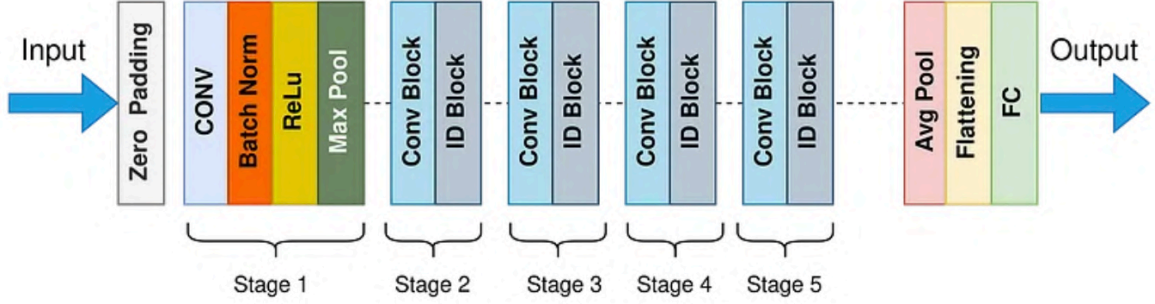


Figure 5.1: *ResNet50 architecture, using skip connections to enable deeper networks and improve performance in tasks like diabetic retinopathy grading. [8]*

convolutions, effectively enabling the network to process both fine and coarse features simultaneously. Inception-V3 addressed the problem of parameter efficiency by using filters of different sizes in parallel, followed by concatenation, which improved the network’s ability to capture diverse feature types without drastically increasing the number of parameters. This advancement allowed deeper networks to extract more complex features, particularly in large and varied datasets.

The most recent advancement is EfficientNet-B5 [72], which takes a novel approach to scaling deep networks through compound scaling. Unlike previous architectures that either increased depth or width, EfficientNet systematically scales depth, width, and resolution in a balanced manner. This model builds on the efficiency innovations of Inception by improving computational performance while maintaining high accuracy. EfficientNet-B5 is particularly effective for large-scale tasks due to its ability to balance the network’s complexity with computational resource demands, offering a solution to the challenge of scaling models without exponentially increasing the number of parameters.

While Convolutional Neural Network (CNN)-based architectures, such as ResNet,

have long dominated the field of image processing, recent advancements have introduced a new approach in deep learning: the Vision Transformer (ViT). The ViT marks a significant change from traditional convolutional approaches, using an architecture originally developed for natural language processing (NLP), the Transformer [114]. The main difference between CNN-based models and ViT models lies in how they handle image data. CNNs, including ResNet, use convolutional layers to scan the image with filters, capturing small, local features like edges, textures, and patterns. As the network goes deeper, these local features are combined, allowing CNNs to effectively model spatial structures and learn detailed features across an image.

In contrast, Vision Transformers do not use convolutions. Instead, they treat an image as a series of fixed-size patches, similar to how words are processed in an NLP sequence. Each image patch is converted into a vector, with positional information added to retain the arrangement of patches in the image. These patches are then passed into a Transformer, which processes them using self-attention mechanisms, Figure 5.2. Unlike CNNs, which rely on scanning small regions of the image, the self-attention in ViTs allows the model to capture relationships between patches, no matter how far apart they are within the image [114]. This global attention helps ViT models capture larger, more complete features from the image, something that CNNs usually struggle with because of their focus on local patterns [114].

Vision Transformer models come in different versions based on their size and the size of the image patches they process, which directly affect the model's capacity and computational requirements. For example, ViT-B-16 and ViT-B-32 are the "Base" models, with the difference being in the patch size: 16x16 patches for ViT-B-16 and

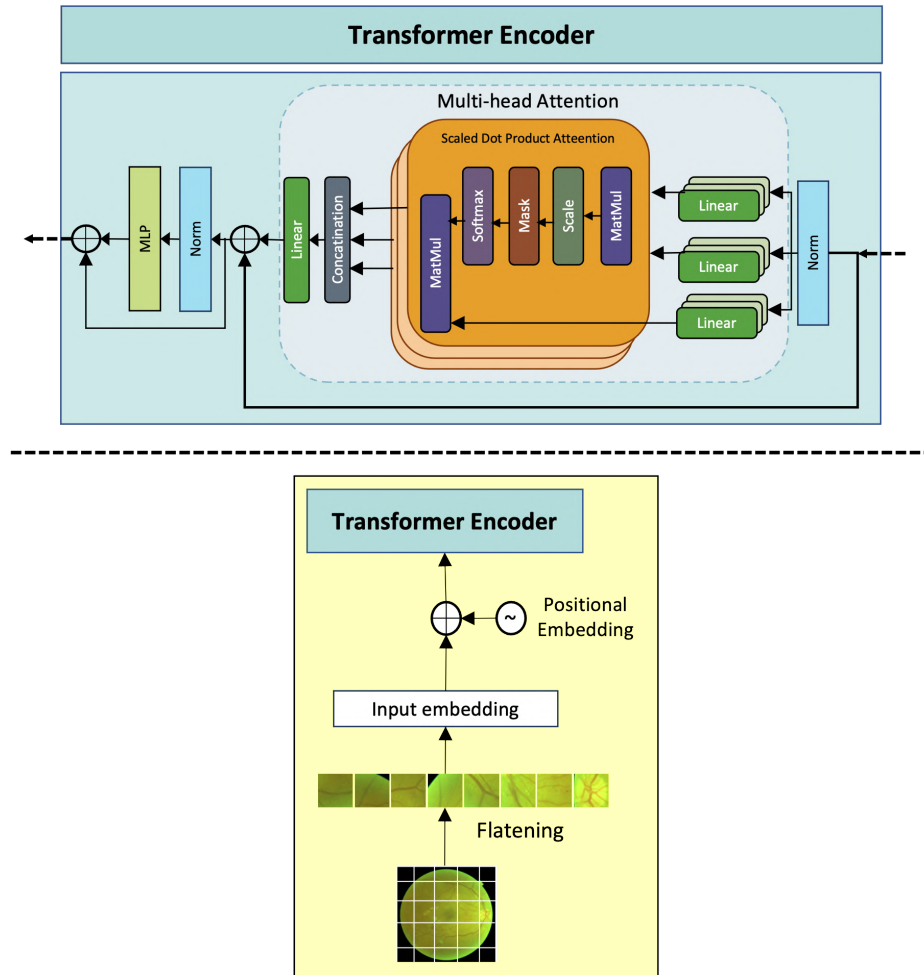


Figure 5.2: *The Vision Transformer architecture. ViT divides images into patches and applies self-attention to capture global context by enabling more comprehensive feature extraction.*

32x32 patches for ViT-B-32. Smaller patch sizes, such as 16x16, allow the model to process more detailed information, while larger patches reduce the computing cost but may lose some detail. Larger models, such as ViT-L-16 and ViT-L-32 (where "L" stands for "Large"), have more layers and greater capacity, which helps them learn more complex patterns from larger datasets. The largest model, ViT-H-14, increases the depth and capacity even further, allowing it to capture more intricate, high-level features.

One of the key advantages of Vision Transformers over CNN-based models is their ability to scale well with large datasets. While CNNs are very effective at detecting local patterns, they have difficulty understanding the overall context of an image, which is important for more complex tasks. ViTs, with their self-attention mechanism, are particularly good at these tasks. However, the global attention in ViTs comes with a trade-off: they generally require much larger datasets and more computational resources to work well. Unlike CNNs, which naturally learn local features, ViTs need to learn both local and global relationships from the data, which makes them less efficient on smaller datasets [115].

5.3 Supervised Transfer Learning in Medical Imaging Diagnosis

Transfer learning is a widely recognized approach that effectively utilizes unlabeled data from either the source domain or the target domain. It is fundamentally a semi-supervised technique that involves domain adaptation [116]. In general terms, transfer

learning can be described as the process of training a model on a large-scale dataset for an initial task and then applying that pre-trained model to learn a new, target task [110]. In most computer vision tasks, the learning process typically leverages models pre-trained on ImageNet [58], the largest publicly available image dataset. These pre-trained models, such as ResNet50, Inception, DenseNet, EfficientNet, and ViT, are fine-tuned to improve performance on specific tasks. However, medical images differ significantly from the general images found in ImageNet, leading to debates regarding the efficacy of these pre-trained models in medical imaging tasks [117], [118], [119].

In supervised learning, pre-training on ImageNet can significantly enhance the performance of models, especially when dealing with large datasets. However, in medical imaging, where data collection and annotation are arduous and time-consuming, the utility of ImageNet pre-trained models is less clear. Although these pre-trained models perform better than training from scratch in low-data regimes, their weights do not always transfer effectively to medical tasks [120], [116], [121].

5.4 Methodology

This section presents the methodology employed to conduct our experiments using CNN and ViT architectures in supervised learning settings for diabetic retinopathy (DR) grading. The primary objective of these experiments is to evaluate the performance of these models when trained for a limited number of epochs across datasets of varying sizes. In particular, we aim to assess how well transfer learning from Im-

ageNet, a large-scale general-purpose dataset, adapts to the domain-specific task of DR grading using medical fundus images. By focusing on short training durations and different dataset scales, we seek to understand the effectiveness of pre-trained CNN and ViT models in extracting relevant features for DR diagnosis. Additionally, we experiment with different model architectures within the same model group to examine how varying the depth and complexity of architectures impacts model performance, providing further insights into their adaptability for DR grading.

5.4.1 Models

We evaluated a diverse set of CNN-based and Vision Transformer (ViT) architectures, all of which were pre-trained on the ImageNet dataset [122]. These models were adapted to our disease classification task using fine-tuning and transfer learning techniques. The following sections provide a detailed overview of the CNN and ViT models utilized in our study, emphasizing the distinctive features and innovations each architecture offers.

CNN-based models: We employed number CNN-based models, each designed to address specific limitations of previous architectures through unique features, Table 5.1. VGG-16, recognized for its straightforward design and reliable performance, uses a consistent structure of multiple convolutional layers with small 3x3 filters, followed by max-pooling layers [111]. ResNet models, particularly ResNet-18 and ResNet-50, introduced skip connections that help overcome the vanishing gradient problem, enabling the network to learn more complex patterns. The key difference between these

models is their depth, with ResNet-18 featuring 18 layers and ResNet-50 featuring 50 layers [8]. EfficientNet-B5 improves performance by balancing the network’s depth, width, and resolution, offering better results but with increased computational demands [72]. Inception-V3, part of Google’s Inception series, employs a complex structure with inception modules that capture a broad range of features at different scales. This model also includes factorized 7x7 convolutions to reduce computational cost and an auxiliary classifier to enhance gradient flow and network regularization [113].

Table 5.1: *Summary of CNN architectures used in the study*

Model	Layers	Param.(M)	Key Features
ResNet-18	18	11.7	Basic residual blocks, efficient for smaller datasets.
ResNet-50	50	25.6	Bottleneck blocks for reduced parameters, good for complex tasks.
EfficientNet-B5	-	30.0	Compound scaling (width, depth, resolution) for high efficiency.
VGG-16	16	138.4	Simple, very deep, but large number of parameters due to fully connected layers.
Inception-V3	48	27.2	Multi-scale feature extraction through Inception modules and auxiliary classifiers.

Transformer-based models: We explored various transformer-based architectures, each tailored to harness the strengths of transformers in computer vision tasks, Table 5.2. The small models, ViT-S-16 and ViT-S-32, utilize patch sizes of 16x16 and 32x32, respectively, and contain around 22 million parameters. The base models, ViT-B-16 and ViT-B-32, utilize patch sizes of 16x16 and 32x32, respectively, and contain around 86 million parameters, showing competitive performance across nu-

merous vision tasks. Larger models, such as ViT-L-16 and ViT-L-32, maintain the same patch sizes but incorporate additional layers and parameters, allowing them to capture more intricate patterns. The ViT-H-14 model, with a patch size of 14x14 and approximately 630 million parameters, represents one of the largest ViTs. All these models build upon the foundational ViT architecture, which includes components like image-to-patch transformation, linear projection, positional encoding, an encoder, and MLP layers [123].

Table 5.2: *Summary of Vision Transformer (ViT) architectures used in the study*

Model	Layers	Hidden	MLP	Heads	Param.(M)	Patch Size
ViT-S (Small)	12	384	1536	6	22.1	16x16
ViT-S (Small)	12	384	1536	6	22.1	32x32
ViT-B/16 (Base)	12	768	3072	12	86.6	16x16
ViT-B/32 (Base)	12	768	3072	12	87.5	32x32
ViT-L/16 (Large)	24	1024	4096	16	307.0	16x16
ViT-L/32 (Large)	24	1024	4096	16	305.5	32x32
ViT-H/14 (Huge)	32	1280	5120	16	632.1	14x14

5.4.2 Datasets

The evaluation of the CNN-based and ViT-based models for Diabetic Retinopathy grading was conducted using three prominent datasets: APTOS2019, DDR, and Messidor-2. These datasets were chosen due to their extensive use in the field and their comprehensive annotations, which are essential for robust model evaluation. The APTOS2019 dataset consists of 3,662 retinal images graded by severity and was part of a global challenge on diabetic retinopathy detection, offering a diverse set of images for thorough evaluation. The DDR (Diabetic Retinopathy Detection) dataset, a large-

scale collection with 13,673 images, provides high-quality retinal images suitable for detailed analysis and model fine-tuning. Messidor-2, a well-established dataset in the medical imaging community, includes 1,748 high-resolution retinal images, which are particularly useful in assessing model generalizability across different imaging standards and modalities. The detailed characteristics and composition of these datasets are thoroughly discussed in Chapter 4. Moreover, the evaluation includes our constructed ODM dataset (Chapter 4) to assess the SL models on a dataset that contains a variety of fundus images views and compare it to the datasets with single-view samples.

5.4.3 Experimental Settings

In this chapter, we conducted experiments on ViT variants with random initialization to compare their performance against CNN-based models from other studies. All models were trained for 25 epochs, using a mini-batch size of 64, on a single NVIDIA RTX 3090 GPU. The training process was optimized using the AdamW optimizer [124], with an initial learning rate of 1×10^{-3} and a weight decay of 0.1. To enhance the training stability and performance, a learning rate warmup [125] was applied during the first 5 epochs, followed by a cosine annealing schedule for the remaining epochs.

5.4.4 Evaluation Metrics

To provide a comprehensive evaluation, we report five key metrics throughout this chapter and the subsequent ones: Kappa, Accuracy, Precision, Recall, and F1-score (Table 3.8).

The quadratic weighted Kappa metric, popularized in the Kaggle DR grading competition and referenced in the APTOS 2019 study [109], is particularly advantageous for handling imbalanced multi-class classification tasks. The κ score ranges from -1 to 1, where a score of 1 represents perfect agreement between the predicted and actual labels, while a score of -1 indicates complete disagreement. The score is calculated as shown in Equation 5.1, where P_0 represents the observed agreement, and P_e is the expected agreement by chance.

$$Kappa\ score(\kappa) = \frac{P_0 - P_e}{1 - P_e} \quad (5.1)$$

$$P_0 = \frac{\sum_{i=1}^r TP_i}{\sum_{i=1}^r (TP_i + FN_i)} \quad (5.2)$$

$$P_e = \frac{\sum_{i=1}^r TP_i + (TP_i + FN_i)}{N^2} \quad (5.3)$$

Here, TP, TN, FP, and FN represent True Positives, True Negatives, False Positives, and False Negatives, respectively, across all r classes, while N is the total number of instances.

It is important to note that DR datasets are often highly imbalanced, with the "normal" grade being the most prevalent [109], [33], [37]. Therefore, following established conventions in previous studies [126], [127], [128], we use the quadratically weighted Kappa score (κ) as our primary evaluation metric, as it better reflects the model's ability to classify all DR grades, not just the majority class.

In such cases, standard metrics like accuracy can be misleading, as they may disproportionately reflect the performance on the majority class, however, we include it to offer insights into the model’s performance, particularly in classifying the majority class.

In addition to the Kappa score and accuracy, we also report Precision, Recall, and F1-score to provide a more detailed evaluation of model performance. These metrics are crucial for understanding how well the model balances its ability to correctly identify positive instances (Recall) and its precision in making those predictions (Precision). The F1-score, which is the harmonic mean of Precision and Recall, offers a balance between these two measures and is particularly useful in imbalanced datasets, where a model might exhibit high precision but low recall, or vice versa. Given the inherent class imbalance in DR datasets, we report the macro average for Precision, Recall, and F1-score. Macro averaging computes the metric independently for each class and then takes the unweighted mean. This ensures that all classes, including minority classes, are given equal importance in the evaluation, which is critical for imbalanced tasks like DR grading. By using macro averages, we avoid the dominance of the majority class in these metrics, thereby providing a fairer and more holistic evaluation of the model’s performance across all DR grades.

5.5 Results and Discussion

This section presents the results of pre-trained models including CNN and ViT models applied to the APTOS, Messidor, and DDR datasets. The performance is eval-

uated across multiple metrics, and a detailed comparison of the models' capabilities in handling different datasets is provided. Additionally, we analyze the class-wise performance through a confusion matrix to understand how these models perform, particularly with the later stages of DR, offering deeper insights into their ability to distinguish severe cases from earlier stages.

5.5.1 Transfer Learning Performance of CNN Models

We analyze the performance of five Convolutional Neural Networks (CNNs) applied to four datasets: ODM, APTOS, Messidor-2, and DDR . The models, ResNet-18, ResNet-50, VGG-16, EfficientNet-B5, and InceptionV3—were pre-trained on ImageNet and fine-tuned for the task of Diabetic Retinopathy (DR) grading. All models were trained for 25 epochs to evaluate the effectiveness of transfer learning from ImageNet to medical imaging, particularly for DR classification. The primary objective of this analysis is to assess how well pre-trained models adapt to a medical context and whether transfer learning can mitigate the challenges associated with training on limited datasets.

The CNN architectures demonstrate moderate performance on the ODM dataset, with EfficientNet-B5 achieving the highest κ value of 0.4130 and an accuracy of 60.82%. ResNet-18 and ResNet-50 show slightly lower kappa values (0.3846 and 0.3973, respectively) and accuracies below 60%. Among the CNNs, VGG-16 and Inception-V3 perform the worst, with VGG-16 having a particularly low kappa of 0.3178 and accuracy of 50.68%. This suggests that while CNNs can provide reasonable performance, their ability to extract fine-grained features and achieve higher

Table 5.3: *Performance comparison of CNN and ViT architectures on the **ODM** dataset across multiple evaluation metrics.*

Architecture	Model	κ	Acc.	Precision	Recall	F1
CNN	ResNet-18	0.3846	57.31%	56.67%	47.10%	47.89%
	ResNet-50	0.3973	59.26%	56.42%	47.59%	48.90%
	EfficientNet-B5	0.4130	60.82%	60.26%	47.44%	49.27%
	VGG-16	0.3178	50.68%	49.30%	42.27%	40.46%
	Inception-V3	0.3770	58.48%	58.20%	42.08%	41.91%
ViT	ViT-S-16	0.8315	67.06%	65.98%	57.18%	58.26%
	ViT-S-32	0.7993	59.06%	58.37%	52.51%	52.89%
	ViT-B-16	0.3880	59.45%	64.44%	42.20%	42.18%
	ViT-B-32	0.3831	59.45%	66.17%	42.60%	41.94%
	ViT-L-16	0.3937	60.04%	67.51%	45.68%	47.05%
	ViT-L-32	0.3117	55.75%	58.42%	39.03%	38.58%
	ViT-H-14	0.3977	60.04%	64.07%	43.73%	44.80%

Table 5.4: *Performance comparison of CNN and ViT architectures on the **APTOS** dataset across multiple evaluation metrics.*

Architecture	Model	κ	Acc.	Precision	Recall	F1
CNN	ResNet-18	0.5899	74.80%	66.01%	42.36%	41.27%
	ResNet-50	0.6478	77.78%	63.80%	51.46%	53.87%
	EfficientNet-B5	0.6553	78.05%	55.99%	49.74%	50.19%
	VGG-16	0.5513	72.63%	28.52%	37.40%	32.03%
	Inception-V3	0.5813	74.53%	49.05%	39.64%	35.70%
ViT	ViT-S-16	0.7321	65.31%	51.50%	35.36%	32.19%
	ViT-S-32	0.7465	65.58%	63.89%	36.42%	34.12%
	ViT-B-16	0.5637	73.17%	52.42%	44.20%	44.93%
	ViT-B-32	0.6222	76.15%	53.42%	49.33%	49.99%
	ViT-L-16	0.6161	74.80%	52.24%	55.52%	52.64%
	ViT-L-32	0.6249	76.42%	53.91%	48.46%	48.91%
	ViT-H-14	0.5959	75.07%	55.63%	45.80%	46.51%

reliability, as indicated by the kappa metric, is limited on the ODM dataset.

On the APTOS dataset, the performance varied across models, with EfficientNet-B5 achieving the best results, with a Kappa score of 0.6553 and an accuracy of 78.05%.

ResNet-50 followed closely with a Kappa of 0.6478 and an accuracy of 77.78%, indi-

cating that deeper architectures perform well in this context. ResNet-18 showed a lower Kappa score of 0.5899 and an accuracy of 74.80%, while InceptionV3 performed slightly better than VGG-16, achieving a Kappa score of 0.5813 and an accuracy of 74.53%. VGG-16, on the other hand, showed the weakest performance among the models, with a Kappa score of 0.5513 and an accuracy of 72.63%. These results suggest that even with a moderate dataset size like APTOS, models can achieve a reasonable level of generalization after 25 epochs of training. However, the relatively lower performance of some models, particularly VGG-16, indicates that further fine-tuning or additional training may be necessary for high-stakes medical tasks such as DR grading.

Table 5.5: *Performance comparison of CNN and ViT architectures on the **Messidor** dataset across multiple evaluation metrics.*

Architecture	Model	κ	Acc.	Precision	Recall	F1
CNN	ResNet-18	0.2544	58.43%	40.77%	29.37%	27.61%
	ResNet-50	0.2776	62.92%	42.69%	32.70%	34.27%
	VGG-16	0.1412	57.87%	47.75%	28.25%	27.41%
	EfficientNet-B5	0.2572	63.48%	30.41%	30.53%	29.37%
	InceptionV3	0.0692	59.55%	35.24%	25.56%	23.25%
ViT	ViT-S-16	0.1757	33.14%	51.42%	24.03%	15.51%
	ViT-S-32	0.1891	39.32%	47.49%	25.06%	17.69%
	ViT-B-16	0.0099	57.87%	21.59%	20.36%	15.68%
	ViT-B-32	0.0441	58.43%	19.86%	20.92%	16.79%
	ViT-L-16	0.1147	60.11%	30.02%	23.13%	20.85%
	ViT-L-32	0.0183	58.43%	31.64%	20.56%	15.80%
	ViT-H-14	0.1357	61.24%	32.41%	25.28%	23.66%

In contrast, the Messidor dataset, which has a significantly smaller sample size, yielded considerably lower performance across most models, as shown in Table 5.5.

EfficientNet-B5 once again outperformed other models, achieving a Kappa score of

0.2572 and an accuracy of 63.48%, though these scores are still much lower than those obtained on the APTOS dataset. ResNet-50 achieved a Kappa score of 0.2776 and an accuracy of 62.92%, reflecting the challenges that arise when applying transfer learning from large-scale datasets like ImageNet to smaller, domain-specific datasets such as Messidor. This decline in performance underscores the limitations of transfer learning when applied to medical imaging tasks with limited data. Even the ViT models, which generally perform well on larger datasets, struggled on Messidor, with ViT-H14 achieving the highest Kappa score among ViT models at only 0.1357. These results suggest that, while transfer learning provides a useful starting point, tasks like DR grading on small datasets require either a greater number of training epochs, data augmentation, or larger, more balanced datasets for the models to effectively capture domain-specific patterns.

Table 5.6: *Performance comparison of CNN and ViT Architectures on the **DDR** dataset across multiple evaluation metrics.*

Architecture	Model	κ	Acc.	Precision	Recall	F1
CNN	ResNet-18	0.3525	66.47%	54.26%	39.92%	41.42%
	ResNet-50	0.5858	76.75%	64.86%	41.91%	44.60%
	VGG-16	0.4736	69.36%	43.18%	38.59%	39.17%
	EfficientNet-B5	0.5506	74.75%	52.10%	43.74%	45.87%
	InceptionV3	0.4543	70.04%	54.10%	40.03%	42.28%
ViT	ViT-S-16	0.2173	22.10%	40.13%	48.71%	35.52%
	ViT-S-32	0.1487	16.22%	38.20%	48.06%	32.51%
	ViT-B-16	0.4577	70.44%	55.38%	36.61%	38.48%
	ViT-B-32	0.4059	67.72%	58.21%	37.51%	40.02%
	ViT-L-16	0.4868	71.51%	44.55%	41.15%	41.39%
	ViT-L-32	0.3717	65.88%	42.87%	36.67%	36.87%
	ViT-H-14	0.4231	68.60%	52.68%	36.96%	38.69%

Surprisingly, the DDR dataset, which has the largest sample size ($\approx 13,000$),

did not demonstrate significantly improved performance compared to the APTOS dataset, Table 5.6. For example, EfficientNet-B5 achieved a κ of 0.5506 with an accuracy of 74.75%, which is only slightly better than the APTOS dataset’s results. This underperformance, despite the larger dataset, could be attributed to the fact that fundus images, such as those in DR datasets, contain features that are vastly different from the natural images in ImageNet. While the models may have more data to learn from, the unique characteristics of fundus images require a more specialized approach than what transfer learning from ImageNet can offer, particularly when training is limited to just 25 epochs.

The results across these datasets indicate that transfer learning from ImageNet is not sufficient for DR grading when the training is constrained to a small number of epochs. The performance of models such as ResNet-18 and VGG-16, which are relatively simpler architectures compared to EfficientNet-B5, further emphasizes this limitation. ResNet-50 generally outperformed ResNet-18, but both models demonstrated that a deeper architecture benefits more from extended training periods. VGG-16, on the other hand, showed the lowest performance across all datasets, which could be attributed to its older architecture and less optimized design for modern tasks like DR grading.

The limitations of transfer learning from ImageNet to medical imaging tasks are evident from these findings. The results suggest that, although transfer learning can provide a starting point by leveraging pre-trained weights, it is not sufficient for accurate DR classification without further fine-tuning or more extensive training. The significant gap in domain relevance between the ImageNet dataset and medical images

used in DR grading contributes to this shortfall. While ImageNet consists of millions of natural images, these images do not capture the complex and nuanced features of retinal images, which are essential for detecting DR. Table 5.7 shows the Grad-CAM visualization of the selected CNN models in this experiment. The selected regions are not consistent with effected areas in the fundus images, Table 5.7.

In conclusion, while transfer learning offers potential for accelerating model training, the results demonstrate that more extensive fine-tuning is needed for the specific task of DR grading. Increasing the number of training epochs could lead to better performance, especially on larger datasets like DDR. Additionally, pre-training on medical image datasets, or using synthetic medical data, may offer a more effective approach for tasks such as DR classification. Future research could explore these alternatives, including domain-specific transfer learning, to overcome the limitations observed when using ImageNet-pretrained models on medical imaging tasks. Furthermore, hybrid approaches that combine CNN architectures with more advanced models like ViT could be explored to enhance the models' capacity to handle the unique challenges of medical image analysis.

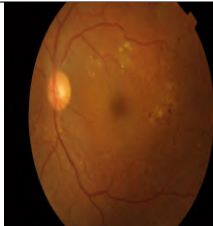
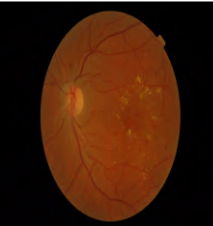
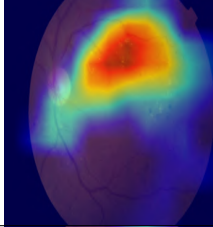
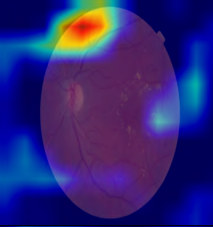
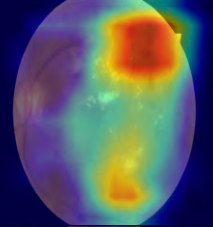
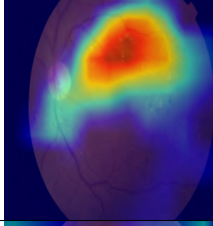
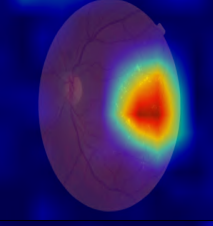
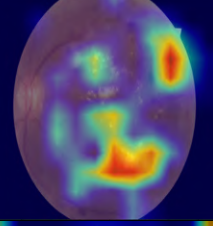
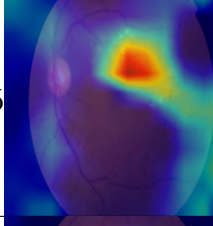
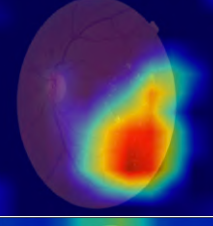
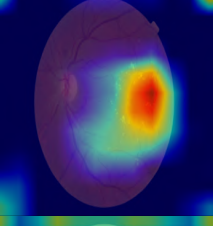
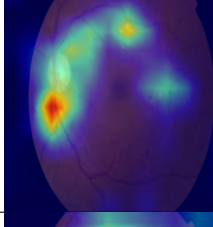
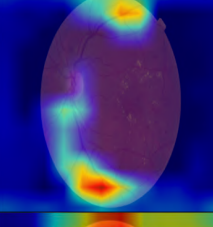
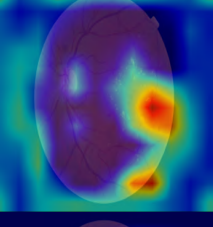
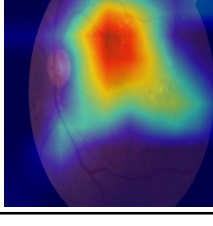
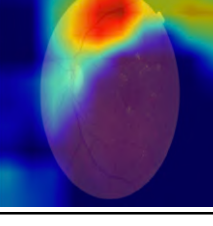
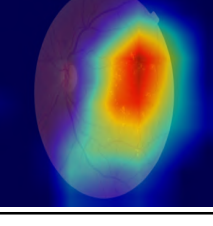
Model	APTOS	Messidor	DDR
Sample			
ResNet-18			
ResNet-50			
EfficientNet-B5			
VGG-16			
Inception-V3			

Table 5.7: *Original samples of severe NPDR (Grade 3) and corresponding Grad-CAM visualizations generated by various CNN models across three datasets. The visualizations highlight the areas of importance used by each model for prediction.*

5.5.2 Transfer Learning Performance of Vision Transformer Models

The ViT models were evaluated across four datasets: ODM, DDR, APTOS, and Mesidor, revealing distinct patterns in performance based on model complexity, dataset size, and transfer learning effectiveness. In the following analysis, we assess the ViT results for each dataset, highlighting the best-performing models and analyzing the relationship between model complexity and dataset size.

The DDR dataset, with approximately 13,000 samples, is the largest dataset used in this study. Across the ViT models, we observe that larger variants, such as ViT-B-16, ViT-B-32 and ViT-L-16, tend to outperform smaller models like ViT-S-16 and ViT-S-32. This is evident from the kappa score and accuracy. For instance, ViT-L-16 achieved a kappa score of 0.4868 and an accuracy of 71.51%, outperforming both smaller and larger models, Table 5.6. This indicates that larger ViT models are better suited for this larger dataset, likely due to their ability to capture more complex features from the data. In addition, ViT-L-16 effectively classified the dominant class (DR0), which makes up the majority of the dataset. However, the recall values were relatively lower across all models, suggesting that while the models learned well from the dominant class, they struggled with minority class detection.

The APTOS dataset, consisting of approximately 3,600 samples, reveals a distinct trend compared to the DDR dataset. In this case, smaller ViT variants, such as ViT-S-16 and ViT-S-32, outperformed their larger counterparts, with ViT-S-32 achieving the highest kappa score of 0.7465. This suggests that smaller ViT models are better suited

for datasets like APTOS, where the limited amount of training data may cause larger models to overfit or struggle to generalize effectively. Despite this, the larger models still demonstrated strong accuracy, indicating they excel in classifying the dominant class (DR0), which constitutes the majority of the dataset, Table 5.4. However, all ViT variants both small and large, exhibited difficulty in accurately identifying minority classes, as reflected in their lower recall and precision scores. This underscores the challenge posed by imbalanced datasets in DR grading, where models are prone to favor the prevalent class at the expense of underrepresented categories.

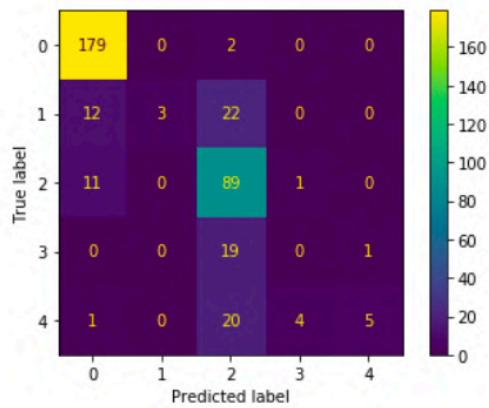
The Messidor dataset, the smallest in this study with approximately 1,700 samples, demonstrated better performance with smaller ViT models. However, the highest accuracy was achieved with the largest model, ViT-H-14, highlighting that even for the dominant class (DR0), the smaller models struggled to perform adequately, Table 5.5. This suggests that the complexity of the ViT-H-14 model was necessary for the model to learn effectively, reflecting the inherent difficulty of transfer learning on small, imbalanced datasets like Messidor. Despite this, the overall performance across all models was notably lower than on the other datasets, underscoring the challenges posed by limited data and class imbalance, where even larger models fail to generalize well without sufficient and diverse training samples. Interestingly, larger ViT models such as ViT-B-16 and ViT-L-32 performed poorly on this dataset, suggesting that model complexity negatively impacts performance on smaller datasets like Messidor. The kappa scores for larger models (e.g., 0.0099 for ViT-B-16) reflect their inability to learn effectively from such a limited sample size.

The Vision Transformer (ViT) models show a stark contrast, with ViT-S-16 achiev-

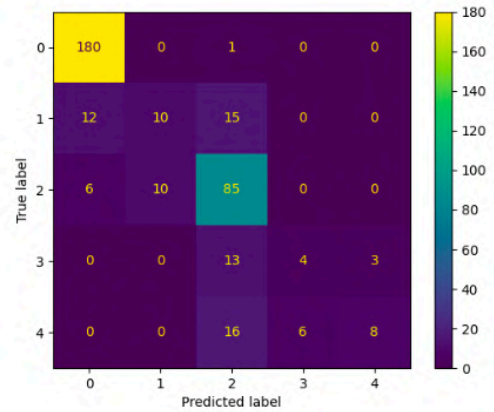
ing the highest kappa score (0.8315) across all datasets, alongside an accuracy of 67.06%. This is a significant improvement over all CNN models, emphasizing the superior capability of ViTs to model global dependencies and capture detailed patterns in the ODM dataset. The ODM dataset, characterized by dual-view fundus images, likely benefits from ViT’s ability to leverage the rich spatial and structural information inherent in these dual-view representations. Compared to datasets like APTOS, Messidor, and DDR, which contain single-view images or lack the comprehensive data richness of ODM, the dual-view nature of ODM provides more context and complementary visual cues, enhancing the model’s ability to differentiate subtle patterns. This explains why the ODM dataset achieves a much higher kappa score, reflecting greater inter-rater agreement and superior classification consistency.

The results across the four datasets, including ODM, clearly demonstrate the effect of model complexity in relation to dataset size and characteristics. On the larger DDR dataset, larger ViT models such as ViT-B-16 and ViT-B-32 outperformed smaller models, leveraging their capacity to learn more complex features from the abundant data. In contrast, on the smaller datasets (APTOS and Messidor), smaller ViT variants (ViT-S-16, ViT-S-32) performed better, as the larger models may have overfitted due to the limited training samples. Interestingly, the ODM dataset stands out, with ViT-S-16 achieving the highest kappa score (0.8315) among all datasets and models. This superior performance highlights the importance of the dual-view fundus images in the ODM dataset, which provide complementary and rich contextual information, enabling ViT-S-16 to effectively capture patterns and relationships without overfitting. This suggests that model complexity needs to be carefully matched to both the

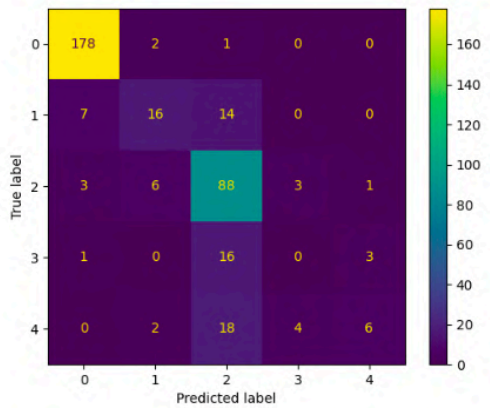
dataset size and its inherent richness, as overly complex models may not generalize well with small datasets, while simpler models can excel when paired with datasets offering high-quality, diverse visual features like ODM.



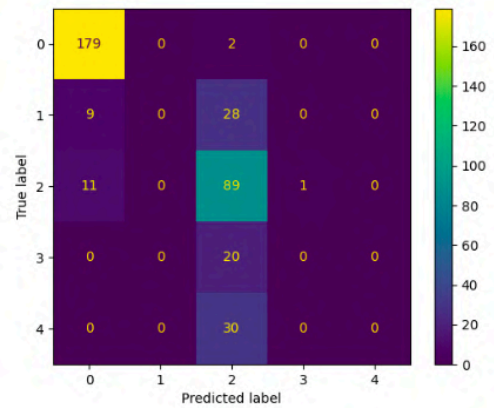
(a) ResNet-18



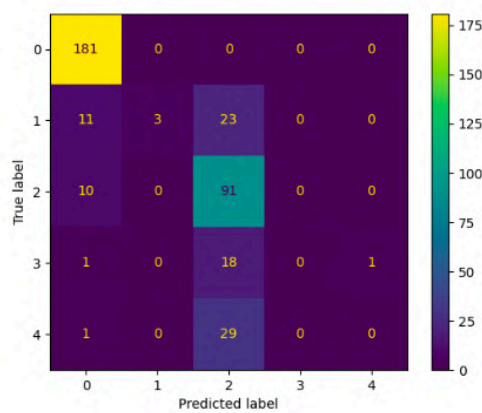
(b) ResNet-50



(c) EfficientNet-B5



(d) VGG-16

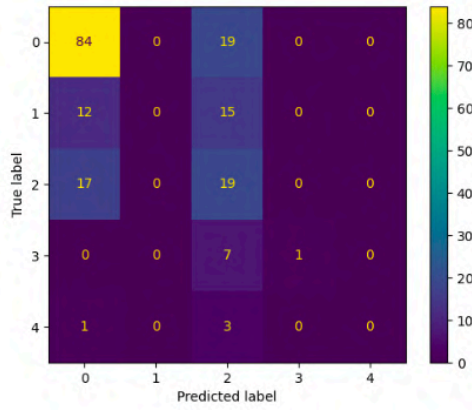


(e) Inception-V3

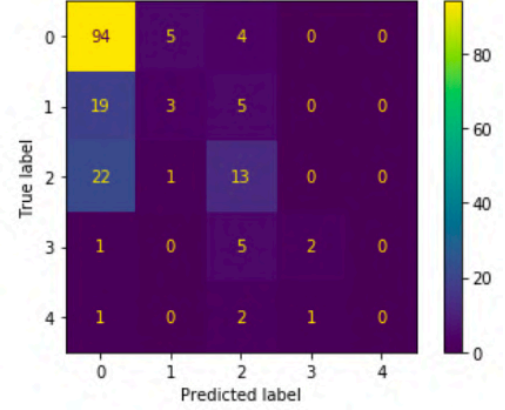
Figure 5.3: *Confusion matrices for different CNN models on the **APTOS** dataset.*

5.5.3 CNN Models Class-Wise Performance

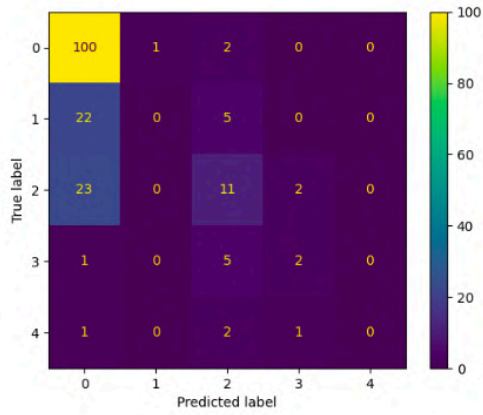
The confusion matrices in Figures 5.3 to 5.5, highlight the class-wise performance of five CNN models: ResNet-18, ResNet-50, EfficientNet-B5, VGG-16, and Inception-V3, on the APTOS dataset for DR grading. Across all models, there is a consistent trend of good performance in detecting the normal class (grade 0), but significant challenges arise in distinguishing between adjacent grades, particularly grades 1 and 2. For example, ResNet-18 and ResNet-50 show substantial misclassification between grades 1 and 2, which share similar lesions such as microaneurysms and hemorrhages. Furthermore, all models struggle to detect the more advanced DR grades (3 and 4), with frequent misclassification. This issue aligns with the observations in the literature, where late-stage DR is often less represented in datasets, making it difficult for models to detect these stages accurately.



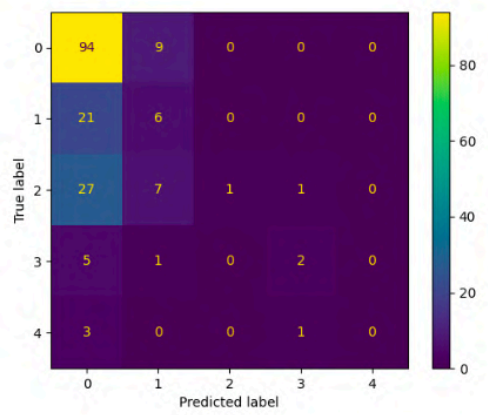
(a) ResNet-18



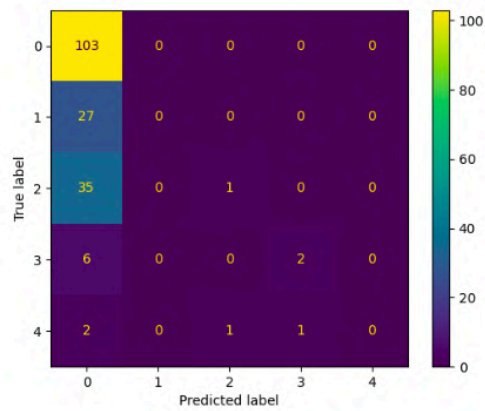
(b) ResNet-50



(c) EfficientNet-B5



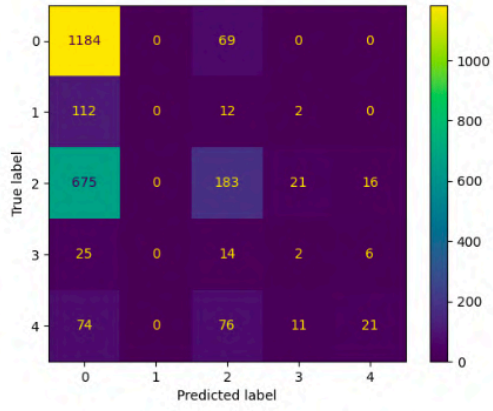
(d) VGG-16



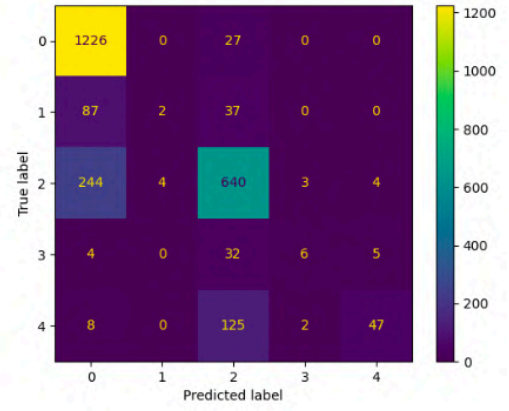
(e) Inception-V3

Figure 5.4: Confusion matrices for different CNN models on the *Messidor* dataset.

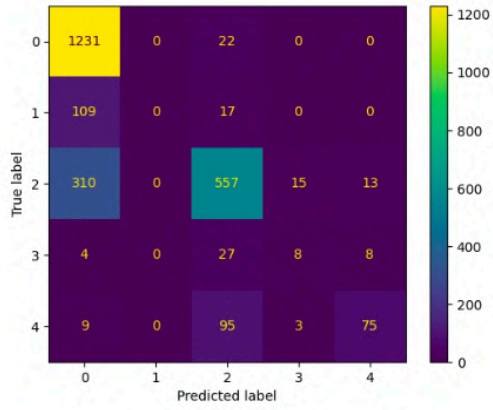
EfficientNet-B5 and Inception-V3 show slight improvements in overall performance, particularly in detecting grade 2, yet the issue of adjacent grade misclassification remains prevalent. VGG-16, with its older architecture, performs the worst in identifying the higher grades, particularly grade 3, where significant misclassification occurs. These results reflect the broader challenge discussed in the literature in Section 2.4: adjacent DR grades often share overlapping features, making it difficult for CNN models to differentiate between them, especially when advanced grades are underrepresented. Addressing these limitations will require strategies such as improving dataset balance or designing more sophisticated models to capture the subtle differences between the grades.



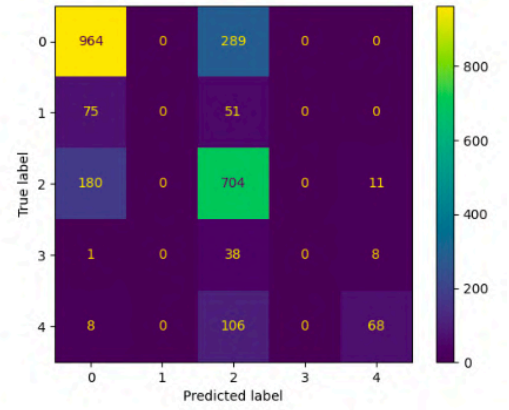
(a) ResNet-18



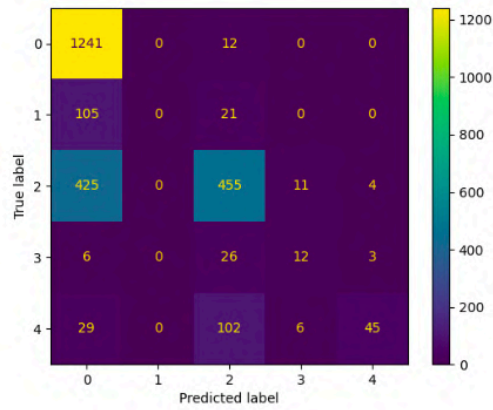
(b) ResNet-50



(c) EfficientNet-B5

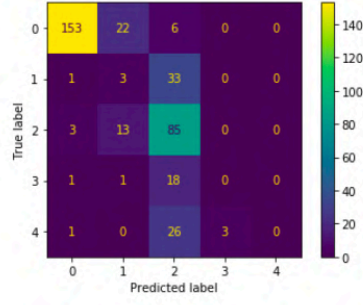


(d) VGG-16

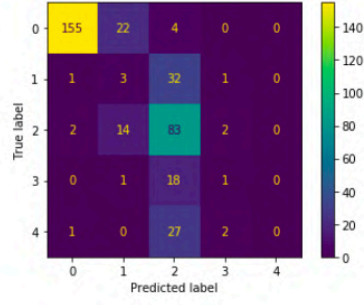


(e) Inception-V3

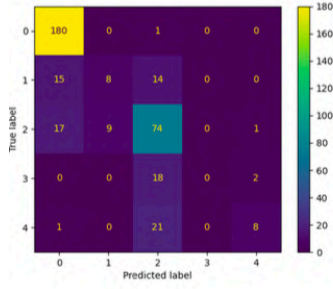
Figure 5.5: *Confusion matrices for different CNN models on the **DDR** dataset.*



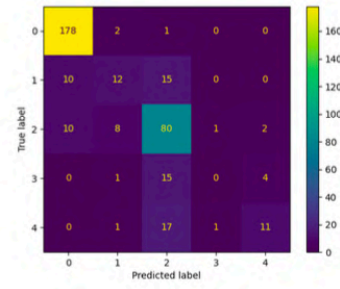
(a) ViT-S16



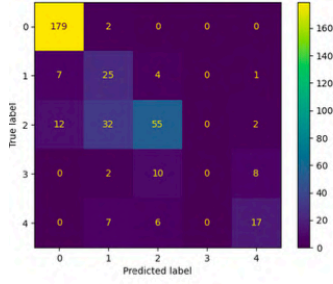
(b) ViT-S32



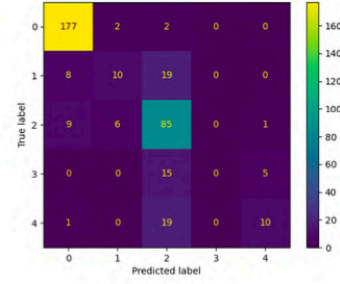
(c) ViT-B16



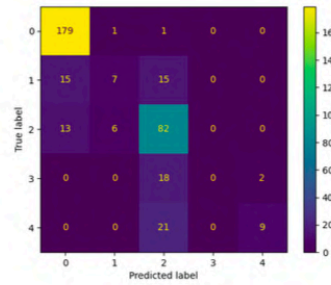
(d) ViT-B32



(e) ViT-L16

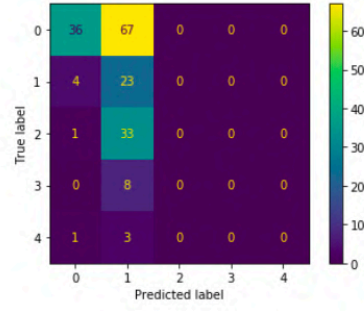


(f) ViT-L32

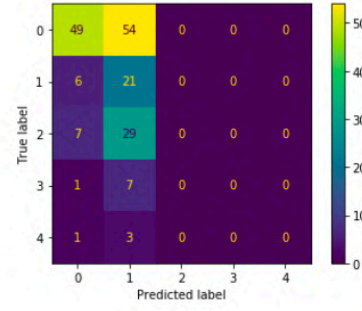


(g) ViT-H16

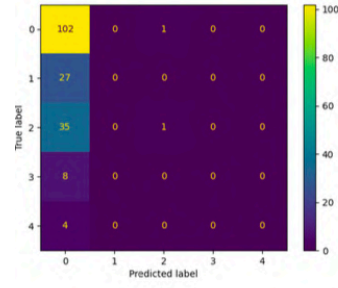
Figure 5.6: *Confusion matrices for different Vision Transformers models on the **APTOS** dataset.*



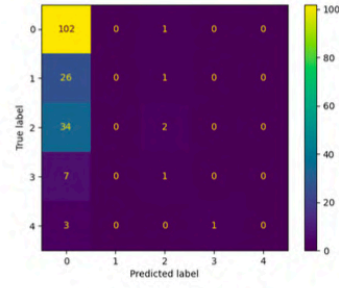
(a) ViT-S16



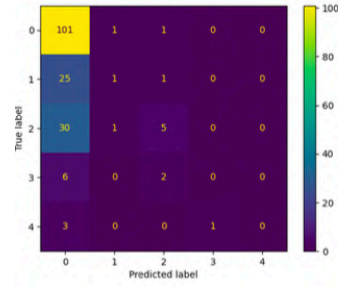
(b) ViT-S32



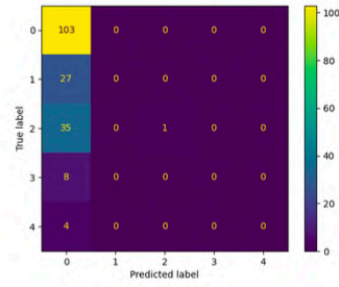
(c) ViT-B16



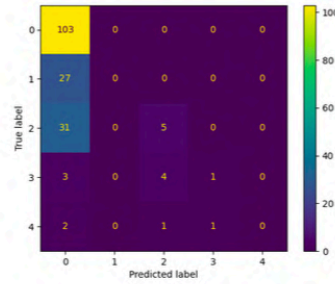
(d) ViT-B32



(e) ViT-L16

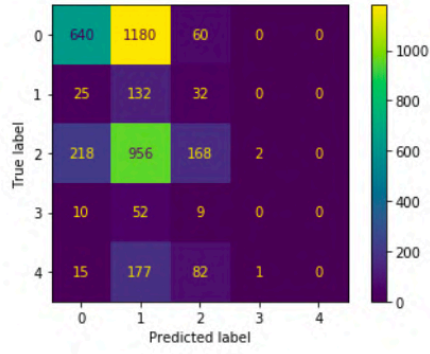


(f) ViT-L32

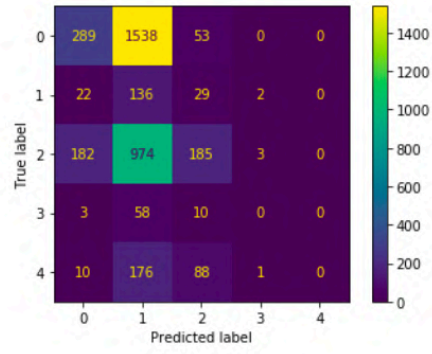


(g) ViT-H16

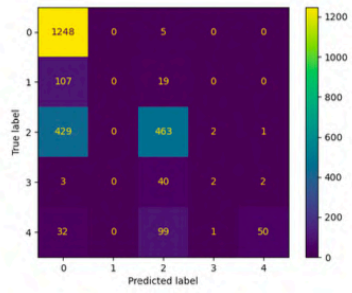
Figure 5.7: *Confusion matrices for different Vision Transformers models on the **Messidor** dataset.*



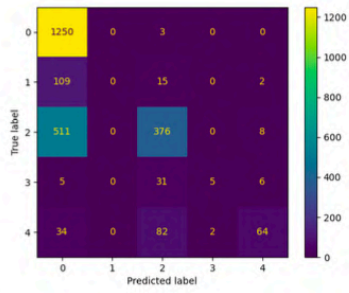
(a) ViT-S16



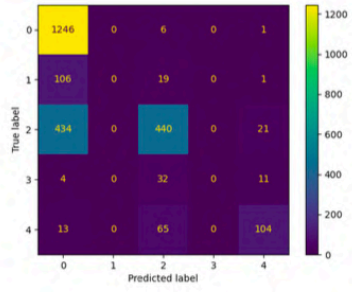
(b) ViT-S32



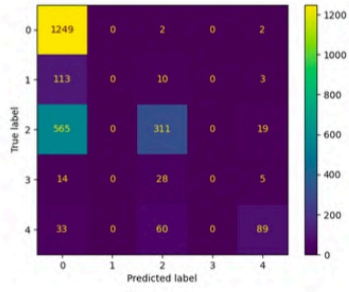
(c) ViT-B16



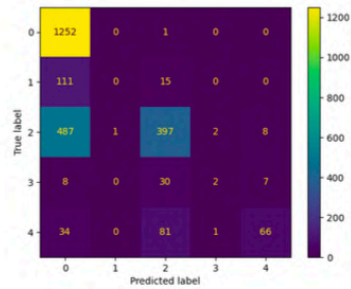
(d) ViT-B32



(e) ViT-L16



(f) ViT-L32



(g) ViT-H16

Figure 5.8: *Confusion matrices for different Vision Transformers models on the DDR dataset.*

5.5.4 Vision Transformer Models Class-Wise Performance

The class-wise performance of the Vision Transformer (ViT) models across the AP-TOS, Messidor, and DDR datasets follows similar patterns to those observed with CNN models. The ViT models consistently excel in identifying the normal class (grade 0) with high accuracy, as seen in all three datasets. For instance, ViT-B16 and ViT-B32 show near-perfect classification for grade 0 across the datasets, reflecting the models' strong ability to detect non-DR cases, Figures 5.6 to 5.8.

However, as with CNN models, the ViT models struggle to differentiate between adjacent DR grades, particularly between grades 1 and 2. This trend is visible in all datasets, where significant misclassification occurs between these stages. The advanced stages (grades 3 and 4) also pose challenges, with frequent confusion between them, particularly in the larger models like ViT-L32 and ViT-H16. These issues align with the CNN findings, where models faced difficulties due to the overlap in visual features between these stages and the under representation of advanced DR grades in the datasets.

Despite slight improvements in handling intermediate stages with the larger ViT models, the overall performance pattern remains consistent with the CNN models. Both architectures exhibit strong performance in detecting early-stage DR and normal cases, but struggle with later stages, where subtle differences are harder to capture. This suggests that while ViT models offer potential, their class-wise performance remains constrained by the same challenges faced by CNN models, highlighting the need for strategies such as better dataset balance or architectural adjustments to

improve the detection of advanced DR stages.

5.5.5 Comparison Results with Previous Studies

In the exploration of deep learning models for diabetic retinopathy (DR) grading on the APTOS dataset, a key factor that consistently emerges is the impact of transfer learning from ImageNet and the architectural complexity of models. Studies evaluating both randomly initialized and ImageNet-pretrained models demonstrate trends similar to our experiments, particularly in terms of the performance benefits conferred by transfer learning.

As shown in the Table 5.8, the number of epochs plays a critical role in the performance of supervised models. For example, models such as ResNet18 and ResNet50, trained with random initialization, required 25 and 50 epochs, respectively, but exhibited relatively low Kappa scores (0.5790 and 0.5927). However, with ImageNet pretraining, the performance of models like ResNet50 improves dramatically, achieving a Kappa score of 0.8057 after just 50 epochs. This underscores the importance of both pretraining and sufficient epochs to fully capture relevant feature representations.

Furthermore, more complex architectures, including Hybrid Inception ResNet-v2, DenseNet, and InceptionV3, consistently show superior performance with accuracy scores surpassing 80%. These models required substantially more epochs, ranging from 100 to 1000, to reach their optimal performance. For instance, E-DenseNet BC-201, which was trained for 100 epochs, achieved an accuracy of 82.20% and a Kappa score of 0.730, highlighting the benefits of densely connected layers in retaining critical feature information for DR severity classification.

The results clearly indicate that more sophisticated architectures, while more computationally intensive, often require extended training to converge to high performance levels. These findings align with our observations of advanced CNN and ViT models, where transfer learning and sufficient training epochs are pivotal for achieving superior results. Our experiments, in conjunction with these studies, confirm that pretraining, especially with ImageNet, remains a key strategy for enhancing model accuracy and robustness in medical imaging tasks, particularly in DR grading.

Table 5.8: *Comparison results of supervised learning studies on APTOS dataset.*

Models	Epochs	Weights	κ	Acc.	Sens.	Spec.
ResNet18[126]	25	Random	0.5790	-	-	-
ResNet50[129]	50		0.5927	-	-	-
Hybrid-Inception	-		-	82.18%		
ResNet-v2[130]						
E-DenseNet-BC-169[131]	100		0.7001	80.20%	73.6	92
E-DenseNet-BC-201[131]	100		0.730	82.20%	74.6	92.3
ResNet50[129]	50	ImageNet	0.8057	-	-	-
Xception[132]	1000		-	79.59%	82.35	86.32
InceptionV3[132]	1000		-	78.72%	63.64	85.37
MobileNet[132]	1000		-	79.01%	76.47	84.62

5.6 Conclusion and Remarks

In this chapter, we investigated the effectiveness of supervised transfer learning using pre-trained ImageNet weights using CNN and ViT models. The limitations of transfer learning from ImageNet to medical imaging tasks are evident from these findings.

The results suggest that, although transfer learning can provide a starting point by leveraging pre-trained weights, it is not sufficient for accurate DR classification without further fine-tuning or more extensive training. The significant gap in domain relevance between the ImageNet dataset and medical images used in DR grading contributes to this shortfall. While ImageNet consists of millions of natural images, these images do not capture the complex and nuanced features of retinal images, which are essential for detecting DR.

Models require some form of modification to achieve better accuracy and reduce false positives. While Vision Transformer (ViT) models can deliver reasonable performance, their effectiveness varies significantly based on dataset size and model complexity. Across all datasets, recall scores were relatively low, particularly for detecting minority classes, which is critical for DR diagnosis. This indicates that the knowledge transferred from ImageNet may not fully capture the specific nuances of DR fundus images, further highlighting the need for domain-specific adaptations.

In conclusion, while transfer learning offers potential for accelerating model training, the results demonstrate that more extensive fine-tuning is needed for the specific task of DR grading. Increasing the number of training epochs could lead to better performance, especially on larger datasets like DDR. Additionally, pre-training on medical image datasets, or using synthetic medical data, may offer a more effective approach for tasks such as DR classification. Future research could explore these alternatives, including domain-specific transfer learning, to overcome the limitations observed when using ImageNet-pretrained models on medical imaging tasks. Furthermore, hybrid approaches that combine CNN architectures with more advanced models

like Vision Transformers (ViT) could be explored to enhance the models' capacity to handle the unique challenges of medical image analysis. In the next chapter, we will delve deeper into transfer learning through contrastive self-supervised learning, with the goal of further improving representation learning for DR grading and addressing the unique challenges posed by medical imaging tasks.

CHAPTER 6

DIABETIC RETINOPATHY

GRADING USING

CONTRASTIVE

SELF-SUPERVISED TRANSFER

LEARNING WITH LIMITED

DATA

6.1 Introduction

Contrastive self-supervised learning has emerged as a highly effective training strategy for medical imaging tasks, particularly in scenarios where annotated datasets are

scarce. This approach enables the extraction of robust and reliable representations from unlabeled data, obviating the need for extensive human annotation. In the context of the classification of diabetic retinopathy, achieving exceptional performance traditionally requires a substantial amount of labeled data, which is costly and time consuming to obtain.

In the previous chapter, we identified key limitations of supervised transfer learning models when applied to domain-specific tasks like diabetic retinopathy (DR) grading. While pretraining on general-purpose datasets such as ImageNet offers a performance boost over random initialization, it does not adequately transfer the specialized knowledge needed for optimal performance in medical imaging tasks. This shortcoming is particularly evident in models like ResNet and ViT, which struggle to capture domain-specific features when their pretraining is not aligned with the task at hand. To address this issue, we turn to self-supervised learning (SSL) as an alternative approach to pretraining that can better leverage the unique characteristics of medical datasets.

To address this issue, in this chapter we propose a novel contrastive self-supervised learning framework, termed CGSSL, designed to maximize the representational capacity of fundus images using a minimal number of unlabeled samples and that can better leverage the unique characteristics of medical datasets. Specifically, this addressing our **RQ4** and **Gap-3, Gap-4 and Gap-6**.

In this chapter, our method leverages the green channel of fundus images to enhance the visibility of critical features through saliency-guided processing. This technique not only improves the discernibility of the fundus characteristics but also facilitates more effective feature extraction. Empirical evidence from our study demon-

strates that a limited set of dual-view fundus images can achieve performance comparable to that of single-view images in a self-supervising learning framework. We further emphasize the importance of the projection head and residual connections in our framework, which play a crucial role in capturing intricate features and steering the training process towards superior performance on downstream tasks. Remarkably, our approach relies on only a few unlabeled fundus images, underscoring the efficiency and strength of our contribution in transferring knowledge representations from limited data samples. This work not only highlights the feasibility of utilizing small, unlabeled datasets for high-performance medical image classification but also sets a new benchmark for future research in the field.

6.2 Background and Motivation

DR is a major contributor to visual impairment and blindness among the working age population around the world [10]. According to the International Diabetes Federation (IDF) [2], around 30% of those diagnosed with diabetes may eventually develop DR, which often goes unnoticed until the retinal abnormalities have progressed to a point where therapy is no longer effective or feasible.

An inherent challenge in medical imaging with deep learning is the high cost and time-consuming nature of annotation, which necessitates the expertise of well-trained medical professionals. Furthermore, the availability of annotated medical images for the computer vision research community is limited due to the need for professional experts in the field, which is not always convenient. Despite the availability of large

quantities of unlabeled samples, researchers want to use them as a valuable resource that may be utilized without the need for labeling. Currently, self-supervised learning (SSL) is considered the most promising method for learning because it takes advantage of unlabeled examples. SSL is renowned for extracting generalized features from vast quantities of unlabeled data [133]. The acquired characteristics are subsequently applied to different tasks using a limited number of labeled samples.

SSL techniques can be classified into two basic categories: generative and discriminative approaches. The generative SSL acquires knowledge of the input data and output predictions, such as picture rotations, jigsaw puzzles, and image colorization, by learning the relationships between them [134], [135]. On the other hand, discriminative techniques, namely contrastive SSL, acquire representations by promoting similarity among positive pairs and dissimilarity among negative pairs via a contrastive loss function [136]. Prior research has employed SSL techniques using single-view fundus images for grading diabetic retinopathy. However, dual-view fundus images provide a comprehensive view of the retina by capturing images from two specific locations: the optic disc and the macula. The limited availability of dual-view fundus images compared to single-view fundus images hinders their widespread utilization. Recently, two newly released datasets of dual-view fundus images are being utilized in supervised learning approaches [108], [107].

Data collection setups in hospitals for single- or dual-view acquisition are taken from different sites and fundus cameras. However, this exposes the images to poor quality conditions such as inadequate illumination, blurriness, and darkness [19]. Hence, pre-processing is a crucial procedure to address the problem of quality and

enhance the level of detail in fundus images. Several methodologies have been suggested, but only a limited number of research employ the green-channel intensity in order to enhance the amount of information provided to the model during supervised training [137], [138], [139]. This is because DR adjacent grades have shared lesions, leading to misclassifying them [140]. Further research is required to concentrate on the distinctive diagnostic features of fundus images during the pre-processing step that differentiate between different stages of DR and improve the model’s overall performance.

This study underscores the critical importance of the green channel in conjunction with salient object detection (SOD) and delves into the significant impact of dual-view fundus images on self-supervised learning (SSL). By leveraging the green channel, we enhance the visibility of crucial features, thereby facilitating more effective salient region detection. The dual-view approach further enriches the representational capacity, capturing comprehensive and nuanced information from the fundus images. Additionally, we highlight the pivotal role of incorporating residual connections within the projection head design, which significantly enhances the model’s ability to capture intricate features and improve learning efficiency.

In summary, CGSSL, a novel framework that synergizes contrastive learning with innovative components: dual green-guided knowledge enhanced by salient object detection, dual-view fundus imaging for enriched feature extraction, and a projection head design featuring residual connections. This comprehensive approach not only advances the state-of-the-art in diabetic retinopathy classification but also sets a new benchmark for the effective use of limited data in medical imaging tasks.

Our experiment in this chapter provides valuable insights into DR grading using SSL through the following contributions:

- Developed the ODM dataset for self-supervised learning, incorporating both single and dual views of fundus images, and capturing a diverse range of variations to support robust model training and evaluation.
- Introduced the CGSSL model, a novel framework pre-trained with a minimal amount of data yet enriched with comprehensive knowledge representations. This model is specifically guided by saliency detection in the green channel and leverages dual-view fundus images to optimize the learning process. The innovative projection head design, featuring residual connections, plays a crucial role in effectively utilizing detailed information and enhancing model performance.
- Highlighted the critical role of dual-view in medical field and how it is beneficial in knowledge representation in self-supervised learning, demonstrating its impact on the efficacy of feature extraction and overall model performance.

6.3 Related Work

The grading of DR has been the subject of extensive research, with studies exploring various methodologies using single- and dual-view approaches. It is noteworthy, however, that the majority of studies in the field have relied on single-view fundus images, primarily due to publicly accessible benchmark datasets.

Supervised learning: CNN-s have been suggested as a solution for the automated

grading of DR [140], [141], [142], [143], [112]. The primary factor contributing to CNN’s success is its capacity to extract highly representative features. Nevertheless, training a network often requires a large, annotated dataset. Ophthalmologists have a significant burden in constructing a well-annotated dataset due to the annotation process’s time-consuming, monotonous, and error-prone nature. Similarly, some studies utilized the ensemble CNNs approach to learn a deeper embedding of fundus images [144], [145], while others introduced attention mechanisms to CNN based-models [146].

Recently, transformer-based models have shown promising performance with medical images [115], such as ViT [123], Class-attention in image Transformers (CaiT) [147], Data-efficient image Transformer (DeiT) [148], and Bidirectional Encoder representation for image Transformer (BEiT) [149]. Unlike CNN-based models, transformer-based models divide the input image into a sequence of patches, process them in parallel, and then apply global attention. The study by [150] adopted four ViT-based models, namely ViT, CaiT, DeiT, and BEiT for DR. CaiT outperformed other transformer-based models. The proposed ensemble image transformer (EiT) achieved the best result among other ensemble experiment setups. When compared to CNN-based models, the transformer-based models demonstrated their strength in leveraging more knowledge representations that contribute to the model’s performance.

Many researchers focused on single-view approaches; a few explored the potential of dual-view fundus images in DR grading. Recognizing its inherent advantages, which capture both macula-centric and optic disc-centric perspectives. Incorporating dual-

view images into the learning process for DR grading remains a relatively nascent area of exploration. However, models such as CrossFit [107] and DL-DeepDRiD [108] that have embraced this approach have demonstrated promising results. By integrating dual-view images into supervised learning frameworks, these studies have showcased the potential benefits of exploiting the added information for enhanced DR grading accuracy and robustness.

Contrastive self-supervised learning: Unlike supervised learning, self-supervised learning utilizes large unlabeled data to learn rich visual representations in the pre-training phase. Then, the knowledge is transferred to a target downstream task such as classification, detection, and segmentation. Recently, the SSL paradigm has gained a lot of attention in the research community, proving its strength in computer vision. Many state-of-the-art SSL models have been introduced for natural image classification, such as SimCLR [151], MoCo-v2 [152], MoCo-v3 [153], and DINO [154].

Contrastive SSL has emerged as a contemporary approach to representation learning, aiming to acquire robust representations from input data by acquiring discriminative capabilities [133]. This is achieved by effectively discerning between similar (positive) and dissimilar (negative) pairs or optimizing the agreement between a pair of positive views per the design principles of the contrastive learning architecture.

Few studies leverage the potential of SSL in DR grading, while most use supervised learning instead. Huang et al. [126] introduced an SSL framework for DR grading named saliency-guided self-supervised image transformer (SSiT) based on MoCo architecture, saliency detection [155], and Transformer [123]. SSiT [126] distinguished

itself from other SSL methods by guiding the learning through saliency maps to focus on salient regions. Saliency detection is vital in steering the SSL to focus on regions of interest. Comparing the results against the state-of-the-art models, the results reveal the strength of incorporating the saliency map and transformer-based models into the SSL framework, which delivers richer semantic information compared to SOTA models [151], [153].

A study conducted by [156] proposed a self-supervised framework called lesion-based contrastive learning for grading diabetic retinopathy. In the conventional contrastive learning approach, rather than using complete images as input, the study utilizes lesion patches to promote the feature extractor’s acquisition of highly discriminative representations for grading DR.

Besides the significant role of the feature extractor in SSL architectures, the projection head has an influential role in driving the loss function to regulate the model weight during the backpropagation process. SimCLR [151] introduced the projection head. Later on, adopted by MoCo-v2 [152], [153], BYOL [157], SimSiam [158], MAE [159], as shown in Figure 6.1. The projection head design is a multi-layer perceptron (MLP), appended after the feature extractor (the backbone) and before the contrastive loss, projecting the extracted features to a lower dimensional space. To understand and expose the role of the projection head even though it is discarded after training, [160] explores the impact of three configurations of the projection head: no projection head, linear projection head and non-linear projection head. Applying the three settings on SimCLR with ResNet as a backbone, t-SNE visualization showed that the non-linear projection head setting is the best choice, confirming the impor-

tance of the projection head as a low-rank mapping. A study conducted by [161] empirically verifies that combining the encoders and the projection head outperforms the individual encoder without a projection head.

Our CGSSL framework distinguished itself from other studies by utilizing small samples during the pretraining of self-supervised learning and achieving promising results compared to SOTA. In fact, incorporating dual-view images into the learning paradigm for DR grading remains a relatively nascent area of exploration. Our CGSSL integrates the knowledge of dual-view fundus images during pretraining, boosting the overall learning by enriching the model with different knowledge views. Furthermore, the utilization of green-saliency guidance contributes to highlighting the region of interest exposed by the green channel. Adding a residual connection to the projection head design enhances the downstream task’s overall performance.

6.4 Dual-view Fundus Images Datasets

Within the current research community, the presence of dual-view DR image datasets is notably scarce. While numerous benchmark datasets are publicly available, it is imperative to highlight that the majority of these resources predominantly provide single-view fundus images. The lack of dual-view datasets hinders the comprehensive analysis and evaluation of DR detection and grading algorithms relying on the combined information from macula-centric and optic disc-centric images. Consequently, researchers encounter limitations in accurately capturing the entirety of the retinal structures and identifying subtle abnormalities that may be obscured in single-view

images.

In this experiment, we used our constructed dual-view dataset ODM, Chapter 4, to train the proposed CGSSL model. The rich information embedded in this dataset in the form of dual-view fundus images, offers a platform to learn rich representation in SSL settings.

6.5 Preliminaries and Notations

6.5.1 Contrastive Self-supervised Learning

Contrastive self-supervised learning (CSSL) is a type of SSL that utilizes contrastive loss to train a model to learn meaningful representations from unlabeled data. The core idea of CSSL is to contrast positive and negative pairs of data samples, with positive pairs representing similar or related samples and negative pairs representing dissimilar or unrelated samples. The basic components of the CSSL architecture mainly contain two encoders for feature extraction of two augmented views and a contrastive loss function to compute the distance between negative and positive pairs. After extracting the representations of the two augmented views, the feature extractor component computes the contrastive loss function to minimize the distance between similar pairings and maximize the distance between dissimilar pairs. Other CSSL architectures, such as SimCLR, MoCo, and BYOL [151], [153], [157], incorporate a projection head after the feature extractor component to enhance the focus on important features. The encoder achieves this by projecting the extracted features. The contrastive loss function utilizes these projected properties to guide the learning

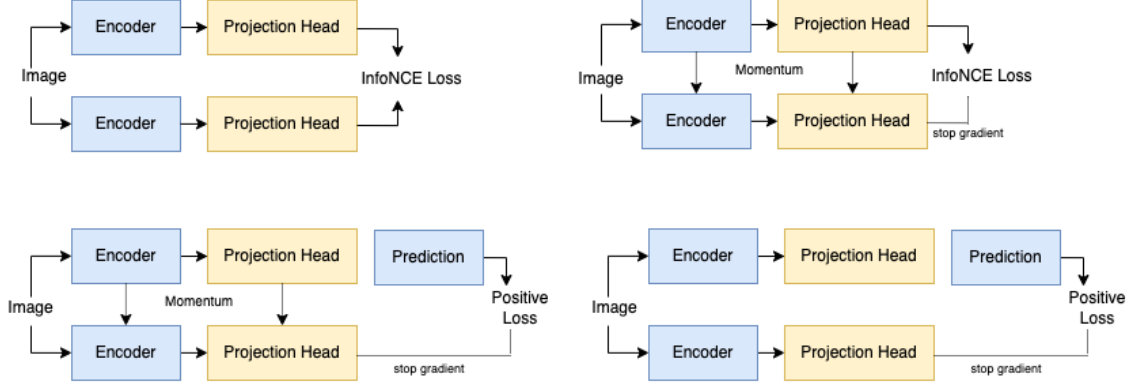


Figure 6.1: *State-of-the-art Self-supervised network architectures. SimCLR, MoCo and BYOL all shared the same projection head design.*

process.

Through learning, contrastive loss plays a vital role in regulating learning by maximizing the similarity between positive pairs and minimizing the similarity between negative pairs. Finally, the model learns to distinguish between similar and dissimilar samples, effectively capturing the underlying structure and patterns within the data. In contrast, two different views are generated by applying compositions of data augmentations (*e.g.*, cropping, flipping, and color jittering). Two encoders named query and key, f_q and f_k both parametrized by θ_q and θ_k .

The query encoder is the base encoder and contains a feature extractor, a projection head, followed by a prediction head. Further, the key encoder is a momentum encoder with the same structure as the base encoder architecture but without a prediction head. The key encoder's primary role is to update its parameters by applying a moving average to the query encoder parameters. The moving average is calculated via equation 6.1, where ϵ is defined as the momentum coefficient,

$$\theta_k = \epsilon\theta_k + (1 - \epsilon)\theta_q, \quad (6.1)$$

the two encoders' output features are z_i and z_j ; contrastive loss uses these two feature maps to maximize the similarity between the positive pairs and minimize the dissimilarity between negative pairs. Here, we adopt the InfoNCE loss [162],

$$\mathcal{L} = -\log \frac{\exp(h_q \cdot z_k^+ / \tau)}{\exp(h_q \cdot z_k^+ / \tau) + \sum_{z_k^-} \exp(h_q \cdot z_k^- / \tau)}, \quad (6.2)$$

where z_k^+ is a positive pair similar to h_q and z_k^- is the negative pair dissimilar to h_q , both at the same training batch. While τ is the temperature parameter that controls the sharpness of the contrastive loss to regulate the overall training.

6.5.2 Saliency Detection

Salient Object Detection (SOD) is a fundamental process in computer vision that aims to localize and segment the most prominent objects or regions within an image. It plays a pivotal role in various applications, including urban surveillance, object re-targeting, semantic segmentation, image synthesis, visual tracking, and image retrieval [163]. Traditional SOD models identify salient regions by analyzing their differences from the surrounding areas, often relying on hand-crafted features such as color, contrast, contour, objectness, focusness, and backgroundness.

Given an image I , the saliency detection process aims to compute a saliency map S , where each pixel value $S(x, y)$ represents the saliency of the corresponding pixel $I(x, y)$ in the original image. The saliency value can be computed based on various

features such as color, intensity, and contrast. A typical equation for saliency detection might be represented as follows:

$$S(x, y) = \sum_{f \in F} w_f \cdot \text{Saliency}_f(I, x, y), \quad (6.3)$$

where $S(x, y)$ is the saliency value at pixel (x, y) , and F is the set of features (e.g., color, intensity, contrast). w_f is the weight assigned to feature f . $\text{Saliency}_f(I, x, y)$ is the saliency value computed using feature f for the pixel (x, y) .

6.6 Methodology

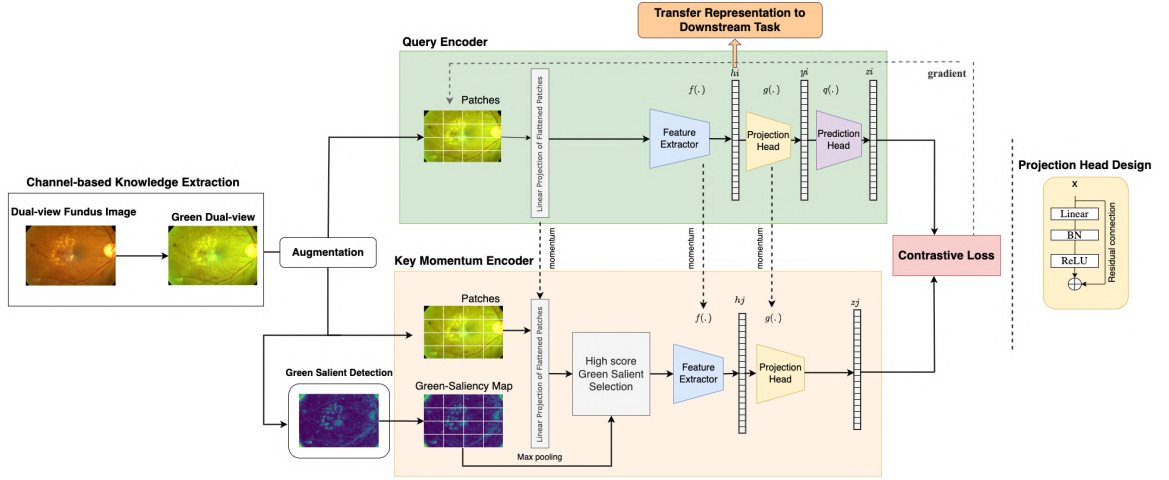


Figure 6.2: The proposed Channel-guided self-supervised learning (CG-SSL) framework. The Green-channel extraction component extracts green-channel knowledge from RGB dual-view fundus image. Two augmented images are derived from the green dual-view image and fed to two encoders. The query encoder is the base encoder where the SSL learning is achieved. The key encoder is a momentum encoder that supports the query encoder by focusing on green-salient guidance. The projection head design consists of three layers with residual connection, batch normalization and ReLU.

6.6.1 CGSSL Framework Construction

CGSSL framework innovatively builds by integrating dual-view fundus images and enhancing the representation learning process with green-channel saliency detection. Given a dual-view fundus image as input x , the green channel is extracted, followed by the generation of two augmented views that are subsequently fed into the two encoders f_q and f_k [153]. In both encoders, the green dual-view fundus image is divided into patches and processed through a linear projection to flatten these patches, as illustrated in Figure 6.2.

Distinctively, the key encoder applies saliency detection on the green channel image, which is then divided into patches. These patches undergo a max-pooling operation, followed by salient patch selection, which refines the input sequence of the key encoder to focus on critical regions and discard less informative patches, thereby enhancing the representation of important features.

We employ a ViT [123] as the backbone to extract embeddings from both encoders h_i and h_j , which are then processed by the projection head. This projection head is a multilayer perceptron with three layers, batch normalization, and ReLU activation, incorporating a residual connection to ensure efficient projection of the features extracted by the encoders. Subsequently, the outputs of the projection heads of the two encoders, y_i and z_j , are extracted, with y_i being further processed by an additional component in the query encoder f_q to produce the embeddings z_i .

The contrastive loss function is employed to maximize the similarity between the positive pairs z_i and z_j while minimizing the similarity between the negative pairs.

Gradient updates are applied exclusively to the query encoder. Upon completion of the training, the embeddings h_i are transferred to the downstream task, showcasing the efficacy of our CGSSL framework in utilizing dual-view fundus images and green-channel saliency for enhanced representation learning.

6.6.2 Channel-Based Knowledge Utilization

We have utilized the inherent benefits of the green channel in fundus images to take advantage of locations that exhibit a higher signal-to-noise ratio (SNR) than the red and blue channels. The use of the green channel is included in our self-supervised framework to direct the detection of salience and improve its capabilities. Our framework stands out from the original contrastive learning approaches by incorporating green-channel knowledge relevant to dual-view fundus images and saliency guidance. Furthermore, the design of the projection head, consisting of a multi-layer perceptron with a residual skip connection, is crucial in harnessing the detailed features that govern the contrastive loss. This combination efficiently produces a robust and complete picture of DR lesions, hence helping to determine the precise DR grade. Incorporating existing knowledge about the dual-view green channel and using a skip connection in the projection head demonstrates a unique and innovative approach, emphasizing the progress achieved in our suggested methodology.

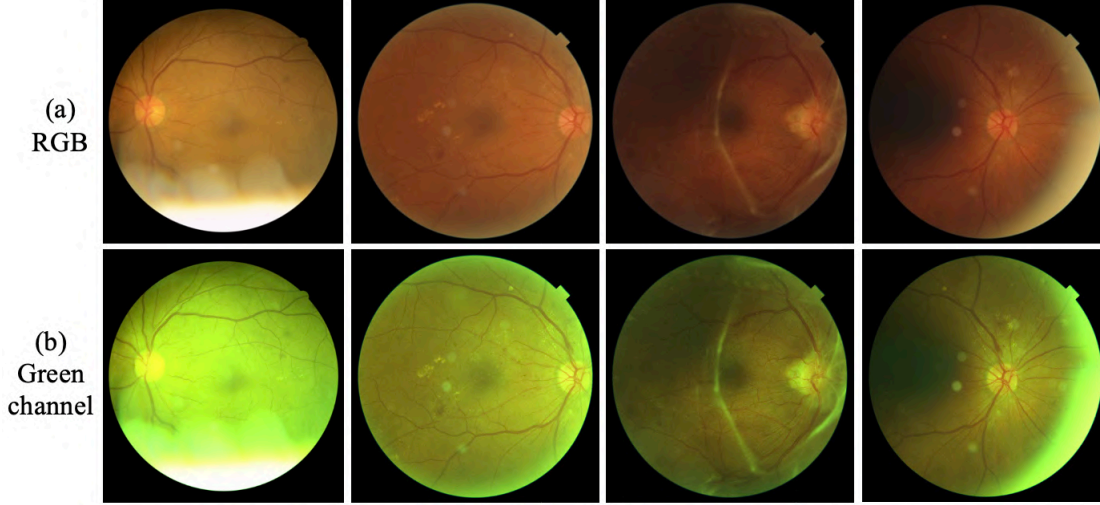


Figure 6.3: Channel-Based Extraction: Selected dual-view fundus images from ODM dataset. (a) Dual-view fundus images in RGB color space. (b) Corresponding green-channel fundus images.

The distinct properties of the green channel compared to the red and blue channels have been extensively documented in the literature due to the twice-sampling rate, resulting in improved sensitivity for capturing visual details [164]; hence, holding more detailed texture information. Based on empirical evidence from clinical practice using fundus cameras, it has been observed that the human visual system shows a higher sensitivity to green color, highlighting the importance of the green channel in analyzing fundus images [138], [165], [137]. As a result, the green channel contains significant essential information that contributes to the model’s performance.

To provide evidence for these claims, we examined 100 fundus images picked randomly from the ODM dataset, as depicted in Figure 6.4. The mean signal-to-noise ratio (SNR) was calculated for each color channel, indicating that the green channel continuously displayed a higher SNR compared to the red and blue channels, especially

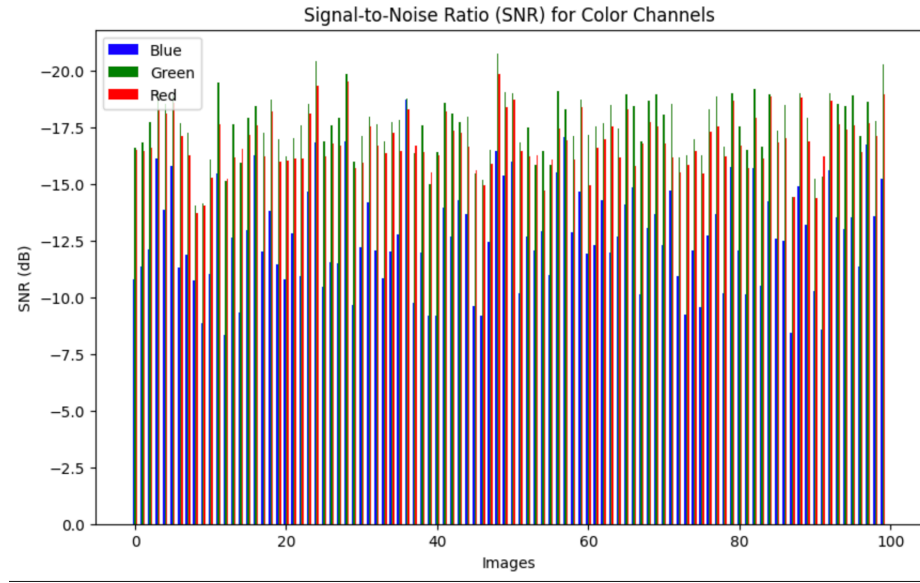


Figure 6.4: *Signal-to-noise ratio (SNR) in different fundus image channels (Red, Green, and Blue) of 100 images randomly chosen from the ODM dataset. The green channel has the highest signal-to-noise ratio compared to other blue and red channels.*

in the presence of noise. Furthermore, because the green channel has a sampling rate that is twice as high as the red and blue channels, it is more effective in maintaining the structure and finer details of the image. Furthermore, it has been observed that the green channel displays a greater abundance of texture patterns than the other channels, especially the vessels. Significantly, the analysis of the fundus images from the ODM dataset demonstrates that the green channel exhibits greater detail complexity than the red and blue channels. Similarly, Figure 6.3 displays samples of ODM fundus images alongside their associated green-channel images. Evidently, the green dual-view samples revealed a greater level of detail compared to the RGB samples.

By leveraging these unique attributes, we designate the pre-existing understanding that the green channel exhibits a greater signal-to-noise ratio (SNR), a superior sampling rate, and, thus, superior channel quality compared to the red and blue chan-

nels. The suggested framework in this paper utilizes valuable knowledge carefully and strategically to use the green channel. This improves our model’s performance when applied to dual-view fundus images. Using dual-view fundus images improves the degree of retinal examination, allowing ophthalmologists to gain a more comprehensive understanding of the DR grade in clinical settings. This, in turn, improves the performance of the DR classification models [107].

6.6.3 Channel-Based Saliency Integration for Enhanced Fundus Analysis

We leveraged saliency object detection (SOD) specifically on green channel images, enabling us to focus on areas of interest and extract more meaningful information. By employing a static fine-grained saliency detection method, as outlined in [155], we constructed saliency maps for all green-channel dual-view fundus images. This technique is especially effective for medical imaging, as it calculates saliency by analyzing the contrasts between the center and surrounding areas of an image. Figure 6.5 presents two saliency maps for comparison: one derived from the RGB fundus image and the other extracted specifically from the green channel of the same image. The saliency map from the green channel highlights critical features with greater clarity, demonstrating the effectiveness of channel-specific saliency detection in fundus imaging. The saliency map derived from the green channel distinctly emphasizes the region of interest, highlighting specific details in the fundus image.

The importance of saliency maps lies in their ability to aid the query encoder.

The momentum encoder evaluates the saliency region of a fundus image, allowing for the exclusion of patches with the lowest saliency scores [126]. Consequently, the most prominent regions with the highest scores are selected to guide the query encoder, ensuring that the significant areas in the fundus image, as revealed by the green channel, are effectively utilized. This strategic use of channel-based saliency detection underscores the strength and innovation of our approach, significantly contributing to the precision and reliability of our framework in detecting diabetic retinopathy lesions.

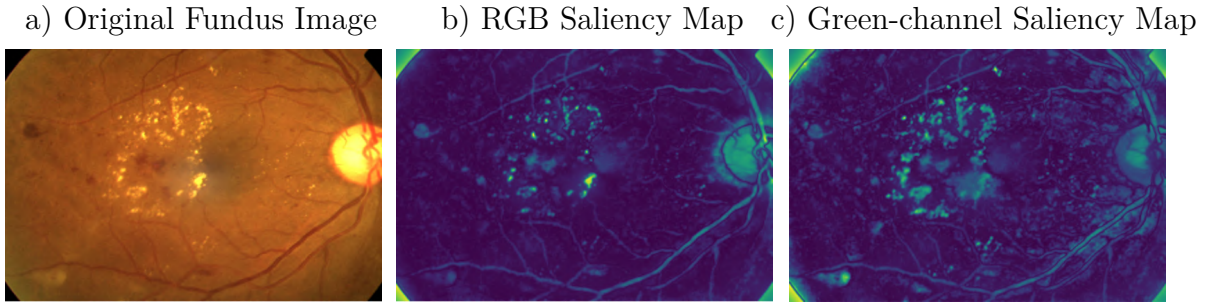


Figure 6.5: *The visualization of the salient maps of a fundus image. a) original fundus image in RGB color space. b) The salient map is detected from the RGB fundus image. c) The salient map is detected from the green-channel fundus image.*

6.7 Experimental setup

6.7.1 Pre-training Dataset

The variants of the ODM dataset with 5100 samples of fundus images in dual-view are used as the pretraining dataset. For different experimental settings in ablation studies, ODM variants OD, MA, ODM-S, and ODM-G are employed to study the effect of the proposed components in our framework, CGSSL. The model is trained with unlabeled samples in the SSL pretraining task (Figure 6.2).

6.7.2 Evaluation Datasets

This study utilizes three evaluation datasets, APTOS, DDR, and Messidor-2, to explore and assess the knowledge transfer from the pretraining task using unlabeled samples to the downstream task using labeled samples. The three datasets consist of single-view data, yet dual-view fundus images serve as the source of transferred knowledge.

6.7.3 Implementation details

Pretraining settings: All pre-training models are trained from scratch. Following the same settings as [126], we used AdamW optimizer [124] with an initial learning rate of 1×10^{-3} and a weight decay of 0.1. In the first 20 epochs, we use a learning rate warmup [125], followed by a cosine annealing schedule during the remaining epochs. We train the model for **300 epochs** with a mini-batch size of 128 on one NVIDIA RTX 3090 GPU. The transferred weights, learned from the proposed SSL framework, were further fine-tuned on the downstream task over the course of **25 epochs** to optimize performance and adapt to the specific characteristics of the task.

Encoders: In this work, we adopted ViT-s as feature extractors in all experiments. The ViT-S contains 12 transformer blocks, each with a six-headed self-attention module and a 384-dimensional hidden size for embedding features. For ViT-S, the patch size employed is 16×16 , yielding a sequence length of 196 for an image with dimensions of 224×224 . Notably, the ViT-S architecture encompasses 21 million trainable parameters, reflecting its complexity and capacity for representation learning.

6.7.4 Self-supervised Evaluation Protocols

This study utilizes a widely used evaluation technique to determine the effectiveness of the acquired representations from each pretraining model in assessing DR. This evaluation technique follows established guidelines by utilizing a standardized evaluation metric, which guarantees that the assessment process is comparable and reliable. Using a rigorous evaluation process, the study thoroughly evaluates and compares the effectiveness of the learned representations created by each pretraining model for the given task. To evaluate the representations learned by our proposed SSL framework, we employed two evaluation protocols: fine-tuning and linear evaluation [126]. Each of these protocols assesses the representations from different aspects, providing a comprehensive evaluation of the model’s effectiveness.

Fine-tuning Evaluation. Fine-tuning assesses the learned representation’s transferability to downstream tasks. Pre-trained parameters of the feature extractor are transferred to models used in these tasks. To evaluate the transfer learning capacity, we unfreeze and fine-tune the feature extractor followed by a linear classifier in supervised downstream tasks. This protocol helps to assess the model’s adaptability and effectiveness in learning new tasks by leveraging the pre-trained knowledge.

Linear Evaluation. Linear evaluation assesses the quality of the learned representation. In this approach, the pretrained feature embedding is kept fixed while a linear classifier is trained on top of it in a fully supervised manner. The performance on downstream tasks is then used as an indicator of the quality of the learned representations. This evaluation protocol primarily focuses on the inherent quality and

generalizability of the learned features without further tuning the feature extractor itself.

6.8 Analysis and Discussion

6.8.1 Transferability and Generalisability

To demonstrate the effectiveness of CGSSL, a comparative analysis was performed between our CGSSL and state-of-the-art (SOTA) SSL methods on three benchmark datasets: , Messidor-2, and DDR. Three evaluation protocols were applied to assess model transferability and representation quality: fine-tuning evaluation (transfer capability), linear evaluation, and k -NN classification (learned representation quality).

Table 6.1 highlights our model’s superior transfer capability in fine-tuning protocol, outperforming SOTA models even with limited data. While SOTA models used the EyePACS dataset (80,000 single-view fundus images) for SSL pretraining, our CGSSL was trained on 5,089 green dual-view fundus images. Our model excelled on the dataset, surpassing ViT-SSL and ResNet-SSL methods by nearly 2%, achieving a kappa score of 0.9252 with limited pretraining samples, similar to results reported in [126] but with larger data. In addition, CGSSL exhibited notable superiority with APTOS in k -NN evaluations, achieving high kappa scores that often surpassed models like DiRA, MoCo-v2, and PCRL.

CGSSL consistently demonstrated strong performance in fine-tuning and linear evaluations with Messidor-2 dataset, reinforcing its adaptability and effectiveness. It achieved a kappa score of 0.6463 in fine-tuning, outperforming MAE (0.5135) and

Table 6.1: Comparison results with state-of-the-art self-supervised models on three evaluation benchmark datasets: APTOS, Messidor, and DDR. Results are in Kappa score. FT stands for Fine-tuning.

Pretraining			No. Images	APTOS		Messidor		DDR					
Methods	Arch.	Dataset		Views	FT	Linear	k -NN	FT	Linear	k -NN			
DiRA [166]	RN50	EyePACS	80K	Single	0.9234	0.698	0.5992	0.7407	0.2444	0.1548	0.7868	0.5186	0.2599
MoCo-v2 [167]	RN50	EyePACS	80K	Single	0.9020	0.8182	0.7434	0.6567	0.4703	0.3396	0.7241	0.6002	0.4312
PCRL [168]	RN18	EyePACS	80K	Single	0.8941	0.7547	0.7127	0.6663	0.3440	0.1935	0.7191	0.5420	0.4121
SimCLR [151]	ViT-S	EyePACS	80K	Single	0.9034	0.8679	0.8329	0.6547	0.5928	0.4087	0.7215	0.6769	0.5353
MoCo-v3 [153]	ViT-S	EyePACS	80K	Single	0.9015	0.8159	0.8068	0.6416	0.5982	0.4041	0.7141	0.6608	0.5605
DINO [154]	ViT-S	EyePACS	80K	Single	0.9020	0.8034	0.7944	0.6430	0.4906	0.4065	0.6849	0.6638	0.4986
MAE [159]	ViT-S	EyePACS	80K	Single	0.9034	0.8324	0.7564	0.5135	0.4350	0.2139	0.7240	0.5931	0.387
SSiT [126]	ViT-S	EyePACS	80K	Single	0.9297	0.8096	0.8465	0.7753	0.6723	0.4942	0.8188	0.7133	0.5889
CGSSL (Ours)	ViT-S	ODM-G	5089	Dual	0.92520.85060.83790	0.64630.39790.37490	0.67850.59120.4502						

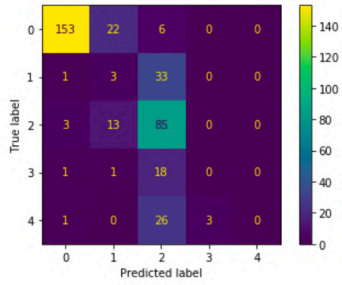
achieving results similar to DINO (0.6430) and MoCo-v3 (0.6416). Notably, CGSSL accomplished this with a significantly smaller sample size compared to the other models.

The overall results highlight CGSSL’s innovative training approach and its ability to generalize well across diverse medical imaging datasets. However, fine-tuning evaluation on the DDR dataset was less satisfactory, possibly due to unique DDR dataset features not captured during pretraining. Despite this, CGSSL remains competitive, transferring robust knowledge from a relatively small dataset.

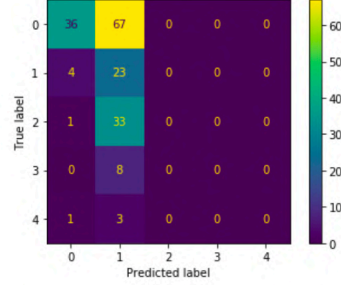
6.8.2 Adjacent DR Grades

The comparative analysis of the ViT-S16 model pretrained with ImageNet weights and our proposed CG-SSL model highlights significant differences in performance, particularly in the context of false positive rates and overall classification accuracy across diabetic retinopathy (DR) stages, Figure 6.6. Pretraining a model on general image datasets like ImageNet often limits its ability to accurately capture domain-specific features, especially in medical imaging tasks where subtle differences between classes can be crucial. This limitation is clearly reflected in the confusion matrices for the APTOS, Messidor, and DDR datasets.

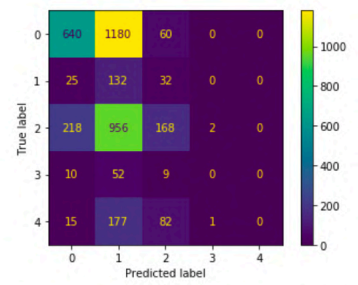
In the first row, the confusion matrices corresponding to the ViT-S16 model pretrained on ImageNet show scattered attention across the DR stages. In the APTOS dataset, for example, while the model shows some ability to classify DR0 (no DR), it misclassifies a notable number of DR1 and DR2 cases into adjacent classes. This results in a higher false positive rate, especially with DR1 cases misclassified as DR0,



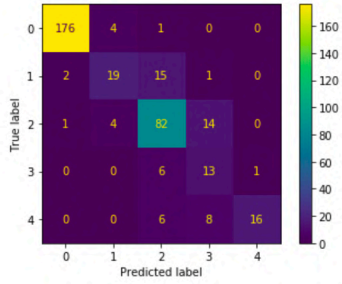
(a) ViT-S16 with APTOS dataset.



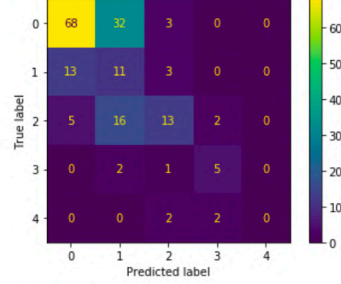
(b) ViT-S16 with Messior dataset.



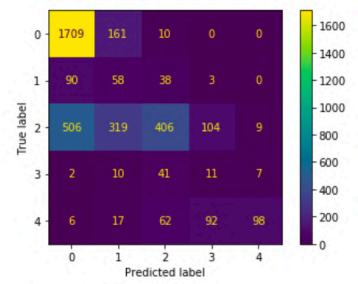
(c) ViT-S16 with DDR dataset.



(d) CG-SSL with APTOS dataset.



(e) CG-SSL with Messior dataset.



(f) CG-SSL with DDR dataset.

Figure 6.6: Fine-tuning the three evaluation datasets for the downstream task using: a,b and c: ViT-S16 model with weights transferred from ImageNet, and d, e and f ViT-S16 with weights transferred from our proposed CG-SSL.

which is a critical concern in medical diagnoses, as it can lead to missed or delayed treatment for patients. Moreover, the model performs poorly in differentiating DR2, with significant overlap between classes. The Messidor dataset reveals similar issues, where DR0 is frequently misclassified as DR1, leading to a high false positive rate for DR1. This misclassification suggests that the ImageNet-pretrained model struggles to capture the unique features distinguishing early stages of DR. The DDR dataset further amplifies these limitations, with large portions of DR0 being incorrectly predicted as DR1, and DR2 predictions dispersed across multiple classes, reflecting an overall lack of precise class separation.

In contrast, the confusion matrices in the second row, representing the ViT-S16 model pretrained with CG-SSL, demonstrate a significant improvement in both accuracy and class separation across all datasets. The CG-SSL model shows enhanced sensitivity to the subtle features that distinguish the different DR stages, resulting in a substantial reduction in false positives. In the APTOS dataset, the CG-SSL-pretrained model not only improves the classification of DR0, but also significantly reduces the misclassification of DR1 and DR2 cases. This improvement is particularly noteworthy given the challenge of classifying minority classes in imbalanced medical datasets. Similarly, in the Messidor dataset, the CG-SSL model exhibits far fewer misclassifications between DR0 and DR1, reflecting its improved ability to distinguish between early DR stages. This reduction in false positives ensures that fewer healthy patients are mistakenly diagnosed with DR, a common issue in traditional models pretrained on non-medical datasets.

The most striking improvement, however, is observed in the DDR dataset, where

the CG-SSL model demonstrates a marked reduction in misclassifications across all DR stages. The model shows a much tighter clustering of predictions around the true labels, with a significant decrease in the number of DR0 cases misclassified as DR1. Additionally, the DR2 predictions are far more concentrated compared to the ImageNet-pretrained model, indicating that the CG-SSL model has a superior ability to recognize and isolate key features associated with intermediate DR stages. This improved focus and precision in classification is critical for medical applications, where even minor improvements in accuracy can have significant clinical implications.

The overall performance gains of the CG-SSL model can be attributed to its ability to leverage domain-specific self-supervised learning, which enables the model to capture more relevant and nuanced features from fundus images. Unlike ImageNet pretraining, which is based on general image categories, CG-SSL adapts to the unique characteristics of medical imaging, where subtle variations between classes, such as the appearance of lesions or abnormal structures, play a vital role in diagnosis. The CG-SSL model’s ability to transfer this learned knowledge effectively to downstream tasks like diabetic retinopathy grading highlights its potential as a more reliable and clinically useful approach compared to traditional supervised pretraining methods.

In conclusion, the CG-SSL-pretrained ViT-S16 model not only reduces the false positive rate across all three datasets but also shows a significant improvement in handling underrepresented DR stages, which are often challenging for models trained with conventional methods. The results clearly indicate that CG-SSL offers a more effective solution for DR classification by capturing critical features with greater precision, making it a valuable tool for advancing medical image analysis in clinical

settings. This analysis suggests that self-supervised learning methods like CG-SSL could play a pivotal role in improving the performance of machine learning models in specialized fields such as ophthalmology, where accurate detection and classification are paramount for early diagnosis and effective treatment.

6.8.3 Vision Transformer Supervised Learning

Table 6.2: *Supervised Learning: performance of Vision Transformer variants (ViT) architectures trained on APTOS dataset with random weights.*

Arch. Size	Variants	Param.	κ	Acc. (%)
Tiny	ViT-T-p16	5.7M	0.7200	64.76
	ViT-T-p32	5.7M	0.6940	62.05
Small	ViT-S-p16	22M	0.7321	65.31
	ViT-S-p32	22M	0.7465	65.58
Base	ViT-B-p16	86M	0.7409	68.83
	ViT-B-p32	86M	0.7533	67.47

CGSSL pre-trained on the ODM-G dataset and fine-tuned on APTOS, achieved a kappa score of 0.9252, outperforming all ViT architectures in Table 6.2. Despite using the small variant as the backbone, CGSSL’s transferred representation knowledge was superior. This improvement highlights the effectiveness of CGSSL’s self-supervised learning approach, leveraging dual-view fundus images to enhance feature extraction and model generalization.

Table 6.3: *Ablation Study: the efficiency of CGSSL components standalone and combined. M_1 to M_6 represent different models with/without a particular component. Results are reported in the Kappa score.*

Models	Dual-view	Green	Residual	APTOS	Messidor-2	DDR
M_1	✓	×	×	0.8904	0.5635	0.6148
M_2	×	✓	×	0.8580	0.3748	0.5562
M_3	×	×	✓	0.8821	0.4317	0.5532
M_4	✓	✓	×	0.9021	0.5148	0.6569
M_5	×	×	×	0.8751	0.5056	0.6640
M_6 (Ours)	✓	✓	✓	0.9252	0.6463	0.6785

6.9 Ablation Studies

This section represents different experiments further to analyze the impact of three components in CGSSL: dual-view fundus images, green-channel prior knowledge, and the projection head with the residual connection.

6.9.1 M_1 : Single View vs. Dual View Fundus Images

We first compared the performance of the single-view and dual-view fundus images using the SSL guided by SOD. The semantic knowledge offered by the two views of a patient’s retina affected by DR exposes more information on the diagnosis case. Consequently, training a model to learn meaningful representation utilizing dual-view fundus images would boost performance in the downstream task with limited samples. Table 6.4 shows gradual improvement in performance from single-view to dual-view.

However, the result shows overall improvement when maintaining the same number of samples as the OD and MA datasets, but with dual-view fundus images. By training the pretraining model on the ODM-S dataset, the acquired result is 0.8866

Table 6.4: *The Fine-tuning performance evaluation illustrates the impact of dual view over single view fundus images on the APTOS dataset.*

Pretraining Model				Fine-tuning
	Model	Arch.	Dataset	Kappa
Single view	SSL+SOD	ViT-S	OD	0.8751
			MA	0.8779
Dual view	SSL+SOD	ViT-S	ODM-S	0.8866
			ODM	0.8904

κ . When we examine these findings about the outcomes from single-view analyses, it becomes evident how integrating dual-view perspectives enriches knowledge representation, leading to a more profound understanding of semantic information. This enrichment, in turn, contributes to a notable enhancement in performance in subsequent tasks. Notably, the augmentation of sample size through training the pretraining model with larger dual-view fundus images (ODM) demonstrates a substantial performance improvement, achieving a kappa coefficient of 0.8904. This discussion highlights the efficacy of leveraging a more extensive dataset for training, underscoring its positive impact on overall model performance.

6.9.2 M_2 : The Sole Impact of Green-channel

The green channel is known for its rich information. We evaluated our CGSSL framework by isolating the dual-view and residual, and focusing solely on the green channel. Table 6.3 shows M_2 model results on APTOS with 0.8580, Messidor-2 with 0.3748,

and DDR with 0.5562. Comparing the results to the M_6 model indicates that the green channel does not have that effect on boosting the model. However, the green channel does embed knowledgeable information and is more efficient than the RGB color space.

6.9.3 M_3 : The Effect of Residual Connection

As previously discussed, the projection head is essential for learning through contrastive loss. This model introduces the residual connection in the projection head design. The projection head contains three fully connected layers, followed by batch normalization and ReLU. Introducing the residual function to the projection head design could mitigate the vanishing gradient problem and facilitate information flow. The M_3 model is trained with RGB single-view fundus images; this setting excludes the green channel and dual-view effect and focuses on the projection head with the residual connection. Table 6.3 shows M_3 's results on APTOS, Messidor-2, and DDR have 0.8821, 0.4317, and 0.5532 scores, respectively.

6.9.4 M_4 : The Impact of Channel-based Knowledge from Dual-view Fundus Images

This model exposes rich knowledge representation by utilizing the green channel of fundus images, either in the pretraining model or the downstream fine-tuning. Different variation settings are used to empirically discover the best utilization and configuration of the green channel. The fine-tuning evaluation was applied to the APTOS

Table 6.5: *The fine-tuning evaluation performance of the SSL baseline on CNN. M_4 model focuses on green-channel with the influence of dual-view fundus images. Results are reported in the Kappa score.*

Pretraining SSL Model					Fine-tuning Evaluation			
Model	Arch.	Dataset	Image Input	Saliency	APTOS - RGB kappa	APTOS - RGB Acc.	APTOS - Green kappa	APTOS - Green Acc.
SSL+SOD	ViT-S	ODM-G	Green	Green	0.9021	78.31	0.7750	89.89
SSL+SOD	ViT-S	ODM	RGB	Green	0.8922	79.4	0.7332	64.22

dataset in two color settings: RGB and green channel.

The green channel serves as the basis for both the image and saliency map inputs, and the second setting is where the input fundus image is represented in RGB color space and the saliency map is extracted from the images' green channel.

The Critical Role of Green Channel Saliency Mapping: The green channel serves as the basis for both the image and saliency map inputs. To explore the effect of the green channel on the downstream task, fine-tuning was performed on two variations of CNN: the original CNN with RGB fundus images and CNN, which contains green-channel fundus images. The best result obtained is by training the pre-trained model with the two inputs both in green-channel format and fine-tuning it on an original CNN with RGB format, achieving 0.9021 κ , Table 6.5. The performance is dramatically decreased when the model is evaluated on the CNN dataset in the green-channel image with 0.7750 κ .

The influence of RGB With Green Channel Saliency Map: In order to discover the influence of RGB images as input to the pretraining model, M_4 receives an input fundus image in RGB color space, and the saliency map is detected from the green

channel 6.5. Results achieved by fine-tuning on CNN-RGB outperformed CNN-G with 0.8922 κ and 0.7332 κ , respectively. In general, results demonstrate that the pre-trained model trained with the green-channel fundus images and fine-tuned on the RGB dataset outperforms other settings in the experiment. Obviously, the green channel exposes more objects to salient detection, which effectively boosts the model under the fine-tuning protocol. This proves the efficacy of green channels in learning representations and regions of interest using the saliency detection method.

6.9.5 M_5 : SSL baseline

The M_5 is the baseline model, where we excluded all the three components introduced in CGSSL. As evidenced in Table 6.4, the kappa score in the context of single-view analyses manifests analogous outcomes. The OD and MA datasets, acquired with OD-centric and MA-centric perspectives, respectively, demonstrate that the pretraining model assimilates embedding, leading to relatively comparable outcomes, attaining a kappa score of 0.8751 on the OD dataset and 0.8779 on the MA dataset. In contrast, when merging the fundus image views, as shown in the dual-view of Table 6.4, the result increased by 1%. This underscores the influence of dual-view pretraining on knowledge acquisition by providing the self-supervised learning model with a broader spectrum of domain information. This observation reveals that, during single-view training, either with OD or MA, the model acquires a similar representation and achieves a similar kappa score of 0.8751 on the OD dataset and 0.8779 on the MA dataset.

6.10 Remarks

Our work highlights the critical role of leveraging green-channel sailent detection and dual-view fundus images for embedding rich, domain-specific knowledge into self-supervised learning models. The projection head design, which captures intricate and detailed features, has proven instrumental in enhancing overall model performance. The proposed CG-SSL framework marks a significant step forward in pretraining contrastive self-supervised learning models on data that closely reflects the structural intricacies of retinal images, enabling a more effective and meaningful learning process. The results from fine-tuning have been highly promising, showing the potential for our approach to generalize to other retinal diseases such as glaucoma, cataracts, and myopia. Importantly, CG-SSL has achieved competitive performance with limited data, outperforming other methods in the domain and demonstrating the power of leveraging task-specific knowledge in self-supervised learning.

One of the key strengths of CG-SSL lies in its ability to address the persistent issue of false positives, particularly in medical image classification. By learning from detailed, representative features embedded in green-channel fundus images, the model is better equipped to distinguish between subtle variations in disease stages, leading to a notable reduction in false positive rates across all datasets. This improvement is critical in clinical settings, where reducing false positives can prevent unnecessary treatments and alleviate patient anxiety. Additionally, the power of CG-SSL is manifest in the self-attention visualizations, where the model consistently focuses on relevant pathological regions such as lesions and abnormal structures. These visual-

izations confirm that CG-SSL effectively captures essential features and emphasizes critical areas in the images, which is vital for accurate disease classification.

The CG-SSL model’s superior performance across various datasets, despite the limitations of available training data, underscores its potential to establish new benchmarks in self-supervised learning for medical imaging. Its robustness, efficiency, and adaptability make it an invaluable tool for both research and practical applications, setting the stage for more efficient and accurate self-supervised learning solutions in the future.

Moving forward, we aim to enhance CGSSL model by extracting high and fine-grained features which are remain a crucial advancement for SSL to capture comprehensive, representative knowledge.

CHAPTER 7

MULTI-SCALE CONTRASTIVE SELF-SUPERVISED LEARNING FOR DIABETIC RETINOPATHY GRADING

7.1 Introduction

Automated deep learning systems for DR grading have shown significant potential in aiding clinicians with diagnosis. However, the development of robust models for this task is challenging due to the complex, domain-specific nature of retinal images and the limited availability of labeled medical data that could boost DL models.

In the field of medical imaging, transfer learning has been widely adopted, often leveraging models pre-trained on large-scale datasets like ImageNet[121]. These

models, such as ResNet, Inception, DenseNet, EfficientNet, and ViTs, are fine-tuned to perform specific medical tasks. While this approach has proven effective in many computer vision tasks, its utility in medical imaging is not as straightforward. The significant differences between medical images and the general images in datasets like ImageNet raise questions about the efficacy of such pre-trained models in the medical domain. Although pre-trained models typically outperform those trained from scratch in low-data scenarios, their weights do not always transfer effectively to medical tasks, limiting their potential in this specialized field[116].

In this chapter, we integrated a novel module into CG-SSL model proposed in Chapter 6 , the Multi-scale Feature Extraction Encoder (MFEE), which enhance the model’s ability to capture and integrate both high-level and fine-grained features, improving the overall quality of learned representations. Followed by another module that learned the fused features, to excel the learning for more robust knowledge representation. Specifically, this addressing our **RQ5** to extract high and fine-grained features for better knowledge representation. In addition, this chapter will fulfill **Gap-3, Gap-4, Gap-5, Gap-6 and Gap-8**.

Our main contributions in this chapter are as follows:

- We introduce two innovative modules within the self-supervised learning (SSL) framework: the **Multi-scale Feature Extraction Encoder (MFEE)** and the **Deep Learner module (DL_e)**. Together, these modules improve the model’s ability to simultaneously capture and integrate both high-level abstract features and fine-grained details, leading to richer and more robust learned knowledge

representations.

- By leveraging the MFEE design, our proposed **MsSSL framework** demonstrates consistent performance gains across multiple model backbones, including **Vision Transformers** and **FPN-ResNet**, within a unified self-supervised learning setting (**MoCo-v3**). This approach ensures more effective feature extraction at different scales and levels of abstraction.
- Our model exhibits superior performance in distinguishing the later stages of **Diabetic Retinopathy (DR)**, showcasing how the **MsSSL framework** effectively learns domain-specific features and transfers relevant knowledge, making it particularly well-suited for the downstream medical image classification task.

7.2 Background and Motivation

Recent advancements in self-supervised learning (SSL) have shown promise in addressing these limitations by enabling models to learn representations directly from the target domain[133], [169], [156]. Studies have demonstrated that SSL can outperform traditional supervised learning, particularly in medical imaging tasks where data is scarce and domain-specific. For example, self-supervised learning frameworks like SimCLR[151], SwAV[170], and DINO[154] investigated in [129], have been shown to produce richer and more transferable embeddings when applied to medical datasets, such as X-rays, dermatology images, and retinal fundus images. These approaches suggest that domain-specific pre-training and advanced model components are crucial for enhancing performance in medical imaging.

The backbone architecture in SSL plays a pivotal role in building scalable and effective solutions. Traditional CNN architectures, such as ResNet[8], often struggle to scale with model and dataset size due to inherent inductive biases, including locality and weight sharing[171]. These biases lead CNNs to focus primarily on local regions within an image, potentially resulting in representations that emphasize local textures while neglecting global contextual information. This limitation is particularly significant in medical imaging, where global contextual understanding is often critical for accurate diagnosis.

In response to the challenges of diabetic retinopathy grading, we propose the Multi-scale Self-Supervised Learning (MsSSL) model, which integrates the strengths of ViTs and CNNs within a single self-supervised learning framework. Specifically, our approach incorporates the detailed multi-scale features extracted by a Feature Pyramid Network (FPN) [172] with a ResNet backbone, along with self-attention mechanisms found in ViTs [123], which excel at capturing global, contextualized representations of input images. These fused features are then fed into the Deep Learner module (DLe), where they undergo systematic upsampling and refinement to enhance their spatial resolution and learn complex patterns and relationships within the data and capture both high-level and fine-grained information. The integration of these components results in more comprehensive and robust feature representations, tailored specifically for accurate DR grading.

7.3 SSL Transfer Learning in Medical Imaging

Transfer learning is a widely recognized approach that effectively utilizes unlabeled data from either the source domain or the target domain. It is fundamentally a semi-supervised technique that involves domain adaptation[116]. In general terms, transfer learning can be described as the process of training a model on a large-scale dataset for an initial task and then applying that pre-trained model to learn a new, target task[110]. In most computer vision tasks, the learning process typically leverages models pre-trained on ImageNet[58], the largest publicly available image dataset. These pre-trained models, such as ResNet50, Inception, DenseNet, EfficientNet, and Vision Transformers (ViTs), are fine-tuned to improve performance on specific tasks. However, medical images differ significantly from the general images found in ImageNet, leading to debates regarding the efficacy of these pre-trained models in medical imaging tasks [117], [118], [119].

In supervised learning, pre-training on ImageNet can significantly enhance the performance of convolutional neural network (CNN) models, especially when dealing with large datasets. However, in medical imaging, where data collection and annotation are arduous and time-consuming, the utility of ImageNet pre-trained models is less clear. Although these pre-trained models perform better than training from scratch in low-data regimes, their weights do not always transfer effectively to medical tasks[120], [116], [121].

Recently, transfer learning with self-supervised learning has shown promise, where the primary task is to learn representations directly from the target domain [129], [126],

[173], [133], [156], [169]. For instance, a study by [120] compared traditional supervised learning models on X-ray and dermatology datasets with the self-supervised learning framework SimCLR [151]. The results demonstrated a significant improvement with self-supervised learning. Similarly, [174] evaluated several self-supervised learning models on various dermatology datasets, reporting superior performance compared to supervised learning approaches.

Further, [129] investigated the transferability of self-supervised ImageNet models in both ensemble and individual settings, including SimCLR, SwAV, and DINO. This study utilized three medical datasets: tumor detection, diabetic retinopathy classification in fundus images, and chest X-ray images. The findings indicated that self-supervised pre-trained models produced richer embeddings than their supervised counterparts, thereby enhancing the performance of downstream tasks in both linear evaluation and fine-tuning evaluation.

In self-supervised learning, encoder is responsible of extracting features that will be learned on by the contrastive loss. Encoders like Vision Transformers (ViTs) have been explored in this context. Their capability to capture global image features, as opposed to the local features typically captured by CNNs, presents a potential advantage in medical imaging. Preliminary studies suggest that ViTs, when fine-tuned with domain-specific data, can outperform traditional CNNs and enhance the effectiveness of self-supervised learning methods[150], [123].

Our work aims to leverage the potential of combining Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) within a single self-supervised learning framework, to capture both high-level and fine-grained features, thereby enhancing

the learned representations in medical image diagnosis. Alongside extracting these features, we also aim to address the challenge of distinguishing adjacent Diabetic Retinopathy (DR) grades, a common issue in conventional transfer learning from ImageNet, where models often struggle with misclassifications between adjacent grades due to their subtle visual similarities. By improving feature extraction and representation learning, we seek to enable the model to better differentiate between these closely related DR grades, improving overall diagnostic accuracy.

7.4 Multi-scale Feature Fusion

Harnessing multi-scale features from various layers significantly boosts feature extraction, leading to more effective learning during the training process as demonstrated by [172], . A good feature extraction is important in the SSL learning, it is the key point that alleviates the representation learning when it is transferred to a specific downstream task. studies confirm that **high level features contain more abstract semantic information, while low-level features contain more fine contextual information** [175]. In supervised learning either CNN-based or ViT-based, the feature representation is conveyed from single-level, and this is cannot accurately represent all image information. To overcome the issue with single- level features, multi-scale feature fusion has been shown to alleviate the representation, which is usually implemented by combinations, such as element-wise addition or channel-wise concatenation [176], [177]. Other studies instead, apply the fusion from different combination path [178], [179] or from different models[129].

7.5 MsSSL overall architecture

Our proposed architecture differentiates itself from the SOTA MoCo-v3, by incorporating a multi-scale feature extraction Encoder designed to extensively capture both high-level and fine-grained features. Additionally, after obtaining high spatial information from two distinct architecture bases, namely ViT and CNN-based extractors, we introduced a deep learner module to effectively learn these concatenated spatial features. We adopted a segmentation objective task, following the methodology outlined in previous work [126], to capture subtle and local information as well. Building on our baseline CGSSL reported in Chapter 6, we utilized the green channel instead of the RGB fundus image, with the entire training framework based on green-channel fundus images. Furthermore, we employed green salient detection from our earlier work to guide the momentum encoder towards salient regions and to generate the ground truth for the segmentation objective.

7.5.1 Multi-scale Features Extraction Encoder

The Multi-scale Feature Extraction Encoder (MFEE) is crucial for extracting both high-level and fine-grained features, which are essential for effective downstream tasks in self-supervised learning (SSL) frameworks. This module leverages the strengths of two prominent architectures in computer vision: ViT and CNNs, each known for their ability to extract both general and subtle features. These architectures process images differently, providing a complementary approach to feature extraction (add citation). The module consists of three main components: the Global Feature Ex-

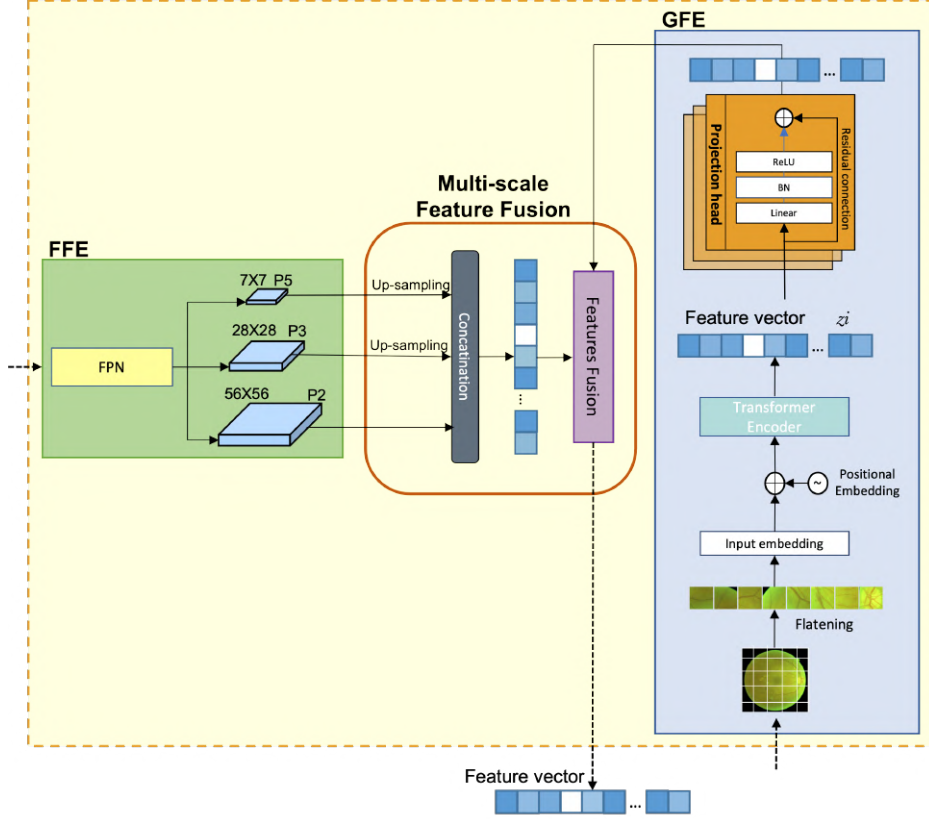


Figure 7.1: Multi-scale Feature Extraction Encoder architecture. GFE: global feature extractor, FFE: fine-grained feature extractor.

tractor (GFE), the Fine-Grained Feature Extractor (FEE) and Multi-scale Feature Fusion (MFF), Figure 7.1. Below, we detail the workings of these components.

Global Feature Extractor The Global Feature Extractor employs a ViT model. In this process, fundus images are divided into patches that are then flattened and processed by the transformer encoder. The resulting feature vectors are passed through an enhanced projection head, following methodologies from our previous work (cite first paper). The final feature vector is then directed to the multi-scale feature fusion component.

Fine-grained Feature Extractor The fine-grained feature extractor uses Feature

Pyramid Network (FPN) based on the ResNet model to capture detailed features, Figure 7.4. The FPN effectively works by constructing a pyramid of features at different scales, enabling the extraction of fine and coarse details from images. In this study, hierarchical features P2, P3, and P5 of the FPN are used, following prior research [180]. The P3 (28x28) and P5 (7x7) features are upsampled to match the P2 (56x56) resolution before being passed to the multi-scale feature fusion component.

Multi-scale Feature Fusion In the multi-scale feature fusion component, the features extracted by the FFE are concatenated. These concatenated features are then fused with the features extracted by the GFE using element-wise addition. This fusion approach is useful when the goal is to emphasize shared information or enhance feature representations with complementary information from multiple scales. This process integrates global and fine-grained features, enhancing the representational power of the combined feature set. The final fused feature vector is then passed to the deep learner component introduced in this work, facilitating further processing within the framework to learn these fused feature representation.

7.5.2 Deep Learner

The Deep Learner (DLe) architecture processes and refines the fused features obtained from two robust feature extractors, which contains: ViT and FPN. The ViT serves as a global feature extractor, leveraging its ability to capture comprehensive, high-level representations from the input images. Simultaneously, the ViT, based on ResNet, functions as a fine-grained feature extractor, adept at identifying intricate and local-

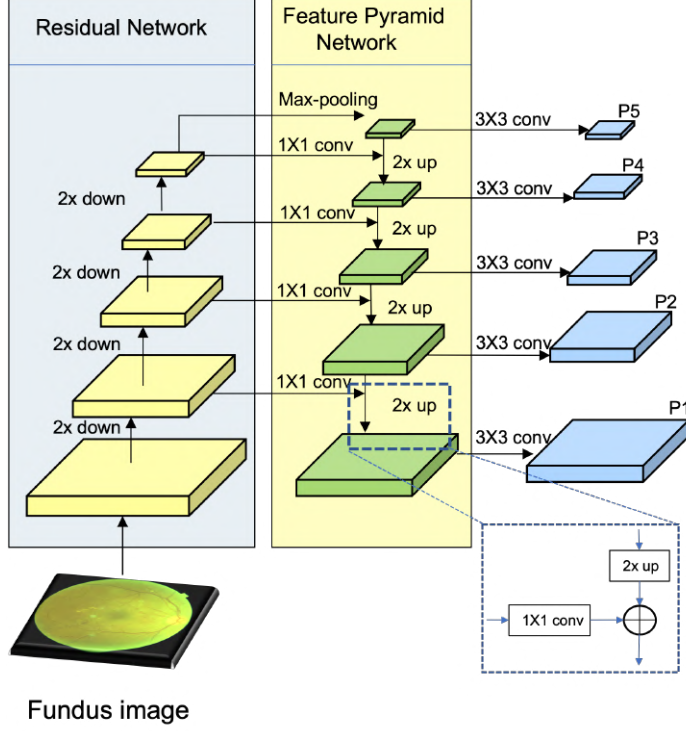


Figure 7.2: *The architecture of Feature Pyramid Network (FPN)*

ized patterns within the data. These features are fused through a multi-scale feature fusion component, combining the strengths of both extractors.

The fused features are then fed into the Deep Learner, where they undergo systematic upsampling through a sequence of ConvTranspose2d layers, also known as deconvolution layers. These layers incrementally increase the spatial dimensions of the input tensor, facilitating the reconstruction of higher-resolution feature maps. Interspersed within this sequence are ReLU activation functions, introducing non-linearity to enable the model to learn complex patterns and relationships within the data.

The final stage of Deep Learner involves the Convolutional Block Attention Module (CBAM) [9], which applies channel and spatial attention mechanisms to upsam-

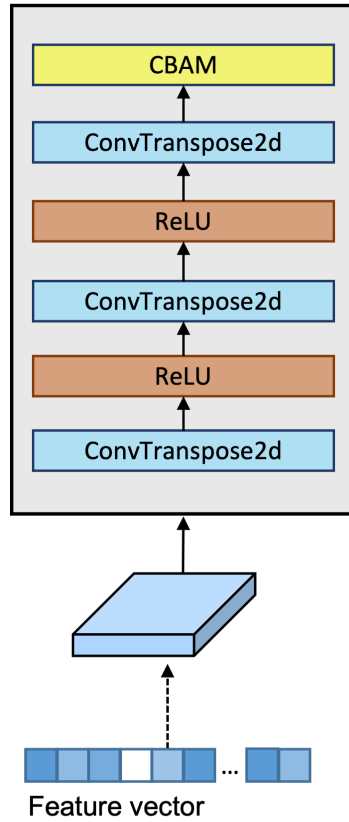


Figure 7.3: *Deep Learner (DLe) architecture refines features through ConvTranspose2d layers with ReLU activations, ending with CBAM for enhanced feature attention and resolution.*

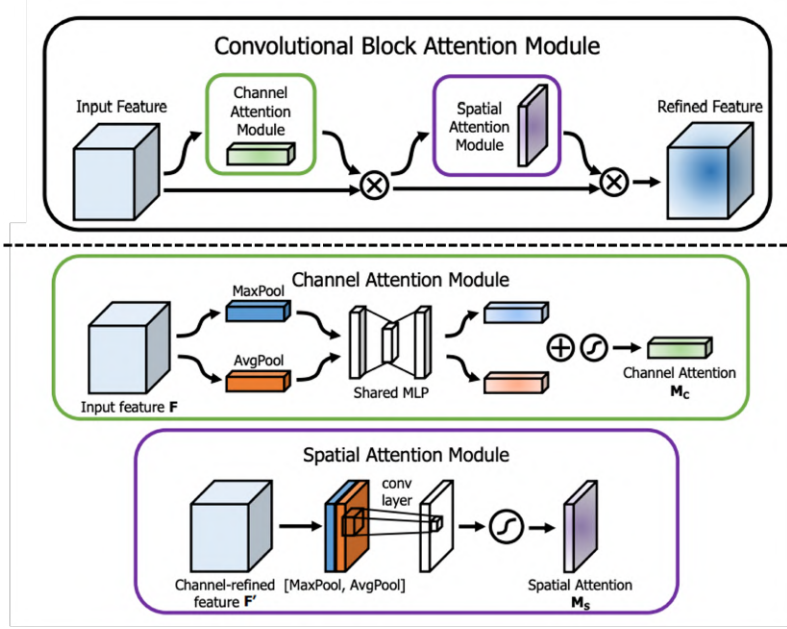


Figure 7.4: The structure of Convolutional Block Attention Module (CBAM) [9].

pled feature maps. This attention mechanism enhances the representational power of the model by focusing on significant features and suppressing less important ones, thereby refining the features before the final output. This architecture, integrating the strengths of ViT and FPN with sophisticated upsampling and attention mechanisms, is highly effective for complex feature extraction and image processing tasks.

7.6 Contrastive loss objectives

Two primary learning objectives are integrated into our MsSSL model: the contrastive loss and the segmentation loss. The contrastive learning objective in this work differs from the segmentation objective in its approach to feature extraction. The contrastive loss aims to maximize the similarity between different augmented views of the same image while minimizing the similarity between views of different images, thus learning

robust and invariant feature representations across various scales. This is achieved using two encoders introduced in this work GFE and FFE, to extract features at multiple scales. The contrastive loss is computed using the InfoNCE formula [162], defined as follows:

$$\mathcal{L}_{\text{contrastive}} = - \sum_{i \in I} \log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k \in A(i)} \exp(\text{sim}(z_i, z_k)/\tau)}, \quad (7.1)$$

where $\text{sim}(z_i, z_j)$ denotes the cosine similarity between two feature vectors z_i and z_j , and τ is a temperature parameter.

Complementing the contrastive loss, the segmentation loss focuses on capturing fine-grained details within fundus images by training the model to accurately segment saliency maps, following the same segmentation loss used in [126]. This task involves generating ground truth saliency masks through thresholding and reconstructing these segmentation maps at full resolution using a lightweight decoder appended to the final transformer block of the query encoder. The segmentation loss is computed using a binary cross-entropy loss function, defined as follows:

$$\mathcal{L}_{\text{seg}}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{\Omega(\mathbf{y})} \sum_{i=1}^{\Omega(\mathbf{y})} -y_i \log(\hat{y}_i), \quad (7.2)$$

where $\Omega(\cdot)$ represents the total number of pixels. To achieve accurate segmentation of the saliency map, the encoder is trained to recognize the shape, color, and texture of the salient regions. This process enhances the model’s ability to identify abnormal regions, which is crucial for the subsequent DR grading task.

By integrating these two learning objectives, the MsSSL model effectively combines

the strengths of both global and local feature learning, thereby enhancing performance in fundus image analysis and DR diagnosis. These two objectives are joint by the following equation,

$$\mathcal{L} = \lambda_{cl} \cdot \mathcal{L}_{cl} + \lambda_{seg} \cdot \mathcal{L}_{seg}, \quad (7.3)$$

where λ_{cl} and λ_{seg} are hyper-parameters that balance the respective contributions of the two objectives.

7.7 Experiment Settings

In this section, we outline the experimental settings used to evaluate the performance of our proposed MsSSL model for diabetic retinopathy (DR) diagnosis.

7.7.1 Encoders.

In this work, we adopted a multi-scale feature extraction encoder that incorporates both ViT and FPN as described in Section 6.6. This proposed encoder is designed to leverage the strengths of both ViT and CNN-based architectures for robust feature extraction.

The ViT module utilizes a series of transformer blocks to process image patches, providing a global feature extraction mechanism. Specifically, the ViT component follows the architecture described previously, with 12 transformer blocks, each containing a six-headed self-attention module and a 384-dimensional hidden size for embedding features. For the ViT module, the patch size employed is 16×16 , yielding a sequence

length of 196 for an image with dimensions of 224×224 .

In parallel, the FPN module is employed to extract fine-grained features. The FPN is based on the ResNet architecture and utilizes hierarchical feature maps from different layers to capture detailed and multi-scale spatial information. The FPN module extracts features at multiple levels (P2, P3, and P5), with the higher-level features (P3 and P5) being upsampled to match the resolution of the lowest level (P2).

The features extracted by the ViT and FPN modules are then fused in the multi-scale feature fusion component, which combines the global and fine-grained features for enhanced representation learning. This multi-scale feature extraction encoder is designed to capture both high-level context and detailed local information, thereby improving the model’s ability to learn robust and discriminative features for diabetic retinopathy grading. This innovative approach aims to enhance the overall performance and interpretability of the model in medical image diagnosis tasks.

7.7.2 Pretraining Datasets.

In this chapter, we utilized our constructed dataset, the ODM dataset, as reported in Chapter 4, for pretraining our proposed model within the CGSSL framework. However, in the enhanced MsSSL model, we expanded our approach by combining the ODM dataset with the DDR dataset. The rationale behind this expansion is due to the increased complexity of the MsSSL framework, where the ODM dataset alone was insufficient to harvest and learn robust features. To ensure the model’s ability to capture diverse and comprehensive feature representations, it was essential to in-

crease the amount of training data. The DDR dataset was selected for this purpose because, after EYEPACS, it is the largest publicly available dataset that we could process with our current computational resources. Furthermore, the DDR dataset includes diverse samples, which are crucial for learning better feature representations, ultimately enhancing the effectiveness of transfer learning for the downstream task. By leveraging the diverse and extensive image data from these datasets without their labels, we aimed to enhance the feature extraction capability of our model, allowing it to effectively generalize across various views and severity levels of diabetic retinopathy.

7.7.3 Downstream Datasets.

To evaluate the features learned during pretraining through our proposed MSSSL framework, we conducted assessments on two datasets: APTOS and Messidor-2.

The Asia Pacific Tele-Ophthalmology Society (APTOS) dataset comprises a total of 5,590 single-view fundus images, with annotations available for 3,662 of these images. For evaluation purposes, we randomly divided the annotated images into training, validation, and testing sets with a split of 70%, 15%, and 15% respectively.

The Messidor-2 dataset, part of the Methods to Evaluate Segmentation and Indexing Techniques in the Field of Retinal Ophthalmology, includes 1,748 single-view fundus images captured with the macula centered. we randomly divided the annotated images into training, validation, and testing sets with a split of 70%, 15%, and 15% respectively.

These datasets allowed us to assess the effectiveness of our pre-trained model on

labeled single-view fundus images, demonstrating the potential for knowledge transfer from the pre-training phase to downstream tasks.

7.7.4 Preprocessing

Channel-based Extraction and Salient Detection. Following our previous experiment in Chapter 6, we extracted the green channel from the fundus images, as it provides higher contrast for the detection of retinal lesions. We then applied salient region extraction to the green channel to emphasize the most critical areas of the images. In addition to these steps, we performed standard preprocessing procedures, including cropping to remove unnecessary borders and normalizing the pixel values to ensure consistent input for the model. These preprocessing steps were crucial in enhancing the quality of the input data, thereby improving the model’s ability to learn meaningful features from the images.

Data Augmentation Operations. Under the self-supervised learning settings, we adopted the data augmentation strategy from SimCLR to enhance the variability and robustness of our training data. This approach includes several key operations: random cropping and resizing to fixed dimensions, ensuring positional invariance; horizontal flipping with a 50% probability to handle orientation changes; color jittering to adjust brightness, contrast, saturation, and hue, simulating various lighting conditions; Gaussian blurring with random kernel sizes to mimic out-of-focus scenarios; and normalization to align pixel value distributions with the pretraining dataset. These augmentations are essential for preventing overfitting and improving the model’s gen-

eralization by exposing it to diverse variations during training.

7.7.5 Pretraining setup

We used AdamW optimizer [124] with an initial learning rate of 1×10^{-3} and a weight decay of 0.1. In the first 20 epochs, we use a learning rate warmup [125], followed by a cosine annealing schedule during the remaining epochs. We train the model for 300 epochs with a mini-batch size of 128 on one NVIDIA RTX 3090 GPU.

7.7.6 Evaluation Protocols

We used as same as the evaluation protocols mentioned in Chapter 6, to evaluate the representations learned by our proposed MsSSL framework, we employed two evaluation protocols: fine-tuning and linear evaluation. Each of these protocols assesses the representations from different aspects, providing a comprehensive evaluation of the model’s effectiveness.

7.8 Results and Analysis

In this section, we present a detailed comparison of the transferability and generalizability of the proposed MsSSL framework. The focus of the analysis is to assess how well the representations learned during the pretraining phase can be transferred to downstream tasks and how effectively the model generalizes across different datasets. The evaluation includes both fine-tuning and linear evaluation protocols, providing a comprehensive understanding of the quality of the learned representations. These

representations are tested on two well-known DR datasets: APTOS and Messidor, allowing us to analyze the model’s ability to generalize to new, unseen data.

7.8.1 Representation Transferability Assessment

On fine-tuning evaluation, Table 7.1 shows the results on the APTOS dataset across various self-supervised learning (SSL) models, highlighting the performance of our proposed model, MsSSL. This table underscores the significance of utilizing weights derived from in-domain images, particularly in the context of medical imaging.

The performance metrics compared include the Kappa (κ), accuracy, and F1 score. The models listed, such as SwAV, SimCLR, S2C2L, DINO, and others, utilize pre-trained weights from different datasets like ImageNet and EyePACS.

Notably, the models employing ImageNet weights, like S2C2L, exhibit reasonable performance with accuract of 83.08 and F1 of 82.27, indicating their potential in general image analysis tasks. However, when the domain shifts to medical imaging, as in the APTOS dataset, the advantage of using in-domain weights becomes apparent. Models like SimCLR pre-trained on EyePACS show improved performance with accuracy and κ values reaching 90.34 and 0.9034, respectively.

Our proposed MsSSL model, utilizing ODM-DDR weights, demonstrates superior performance with the highest κ value of 0.9324, accuracy of 84.55, and an F1 of 72.07. These results affirm that learning representations from domain-specific images significantly enhances the model’s ability to generalize and perform well on downstream

medical imaging tasks. This analysis substantiates the importance of domain-adaptive pre-training for achieving state-of-the-art performance in medical image analysis.

Table 7.1: *Comparison results of fine-tuning evaluation on APTOS dataset.*

Model	Weights	κ	Acc.%	F1%
SwAV[129]	ImageNet	0.8293	-	-
SimCLR [129]		0.8264	-	-
S2C2L[181]		0.8907	83.08	82.27
DINO[154]		0.8365	-	-
DVME[129]		0.8242	-	-
SimCLR-RT[182]		-	83.10	-
SimCLR-FT[182]		-	83.32	-
MoCo-v2	EyePACS	0.9020	-	-
PCRL		0.8941	-	-
SimCLR		0.9034	-	-
MoCo-v3		0.8824	-	-
DINO[114]		0.8824	-	-
MAE		0.9034	-	-
MSN[114]		0.8887	-	-
MsSSL(Ours)	ODM-DDR	0.9324	84.55	72.07

7.8.2 Representation Quality Assessment

Various self-supervised learning (SSL) models evaluated on the APTOS dataset through linear evaluation represented in Table 7.2, which assesses the quality of the learned representations. The table underscores the significance of using in-domain weights for medical image analysis by comparing the proposed MsSSL model with other SSL approaches, using (κ) as the evaluation metric.

The MsSSL model, utilizing weights derived from the ODM-DDR dataset, achieves

a superior κ value of 0.8698, demonstrating the highest performance among the compared models. Other models, such as SwAV, SimCLR, DINO, and DVME, which employ ImageNet weights, report κ values of 0.7617, 0.6989, 0.7790, and 0.8242 respectively. Additionally, models leveraging alternative pretraining datasets, like EyePACS for PCRL, show a κ value of 0.7547.

Further comparisons include DiRA and MoCo-v2, which attain κ values of 0.6980 and 0.8182 respectively, as well as MoCo-v3 and DINO, which record values of 0.8159 and 0.8034. The MAE model also performs well, with a κ of 0.8324. These results highlight the critical role of using in-domain images for pretraining weights, particularly in the context of medical image analysis. The superior performance of the MsSSL model suggests that domain-specific SSL approaches significantly enhance the quality of learned representations, thereby improving model accuracy and reliability in medical datasets.

Table 7.2: *Comparison results of linear evaluation on APTOS dataset.*

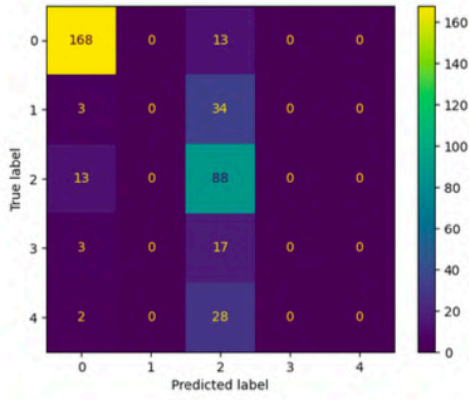
Model	Weights	κ
SwAV[129]	ImageNet	0.7617
SimCLR[129]		0.6989
DINO[129]		0.7790
DVME[129]		0.8242
DiRA	EyePACS	0.6980
MoCo-v2		0.8182
PCRL		0.7547
MoCo-v3		0.8159
DINO		0.8034
MAE		0.8324
MsSSL(Ours)	ODM-DDR	0.8698

7.8.3 Comparative Analysis of Self-Supervised vs. Supervised Learning Approaches

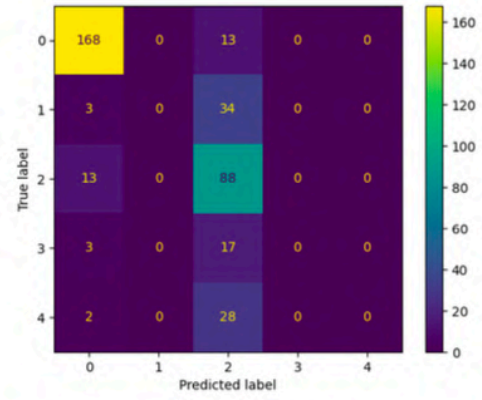
In this subsection, we present a detailed comparison between our proposed SSL frameworks: CGSSL and MsSSL, and the best traditional supervised learning approaches from our supervised learning experiments reported in Chapter 5, utilizing two prominent architectures: CNN and ViT. The goal of this comparison is to evaluate the effectiveness of SSL in enhancing feature extraction and overall performance for DDR grading tasks, relative to well-established supervised learning techniques. The results of this comparison highlight the advantages of SSL in scenarios with limited labeled data, showcasing its ability to learn robust and transferable features with APTOS and Messidor datasets. Additionally, we compare their class-wise performance to showcasing how well SSL in distinguishing adjacent classes especially the late ones, Figure 7.5 and Figure 7.6.

Table 7.3: *Performance Comparison of best CNN and best ViT in SL settings based on kappa score, to our proposed models **CGSSL** and **MsSSL** on the **APTOS** dataset across multiple evaluation metrics.*

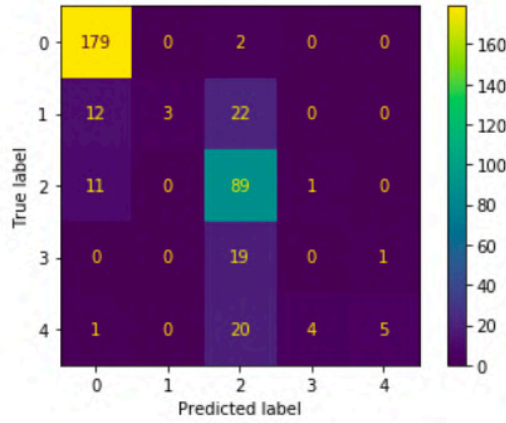
Learning	Model	Weights	κ	Acc.	Precision	Recall	F1
Supervised	ResNet-50	Random	0.5023	69.38%	27.56%	35.99%	30.69%
	ViT-S-32		0.6201	60.76%	33.65%	32.15%	32.88%
Supervised	ResNet-50	ImageNet	0.2776	62.92%	42.69%	32.70%	34.27%
	ViT-S-32		0.7465	65.58%	63.89%	36.42%	34.12%
SSL(Ours)	ViT-S-16	CGSSL	0.9252	82.92%	74.69%	69.62%	69.89%
	ViT-S-16	MsSSL	0.9324	84.55%	77.44%	72.28%	72.12%



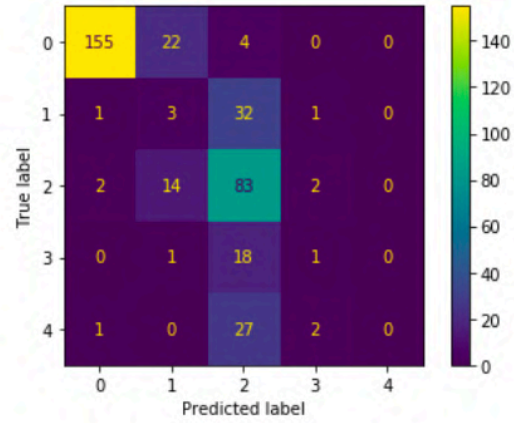
(a) ResNet-50(Random)



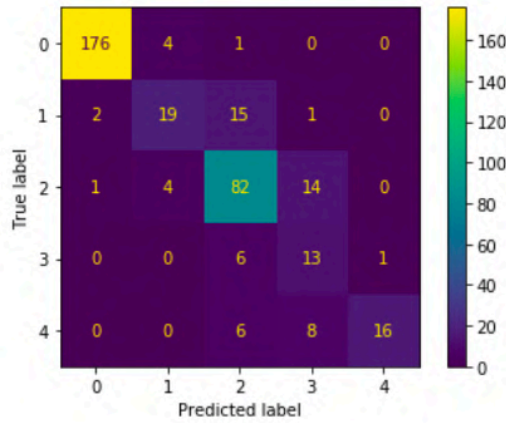
(b) ViT-S-32(Random)



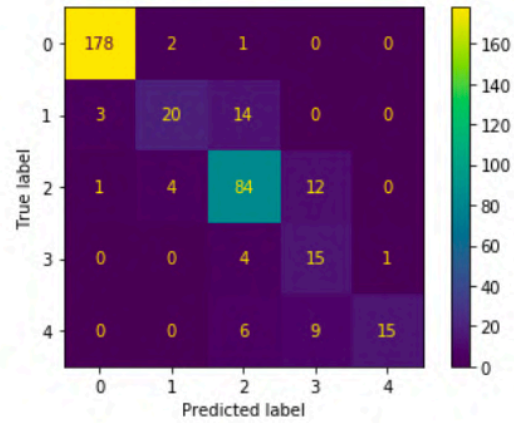
(c) ResNet-18(ImageNet)



(d) ViT-S-32(ImageNet)



(e) ViT-S-16(CGSSL)



(f) ViT-S-16(MsSSL)

Figure 7.5: Confusion matrices for the best CNN, ViT models in supervised learning, compared to our proposed SSL models CGSSL and MsSSL on the **APTOS** dataset.

APTOS Analysis

We analyze the results of applying both supervised learning models and our proposed self-supervised learning (SSL) models on the APTOS dataset. The comparison includes two supervised models, ResNet-50 (CNN-based) and ViT-S-32 (ViT-based), which were pre-trained on ImageNet in Chapter 5 and randomly in this chapter. As well as our proposed models in SSL, CGSSL and MsSSL, which utilize self-supervised pretraining to learn representations without labels from both our constructed dataset ODM and the DDR dataset and then fine-tuned the in-domain trained model ViT-S-16 on the APTOS dataset.

The performance clearly shows that our SSL models significantly outperform the SL models. The best supervised model, ViT-S-32, achieved a kappa score of 0.7465 and an accuracy of 65.58%. In contrast, CGSSL and MsSSL reached kappa scores of 0.9252 and 0.9324, respectively, with MsSSL achieving the highest accuracy at 84.55%.

This notable improvement, particularly in Kappa, suggests that the representations learned through SSL are more effective for DR grading tasks, as Kappa reflects the model's ability to accurately differentiate between the multiple DR classes, especially the minority and more challenging stages. The precision, recall, and F1-scores further support this, with MsSSL achieving a significantly higher precision (77.44%) and recall (72.28%) compared to the SL models, indicating its ability to balance sensitivity across all classes.

The confusion matrices in Figure 7.5 offer a visual representation of how each model performs in distinguishing between DR grades. Starting with the super-

vised models, ResNet-18 and ViT-S-32 struggle with accurately identifying higher DR grades, particularly grades 3 and 4. As demonstrated in the confusion matrices, ResNet-18 misclassifies higher DR grades as lower ones, especially in grade 4, where 20 cases are predicted as grade 2. Similarly, ViT-S-32 also exhibits significant misclassification, with 27 instances of grade 4 misclassified as grade 2. These results reveal the challenges faced by CNN-based models and standard ViT variants when handling imbalanced classes, especially in later stages of DR, which are critical for clinical diagnosis.

In contrast, the self-supervised learning models, particularly CGSSL and MsSSL, show far better performance in recognizing the more severe DR stages. The confusion matrices for these models demonstrate fewer misclassifications in higher DR grades. For example, MsSSL correctly classifies 15 out of 30 cases of grade 4 DR, with only 6 misclassifications as grade 2. This improvement highlights the effectiveness of the SSL approach in learning richer feature representations from unlabelled data, which enables the model to better differentiate between the subtle variations in DR stages.

When comparing ViT-S-16 (MsSSL) to ResNet-50 models (both trained from scratch and fine-tuned on ImageNet), the MsSSL approach significantly outperforms these models across all classes, particularly in the identification of minority classes like grade 4. The SSL framework allows ViT-S-16 to learn more meaningful features through contrastive learning on unlabelled data, resulting in improved alignment with the target domain. This leads to superior generalization and classification consistency compared to both scratch-trained and ImageNet-pretrained models. Unlike CNN models and standard ViT variants which often struggle with inter-class confusion

and imbalance in prediction accuracy, MsSSL leverages domain-specific pretraining to maximize the potential of ViT’s global feature extraction capabilities. This enables ViT-S-16 (MsSSL) to better capture the nuances of the DR dataset, resulting in more accurate and consistent predictions, especially for the more challenging higher grades.

Messidor Analysis

We analyze the performance of both supervised learning (SL) models and our proposed self-supervised learning (SSL) models on the Messidor dataset. The comparison includes two supervised models—ResNet-50 (CNN-based) and ViT-S-32 (ViT-based)—which were pre-trained on ImageNet and randomly, as well as our proposed models, CGSSL and MsSSL, which utilize self-supervised pretraining to learn feature representations without labels before being fine-tuned for diabetic retinopathy (DR) grading.

The results on the Messidor dataset, which is smaller in size compared to AP-TOS, reveal critical insights into the performance of supervised learning (SL) and self-supervised learning (SSL) models in DDR classification. Despite the challenges posed by the smaller dataset, our SSL models: CGSSL and MsSSL, significantly outperform the best performing SL models, ResNet-50 and ViT-S-32, particularly in their ability to generalize across multiple DR grades.

As shown in Table 7.4, MsSSL leads with a remarkable kappa score of 0.7431, compared to the best SL model both trained randomly or with ImageNet. Importantly, the precision (46.03%) and recall (49.59%) of MsSSL demonstrate its ability to provide balanced predictions across all DR grades, compared to the more skewed performance

of SL models that tend to focus on the dominant class. Although the SL models achieve reasonable accuracy, this does not fully reflect their ability to handle the more nuanced and difficult-to-classify stages of DR, as indicated by their lower Kappa scores and poor recall values. MsSSL, with its higher precision (46.03%) and recall (49.59%), demonstrates superior overall performance, especially in classifying the more severe stages of DR, which is crucial for clinical decision-making.

The confusion matrix results highlight the significant differences in model performance when training ResNet-50 with random initialization versus pre-trained ImageNet weights, as well as the substantial improvements introduced by our SSL models, CGSSL and MsSSL. When trained with random initialization, ResNet-50 struggles to classify diabetic retinopathy (DR) grades accurately. The majority of predictions are concentrated in grade 0 (healthy), with minimal success in distinguishing higher grades. Specifically, the model achieves 98 correct predictions for grade 0, while misclassifying the vast majority of grades 1 through 4 into grade 0. This outcome reflects the inability of the model to learn meaningful features from the limited dataset without pre-training, as it is heavily biased toward the majority class. For grades 1 through 4, the model’s predictions are sparse, indicating a failure to recognize the nuances of different DR stages.

In contrast, ResNet-50 pre-trained on ImageNet demonstrates considerable improvement across all DR grades. For grade 0, the model maintains high accuracy, correctly identifying 94 instances. For grades 1 and 2, the model shows significant progress, correctly classifying 19 and 22 instances, respectively, compared to its random counterpart. Similarly, for grade 3, the pre-trained model identifies 13 instances

correctly. However, the classification of grade 4 remains a notable challenge, as the model struggles to correctly predict this class, with most grade 4 samples being misclassified into lower grades. This reflects the limitations posed by the imbalance in the Messidor dataset, where the severe cases (grade 4) are significantly underrepresented, hindering the model’s ability to generalize to rare but critical cases.

Our SSL models, CGSSL and MsSSL Figure 7.6, demonstrate superior performance in classifying the higher DR grades compared to both ResNet-50 and ViT-S-32. CGSSL achieves significant improvements in the classification of late-stage DR grades, correctly identifying 13 out of 36 grade 2 cases and 5 out of 8 grade 3 cases. However, like ResNet-50, CGSSL struggles with grade 4, failing to make any correct predictions. Instead, grade 4 cases are misclassified as grade 3 or grade 2, although the model avoids misclassifying them as grade 1, unlike ResNet-50 and ViT-S-32. This indicates the model’s ability to focus its misclassifications within the more clinically relevant range of severe DR stages, reducing errors that would classify advanced DR as healthy or minimally referable.

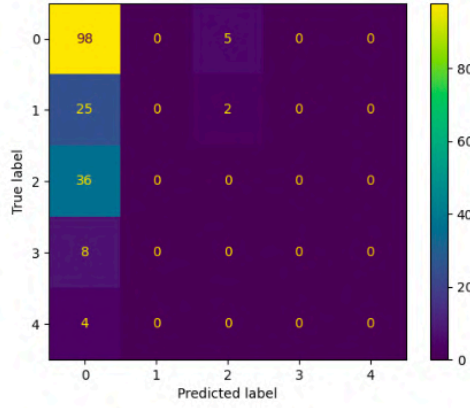
MsSSL builds upon the strengths of CGSSL, showing marked improvements in overall performance, particularly for grades 2 and 3. The model correctly identifies 16 out of 36 grade 2 cases and 6 out of 8 grade 3 cases, representing an advancement over CGSSL. Additionally, MsSSL demonstrates an ability to reduce false negatives in the most critical clinical contexts. For example, it successfully classifies four previously unreferable cases as referable, aligning with the needs of real-world medical practice. However, like all models examined, MsSSL continues to face challenges with grade 4 classification, misclassifying severe cases as grades 2 or 3 due to the scarcity of grade

4 samples in the dataset. This consistent difficulty highlights the impact of data imbalance in the training set, where lower-grade cases dominate, limiting the models' ability to generalize to rare but clinically crucial severe DR stages.

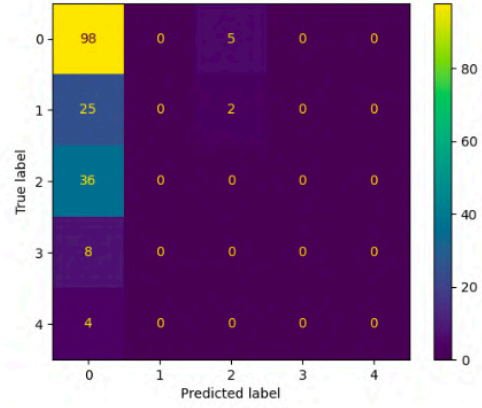
Overall, the results indicate that while ImageNet pre-training provides a significant performance boost for ResNet-50, our SSL approaches, particularly MsSSL, outperform traditional supervised methods in classifying critical DR stages. The ability of MsSSL to reduce false negatives and focus on clinically relevant misclassifications underscores its potential for practical deployment in medical screening tasks. However, the persistent challenges with grade 4 classification across all models point to the need for further exploration of data augmentation or alternative training strategies to address dataset imbalance and improve performance on rare cases.

Table 7.4: *Performance comparison of best CNN and best ViT in SL settings, to our proposed models **CGSSL** and **MsSSL** on the **Messidor** dataset across multiple evaluation metrics.*

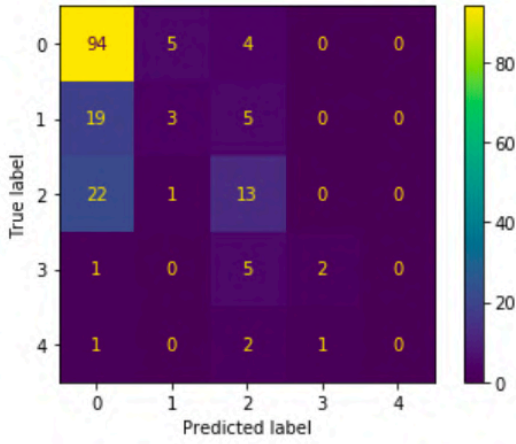
Learning	Model	Weights	κ	Acc.	Precision	Recall	F1
Supervised	ResNet-50	Random	0.0305	55.06%	11.46%	19.03%	14.31%
	ViT-S-32		0.0236	30.43%	20.78%	18.39%	19.52%
Supervised	ResNet-50	ImageNet	0.2776	62.92%	42.69%	32.70%	34.27%
	ViT-S-32		0.1891	39.32%	47.49%	25.06%	32.78%
SSL(Ours)	ViT-S-16	CGSSL	0.6463	54.49%	46.03%	47.93%	45.28%
		MsSSL	0.7431	60.67%	46.03%	49.59%	45.92%



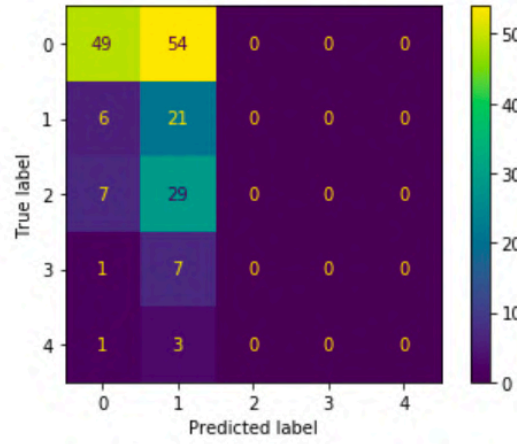
(a) ResNet-50(Random)



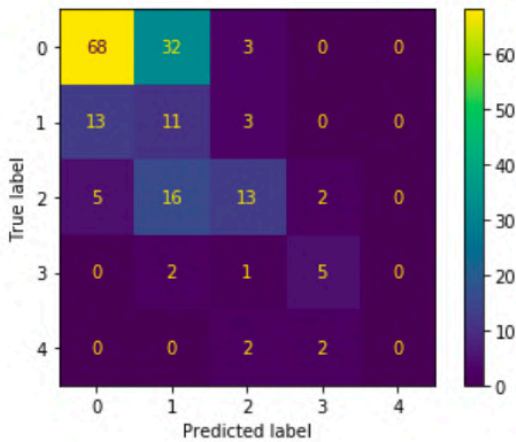
(b) ViT-S-32(Random)



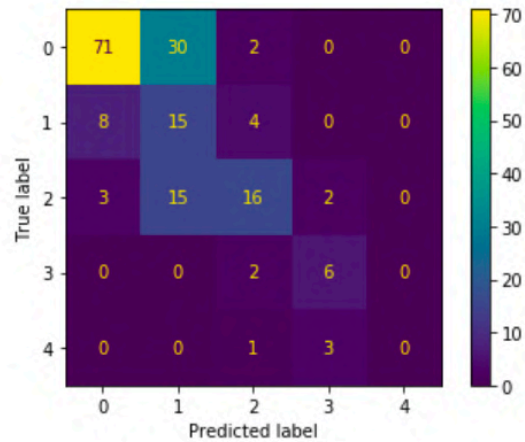
(c) ResNet-50(ImageNet)



(d) ViT-S-32(ImageNet)



(e) ViT-S-16(CGSSL)



(f) ViT-S-16(MsSSL)

Figure 7.6: Confusion matrices for the best CNN, ViT models in supervised learning, compared to our proposed SSL models CGSSL and MsSSL on the **Messidor** dataset.

7.8.4 Comparative Analysis of the Best ViT Backbone in Supervised and Self-Supervised Learning Settings

In this subsection, we present an analysis of the ViT-S-16 backbone’s performance across supervised learning (SL) and self-supervised learning (SSL) settings, highlighting key findings and insights. The experiments compare supervised and self-supervised approaches across multiple datasets, including APTOS, Messidor, DDR, ODM, and ODM-Fused. For the supervised experiments, the ViT-S-16 backbone utilized pre-trained weights from ImageNet, while for APTOS(MsSSL) and Messidor(MsSSL), the model leveraged knowledge learned through our proposed self-supervised learning framework, MsSSL.

This analysis provides evidence of the clear advantages of MsSSL in transferring domain-specific knowledge, enabling superior performance on downstream tasks. Furthermore, it underscores the limitations of traditional supervised learning with ImageNet pre-trained weights, especially when dealing with datasets exhibiting varying distributions. A particular focus is given to the significant impact of treating paired images in the ODM dataset as single cases, which leads to a notable decline in performance, further emphasizing the importance of leveraging the complementary information from distinct views in fundus imaging.

Performance in Supervised Learning

Under supervised learning, the ViT-S-16 backbone exhibited inconsistent performance across the datasets. While it performed relatively well on the APTOS dataset ($\kappa =$

Table 7.5: *Performance Comparison of the ViT-S-16 Backbone in Supervised and Self-Supervised Learning Settings across Various Datasets.*

Datasets	Kappa	Acc.	Precision	Recall	F1
APTOS	0.7321	65.31%	51.50%	35.36%	32.19%
Messidor	0.1757	33.14%	51.42%	24.03%	15.51%
DDR	0.2173	22.10%	40.13%	48.71%	35.52%
ODM	0.8315	67.06%	65.98%	57.18%	58.26%
ODM-Fused	0.7140	48.75%	64.04%	48.05%	44.53%
APTOS (MsSSL)	0.9324	84.55%	77.44%	72.28%	72.12%
Messidor (MsSSL)	0.7431	60.67%	46.03%	49.59%	45.92%

0.7321, F1 = 32.19%), its performance on the other datasets was subpar, with Messidor achieving $\kappa = 0.1757$, accuracy = 33.14%, and F1 = 15.51%, and DDR achieving $\kappa = 0.2173$, accuracy = 22.10%, and F1 = 35.52%. On the ODM dataset, the supervised approach achieved its highest performance across these experiments, with $\kappa = 0.8315$ and F1 = 58.26%, highlighting the benefit of leveraging paired images as distinct samples. However, treating paired images as single cases in ODM-Fused led to a marked decline in performance, with $\kappa = 0.7140$ and F1 = 44.53%.

The decline in performance in the ODM-Fused setting illustrates the limitations of supervised learning with ImageNet pre-trained weights. Treating paired images as single cases dilutes the complementary information provided by each view, resulting in suboptimal feature extraction and reduced classification accuracy.

Impact of MsSSL Knowledge Transfer

The experiments with MsSSL underscore the critical importance of effective knowledge transfer for the ViT-S-16 backbone, particularly in leveraging the rich information

embedded in dual-view fundus images. On the APTOS dataset, the use of MsSSL significantly improved all metrics, with $\kappa = 0.9324$, accuracy = 84.55%, and F1 = 72.12%, representing a dramatic increase compared to the supervised learning results ($\kappa = 0.7321$, accuracy = 65.31%, F1 = 32.19%). Similarly, for Messidor, MsSSL achieved $\kappa = 0.7431$, accuracy = 60.67%, and F1 = 45.92%, outperforming the corresponding supervised metrics ($\kappa = 0.1757$, accuracy = 33.14%, F1 = 15.51%).

These improvements can be attributed not only to the self-supervised learning approach but also to the rich dual-view knowledge embedded in the ODM dataset used during the pretraining phase. The dual-view setup provides complementary perspectives of the retina, enabling the model to learn more comprehensive and nuanced feature representations. This embedded richness serves as a key factor in enabling MsSSL to generalize effectively to downstream tasks, even on datasets with varying distributions like APTOS and Messidor.

Furthermore, these findings highlight an essential consideration for self-supervised learning: the choice of the source dataset for pretraining is as important as the size of the dataset. While larger datasets are typically favored in SSL to capture diverse features, the quality and relevance of the knowledge contained within the dataset are equally, if not more, crucial. The dual-view nature of the ODM dataset offered a unique advantage, allowing MsSSL to extract clinically meaningful and transferable features that align with real-world diagnostic practices in ophthalmology. This demonstrates that a well-curated, domain-specific dataset, such as ODM, can be more impactful than merely increasing dataset size, as it provides the model with the right context for learning.

The success of MsSSL lies not only in the methodology but also in the strategic selection of a knowledge-rich and clinically relevant source dataset. The ability of MsSSL to effectively utilize dual-view fundus images exemplifies the importance of choosing the right source of knowledge for self-supervised pretraining, enabling robust generalization and adaptability to downstream medical imaging tasks.

Analysis of Paired Fundus Image

The experiments on ODM and ODM-Fused further underscore the importance of how paired images are treated in both supervised and self-supervised settings. When paired images in the ODM dataset are treated as distinct samples, the model benefits from the rich complementary information available in each view, achieving the highest performance ($\kappa = 0.8315$, accuracy = 67.06%, F1 = 58.26%) among the supervised learning experiments. However, when paired images are fused into single cases in ODM-Fused, there is a significant drop in performance, with $\kappa = 0.7140$, accuracy = 48.75%, and F1 = 44.53%. This decline highlights the limitations of combining paired images, as it reduces the granularity of information available to the model, hindering its ability to distinguish subtle features critical for accurate classification.

From a clinical perspective, each view of a fundus image typically captured from different angles or regions of the retina is treated as an independent source of information to diagnose the case. This approach allows clinicians to analyze specific features unique to each view, leading to a more comprehensive and accurate diagnosis. Similarly, treating paired images as distinct samples in the model preserves this clinical practice by allowing the algorithm to learn and utilize the complementary fea-

tures present in each view. The fusion of paired images, in contrast, results in a loss of this critical information, as the distinctive features from each view are averaged, diminishing the model’s diagnostic capability.

Furthermore, the consistent superior performance of MsSSL compared to supervised learning underscores the framework’s ability to learn effectively from different views of fundus images during pretraining. By leveraging diverse, unlabeled views in the ODM dataset, MsSSL captures a rich and generalized representation of the retina. This enables the model to transfer its knowledge effectively to downstream tasks, regardless of the specific view captured by the fundus camera. This flexibility is critical in clinical settings, where variations in image capture protocols, angles, and equipment can significantly influence the quality and type of data collected. The ability of MsSSL to generalize across such variations ensures robust performance and adaptability to real-world clinical scenarios, where models may encounter diverse and unseen data.

The overall analysis highlights the transformative impact of self-supervised learning through MsSSL on the performance of the ViT-S-16 backbone, consistently outperforming supervised learning with ImageNet pre-trained weights across all datasets. This advantage is particularly evident on APTOS and Messidor, where domain-specific knowledge transfer is critical for achieving robust performance. The findings also emphasize the detrimental effect of treating paired fundus images as single cases, as seen in the ODM-Fused experiments, which dilute complementary information and lead to reduced accuracy. By leveraging the rich and diverse features from distinct paired images and employing self-supervised learning, MsSSL addresses the inherent limi-

tations of supervised learning, setting a new benchmark for adaptable and effective feature extraction in medical imaging tasks. Moreover, MsSSL’s ability to generalize across multiple views of fundus images aligns with clinical practices, where each view is treated independently for diagnosis, further enhancing its relevance and reliability for automated medical diagnosis in ophthalmology.

7.8.5 Qualitative Evaluation

The qualitative evaluation focuses on analyzing the visual attention maps generated by our proposed models CGSSL and MsSSL comparing them to the ViT-s16 model pre-trained on ImageNet. By examining these self-attention maps based on [183], we gain insights into the decision-making process of the model, particularly by identifying which regions of the input image the model focuses on when making predictions.

The attention mechanism in transformer-based models assigns varying degrees of importance to different parts of the image, effectively highlighting the areas that contribute most to the final output. Visualizing these attention patterns allows us to better understand how the model processes information and forms associations between image patches. This interpretability is critical in assessing whether the model’s focus aligns with human intuition, and in identifying any potential biases or failure modes within the model’s reasoning.

We present visualizations of self-attention maps of supervised and self-supervised ViT-S16. The fundus images are randomly sampled from the APTOS dataset, which was not used during pre-training. Recently, DINO [183] demonstrated that the attention maps generated by self-distillation-based pre-trained ViTs capture semantic

information from natural images. Following their approach to produce detailed attention maps, the images are resized to a higher resolution of 1024×1024 , resulting in 64×64 patches with a patch size of 16 used as the input sequence. These images are then processed by self-supervised ViTs, and the self-attention maps from the final layer are displayed. We visualize the attention by averaging the normalized multi-headed self-attention maps for the global token, offering insights into how these models attend to different regions of the input.

The following analysis examines the self-attention maps produced across different heads of the model, providing a deeper understanding of the spatial regions the model prioritizes during inference.

In the case of MsSSL, we observe that the attention maps consistently focus on clinically relevant regions, particularly those with visible pathological signs such as lesions, microaneurysms, and exudates, Figure 7.7. The MsSSL model demonstrates a greater ability to highlight these affected areas compared to CGSSL, suggesting that it is more adept at capturing subtle disease markers. This improvement can be attributed to the integration of the Multi-Feature Extraction Encoder (MFEE), which enhances feature representation by extracting more granular details. In contrast, while CGSSL also localizes key regions, its attention maps appear less precise, occasionally spreading focus to irrelevant areas. Overall, the qualitative results demonstrate that MsSSL not only identifies disease-affected regions more effectively but also improves the interpretability of the model’s predictions, which is critical for clinical applications such as medical image diagnosis.

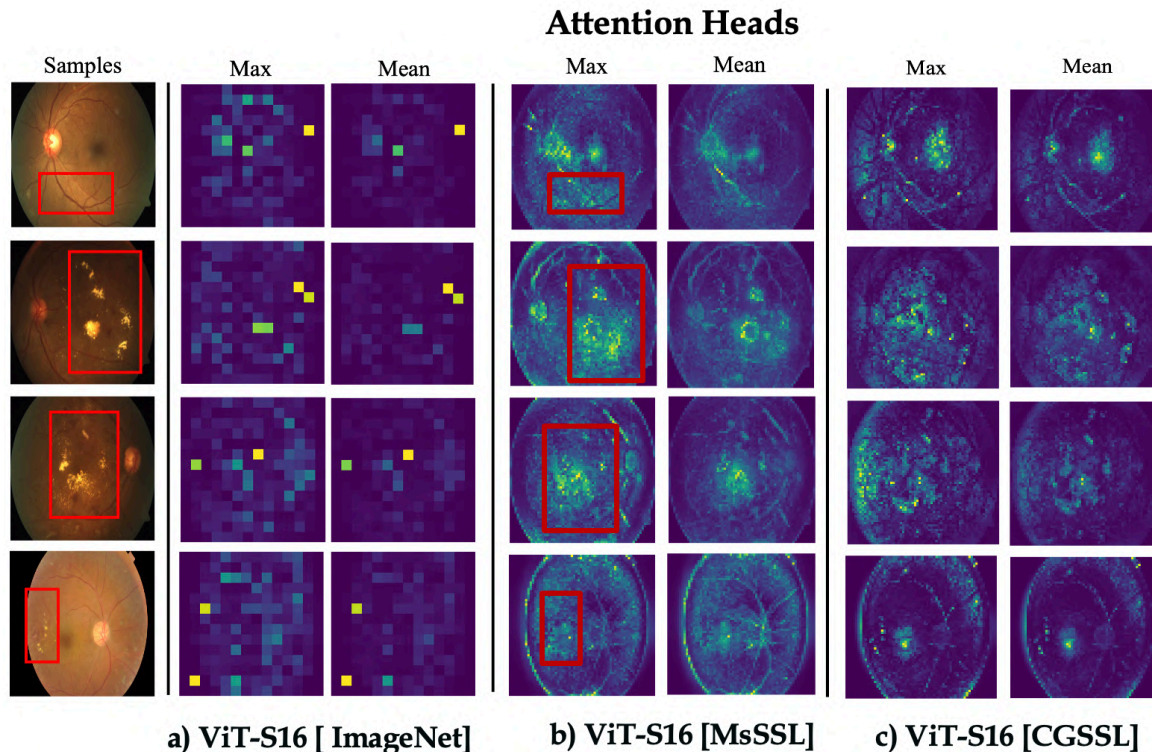


Figure 7.7: *Self-attention map visualization illustrating model decisions. a) ViT-S16 with weights pre-trained on ImageNet. b) ViT-S16 with weights from MsSSL. c) ViT-S16 with weights from CGSSL.*

The ViT-S16 model, trained under a supervised learning regime, produces relatively sparse and less focused attention maps, with minimal emphasis on pathological regions, Figure 7.7,a. This scattered attention suggests that the model struggles to localize important features associated with the disease, which may reduce its effectiveness in medical diagnosis tasks. In contrast, the attention maps generated by MsSSL exhibit a clear advantage, as the model consistently highlights critical regions, including lesions and abnormal structures, across all the samples. This superior performance can be attributed to the self-supervised learning approach, which enables MsSSL to capture more detailed and informative feature representations, thereby enhancing its

ability to focus on disease-relevant areas. The improved interpretability and localization precision of MsSSL demonstrate its potential for more accurate and reliable predictions in medical imaging.

7.8.6 Error Analysis of Challenging Cases

The aim of this subsection is to provide an in-depth analysis of the errors made by the models with a particular focus on understanding the limitations of the proposed approach. Error analysis is crucial for gaining insights into the specific challenges posed by difficult-to-classify samples, especially in the context of medical imaging tasks where accurate predictions are critical.

For this purpose, we selected a subset of samples from the Messidor dataset that were consistently misclassified across all experiments, including those involving our proposed model, MsSSL. These selected samples include four cases belonging to DR grade 4, the most severe stage of diabetic retinopathy, Table 7.6. Despite the enhanced generalization capabilities of MsSSL, these grade 4 cases remained challenging for the model to classify correctly, highlighting the inherent difficulties associated with rare and severe cases in the dataset. This analysis focuses on examining these misclassified samples in detail to uncover patterns or factors that contributed to the model’s inability to correctly predict their labels.

The analysis of predictions for DR grade 4 fundus images highlights key performance differences between the supervised learning (SL) model and the proposed MsSSL model. The SL model underperformed significantly, consistently predicting a lower severity grade (1 or 0) for all four samples. This indicates that the SL

Table 7.6: *Prediction results for samples of grade 4 fundus images using SL and MsSSL models.*

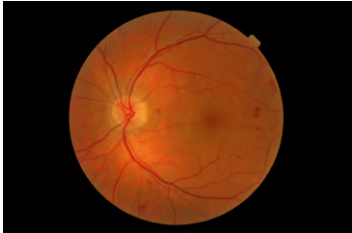
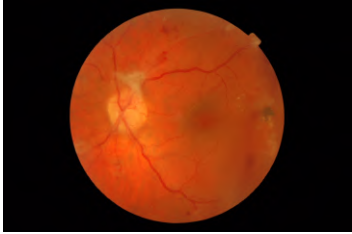
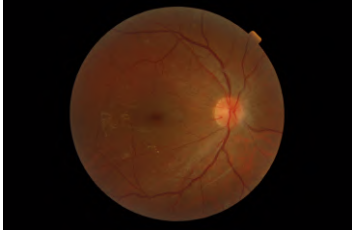
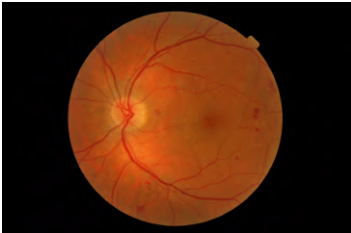
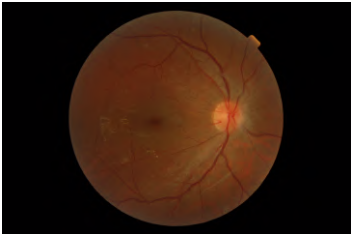
Image	SL Prediction ViT-S-32	MsSSL Prediction ViT-S-16
	1	3
	1	3
	1	3
	0	2

Table 7.7: *Ophthalmologist Assessment.*

Image	Medical Assessment
	Moderate (grade 3): There are diabetic retinopathy changes in two quadrants (Dot, Blot hemorrhages)
	Mild (grade 2): There is minimal changes in one quadrant (Temporal macula)
	PDR (grade 4): Neovascularization on optic disc which is a sign of proliferative diabetic retinopathy changes
	PDR (grade 4): Media opacity temporally due to vitreous hemorrhage associated with development of proliferative membrane on optic disc, both are signs of advanced diabetic retinopathy

model struggled to recognize advanced pathological features of grade 4 DR, such as widespread hard exudates, hemorrhages, and neovascularization. This misclassification suggests that the SL model primarily focuses on less severe indicators, failing to capture the complexity and severity of advanced DR cases.

Conversely, the MsSSL model demonstrated a more robust ability to identify features associated with severe DR, as reflected in its predictions of grade 3 for three samples and grade 2 for one sample. Although the model did not achieve the correct grade 4 classification, its consistent prediction of higher severity grades compared to the SL model indicates its capability to capture advanced retinal abnormalities. Features such as extensive exudates, prominent vascular changes, and retinal damage were partially recognized by the MsSSL model, enabling it to differentiate these images from lower-grade cases. However, the inability to correctly predict grade 4 suggests challenges in identifying neovascularization and distinguishing finer details that separate grade 3 and grade 4 severity levels.

The insights provided by the ophthalmologist in Table 7.7, reinforce this analysis by offering a detailed medical perspective on the selected cases. For instance, the ophthalmologist’s diagnosis of moderate (grade 3) and mild (grade 2) for two cases underscores the overlap in features between adjacent grades, such as hemorrhages in two quadrants or minimal changes in one quadrant. The other two cases were accurately identified as PDR (grade 4), highlighting advanced pathological features like neovascularization on the optic disc and media opacity from vitreous hemorrhages. These findings emphasize the nuanced nature of severe DR cases and the difficulty in distinguishing between overlapping or ambiguous features, particularly for grade 4

cases.

In medical practice, DR cases are categorized as non-referable (grades 0 and 1) or referable (grades 2, 3, and 4) to prioritize patients requiring immediate medical attention. The MsSSL model’s consistent ability to classify all four cases as referable (grades 2 or 3) reflects its clinical relevance in reducing false negatives and improving patient outcomes. In contrast, the SL model misclassified all cases as non-referable, posing significant risks in medical decision-making. Furthermore, the MsSSL model demonstrated its ability to detect more pathological features associated with advanced DR grades, enabling it to provide predictions closer to the ground truth.

In summary, the combined analysis of the models’ predictions and ophthalmologist feedback highlights the limitations of the SL model and the significant advancements offered by the MsSSL model. While the MsSSL model did not achieve precise grade 4 predictions, its ability to identify advanced pathological features and consistently classify referable cases demonstrates its potential to minimize false negatives and enhance clinical decision-making in DR screening. This underscores the importance of incorporating clinical expertise and developing robust models capable of addressing the nuanced challenges of medical imaging tasks.

7.9 Discussion

The comparative analysis in Section 7.8 clearly demonstrates the superior performance of our proposed self-supervised learning model, MsSSL, across various benchmarks. Through thorough experimentation, it is evident that our model consistently achieves

a higher quality of learning representation and transferability, particularly when fine-tuned on medical imaging datasets like APTOS. This performance not only surpasses other state-of-the-art models but also highlights the advantages of integrating specific architectural enhancements and domain-specific pre-training data into the learning process.

In Table 7.1, we present a focused comparison on the APTOS dataset, emphasizing the critical role of utilizing in-domain pre-trained weights for fine-tuning tasks. Models that were pre-trained on medical image datasets, such as EyePACS, show substantial improvements in performance. For instance, SimCLR pre-trained on EyePACS achieves a notable increase in accuracy and Kappa coefficient (κ), underscoring the importance of domain-specific feature extraction for transfer learning in medical applications. However, the MsSSL model, leveraging combined ODM-DDR weights, outperforms these models, achieving the highest κ value of 0.9324, an accuracy of 84.55%, and an F1 score of 72.07. This significant improvement highlights the effectiveness of our approach in capturing and transferring relevant features that are crucial for medical imaging tasks, especially in diabetic retinopathy (DR) grading.

Despite the ODM-DDR dataset being smaller than the EyePACS dataset, the thoughtful combination of dual-view and single-view samples within our MsSSL framework has proven highly effective in maximizing the utility of limited data. By carefully curating these diverse samples and leveraging the proposed components in the study, we were able to achieve results comparable to those obtained with much larger datasets. This approach not only optimizes the learning process in data-constrained environments but also ensures robust feature extraction and transferability.

Crucially, the success of MsSSL in the qualitative evaluation can be attributed to the MFEE, Figure 7.7,b. The MFEE performed exactly as intended by extracting both high-level and fine-grained features, which is clearly reflected in the attention maps. The visualizations show that the MFEE successfully captures a broad range of feature details, from general image structure to small lesion-specific characteristics. Additionally, the Deep Learner module (DLe) refines these features, enabling the model to adapt to the nuances of medical imaging tasks such as diabetic retinopathy (DR) grading. These two novel components work together to make MsSSL highly effective at both feature extraction and transfer learning, resulting in a more precise focus on disease-affected areas, as seen in the attention map visualizations.

The qualitative and class-wise evaluations highlight the pivotal role of the MFEE in MsSSL’s superior performance. The attention maps clearly showcase the model’s ability to effectively capture and focus on relevant image regions, demonstrating the MFEE’s success in extracting both high-level and fine-grained features. This reinforces the effectiveness of our design choices. Furthermore, the class-wise confusion matrix offers additional validation, showing a significant reduction in false positives compared to models trained in supervised settings. This reduction not only enhances the model’s precision but also emphasizes the overall reliability of MsSSL in accurately identifying disease-affected areas, further solidifying its advantages over traditional supervised approaches.

The ability to fine-tune the knowledge transferred from MsSSL in just 25 epochs with superior performance highlights the advantages of SSL pre-training for domain-specific tasks like DR grading. The strong representations learned during the SSL

pre-training phase enable the model to adapt quickly and effectively with minimal additional training. In contrast, models using ImageNet pre-trained weights require more fine-tuning to adjust to the domain-specific features of DR images as we reported in Chapter 5. This highlights the benefits of using an SSL framework that is more closely aligned with the target task.

This efficiency not only saves time and resources but also leads to models that generalize better with less risk of overfitting. The success of your MsSSL approach demonstrates the importance of task-specific pre-training for medical imaging tasks, and it could set a new benchmark for future research in DR grading and other similar medical diagnosis tasks.

7.10 Ablation Studies

To validate the effectiveness of the proposed components in our MsSSL framework, we conducted a series of ablation studies using the APTOS dataset. The objective of these experiments was to evaluate how each enhancement contributes to the overall performance of the model and to identify the optimal settings for diabetic retinopathy (DR) grading. Each subsection in this section focuses on a specific component, analyzing its impact on the model’s predictive performance.

For each experiment, the respective settings were trained for **300 epochs**, allowing the model to thoroughly learn feature representations before transferring this knowledge and fine-tuning it on the APTOS dataset. This extended training period ensures that the model had sufficient time to generalize across various feature hierarchies and

properly adapt to the downstream DR classification task.

These ablation studies are critical for understanding how individual modifications, such as changes in the backbone architectures, feature fusion techniques, and attention mechanisms, affect the model’s ability to extract meaningful features from medical images. By isolating each component, we aim to demonstrate the incremental value brought by each enhancement, ensuring that the final configuration of the MsSSL model is both efficient and effective for the DR classification task.

As shown in Table 7.8, the best-performing configuration includes fine-grained saliency detection, ResNet50 as the FPN backbone, ViT-S-p16 as the Vision Transformer backbone, element-wise addition for feature fusion, and the Convolutional Block Attention Module (CBAM) for enhanced attention. These components were selected based on their ability to improve key metrics such as Kappa coefficient, accuracy, and F1 score, making them the optimal choices for our MsSSL framework.

Table 7.8: *MsSSL Best Settings.*

Component	Value
FPN Backbone	ResNet50
ViT Backbone	ViT-S-p16
Fusion	Addition
Saliency Detection	Fine grained
DLe Layers	Conv2DTransposed
DLe Attention	CBAM at the end

This ablation analysis not only highlights the importance of each module within the MsSSL framework but also reinforces the necessity of carefully selecting and combining these components to achieve optimal performance in medical image analysis. The extended training of **300 epochs** allowed the model to fully explore the feature space,

resulting in a robust transfer of knowledge that was effectively fine-tuned for the APTOS dataset, enhancing the model’s accuracy in detecting diabetic retinopathy.

7.10.1 FPN-backbone

To assess the influence of different Feature Pyramid Network (FPN) backbones on the performance of the MsSSL model, we compared the performance of ResNet18 and ResNet50 as backbone architectures. FPN is commonly used to enhance feature extraction by integrating multi-scale features, which is crucial for tasks like diabetic retinopathy (DR) grading, where both global and fine-grained information are essential.

The results reveal a clear advantage of using deeper backbones, with ResNet50 outperforming ResNet18 in all key metrics, Table 7.9. Specifically, the ResNet50 backbone achieved a Kappa coefficient of 0.9324, an accuracy of 84.55%, and an F1 score of 72.07%. In contrast, ResNet18, while still performing well, reached a Kappa score of 0.92216, an accuracy of 82.38%, and an F1 score of 69.49%. These differences indicate that the greater depth of ResNet50 allows for more robust feature extraction, enabling the model to capture more detailed information across multiple scales, which is particularly important in detecting subtle variations in DR severity.

The superior performance of ResNet50 highlights the importance of utilizing deeper FPN backbones for medical image analysis tasks. While ResNet18 provides a more lightweight option, ResNet50’s additional layers contribute to improved feature representation and more accurate predictions. This finding underscores the need for deeper architectures when aiming for high performance in complex tasks like DR grad-

ing, where small differences in feature details can significantly impact classification outcomes. The deeper ResNet50 backbone integrates multi-scale features more effectively, resulting in better downstream task performance within the MsSSL framework.

Table 7.9: *MsSSL model with different FPN backbones.*

FPN backbone	κ	Acc.%	F1%
ResNet18	0.92216	82.38	69.49
ResNet50	0.93240	84.55	72.07

7.10.2 ViT Variants as Backbone

The choice of Vision Transformer (ViT) backbone plays a pivotal role in the performance of the MsSSL model, directly impacting its ability to effectively learn and generalize from medical imaging data. In our investigation, we evaluated several ViT variants, including ViT-S-p16, ViT-S-p32, ViT-B-p16, and ViT-B-p32, each differing in patch size and architecture complexity. These variants were chosen to explore how different configurations of patch size (16x16 vs. 32x32) and model scale (Small vs. Base) influence the model’s capability to capture relevant features for diabetic retinopathy (DR) grading, Table 7.10 .

Among the tested variants, the ViT-S-p16 backbone emerged as the most effective, delivering the highest performance with a Kappa coefficient of 0.9324, an accuracy of 84.55%, and an F1 score of 72.07%. This indicates that the smaller patch size of 16x16, combined with the relatively lighter ViT-S architecture, provides an optimal balance between feature granularity and model complexity. The smaller patches allow

Table 7.10: *MsSSL model with different ViT models backbones.*

ViT backbone	κ	Acc. %	F1 %
ViT-S-p32	0.9089	79.67	65.17
ViT-B-p16	0.9300	82.66	69.67
ViT-B-p32	0.9102	78.86	64.52
ViT-S-p16	0.9324	84.55	72.07

for finer feature extraction, which is particularly important in medical imaging tasks where subtle differences in visual features can significantly affect diagnosis.

Other backbones, such as ViT-B-p16 and ViT-S-p32, while still competitive, demonstrated lower performance. The ViT-B-p16 achieved a Kappa score of 0.9300, slightly trailing the ViT-S-p16 but still showing the effectiveness of the 16x16 patch size. However, the larger ViT-B variant introduces more parameters, potentially leading to higher computational demands without substantial gains in performance. On the other hand, ViT-S-p32 and ViT-B-p32, which utilize larger 32x32 patches, performed noticeably worse. The reduced resolution from the larger patches likely hinders the model’s ability to capture fine-grained details crucial for accurate DR grading, as evidenced by their Kappa scores of 0.9089 and 0.9102, respectively.

These findings underscore the significance of choosing an appropriate ViT backbone for the task at hand. The superior performance of the ViT-S-p16 variant highlights that, for complex medical imaging tasks like DR detection, a smaller patch size and a simpler architecture offer the best trade-off between performance and efficiency. As a result, the ViT-S-p16 backbone proves to be the most suitable for the MsSSL framework, enhancing its capability to extract relevant features while maintaining

computational feasibility.

7.10.3 Feature Fusing Approaches

The selection of feature fusion methods plays a critical role in determining the overall performance of the MsSSL model, particularly in integrating features extracted from multiple views within the model architecture. To evaluate the most effective approach, we experimented with three different fusion methods: concatenation, element-wise multiplication, and element-wise addition, Table 7.11. Each method combines the feature maps in a unique way, potentially impacting the model’s ability to capture both high-level and fine-grained details crucial for accurate diabetic retinopathy (DR) grading.

Table 7.11: *MsSSL model with different fusion methods.*

Fusion	κ	Acc.%	F1%
Concatenation	0.9198	81.84	68.19
Multiplication	0.9262	84.01	71.86
Addition	0.9324	84.55	72.07

Among these approaches, element-wise addition demonstrated the best performance, achieving a Kappa coefficient of 0.9324, an accuracy of 84.55%, and an F1 score of 72.07%. This indicates that element-wise addition allows for more efficient integration of complementary information from the fused feature maps, enhancing the model’s ability to learn discriminative features across different DR stages.

In contrast, element-wise multiplication, while often effective in other applications,

yielded the lowest Kappa score of 0.9034. This suggests that multiplying feature maps may overly constrain the model by amplifying certain features while suppressing others, limiting the diversity of information being captured. Concatenation, which combines features by appending them along a new dimension, performed moderately well with a Kappa score of 0.9198 but still fell short compared to the addition method.

These results highlight the importance of selecting an appropriate fusion strategy in multi-scale or multi-view learning frameworks like MsSSL. Element-wise addition, by balancing and integrating different feature representations more effectively, proved to be the most suitable fusion method in this context, resulting in superior model performance.

7.10.4 Salient detection

Effective saliency detection is a crucial aspect of medical image analysis, particularly in tasks like diabetic retinopathy (DR) grading, where the ability to accurately identify and focus on the most relevant features can significantly improve model performance. To assess the impact of different saliency detection methods, we experimented with two approaches within our MsSSL framework: spectral saliency detection and fine-grained saliency detection.

The results, summarized in Table 7.12, show a clear advantage for the fine-grained method, which outperformed the spectral approach across all key metrics. Specifically, the fine-grained method achieved a Kappa coefficient of 0.9324, compared to 0.9152 for the spectral method. In terms of accuracy and F1 score, the fine-grained approach also delivered superior results, with an accuracy of 84.55% and an F1 score of 72.07%,

in contrast to the spectral method’s 79.13% accuracy and 62.53% F1 score.

Table 7.12: *MsSSL model with different saliency detection methods*

Salient detection	κ	Acc.%	F1%
Spectral	0.9152	79.13	62.53
fine-grained	0.9324	84.55	72.07

These findings demonstrate that fine-grained saliency detection is significantly more effective at capturing critical details in medical images. By focusing on more localized and intricate features, this method enhances the model’s ability to distinguish between different stages of DR, contributing to better overall performance in the classification task. The superiority of the fine-grained approach underscores the importance of using advanced saliency detection techniques in medical image analysis, where subtle variations in visual features often hold the key to accurate diagnosis and grading.

7.10.5 The Impact of Deep Learner Module

The integration of the Deep Learner (DLe) module into the MsSSL framework demonstrates a clear improvement in model performance, as highlighted in Table 7.13. When the DLe module is absent from the MsSSL framework, the model achieves a Kappa score of 0.9260, an accuracy of 82.92%, and an F1 score of 69.21%. However, with the inclusion of the DLe module, the performance metrics increase significantly. The Kappa score improves to 0.9324, the accuracy rises to 84.55%, and the F1 score reaches 72.07%.

Table 7.13: *MsSSL model with and without the Deep Learner (DLe) module.*

Model	κ	Acc. %	F1 %
MsSSL w/o DLe	0.9260	82.92	69.21
MsSSL w. DLe	0.9324	84.55	72.07

These results indicate that the DLe module enhances the model’s ability to extract and integrate more meaningful representations, particularly in distinguishing between more challenging classes. The module’s contribution is crucial for improving the quality of learned representations and overall model performance, further validating its importance in the MsSSL framework.

Achieving optimal feature extraction in medical image analysis requires not only robust models but also careful consideration of how each component interacts within the broader framework. In our exploration of the Deep Learner (DLe) module, we investigated five distinct configurations to identify the most effective design for diabetic retinopathy (DR) grading. The goal was to determine which design would best complement the other components in the Multi-scale Self-supervised Learning (MsSSL) framework, enhancing its overall capacity to capture high-level and fine-grained features.

DLe layers configurations : These configurations explored various convolutional layers and the integration of the CBAM, revealing critical insights into how design choices influence model performance, Table 7.14. By assessing different combinations of layers and attention mechanisms, we sought the optimal DLe design that could seamlessly integrate with the MsSSL framework, contributing to its ability to learn

more robust and discriminative feature representations for DR classification.

Table 7.14: *Deep Learner (DLe) module with different layers design.*

Layers	Kernel	κ	Acc.%	F1%
Conv1D	3X3	0.9149	80.75	66.60
Conv2D	3X3	0.9277	82.38	69.31
Conv2DTransposed	3X3	0.9288	83.46	69.48
Conv2DTransposed + CBAM First	3X3	0.9189	82.65	68.35
Conv2DTransposed + CBAM Last	3X3	0.9324	84.55	72.07

The first design employed a 1D convolutional layer with a 3x3 kernel, resulting in a Kappa score of 0.9149, accuracy of 80.75%, and an F1 score of 66.60%. While this approach captured sequential patterns, it fell short of effectively extracting the complex spatial features required for DR grading. This underscored the limitations of 1D convolutions for high-dimensional image data, which led us to explore more sophisticated alternatives.

Shifting to a 2D convolutional layer with the same kernel size marked a significant improvement. The model’s Kappa score increased to 0.9277, with accuracy reaching 82.38% and the F1 score rising to 69.31%. This shift demonstrated the power of 2D convolutions in capturing intricate spatial relationships, essential for distinguishing between the subtle features of different DR stages.

Building on this, the next configuration used Conv2DTransposed layers, a technique known for its up-sampling capabilities. This adjustment further enhanced performance, boosting the Kappa score to 0.9288, accuracy to 83.46%, and the F1 score to 69.48%. The ability of transposed convolutions to reconstruct finer details from

the learned features proved beneficial in refining the model’s predictions.

To improve feature focus, we introduced CBAM in two different placements within the DLe module. When CBAM was applied at the start, the results were less impressive, yielding a Kappa score of 0.9189 and accuracy of 82.65%. However, when the attention mechanism was placed after the Conv2DTransposed layers, it led to a substantial performance boost. This final design achieved the highest Kappa score of 0.9324, accuracy of 84.55%, and F1 score of 72.07%, demonstrating that applying attention mechanisms after core feature extraction allows the model to better focus on the most critical aspects of the image.

These results underscore the critical role of strategic design in convolutional layers and attention mechanisms. By placing the CBAM module after feature extraction, the model achieved superior performance, demonstrating how thoughtful choices like the use of Conv2DTransposed with CBAM can significantly enhance the learning of robust features for DR grading. This highlights the importance of carefully structuring layers and attention modules to optimize deep learning models in medical image analysis.

7.10.6 Preprocessing Ablation Studies

To evaluate the impact of individual preprocessing techniques on the performance of the proposed model, we conducted a series of ablation experiments. The preprocessing pipeline originally included multiple augmentations, such as random horizontal and vertical flips, random resized cropping, color jitter, random rotation, and affine transformations, followed by normalization. In each ablation experiment, one augmentation technique was removed while keeping the others unchanged, allowing us

to isolate its contribution to the overall model performance. This analysis provides valuable insights into which preprocessing steps are most critical for achieving optimal performance and highlights the robustness of the model under varying levels of augmentation. The findings are summarized in Table 7.15.

Table 7.15: *Ablation study on the effect of preprocessing techniques on the model’s performance.*

Preprocessing Configuration	Kappa	Acc.(%)	F1(%)
Baseline (Full Preprocessing)	0.9324	84.55	72.07
Without preprocessing	0.8929	79.94	64.86
Without Horizontal/Vertical Flip	0.9135	81.57	67.16
Without Random Resized Crop	0.9207	81.02	67.95
Without Color Jitter	0.9132	80.48	64.83
Without Random Rotation	0.9267	80.75	66.85
Without Random Affine	0.9243	80.75	64.27

The results presented in Table 7.15 indicate that the baseline configuration, which includes full preprocessing, achieves the highest Kappa (0.9324), accuracy (84.55%), and F1-score (72.07%), demonstrating the effectiveness of the complete preprocessing pipeline. When preprocessing is entirely excluded, there is a notable decline in performance, with the Kappa dropping to 0.8929, accuracy to 79.94%, and F1-score to 64.86%. This highlights the critical role of preprocessing in enhancing model performance.

Among the individual preprocessing techniques, the omission of horizontal/vertical flipping, random resized cropping, color jittering, random rotation, or random affine transformations each results in a reduction in metrics, albeit to varying extents. For example, removing horizontal/vertical flipping yields a slight drop in Kappa (0.9135),

accuracy (81.57%), and F1-score (67.16%). Similarly, the absence of random resized cropping leads to Kappa, accuracy, and F1-score values of 0.9207, 81.02%, and 67.95%, respectively. The results suggest that each component contributes uniquely to the model's overall performance, with the baseline demonstrating the synergistic effect of combining these preprocessing techniques.

7.10.7 Evaluation of Data Splitting Strategies

To evaluate the robustness of the proposed model across different data splitting strategies, we implemented four distinct approaches. The first approach, referred to as the official split, follows a 70%-15%-15% ratio for training, validation, and testing, respectively, using a fixed random seed to ensure reproducibility. This serves as the baseline for comparison. The second approach introduces an alternative random split, which maintains the same 70%-15%-15% ratio but utilizes a different random seed to evaluate the impact of data shuffling on model performance. The third approach employs a stratified split, ensuring that the class distributions are proportionally represented across the training, validation, and testing sets. This method mitigates potential imbalances and allows for a more equitable evaluation of performance across all classes. Finally, a custom split modifies the dataset allocation to a 60%-20%-20% ratio for Mesidor dataset only, inspired by data partitioning strategies used in related study [126]. We did not perform alternative splitting strategies for the APTOS dataset, as it follows the same splitting methodology as a previous study to ensure consistency [126]. These diverse splitting strategies aim to investigate the extent to which the choice of data partitioning impacts model performance, providing insights into the stability

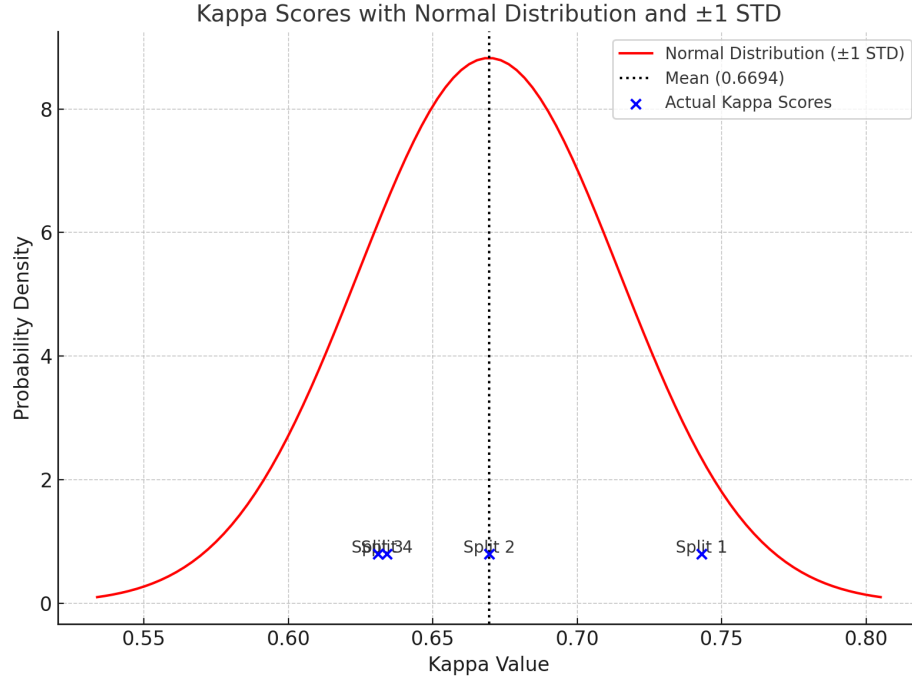


Figure 7.8: *The mean and the STD Distribution of different Splittings Strategies on **Messidor** Dataset.*

and generalizability of the proposed model.

Table 7.16: *Comparison of Model Performance Across Different Data Splitting Strategies on **Messidor** Dataset.*

Splitting Strategy	Kappa
Our Splitting 1	0.7431
Splitting 2	0.6695
Splitting 3	0.6309
Splitting 4	0.6341

Messidor Dataset

The results on Messidor dataset indicate a mean Kappa value of 0.6694 across the four splitting strategies, with a standard deviation of 0.0452, Figure 7.8. The relatively high mean Kappa score demonstrates that the proposed model achieves consistent

agreement with the ground truth, regardless of the splitting strategy. Moreover, the low standard deviation highlights the stability of the model’s performance, suggesting that variations in data splitting have minimal impact on its ability to generalize. This consistency is particularly important in validating the robustness of the model under different experimental conditions. The slightly higher Kappa score observed in our splitting strategy (0.7431) compared to the alternative splits reinforces the appropriateness of this approach for evaluating the model, Table 7.16. However, the limited variation between splitting strategies further supports the claim that the choice of splitting strategy does not significantly influence the model’s performance, underscoring its generalizability and reliability.

Table 7.17: *Comparison of Model Performance Across Different Data Splitting Strategies on **APTOS** Dataset.*

Splitting Strategy	Kappa
Our Splitting 1	0.9324
Splitting 2	0.8943
Splitting 3	0.8925

APTOS Dataset

The results presented in Table 7.17 on APTOS dataset, demonstrate the performance of the proposed model across three splitting strategies for the APTOS dataset. The highest Kappa score of 0.9324 was achieved using "Our Splitting 1," which follows the splitting methodology used in prior studies to ensure consistent evaluation. Splitting strategies 2 and 3 yielded slightly lower Kappa scores of 0.8943 and 0.8925, respectively.

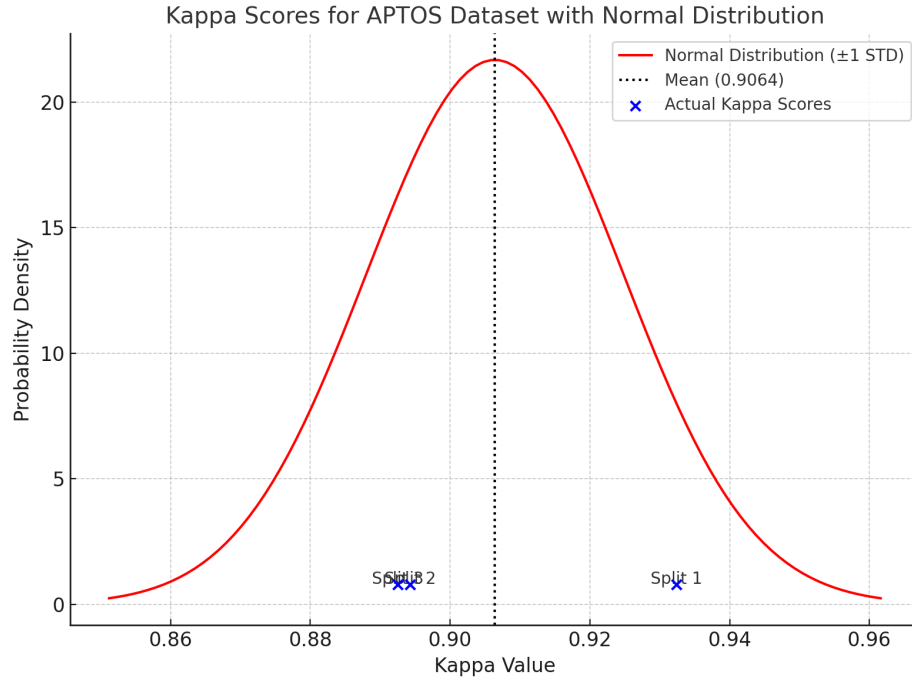


Figure 7.9: *The mean and the STD Distribution of different Splittings Strategies on **APTOS** Dataset.*

The overall mean Kappa value of 0.9064, combined with a low standard deviation of 0.0184, highlights the model’s stability and robustness across different splitting strategies. The minor variation in performance, as illustrated by the visualization of the normal distribution Figure 7.9, suggests that the model is resilient to changes in data splits and maintains strong agreement with the ground truth.

These findings reinforce the claim that the model is resilient to variability in data splits, further validating its generalizability on the APTOS dataset.

7.11 Conclusion and Remarks

Overall, the results of this chapter highlight the importance of both the choice of pre-training data and the architectural design of SSL models, particularly in domain-specific applications like medical imaging. The success of our MsSSL model, achieved through its innovative use of in-domain pre-training with ODM-DDR weights, represents a significant step forward in the quality of learned representations for medical tasks. This improvement is reflected in the model’s performance across multiple datasets and tasks, demonstrating its strong potential for real-world applications. The ability of MsSSL to generalize across different medical tasks and datasets makes it a valuable tool for practitioners, further validating the utility of self-supervised learning in medical image analysis.

The results clearly demonstrate that our SSL-based models, particularly MsSSL, provide superior performance in DR grading compared to the supervised models. The key insights are as follows:

- **Improved Representation Learning:** The SSL models significantly outperform SL models in Kappa, precision, recall, and F1-score, suggesting that the learned representations in SSL are more discriminative and better suited for multi-class DR grading.
- **Performance in Late-Stage DR:** The confusion matrices highlight that SSL models are more effective in classifying the more severe stages of DR (grades 3 and 4), which are critical for timely intervention. This is a significant advantage over the SL models, which tend to misclassify these cases as lower grades.

- **Transfer Learning from ImageNet:** The SL models, which transferred knowledge from ImageNet, struggle with late-stage DR and produce lower Kappa scores. This suggests that the general features learned from ImageNet are insufficient for fine-grained DR diagnosis, where domain-specific information is crucial. In contrast, our SSL approach, which learns from unlabeled medical data, provides a more effective transfer of knowledge to the DR task.
- **Balanced Sensitivity Across Classes:** The higher recall scores in CGSSL and MsSSL indicate that our SSL models are better at identifying DR cases across all grades, addressing the issue of class imbalance more effectively than the SL models.

Overall, these findings underscore the effectiveness of self-supervised learning for medical image analysis tasks, especially when handling complex and imbalanced datasets like APTOS. The significant improvement in later-stage DR classification demonstrates that SSL models, particularly MsSSL, are capable of learning more robust and generalizable features that better capture the nuances of diabetic retinopathy. This points to the limitation of relying on supervised models pre-trained on general datasets like ImageNet and highlights the potential of SSL approaches tailored for medical imaging tasks.

CHAPTER 8

CONCLUSION

8.1 Contributions and Findings

In the previous chapters, we identified key limitations of supervised transfer learning models when applied to domain-specific tasks like diabetic retinopathy (DR) grading. While pretraining on general-purpose datasets such as ImageNet offers a performance boost over random initialization, it does not adequately transfer the specialized knowledge needed for optimal performance in medical imaging tasks. This shortcoming is particularly evident in models like ResNet and ViT, which struggle to capture domain-specific features when their pretraining is not aligned with the task at hand. To address this issue, we turn to self-supervised learning (SSL) as an alternative approach to pretraining that can better leverage the unique characteristics of medical datasets.

Self-supervised learning offers a compelling solution to the challenges posed by supervised learning in specialized fields like medical imaging. In SSL, models are trained to learn representations without requiring labeled data. This is especially advantageous in medical domains where obtaining labeled data can be expensive,

time-consuming, and often subjective. SSL models learn to extract valuable features from the input data itself by solving pretext tasks, such as contrasting positive and negative image pairs, which can be leveraged for downstream tasks like DR grading.

Contrastive learning, a popular SSL paradigm, encourages the model to pull together representations of similar images (positive pairs) while pushing apart representations of dissimilar images (negative pairs). This allows the model to learn robust, generalizable features that are not reliant on manually curated labels. In medical imaging, contrastive learning can be particularly effective in capturing subtle yet important variations between healthy and pathological retinal images, which might otherwise be overlooked by models pre-trained on non-medical datasets like ImageNet.

One of the key advantages of SSL is its ability to harness the rich, domain-specific information inherent in the data itself. By pretraining on medical datasets, SSL models can better capture the nuances and intricacies of medical images, such as retinal abnormalities in DR, which general-purpose datasets like ImageNet cannot offer. This approach directly addresses the knowledge transfer issue identified in the previous chapter, where models pretrained on ImageNet struggled to adapt to the APTOS dataset for DR grading.

Unlike traditional supervised learning, which relies on manually labeled datasets, SSL models can be pretrained on large volumes of unlabeled medical images, allowing them to extract more meaningful features from the data. This leads to more effective knowledge transfer when the model is fine-tuned for domain-specific tasks, such as DR classification. The use of contrastive learning, in particular, ensures that the model focuses on learning differences between healthy and diseased retinal images, resulting

in more accurate and robust predictions.

To fully exploit the potential of SSL, we designed an experiment to pretrain our models on two retinal-specific datasets: DDR and the ODM-DDR dataset, the latter of which was constructed specifically for this study. DDR is a well-known dataset used for DR detection, containing a large variety of retinal images with varying levels of diabetic retinopathy. Pretraining on DDR allows the model to develop a deep understanding of retinal features associated with DR. However, to push the boundaries of feature learning, we also developed the ODM-DDR dataset, carefully curated to capture additional retinal abnormalities not present in DDR. By using these datasets for contrastive learning, our goal was to enhance the model’s ability to recognize subtle and complex features in retinal images that are indicative of DR.

The MsSSL model, pretrained using contrastive learning on the ODM-DDR dataset, aims to not only outperform previous supervised learning models but also set a new benchmark for domain-specific feature extraction. Through contrastive learning, MsSSL is expected to harvest representations that are more aligned with medical imaging needs, leading to better performance when fine-tuned on the APTOS dataset for DR grading.

The shift from supervised to self-supervised learning in this research brings several notable advantages:

- **Reduced Reliance on Labeled Data:** Medical image labeling is often expensive and subjective. SSL eliminates the need for large amounts of labeled data during pretraining, using the images themselves to extract meaningful rep-

representations.

- **Better Domain-Specific Feature Learning:** SSL models pretrained on medical datasets like EyePACS and ODM-DDR learn features that are highly relevant to DR grading, addressing the limitations seen with models pretrained on ImageNet.
- **Improved Transfer Learning:** Models pretrained using contrastive SSL transfer knowledge more effectively to downstream tasks, leading to better performance in domain-specific tasks such as DR grading.
- **Robustness in Complex Medical Tasks:** By learning to distinguish between subtle variations in retinal images, SSL models become more robust to the complexities of medical imaging tasks, such as detecting early signs of DR.

This study highlights the importance of both pretraining strategies and model architecture in developing effective models for diabetic retinopathy (DR) classification. Supervised learning models trained from scratch struggled to capture domain-specific knowledge, underscoring the limitations of random initialization in medical imaging tasks. While pretraining on a general-purpose dataset like ImageNet provided some performance improvements, it became evident that such pretraining does not fully transfer the specialized features required for accurate DR detection.

Self-supervised learning (SSL) models demonstrated a more effective approach to transfer learning, particularly when pretrained on retinal-specific datasets like EyePACS. These models were able to better capture relevant features, resulting in improved adaptability for medical tasks. This reinforces the value of domain-specific

pretraining in medical imaging, where generalized knowledge from large datasets like ImageNet may fall short.

Our proposed MsSSL model exemplifies the power of combining a carefully constructed, domain-specific dataset (ODM-DDR) with a robust model design to optimize knowledge transfer for downstream tasks like DR grading. The tailored feature extraction from ODM-DDR, paired with the model’s ability to effectively learn and transfer knowledgeable representations, resulted in superior performance. These findings emphasize that both specialized datasets and innovative SSL model designs are crucial for advancing the state of the art in medical imaging and achieving reliable, real-world applications in disease detection.

In summary, contrastive self-supervised learning addresses the limitations of supervised transfer learning by leveraging domain-specific datasets and robust pretraining strategies. The use of SSL, particularly when coupled with carefully curated datasets like ODM-DDR, provides a powerful framework for medical image analysis, offering more effective and scalable solutions for DR detection and beyond.

8.2 Thesis Meets the Research Objectives and Questions

In this final chapter, we demonstrate how the research questions and objectives outlined at the beginning of this thesis have been addressed throughout our work. Each **research question (RQ)** corresponds to specific **research objectives (RO)**, and together they form the foundation for our investigation into the use of advanced deep

learning methods, particularly ViT and SSL, for DR. Below, we discuss how our contributions have successfully met these questions and objectives, solidifying the impact of this research on the field of computer vision in medical diagnosis.

1. **Comprehensive Systematic Literature Review (RO1, RQ1):** Through the literature review conducted in Chapter 3, we provided a thorough evaluation of the current methodologies used for DR severity grading, including an analysis of their effectiveness and limitations. This review identified emerging trends in both machine learning and deep learning approaches, highlighting the areas where our research could contribute. By focusing on gaps in the literature, such as the limited application of supervised learning and the reliance on ImageNet for transfer learning, we laid the groundwork for the innovations presented in this thesis.
2. **Evaluation of DR Datasets (RO2, RQ2):** In this thesis, we successfully addressed RO2 by conducting a thorough evaluation of existing Diabetic Retinopathy (DR) datasets to assess their relevance, availability, and suitability for our research goals through the literature review conducted in Chapter 3. In response to our research direction in SSL, we constructed and publicly released the ODM dataset, specifically designed to facilitate SSL by providing dual-view fundus images. This dataset consists of 5,089 dual-view images, offering a unique opportunity to explore how different views (single and dual views) alongside with different color space (both RGB and green-channel) can enhance the learning of

robust representations for DR grading.

Additionally, recognizing that the complexity of our proposed MsSSL framework required a larger dataset, we combined the ODM dataset with the DDR dataset, which is one of the largest publicly available DR datasets after EyePACS. The DDR dataset was selected for its high-quality and diverse samples, which were crucial for improving the SSL pretraining process and achieving robust model performance. This integration allowed us to better address the challenges posed by limited data in medical imaging, particularly for the task of distinguishing adjacent DR grades. This approach provided a comprehensive dataset foundation that supported the development and fine-tuning of models throughout the thesis, meeting the specific needs of both SSL pretraining and supervised evaluation in the context of DR grading.

3. Comparison of ViT and CNN Architectures (RO3, RQ3): In Chapter 5, we presented a comprehensive analysis comparing the scalability and effectiveness of ViT-based and CNN-based models for DR grading in a supervised learning context. Our results demonstrated that transfer learning from ImageNet could enhance model performance, but highlighted the limitations of using out-domain data, particularly when fine-grained domain-specific features are required. This fulfilled our objective of investigating the role of transfer learning and assessing the performance of these cutting-edge models in medical imaging.

4. Exploring Self-Supervised Learning for DR Grading (RO4, RQ4): To

address this objective, we explore the potential of SSL for DDR grading, specifically focusing on how it could enhance representation learning for medical image analysis. In addressing this objective, we developed the **CGSSL** framework, as detailed in Chapter 6. This novel SSL model was designed to learn robust feature representations by leveraging the detailed held in green-channel and applying salient detection to guide the learning.

The results of our experiments clearly demonstrated the advantages of SSL over conventional transfer learning techniques. Through rigorous evaluation, the **CGSSL** model showed that SSL can significantly improve the quality of learned representations, even when faced with the challenges of limited labeled data. This was a critical finding, as the limited availability of annotated medical images often hampers the performance of supervised models. The **CGSSL** framework addressed this challenge by using unlabeled data of dual view fundus images constructed in Chapter 4, to pre-train models and improve their ability to generalize in downstream DR grading tasks.

Our experiments confirmed the effectiveness of SSL in improving model generalization and feature extraction capabilities. By using green channel of fundus images and detection of the salient regions that highlights the region of interest, for pretraining, **CGSSL** was able to capture representative features, resulting in more accurate and robust classification of DR severity.

These findings not only validate the viability of SSL for DR grading but also highlight its superiority over traditional ImageNet-based transfer learning.

While transfer learning can accelerate model development by leveraging pre-trained weights, the domain gap between natural images and medical images limits its effectiveness in specialized tasks like DR classification. In contrast, SSL-based pretraining on medical datasets allows the model to learn more relevant, domain-specific features, enhancing its performance in diagnosing and grading DR.

By successfully developing and testing the **CGSSL** framework, we effectively achieved **RO4** and **RQ4**, providing new insights into the use of self-supervised learning in medical imaging. Our work demonstrates that SSL is a powerful tool for addressing the limitations of traditional transfer learning, offering a pathway toward more accurate, data-efficient, and robust models for DR grading and other medical image analysis tasks.

5. **Enhancing SSL Framework for Domain Adaptation (RO5, RQ5):** In Chapter Chapter 7, we introduced novel strategies for enhancing the **CGSSL** framework by integrating modules such as the MFEE and Deep Learner module (DLe), to introduce MsSSL framework. These enhancements were specifically designed to address the challenge of distinguishing adjacent DR grades, a key issue in DR grading. Our experiments showed that these modules significantly improved the model’s ability to extract high-level and fine-grained features, resulting in more accurate classification, and handling the complexities of DR classification, especially in the more difficult to distinguish late stages of the disease.

6. Disseminating Research Findings (RO6): Throughout the course of this research, we have published one paper in Q1 journal and the rest of papers are submitted waiting for publication. Through these papers we are contributing to the broader research community, ensuring that our results are accessible to other researchers and practitioners in the field.

In summary, this thesis successfully meets the research objectives and answers all research questions. Through our novel MsSSL framework, we have made significant contributions to the field of medical image analysis, particularly in the context of diabetic retinopathy grading, by pushing the boundaries of self-supervised learning, transfer learning, and feature extraction. Our work sets the stage for further research and practical application of SSL techniques in medical diagnosis.

8.3 Limitations and Future Work

While this thesis presents several advancements in self-supervised learning (SSL) and diabetic retinopathy (DR) grading, there are limitations that highlight areas for future research. Addressing these limitations will push the field forward, enabling more robust and accurate solutions for DR diagnosis.

Large Vision Transformer (ViT) Models: While we were able to run up to the H-14 configuration in supervised learning, we could not use larger ViT configurations, such as the L or H variants, as feature extractor encoders within our proposed MsSSL model due to limited computational resources. These larger models have shown significant potential in capturing complex patterns in other domains. Future work could

focus on utilizing distributed computing or cloud-based solutions to train these larger ViT architectures, potentially leading to improved performance in medical image analysis by leveraging their advanced capacity for feature extraction.

Attention Mechanism: In our Dual-view Learning Encoder (DLe) design, we only employed the Convolutional Block Attention Module (CBAM). While CBAM proved effective, it is possible that other attention mechanisms, such as self-attention or squeeze-and-excitation, could perform better. Future work should explore and compare different attention mechanisms to further optimize feature extraction for medical imaging.

Evaluation Datasets: We initially evaluated transferred knowledge on three datasets: APTOS2019, Messidor-2, and DDR. However, toward the end of the experiments, we reduced the evaluation to only two datasets (APTOS and Messidor). Future studies could expand this evaluation by including additional datasets, even with fewer samples, to test model generalization across a broader range of data sources.

Loss Objectives: In our multi-scale self-supervised learning (MsSSL) framework, we focused on two objectives: contrastive loss and a segmentation objective. Future work could explore multi-objective learning by leveraging multi-scale features, where each scale serves as an independent objective, and using weighted loss techniques to balance learning and improve performance.

Knowledge Transfer for Downstream Task: Our study focused on transferring knowledge from SSL to a classification task. Future research should explore transferring learned knowledge to tasks such as segmentation or object detection, which could

enhance localization and diagnostic capabilities in medical imaging.

DR Classification: We concentrated on DR grading, which involves multi-class classification of disease severity. Future work could apply these techniques to binary classification (disease vs. no disease), which may simplify the task and improve performance when data is limited.

Hyperparameter Tuning: In this study, we did not perform extensive hyperparameter tuning, relying on default parameters. Future work should conduct a thorough hyperparameter search to identify optimal settings that can improve model performance further.

Augmentation Strategies: We used the default augmentations provided by the SimCLR framework for SSL. Future work could explore augmentation strategies tailored specifically to medical imaging, such as vessel enhancement or contrast adjustments, to improve model robustness in DR grading.

Training Epochs in Supervised Learning: In our supervised learning experiments, we limited training to 25 epochs to ensure controlled comparisons. Future research could implement early stopping or longer training schedules, potentially achieving optimal performance without overfitting.

Preprocessing: Our preprocessing approach was minimal, primarily relying on the raw images from the datasets. Future research could explore advanced preprocessing techniques, such as image quality assessment, lesion enhancement, or noise and illumination reduction, to improve input quality and model performance.

Pre-training Duration: We trained our proposed models, CGSSL and MsSSL, for 300 epochs. Future work could explore the effects of extending training duration, investigating whether more epochs lead to better performance or if early stopping could be used more effectively.

In addition to addressing the limitations identified in this research, several promising avenues for future work have emerged, which could significantly advance the field of diabetic retinopathy (DR) diagnosis. These are not constraints of the current study, but rather opportunities to expand on the methodologies presented and push the boundaries of what is possible in DR detection and grading. Exploring these future directions could introduce innovative approaches and enhance the existing framework, offering new insights and capabilities for DR diagnosis. The following suggested future works could shed light on new directions to further enhance DR grading and improve model performance.

Exploration of Other SSL Frameworks: While we used contrastive SSL in our study, other SSL frameworks such as SimSiam, BYOL, or SwAV may also offer advantages. Future work should compare these frameworks to determine which is best suited for the complexities of medical imaging and DR diagnosis.

Exploration of Few-Shot Learning: We did not explore few-shot learning, which could be highly beneficial in medical imaging where labeled data is scarce. Future studies should investigate how few-shot learning can enhance DR grading models, allowing them to perform well even with minimal training data.

In-Domain Pretext Tasks: While contrastive SSL has been widely used, there is

still a need for pretext tasks specifically tailored to medical imaging. Medical images, particularly in DR, are characterized by subtle lesions and high variability. Future research should focus on designing domain-specific pretext tasks, incorporating clinical knowledge to enhance SSL feature extraction for DR and other medical conditions.

Generalizability Across Demographics and Imaging Systems: An important direction for future work is to evaluate the performance of the MsSSL model across different demographic groups, as variations in medical data may impact model generalizability. Furthermore, testing the model with retinal images captured using different camera systems would help assess its robustness and adaptability to varying image quality and technical settings, ensuring broader applicability in real-world medical environments.

Domain Expert Evaluation: The proposed MsSSL model could be further validated through evaluation by domain experts. Involving medical professionals in the assessment process would provide critical insights into the model’s practical utility and accuracy in real-world clinical settings. This expert evaluation would help ensure that the model’s predictions align with clinical interpretations, offering a deeper understanding of its strengths and limitations in diabetic retinopathy grading.

REFERENCES

- [1] C. Wilkinson, F. L. Ferris III, R. E. Klein, P. P. Lee, C. D. Agardh, M. Davis, D. Dills, A. Kampik, R. Pararajasegaram, J. T. Verdaguer *et al.*, “Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales,” *Ophthalmology*, vol. 110, no. 9, pp. 1677–1682, 2003.
- [2] D. Atlas *et al.*, “International diabetes federation,” *IDF Diabetes Atlas, 7th edn. Brussels, Belgium: International Diabetes Federation*, 2015.
- [3] “Desktop fundus camera,” <https://www.insighteye2020.com/product/topcon-nw8-non-myd-fundus-camera/>, accessed: 2022-06-15.
- [4] “Portable fundus camera,” <https://www.zenroxmed.com/product/portable-fundus-camera/>, accessed: 2022-06-15.
- [5] G. E. Associates. (2022) Gadsden eye associates. [Online]. Available: <https://gadsdeneye.com/diabetic-retinopathy/>
- [6] H. Liu, K. Yue, S. Cheng, C. Pan, J. Sun, and W. Li, “Hybrid model structure for diabetic retinopathy classification,” *Journal of Healthcare Engineering*, vol. 2020, 2020.

- [7] J. Lee, B. C. Y. Zee, and Q. Li, “Detection of neovascularization based on fractal and texture analysis with interaction effects in diabetic retinopathy,” *PloS one*, vol. 8, no. 12, p. e75699, 2013.
- [8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [9] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
- [10] “World health organization.” [Online]. Available: <https://www.who.int/>
- [11] S. Wild, G. Roglic, A. Green, R. Sicree, and H. King, “Global prevalence of diabetes: estimates for the year 2000 and projections for 2030,” *Diabetes care*, vol. 27, no. 5, pp. 1047–1053, 2004.
- [12] A. Sugeno, Y. Ishikawa, T. Ohshima, and R. Muramatsu, “Simple methods for the lesion detection and severity grading of diabetic retinopathy by image processing and transfer learning,” *Computers in Biology and Medicine*, vol. 137, 10 2021.
- [13] C. Suedumrong, K. Leksakul, P. Wattana, and P. Chaopaisarn, “Application of deep convolutional neural networks vgg-16 and googlenet for level diabetic retinopathy detection,” *Lecture Notes in Networks and Systems*, vol. 359 LNNS, pp. 56–65, 2022.

- [14] E. Elyan, P. Vuttipittayamongkol, P. Johnston, K. Martin, K. McPherson, C. F. Moreno-García, C. Jayne, M. K. Sarker *et al.*, “Computer vision and machine learning for medical image analysis: recent advances, challenges, and way forward.” *Artificial Intelligence Surgery*, vol. 2, 2022.
- [15] M. M. Islam, H.-C. Yang, T. N. Poly, W.-S. Jian, and Y.-C. J. Li, “Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis,” *Computer Methods and Programs in Biomedicine*, vol. 191, p. 105320, 2020.
- [16] M. C. S. Tang, S. S. Teoh, H. Ibrahim, and Z. Embong, “Neovascularization detection and localization in fundus images using deep learning,” *Sensors*, vol. 21, no. 16, p. 5327, 2021.
- [17] C. N. K. Michael Abràmoff, “Chapter 6 - image processing,” *Retina (Fifth Edition)*, vol. 1, no. -, pp. 151–176, 2013.
- [18] M. D. Abràmoff, M. K. Garvin, and M. Sonka, “Retinal imaging and image analysis,” *IEEE reviews in biomedical engineering*, vol. 3, pp. 169–208, 2010.
- [19] Z. Shen, H. Fu, J. Shen, and L. Shao, “Understanding and correcting low-quality retinal fundus images for clinical analysis,” *arXiv preprint arXiv:2005.05594*, 2020.
- [20] A. D. Association. (2022) American diabetes association. [Online]. Available: <https://www.diabetes.org/>

- [21] M. Aouf, D. Ali, and G. Kareem, “Optic disc and optic cup segmentation methodology for glaucoma detection,” in *Conf. Paper*, 2015.
- [22] U. Ishtiaq, S. Abdul Kareem, E. R. M. F. Abdullah, G. Mujtaba, R. Jahangir, and H. Y. Ghafoor, “Diabetic retinopathy detection through artificial intelligent techniques: a review and open issues,” *Multimedia Tools and Applications*, vol. 79, no. 21, pp. 15 209–15 252, 2020.
- [23] N. Tsiknakis, D. Theodoropoulos, G. Manikis, E. Ktistakis, O. Boutsora, A. Berto, F. Scarpa, A. Scarpa, D. I. Fotiadis, and K. Marias, “Deep learning for diabetic retinopathy detection and classification based on fundus images: A review,” *Computers in Biology and Medicine*, vol. 135, p. 104599, 2021.
- [24] W. L. Alyoubi, W. M. Shalash, and M. F. Abulkhair, “Diabetic retinopathy detection through deep learning techniques: A review,” *Informatics in Medicine Unlocked*, vol. 20, p. 100377, 2020.
- [25] I. Kandel and M. Castelli, “Transfer learning with convolutional neural networks for diabetic retinopathy image classification. a review,” *Applied Sciences*, vol. 10, no. 6, p. 2021, 2020.
- [26] N. Asiri, M. Hussain, F. Al Adel, and N. Alzaidi, “Deep learning based computer-aided diagnosis systems for diabetic retinopathy: A survey,” *Artificial intelligence in medicine*, vol. 99, p. 101701, 2019.

- [27] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. Van Ginneken, “Ridge-based vessel segmentation in color images of the retina,” *IEEE transactions on medical imaging*, vol. 23, no. 4, pp. 501–509, 2004.
- [28] B. Al-Diri, A. Hunter, D. Steel, M. Habib, T. Hudaib, and S. Berry, “A reference data set for retinal vessel profiles,” in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2008, pp. 2262–2265.
- [29] M. M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen, and S. A. Barman, “An ensemble classification-based approach applied to retinal blood vessel segmentation,” *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 9, pp. 2538–2548, Sep. 2012.
- [30] T. Köhler, A. Budai, M. F. Kraus, J. Odstrčilík, G. Michelson, and J. Hornegger, “Automatic no-reference quality assessment for retinal fundus images using vessel segmentation,” in *Proceedings of the 26th IEEE international symposium on computer-based medical systems*. IEEE, 2013, pp. 95–100.
- [31] Y. Zhou, B. Wang, L. Huang, S. Cui, and L. Shao, “A benchmark for studying diabetic retinopathy: segmentation, grading, and transferability,” *IEEE Transactions on Medical Imaging*, vol. 40, no. 3, pp. 818–828, 2020.
- [32] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabudhe, and F. Meriaudeau, “Indian diabetic retinopathy image dataset (idrid): a

- database for diabetic retinopathy screening research,” *Data*, vol. 3, no. 3, p. 25, 2018.
- [33] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang, “Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening,” *Information Sciences*, vol. 501, pp. 511–522, 2019.
- [34] E. Decenciere, G. Cazuguel, X. Zhang, G. Thibault, J.-C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, R. Danno *et al.*, “Teleophtha: Machine learning and image processing methods for teleophthalmology,” *Irbm*, vol. 34, no. 2, pp. 196–203, 2013.
- [35] the American Society of Retina Specialist. (2019) Retina image bank dataset. [Online]. Available: <https://imagebank.asrs.org/>
- [36] Kaggle2015. (2012) Diabetic retinopathy detection,. [Online]. Available: <https://www.kaggle.com/c/diabetic-retinopathy-detection>
- [37] E. Decenci re, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay *et al.*, “Feedback on a publicly distributed image database: the messidor database,” *Image Analysis & Stereology*, vol. 33, no. 3, pp. 231–234, 2014.
- [38] DeepDR, “Deep diabetic retinopathy (deepdr),” 2019.
- [39] A. Hoover, “Structured analysis of the retina stare,” 2015.

- [40] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, A. Raninen, R. Voutilainen, H. Uusitalo, H. Kälviäinen, and J. Pietilä, “Evaluation database and methodology for diabetic retinopathy algorithms,” 2007.
- [41] —, “The diaretdb1 diabetic retinopathy database and evaluation protocol.” in *BMVC*, vol. 1, 2007, pp. 1–10.
- [42] Kaggle2019. (2019) Diabetic retinopathy detection,. [Online]. Available: <https://www.kaggle.com/competitions/aptos2019-blindness-detection/overview>
- [43] C. Mahiba and A. Jayachandran, “Severity analysis of diabetic retinopathy in retinal images using hybrid structure descriptor and modified cnns,” *Measurement: Journal of the International Measurement Confederation*, vol. 135, pp. 762–767, 3 2019.
- [44] T. Nazir, A. Irtaza, Z. Shabbir, A. Javed, U. Akram, and M. T. Mahmood, “Diabetic retinopathy detection through novel tetragonal local octa patterns and extreme learning machines,” *Artificial Intelligence in Medicine*, vol. 99, 8 2019.
- [45] A. S. Jadhav, P. B. Patil, and S. Biradar, “Optimal feature selection-based diabetic retinopathy detection using improved rider optimization algorithm enabled with deep learning,” *Evolutionary Intelligence*, vol. 14, pp. 1431–1448, 12 2021.
- [46] E. Abdelmaksoud, S. El-Sappagh, S. Barakat, T. Abuhmed, and M. Elmogy, “Automatic diabetic retinopathy grading system based on detecting multiple retinal lesions,” *IEEE Access*, vol. 9, pp. 15 939–15 960, 2021.

- [47] M. K. Yaqoob, S. F. Ali, I. Kareem, and M. M. Fraz, "Feature-based optimized deep residual network architecture for diabetic retinopathy detection," *IEEE International Multi-Topic Conference, INMIC 2020*, 11 2020.
- [48] S. S. Rahim, V. Palade, I. Almakky, and A. Holzinger, "Detection of diabetic retinopathy and maculopathy in eye fundus images using deep learning and image augmentation," *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, vol. 11713 LNCS, pp. 114–127, 2019.
- [49] M. S. Chowdhury, F. R. Taimy, N. Sikder, and A.-A. Nahid, "Diabetic retinopathy classification with a light convolutional neural network," *International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering*, 2019.
- [50] S. N. Chakravarthy, H. Singhal, and N. R. Yadav, "Dr-net: A stacked convolutional classifier framework for detection of diabetic retinopathy," *International Joint Conference on Neural Networks.*, 2019. [Online]. Available: <http://www.ieee.org/publications>
- [51] M. Arora and M. Pandey, "Deep neural network for diabetic retinopathy detection," *International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con)*, 2019.
- [52] E. Abdelmaksoud, S. Barakat, and M. Elmogy, "Diabetic retinopathy grading based on a hybrid deep learning model," *International Conference on Data*

Analytics for Business and Industry: Way Towards a Sustainable Economy, ICDABI 2020, 10 2020.

- [53] D. J. Hemanth, O. Deperlioglu, and U. Kose, “An enhanced diabetic retinopathy detection and classification approach using deep convolutional neural network,” *Neural Computing and Applications*, vol. 32, pp. 707–721, 2 2020.
- [54] Z. Qian, C. Wu, H. Chen, and M. Chen, “Diabetic retinopathy grading using attention based convolution neural network,” *IEEE Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pp. 2652–2655, 2021.
- [55] F. Saeed, M. Hussain, S. Member, H. A. Aboalsamh, F. A. Adel, and A. M. A. Owaifeer, “Diabetic retinopathy screening using custom-designed convolutional neural network,” *arXiv preprint arXiv:2110.03877*, 2021.
- [56] Y.-P. Liu, Z. Li, C. Xu, J. Li, and R. Liang, “Referable diabetic retinopathy identification from eye fundus images with weighted path for convolutional neural network,” *Artificial intelligence in medicine*, vol. 99, p. 101694, 2019.
- [57] W. L. Alyoubi, M. F. Abulkhair, and W. M. Shalash, “Diabetic retinopathy fundus image classification and lesions localization system using deep learning,” *Sensors*, vol. 21, 6 2021.
- [58] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

- [59] S. Wan, Y. Liang, and Y. Zhang, “Deep convolutional neural networks for diabetic retinopathy detection by image classification,” *Computers & Electrical Engineering*, vol. 72, pp. 274–282, 2018.
- [60] C. Lian, Y. Liang, R. Kang, and Y. Xiang, “Deep convolutional neural networks for diabetic retinopathy classification,” *ACM International Conference Proceeding Series*, pp. 68–72, 6 2018.
- [61] Z. Zhao, K. Zhang, X. Hao, J. Tian, M. C. H. Chua, L. Chen, and X. Xu, “Bira-net: Bilinear attention net for diabetic retinopathy grading,” *IEEE International Conference on Image Processing (ICIP)*, 2019.
- [62] A. Lands, A. J. Kottarathil, A. Biju, E. M. Jacob, and S. Thomas, “Implementation of deep learning based algorithms for diabetic retinopathy classification from fundus images,” *International Conference on Trends in Electronics and Informatics (ICOEI 2020)*, 2020.
- [63] L. Dai, L. Wu, H. Li, C. Cai, Q. Wu, H. Kong, R. Liu, X. Wang, X. Hou, Y. Liu, X. Long, Y. Wen, L. Lu, Y. Shen, Y. Chen, D. Shen, X. Yang, H. Zou, B. Sheng, and W. Jia, “A deep learning system for detecting diabetic retinopathy across the disease spectrum,” *Nature Communications*, vol. 12, 12 2021.
- [64] C. H. Hua, K. Kim, T. Huynh-The, J. I. You, S. Y. Yu, T. Le-Tien, S. H. Bae, and S. Lee, “Convolutional network with twofold feature augmentation for diabetic retinopathy recognition from multi-modal images,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, pp. 2686–2697, 7 2021.

- [65] A. C. Metan, A. Lambert, and M. Pickering, “Small scale featurepropagation using deep residual learningfor diabetic retinopathy classification,” *IEEE International Conference on Image, Vision and Computing*, 2019.
- [66] S. Qummar, F. G. Khan, S. Shah, A. Khan, S. Shamshirband, Z. U. Rehman, I. A. Khan, and W. Jadoon, “A deep learning ensemble approach for diabetic retinopathy detection,” *IEEE Access*, vol. 7, pp. 150 530–150 539, 2019.
- [67] W. Zhang, J. Zhong, S. Yang, Z. Gao, J. Hu, Y. Chen, and Z. Yi, “Automated identification and grading system of diabetic retinopathy using deep neural networks,” *Knowledge-Based Systems*, vol. 175, pp. 12–25, 7 2019.
- [68] K. Harihanth and B. Karthikeyan, “Diabetic retinopathy detection using ensemble deep learning and individual channel training,” *Proceedings of the 3rd International Conference on Intelligent Sustainable Systems, ICISS 2020*, pp. 1042–1049, 12 2020.
- [69] M. R. Islam, M. A. M. Hasan, and A. Sayeed, “Transfer learning based diabetic retinopathy detection with a novel preprocessed layer,” *IEEE Region 10 Symposium (TENSYP)*, 2020.
- [70] N. Doshi, U. Oza, and P. Kumar, “Diabetic retinopathy classification using downscaling algorithms and deep learning,” *nternational Conference on Signal Processing and Integrated Networks (SPIN)*, 2020.

- [71] A. Samanta, A. Saha, S. C. Satapathy, S. L. Fernandes, and Y. D. Zhang, “Automated detection of diabetic retinopathy using convolutional neural networks on a small dataset,” *Pattern Recognition Letters*, vol. 135, pp. 293–298, 7 2020.
- [72] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.
- [73] M. Toğaçar, “Detection of retinopathy disease using morphological gradient and segmentation approaches in fundus images,” *Computer Methods and Programs in Biomedicine*, vol. 214, 2 2022.
- [74] G. Çınarar and K. Kiliç, “Diabetic retinopathy detection with deep transfer learning methods,” *Lecture Notes in Networks and Systems*, vol. 308, pp. 147–154, 2022.
- [75] V. Vives-Boix and D. Ruiz-Fernández, “Diabetic retinopathy detection through convolutional neural networks with synaptic metaplasticity,” *Computer Methods and Programs in Biomedicine*, vol. 206, 7 2021.
- [76] F. J. Martinez-Murcia, A. Ortiz, J. Ramírez, J. M. Górriz, and R. Cruz, “Deep residual transfer learning for automatic diagnosis and grading of diabetic retinopathy,” *Neurocomputing*, vol. 452, pp. 424–434, 9 2021.
- [77] F. Saeed, M. Hussain, and H. A. Aboalsamh, “Automatic diabetic retinopathy diagnosis using adaptive fine-tuned convolutional neural network,” *IEEE Access*, vol. 9, pp. 41 344–41 359, 2021.

- [78] J. He, L. Shen, X. Ai, and X. Li, “Diabetic retinopathy grade and macular edema risk classification using convolutional neural networks,” *IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS)*, 2019.
- [79] K. Zhou, Z. Gu, W. Liu, W. Luo, J. Cheng, S. Gao, and J. Liu, “Multi-cell multi-task convolutional neural networks for diabetic retinopathy grading,” *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 8 2018. [Online]. Available: <http://arxiv.org/abs/1808.10564>
- [80] Y. X. L. R. K. Y. M. Chun Yan, Lian, Y. Xiang, and L. Gao, “A novel diabetic retinopathy classification scheme based on compact bilinear pooling cnn and gradient boosted decision tree,” *Journal of Information Hiding and Multimedia Signal Processing c*, vol. 9, 2018.
- [81] E. Bhatti and P. Kaur, “Draodm: Diabetic retinopathy analysis through optimized deep learning with multi support vector machine for classification,” *Communications in Computer and Information Science*, vol. 1036, pp. 174–188, 2019.
- [82] S. Taufiqurrahman, A. Handayani, B. R. Hermanto, and T. L. E. R. Mengko, “Diabetic retinopathy classification using a hybrid and efficient mobilenetv2-svm model,” *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, vol. 2020-November, pp. 235–240, 11 2020.
- [83] C. Bhardwaj, S. Jain, and M. Sood, “Transfer learning based robust automatic detection system for diabetic retinopathy grading,” *Neural Computing and Applications*, 2021.

- [84] S. Gayathri, V. P. Gopi, and P. Palanisamy, “Diabetic retinopathy classification based on multipath cnn and machine learning classifiers,” *Physical and Engineering Sciences in Medicine*, vol. 44, pp. 639–653, 9 2021.
- [85] —, “A lightweight cnn for diabetic retinopathy classification from fundus images,” *Biomedical Signal Processing and Control*, vol. 62, 9 2020.
- [86] V. Deepa, C. S. Kumar, and T. Cherian, “Ensemble of multi-stage deep convolutional neural networks for automated grading of diabetic retinopathy using image patches,” *Journal of King Saud University - Computer and Information Sciences*, 2021.
- [87] M. K. Yaqoob, S. F. Ali, M. Bilal, M. S. Hanif, and U. M. Al-Saggaf, “Resnet based deep features and random forest classifier for diabetic retinopathy detection†,” *Sensors*, vol. 21, 6 2021.
- [88] M. Niemeijer, B. Van Ginneken, M. J. Cree, A. Mizutani, G. Quellec, C. I. Sánchez, B. Zhang, R. Hornero, M. Lamard, C. Muramatsu *et al.*, “Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs,” *IEEE transactions on medical imaging*, vol. 29, no. 1, pp. 185–195, 2009.
- [89] X. Zeng, H. Chen, Y. Luo, and W. Ye, “Automated diabetic retinopathy detection based on binocular siamese-like convolutional neural network,” *IEEE Access*, vol. 7, pp. 30 744–30 753, 2019.

- [90] G. Kalyani, B. Janakiramaiah, A. Karuna, and L. V. N. Prasad, “Diabetic retinopathy detection and classification using capsule networks,” *Complex Intelligent Systems*, 3 2021.
- [91] G. Kumar, S. Chatterjee, and C. Chattopadhyay, “Dristi: a hybrid deep neural network for diabetic retinopathy diagnosis,” *Signal, Image and Video Processing*, 2021.
- [92] Z. Khan, F. G. Khan, A. Khan, Z. U. Rehman, S. Shah, S. Qummar, F. Ali, and S. Pack, “Diabetic retinopathy detection using vgg-nin a deep learning architecture,” *IEEE Access*, vol. 9, pp. 61 408–61 416, 2021.
- [93] C. Matsoukas, J. F. Haslum, M. Söderberg, and K. Smith, “Is it time to replace cnns with transformers for medical images?” *arXiv e-prints*, 8 2021.
[Online]. Available: <http://arxiv.org/abs/2108.09038>
- [94] Y. Li, Z. Song, S. Kang, S. Jung, and W. Kang, “Semi-supervised auto-encoder graph network for diabetic retinopathy grading,” *IEEE Access*, vol. 9, pp. 140 759–140 767, 2021.
- [95] G. Zhang, J. Pan, Z. Zhang, H. Zhang, C. Xing, B. Sun, and M. Li, “Hybrid graph convolutional network for semi-supervised retinal image classification,” *IEEE Access*, vol. 9, pp. 35 778–35 789, 2021.
- [96] Y. Luo, J. Pan, S. Fan, Z. Du, and G. Zhang, “Retinal image classification by self-supervised fuzzy clustering network,” *IEEE Access*, vol. 8, pp. 92 352–92 362, 2020.

- [97] S. Lal, S. U. Rehman, J. H. Shah, T. Meraj, H. T. Rauf, R. Damaševičius, M. A. Mohammed, and K. H. Abdulkareem, “Adversarial attack and defence through adversarial training and feature fusion for diabetic retinopathy recognition,” *Sensors*, vol. 21, 6 2021.
- [98] R. E. Hacisoftoglu, M. Karakaya, and A. B. Sallam, “Deep learning frameworks for diabetic retinopathy detection with smartphone-based retinal imaging systems,” *Pattern Recognition Letters*, vol. 135, pp. 409–417, 7 2020.
- [99] Y. Zhou, X. He, L. Huang, L. Liu, F. Zhu, S. Cui, and L. Shao, “Collaborative learning of semi-supervised segmentation and classification for medical images,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2079–2088.
- [100] N. Barhate, S. Bhawe, R. Bhise, R. G. Sutar, and D. C. Karia, “Reducing overfitting in diabetic retinopathy detection using transfer learning,” *IEEE 5th International Conference on Computing Communication and Automation, ICCCA 2020*, pp. 298–301, 10 2020.
- [101] Z. Zhao, K. Chopra, Z. Zeng, and X. Li, “Sea-net: Squeeze-and-excitation attention net for diabetic retinopathy grading,” *IEEE International Conference on Image Processing (ICIP)*, 2020.
- [102] P. Burlina, W. Paul, P. Mathew, N. Joshi, K. D. Pacheco, and N. M. Bressler, “Low-shot deep learning of diabetic retinopathy with potential applications to

- address artificial intelligence bias in retinal diagnostics and rare ophthalmic diseases,” *JAMA Ophthalmology*, vol. 138, pp. 1070–1077, 10 2020.
- [103] H. Jiang, J. Xu, R. Shi, K. Yang, D. Zhang, M. Gao, H. Ma, and W. Qian, “A multi-label deep learning model with interpretable grad-cam for diabetic retinopathy classification,” *Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2020.
- [104] A. Foo, W. Hsu, M. L. Lee, G. Lim, and T. Y. Wong, “Multi-task learning for diabetic retinopathy grading and lesion segmentation,” *The Thirty-Second Innovative Applications of Artificial Intelligence Conference (IAAI-20)*, 2020.
[Online]. Available: www.aaai.org
- [105] S. S. Karki and P. Kulkarni, “Diabetic retinopathy classification using a combination of efficientnets,” *International Conference on Emerging Smart Computing and Informatics, ESCI 2021*, pp. 68–72, 3 2021.
- [106] L. Alzubaidi, J. Santamaría, M. Manoufali, B. Mohammed, M. A. Fadhel, J. Zhang, A. H. Al-Timemy, O. Al-Shamma, and Y. Duan, “Mednet: Pre-trained convolutional neural network model for the medical imaging tasks,” *arXiv preprint arXiv:2110.06512*, 2021.
- [107] J. Hou, J. Xu, F. Xiao, R.-W. Zhao, Y. Zhang, H. Zou, L. Lu, W. Xue, and R. Feng, “Cross-field transformer for diabetic retinopathy grading on two-field fundus images,” in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE Computer Society, 2022, pp. 985–990.

- [108] R. Liu, X. Wang, Q. Wu, L. Dai, X. Fang, T. Yan, J. Son, S. Tang, J. Li, Z. Gao, A. Galdran, J. Poorneshwaran, H. Liu, J. Wang, Y. Chen, P. Porwal, G. S. Wei Tan, X. Yang, C. Dai, H. Song, M. Chen, H. Li, W. Jia, D. Shen, B. Sheng, and P. Zhang, “Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge,” *Patterns*, p. 100512, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666389922001040>
- [109] APTOS2019, “Diabetic retinopathy detection,,” 2019. [Online]. Available: <https://www.kaggle.com/competitions/aptos2019-blindness-detection/overview>
- [110] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, “A brief review of domain adaptation,” *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pp. 877–894, 2021.
- [111] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014. [Online]. Available: <https://arxiv.org/abs/1409.1556>
- [112] A. C. Metan, A. Lambert, and M. Pickering, “Small scale featurepropagation using deep residual learningfor diabetic retinopathy classification,” *IEEE International Conference on Image, Vision and Computing*, 2019.
- [113] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” 2015. [Online]. Available: <https://arxiv.org/abs/1512.00567>

- [114] K. Ohri, M. Kumar, and D. Sukheja, “Self-supervised approach for diabetic retinopathy severity detection using vision transformer,” *Progress in Artificial Intelligence*, pp. 1–19, 2024.
- [115] P. Celard, E. L. Iglesias, J. M. Sorribes-Fdez, R. Romero, A. S. Vieira, and L. Borrajo, “A survey on deep learning applied to medical images: from simple artificial neural networks to generative models,” *Neural Computing and Applications*, vol. 35, no. 3, pp. 2291–2323, 2023.
- [116] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, “Transfusion: Understanding transfer learning for medical imaging,” *Advances in neural information processing systems*, vol. 32, 2019.
- [117] ———, “Transfusion: Understanding transfer learning for medical imaging,” *Advances in neural information processing systems*, vol. 32, 2019.
- [118] A. Ke, W. Ellsworth, O. Banerjee, A. Y. Ng, and P. Rajpurkar, “Chextransfer: performance and parameter efficiency of imagenet models for chest x-ray interpretation,” in *Proceedings of the conference on health, inference, and learning*, 2021, pp. 116–124.
- [119] K. He, R. Girshick, and P. Dollár, “Rethinking imagenet pre-training,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4918–4927.
- [120] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen *et al.*, “Big self-supervised models

- advance medical image classification,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3478–3488.
- [121] B. VanBerlo, J. Hoey, and A. Wong, “A survey of the impact of self-supervised pretraining for diagnostic tasks in medical x-ray, ct, mri, and ultrasound,” *BMC Medical Imaging*, vol. 24, no. 1, p. 79, 2024.
- [122] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [123] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [124] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.
- [125] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, “Accurate, large minibatch sgd: Training imagenet in 1 hour,” *arXiv preprint arXiv:1706.02677*, 2017.
- [126] Y. Huang, J. Lyu, P. Cheng, R. Tam, and X. Tang, “Ssit: Saliency-guided self-supervised image transformer for diabetic retinopathy grading,” *IEEE TRANSACTIONS ON MEDICAL IMAGING*, 2022.

- [127] A. He, T. Li, N. Li, K. Wang, and H. Fu, “Cabnet: Category attention block for imbalanced diabetic retinopathy grading,” *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 143–153, 2020.
- [128] Y. Huang, L. Lin, P. Cheng, J. Lyu, R. Tam, and X. Tang, “Identifying the key components in resnet-50 for diabetic retinopathy grading from fundus images: a systematic investigation,” *Diagnostics*, vol. 13, no. 10, p. 1664, 2023.
- [129] T. Truong, S. Mohammadi, and M. Lenga, “How transferable are self-supervised features in medical image classification tasks?” in *Machine Learning for Health*. PMLR, 2021, pp. 54–74.
- [130] A. K. Gangwar and V. Ravi, “Diabetic retinopathy detection using transfer learning and deep learning,” in *Evolution in Computational Intelligence: Frontiers in Intelligent Computing: Theory and Applications (FICTA 2020), Volume 1*. Springer, 2021, pp. 679–689.
- [131] E. AbdelMaksoud, S. Barakat, and M. Elmogy, “A computer-aided diagnosis system for detecting various diabetic retinopathy grades based on a hybrid deep learning technique,” *Medical & Biological Engineering & Computing*, vol. 60, no. 7, pp. 2015–2038, 2022.
- [132] S. H. Kassani, P. H. Kassani, R. Khazaeinezhad, M. J. Wesolowski, K. A. Schneider, and R. Deters, “Diabetic retinopathy classification using a modified xception architecture,” in *2019 IEEE international symposium on signal processing and information technology (ISSPIT)*. IEEE, 2019, pp. 1–6.

- [133] D. R. Shurrab S, “Self-supervised learning methods and applications in medical imaging analysis: a survey,” *PeerJ Computer Science* 8:e1045, 2022.
- [134] C. Doersch, A. Gupta, and A. A. Efros, “Unsupervised visual representation learning by context prediction,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1422–1430.
- [135] D. Kim, D. Cho, D. Yoo, and I. S. Kweon, “Learning image representations by completing damaged jigsaw puzzles,” in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 793–802.
- [136] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.
- [137] S. N. Chakravarthy, H. Singhal, and N. R. Yadav, “Dr-net: A stacked convolutional classifier framework for detection of diabetic retinopathy,” 2019. [Online]. Available: <http://www.ieee.org/publications>
- [138] G. Kalyani, B. Janakiramaiah, A. Karuna, and L. V. N. Prasad, “Diabetic retinopathy detection and classification using capsule networks,” *Complex & Intelligent Systems*, 3 2021.
- [139] M. S. Chowdhury, F. R. Taimy, N. Sikder, and A.-A. Nahid, “Diabetic retinopathy classification with a light convolutional neural network,” in *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*. IEEE, 2019, pp. 1–4.

- [140] W. Almattar, H. Luqman, and F. A. Khan, “Diabetic retinopathy grading review: Current techniques and future directions,” *Image and Vision Computing*, p. 104821, 2023.
- [141] C. Suedumrong, K. Leksakul, P. Wattana, and P. Chaopaisarn, “Application of deep convolutional neural networks vgg-16 and googlenet for level diabetic retinopathy detection,” in *Proceedings of the Future Technologies Conference (FTC) 2021, Volume 2*. Springer, 2022, pp. 56–65.
- [142] C. Lian, Y. Liang, R. Kang, and Y. Xiang, “Deep convolutional neural networks for diabetic retinopathy classification,” *ACM International Conference Proceeding Series*, pp. 68–72, 6 2018.
- [143] A. Lands, A. J. Kottarathil, A. Biju, E. M. Jacob, and S. Thomas, “Implementation of deep learning based algorithms for diabetic retinopathy classification from fundus images,” in *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*. IEEE, 2020, pp. 1028–1032.
- [144] K. Harihanth and B. Karthikeyan, “Diabetic retinopathy detection using ensemble deep learning and individual channel training,” in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*. IEEE, 2020, pp. 1042–1049.
- [145] S. Qummar, F. G. Khan, S. Shah, A. Khan, S. Shamshirband, Z. U. Rehman, I. A. Khan, and W. Jadoon, “A deep learning ensemble approach for diabetic retinopathy detection,” *IEEE Access*, vol. 7, pp. 150 530–150 539, 2019.

- [146] M. Canayaz, S. Şehribanoğlu, M. Özgökçe, and M. B. Akıncı, “A comprehensive exploration of deep learning approaches for pulmonary nodule classification and segmentation in chest ct images,” *Neural Computing and Applications*, vol. 36, no. 13, pp. 7245–7264, 2024.
- [147] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, “Going deeper with image transformers,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 32–42.
- [148] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, “Training data-efficient image transformers distillation through attention,” in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 10 347–10 357. [Online]. Available: <https://proceedings.mlr.press/v139/touvron21a.html>
- [149] H. Bao, L. Dong, S. Piao, and F. Wei, “Beit: Bert pre-training of image transformers,” in *International Conference on Learning Representations*, 2021.
- [150] C. Adak, T. Karkera, S. Chattopadhyay, and M. Saqib, “Detecting severity of diabetic retinopathy from fundus images using ensembled transformers,” *arXiv preprint arXiv:2301.00973*, 2023.
- [151] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

- [152] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.
- [153] X. Chen, S. Xie, and K. He, “An empirical study of training self-supervised vision transformers,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9640–9649.
- [154] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging properties in self-supervised vision transformers,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.
- [155] S. Montabone and A. Soto, “Human detection using a mobile platform and novel features derived from a visual saliency mechanism,” *Image and Vision Computing*, vol. 28, no. 3, pp. 391–402, 2010.
- [156] Y. Huang, L. Lin, P. Cheng, J. Lyu, and X. Tang, “Lesion-based contrastive learning for diabetic retinopathy grading from fundus images,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:236088105>
- [157] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, “Bootstrap your own latent-a new approach to self-supervised learning,” *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

- [158] X. Chen and K. He, “Exploring simple siamese representation learning,” *arXiv preprint arXiv:2011.10566*, 2020.
- [159] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.
- [160] K. Gupta, T. Ajanthan, A. v. d. Hengel, and S. Gould, “Understanding and improving the role of projection head in self-supervised learning,” *arXiv preprint arXiv:2212.11491*, 2022.
- [161] J. Ma, T. Hu, and W. Wang, “Deciphering the projection head: Representation evaluation self-supervised learning,” *arXiv preprint arXiv:2301.12189*, 2023.
- [162] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.
- [163] A. Borji and L. Itti, “State-of-the-art in visual attention modeling,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 185–207, 2012.
- [164] X. Tan, S. Lai, Y. Liu, M. Zhang *et al.*, “Green channel guiding denoising on bayer image,” *The Scientific World Journal*, vol. 2014, 2014.
- [165] E. Abdelmaksoud, S. El-Sappagh, S. Barakat, T. Abuhmed, and M. Elmogy, “Automatic diabetic retinopathy grading system based on detecting multiple retinal lesions,” *IEEE Access*, vol. 9, pp. 15 939–15 960, 2021.

- [166] F. Haghighi, M. R. H. Taher, M. B. Gotway, and J. Liang, “Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 824–20 834.
- [167] X. Chen, H. Fan, R. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” *arXiv preprint arXiv:2003.04297*, 2020.
- [168] H.-Y. Zhou, C. Lu, S. Yang, X. Han, and Y. Yu, “Preservational learning improves self-supervised medical image models by reconstructing diverse contexts,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3499–3509.
- [169] A. Newell and J. Deng, “How useful is self-supervised pretraining for visual tasks?” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7345–7354.
- [170] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, “Unsupervised learning of visual features by contrasting cluster assignments,” *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020.
- [171] Y.-H. Cao and J. Wu, “A random cnn sees objects: One inductive bias of cnn and its applications,” in *Proceedings Of The AAAI Conference On Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 194–202.

- [172] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.
- [173] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox, “Discriminative unsupervised feature learning with convolutional neural networks,” *Advances in neural information processing systems*, vol. 27, 2014.
- [174] L. Chaves, A. Bissoto, E. Valle, and S. Avila, “An evaluation of self-supervised pre-training for skin-lesion analysis,” in *European Conference on Computer Vision*. Springer, 2022, pp. 150–166.
- [175] Y. Chen, X. Zhu, Y. Li, Y. Wei, and L. Ye, “Enhanced semantic feature pyramid network for small object detection,” *Signal Processing: Image Communication*, vol. 113, p. 116919, 2023.
- [176] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.
- [177] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

- [178] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, “Panet: Few-shot image semantic segmentation with prototype alignment,” in *proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9197–9206.
- [179] N. Wang, Y. Gao, H. Chen, P. Wang, Z. Tian, C. Shen, and Y. Zhang, “Nas-fcos: Fast neural architecture search for object detection,” in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 943–11 951.
- [180] C. Zhang, Y. Chen, L. Liu, Q. Liu, and X. Zhou, “Hico: Hierarchical contrastive learning for ultrasound video model pretraining,” in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 229–246.
- [181] Q. Bi, H. Zheng, X. Sun, J. Yi, W. Zhang, Y. Huang, Y. Li, and Y. Zheng, “Self-supervised cross-level consistency learning for fundus image classification,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1781–1785.
- [182] P. C. Chhipa, M. Chopra, G. Mengi, V. Gupta, R. Upadhyay, M. S. Chippa, K. De, R. Saini, S. Uchida, and M. Liwicki, “Functional knowledge transfer with self-supervised representation learning,” in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 3339–3343.
- [183] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging properties in self-supervised vision transformers,” in *Pro-*

ceedings of the IEEE/CVF international conference on computer vision, 2021,
pp. 9650–9660.

VITAE

- **Name:** Wadha Mutaib Mohammad Almattar
- **Nationality:** Kingdom of Saudi Arabia
- **Date of Birth:** 28/05/1982
- **Email:** *almatarwadha@gmail.com*
- **Permenant Address:** Villa, Abu Baker Alsiddiq Street, Ash Shulah Neighborhood, Dammam city, Eastern Province, Kingdom of Saudi Arabia, Zip code .
- **Academic background:**
 - **2019-2024:** PhD (Computer Science), King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia.
 - **2010-2011:** MS (Software System and Internet Technology), University of Sheffield, Sheffield, United Kingdom.
 - **2001-2004:** B.Sc.(Computer Science), Dammam University, Dammam, Saudi Arabia.

- **Teaching experience:**

- **2011-2020:** Lecturer, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia.
- **2005-2010:** Teaching Assistant, Dammam Univesity, Dammam, Saudi Arabia.

- **Journal papers:**

- Almattar, Wadha, Hamzah Luqman, and Fakhri Alam Khan. Diabetic retinopathy grading review: Current techniques and future directions. *Image and Vision Computing* (2023): 104821.
- Alhaidari, Fahd A., Saleh A. Dossary, Wadha M. Al-Mattar, Mahi M. Kamaleldin, Areej H. Abed, Atheer A. Alshaikhhussain, Heba K. Motawei, and Nadeen S. Alamoudi. Comparison of different edge detectors on general public images and two-dimensional seismic data. *Journal of Computational and Theoretical Nanoscience* 16, no. 5-6 (2019): 1794-1805.

- **Conferences papers:**

- Almattar, Wadha, and Atheer Algherairy. Investigating Image Augmentation for Classification of Chest X-Ray Images. 2022 14th International Conference on Computational Intelligence and Communication Networks (CICN). IEEE, 2022.
- Almubairik, Norah, Wadha Almattar, and Amani Alqarni. WormHex: A Volatile Memory Analysis Tool for Retrieval of Social Media Evidence.

International Journal of Information and Communication Engineering 16.6
(2022): 233-238.

- Algherairy, Atheer, et al. The impact of feature selection on different machine learning models for breast cancer classification. 2022 7th International Conference on Data Science and Machine Learning Applications (CDMA). IEEE, 2022.

Wadha Almattar received her B.S. degree in Computer Science from Imam Abdulrahman Bin Faisal University, Saudi Arabia, in 2005, and her MSc in Software System and Internet Technology from the University of Sheffield, UK, in 2011. She submitted her dissertation to fulfill the requirements for a Ph.D. in Computer Science at KFUPM. Since 2005, her academic journey began as a teaching assistant, and she advanced to the role of lecturer in 2011 at Imam Abdulrahman Bin Faisal University (IAU). Over these years, she has taught various undergraduate courses, co-supervised senior projects, and led numerous course-related initiatives, including supervising students and taking on diverse committee responsibilities.

Her research interests focus on Computer Vision (CV), particularly machine learning and deep learning, including supervised and self-supervised learning. Wadha has published her work in top-tier journals. She holds two TOT diplomas, certified by the Global Academy for Training and Consulting and the Oxford Leading Training Organization. Since 2016, she has been actively involved with King Abdulaziz and His Companions Foundation for Giftedness and Creativity (Mawhiba), where she has taken on multiple roles. In addition to her research and teaching, Wadha is deeply en-

gaged in academic outreach, contributing to various initiatives such as Mawhiba-IAU, Mawhiba-KFUPM, Misbar-KFUPM, and Women in Data Science-KFUPM (Stanford), and Black-hat, where she serves in different roles including: delivering talks, trainer, evaluator, conference organizer, and initiative supporter.