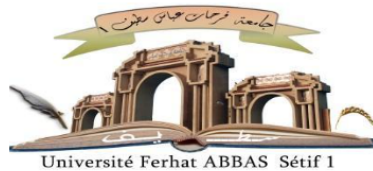


People's Democratic Republic of Algeria  
Ministry of Higher Education and Scientific Research  
Ferhat Abbas University Setif-1  
Faculty of Sciences  
Department of Computer Science



Ph.D Dissertation  
Presented by:  
**Ayoub LAOUAREM**

As a requirement for obtaining the Doctoral degree in Computer Science  
Option: Smart Systems and Machine Learning

---

# **Toward Advanced Deep Learning Techniques for Medical Image Analysis**

---

Ph.D defense board:

|                         |                                   |               |
|-------------------------|-----------------------------------|---------------|
| Pr. Salim BOUAMAMA      | Univ. Ferhat Abbes Setif 1        | President     |
| Dr. Chafia KARA-MOHAMED | Univ. Ferhat Abbes Setif 1        | Supervisor    |
| Pr. El-Bay BOURENNANE   | Univ. Bourgogne, Dijon, France    | Co-Supervisor |
| Pr. Farid NOUIOUA       | Univ. Mohamed Elbachir Elibrahimi | Examiner      |
| Dr. Moussa SEMCHEDINE   | Univ. Ferhat Abbes Setif 1        | Examiner      |
| Dr. Yacine SLIMANI      | Univ. Ferhat Abbes Setif 1        | Examiner      |

Publicly defended on: December 08, 2024

# Abstract

Deep learning (DL) has revolutionized medical diagnosis by enabling the analysis of diverse data modalities, particularly medical images, from a wide range of sources. This transformative capability has made DL approaches an increasingly vital tool in modern healthcare. Indeed, the potential of DL in medical image analysis is undeniably promising, with significant advancements already being made. However, only a few DL-based approaches have successfully transitioned into clinical practice. This may be due to various factors such as overfitted models, selection bias, and the extensive preprocessing of datasets, which fail to accurately represent clinical diversity and local variations.

This thesis aims to explore the transformative potential of DL in medical image analysis, focusing on the development of novel models that enhance diagnostic accuracy, efficiency, and personalization.

In this thesis, we address several key medical imaging modalities—X-rays, computed tomography (CT) scans, optical coherence tomography (OCT), and dermatoscopic imaging—to tackle critical diagnostic challenges in diseases such as COVID-19, retinal disorders, and skin lesions. This work offers a thorough exploration of advanced DL architectures and hybrid methodologies tailored for medical diagnostics. The first contribution underscores the effectiveness of attention mechanisms in achieving high-accuracy diagnoses for COVID-19 and other pulmonary diseases. Specifically, the integrated attention mechanisms resulted in notable improvements, with quantified performance metrics indicating high sensitivity and precision as high as 99% in COVID-19 diagnosis. The second introduces HTC-Retina, a cutting-edge hybrid approach that combines vision transformers (ViTs) with convolutional neural networks (CNNs). This approach, applied to OCT image classification, showed enhanced performance in identifying retinal disorders, with significant improvements in accuracy over traditional methods, achieving classification rates ranging from 97% to 99%. The third presents JILDYA-Net, a novel lightweight model optimized for skin lesion classification. This model employs an efficient architecture designed for resource-constrained clinical settings. It was evaluated on dermatoscopic images and demonstrated superior classification accuracy. The final contribution is the development of DA-UNet-Plus, a novel hybrid segmentation model. Initially applied to detect intraretinal fluid in OCT images with notable performance, DA-UNet-Plus was later adapted for skin lesion localization in dermatoscopic images, achieving improved lesion boundary delineation.

Overall, these contributions demonstrate the powerful potential of DL in automating complex diagnostic tasks and fostering personalized healthcare while paving the way for future innovations in medical image analysis.

## ملخص

لقد أحدث التعلم العميق ثورة في تشخيص الأمراض من خلال تمكين تحليل أنواع متعددة من البيانات، وخاصة الصور الطبية، من مجموعة واسعة من المصادر. وقد جعلت هذه القدرة التحويلية أساليب التعلم العميق أداة حيوية بشكل متزايد في الرعاية الصحية الحديثة. إن الإمكانيات التي يقدمها التعلم العميق في تحليل الصور الطبية واعدة بلا شك، مع تحقيق تقدم كبير بالفعل. ومع ذلك، فإن القليل من الأساليب المعتمدة على التعلم العميق قد انتقلت بنجاح إلى الممارسة السريرية. قد يرجع ذلك إلى عوامل متعددة مثل النماذج المفرطة التكيف، والانحياز في الاختيار، والإعداد المفرط لمجموعات البيانات، والتي لا تعكس التنوع السريري والمحلي بدقة.

تسعى هذه الأطروحة إلى استكشاف الإمكانيات التحويلية للتعلم العميق في تحليل الصور الطبية، مع التركيز على تطوير نماذج جديدة تعزز دقة التشخيص وكفاءته وتخصيصه. في هذه الأطروحة، نتناول العديد من طرق التصوير الطبي الرئيسية: الأشعة السينية، التصوير المقطعي المحوسب، التصوير المقطعي البصري، والتصوير الجليدي، لمعالجة التحديات التشخيصية الحاسمة في الأمراض مثل، فيروس كورونا ٢٠١٩، واضطرابات الشبكية، والآفات الجلدية. يقدم هذا العمل استكشافاً شاملاً لهماكل التعلم العميق والمنهجيات الهجينة المصممة خصيصاً للتشخيص الطبي. تسلط المساهمة الأولى الضوء على فعالية آليات الانتباه في تحقيق تشخيصات عالية الدقة لكوفيد ١٩ وأمراض الرئة الأخرى. على وجه التحديد، أدت آليات الانتباه المدججة إلى تحسينات ملحوظة، مع مؤشرات أداء تشير إلى حساسية ودقة تصل إلى ٩٩ % في تشخيص كوفيد ١٩. المساهمة الثانية تقدم HTC-Retina ، وهو نهج هجين متقدم يجمع بين المحولات البصرية والشبكات العصبية التلافيفية. أظهر هذا النهج، عند تطبيقه على تصنيف صور OCT ، أداءً معززاً في تحديد اضطرابات الشبكية، مع تحسينات كبيرة في الدقة مقارنة بالطرق التقليدية، محققاً معدلات تصنيف تتراوح بين ٩٧ % و ٩٩ %. المساهمة الثالثة تقدم JILDYA-Net ، وهو نموذج خفيف الوزن جديد مُحسن لتصنيف الآفات الجلدية. يستخدم هذا النموذج بنية فعالة مصممة للأوضاع السريرية المحدودة الموارد. تم تقييمه على الصور الجلدية وأظهر دقة تصنيف متفوقة. المساهمة الأخيرة هي تطوير DA-UNet-Plus ، وهو نموذج هجين جديد للتقطيع. تم تطبيق DA-UNet-Plus في البداية لاكتشاف السوائل داخل الشبكية في صور OCT مع أداء ملحوظ، ثم تم تعديله لاحقاً لتحديد موقع الآفات في الصور الجلدية، محققاً تحسينات في تحديد حدود الآفات.

بوجه عام، تُظهر هذه المساهمات الإمكانيات القوية للتعلم العميق في أتمتة المهام التشخيصية المعقدة وتعزيز الرعاية الصحية الشخصية، مما يمهّد الطريق للابتكارات المستقبلية في تحليل الصور الطبية.

# Résumé

L'apprentissage profond a révolutionné le diagnostic médical en permettant l'analyse de diverses modalités de données, en particulier les images médicales, provenant d'un large éventail de sources. Cette capacité transformatrice a fait des approches basées sur l'apprentissage profond un outil de plus en plus important dans les soins de santé. En effet, le potentiel de l'apprentissage profond dans l'analyse d'images médicales est indéniablement prometteur, avec des avancées significatives déjà réalisées. Cependant, seules quelques approches basées sur l'apprentissage profond ont réussi à être intégrées dans la pratique clinique. Cela peut être dû à divers facteurs tels que les modèles sur-ajustés, les biais de sélection et le prétraitement extensif des ensembles de données, qui ne représentent pas fidèlement la diversité clinique et les variations locales.

Cette thèse vise à explorer le potentiel transformateur de l'apprentissage profond dans l'analyse d'images médicales, en se concentrant sur le développement de nouveaux modèles améliorant la précision diagnostique, l'efficacité et la personnalisation. Nous abordons ici plusieurs modalités d'imagerie médicale clés—radiographies, tomodensitométrie (CT), tomographie par cohérence optique (OCT) et imagerie dermatoscopique—afin de relever des défis diagnostiques critiques dans des maladies telles que le COVID-19, les troubles rétinien et les lésions cutanées. Ce travail offre une exploration des architectures avancées basées sur l'apprentissage profond et des méthodologies hybrides adaptées au diagnostic médical. La première contribution souligne l'efficacité des mécanismes d'attention dans l'obtention de diagnostics très précis pour le COVID-19 et d'autres maladies pulmonaires. En particulier, les mécanismes d'attention intégrés ont conduit à des améliorations notables, avec des performances quantifiées indiquant une sensibilité et une précision élevées atteignant jusqu'à 99% dans le diagnostic du COVID-19. La deuxième contribution introduit HTC-Retina, une approche hybride qui combine les transformateurs de vision (ViTs) avec les réseaux de neurones convolutifs (CNNs). Cette approche, appliquée à la classification d'images OCT, a montré des performances accrues dans l'identification des troubles rétinien, avec des améliorations significatives en termes de précision par rapport aux méthodes traditionnelles, atteignant des taux de classification allant de 97% à 99%. La troisième contribution présente JILDYA-Net, un nouveau modèle léger optimisé pour la classification des lésions cutanées. Ce modèle utilise une architecture efficace conçue pour des environnements cliniques à ressources limitées. Il a été évalué sur des images dermatoscopiques et a démontré une précision de classification supérieure. La dernière contribution est le développement de



DA-UNet-Plus, un nouveau modèle de segmentation hybride. Initialement appliqué à la détection de liquide intra-rétinien dans les images OCT avec des performances remarquables, DA-UNet-Plus a ensuite été adapté pour la localisation des lésions cutanées dans les images dermatoscopiques, atteignant une meilleure délimitation des contours des lésions.

Dans l'ensemble, ces contributions démontrent le puissant potentiel de l'apprentissage profond dans l'automatisation des tâches diagnostiques complexes et la promotion de soins de santé personnalisés, tout en ouvrant la voie à de futures innovations dans l'analyse d'images médicales.

I dedicate this thesis to my late father, who was the person who most wanted me to get my PhD, and his memory will remain dear to me forever. I also dedicate this work to my beloved mother, for whom I would do anything . . .

## **Declaration**

I hereby declare that except where specific reference is made to the work of others, the contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others.

Ayoub Laouarem  
December 2024

## Acknowledgements

الحمد لله الذي تتم بتوفيقه الأعمال الصالحة

First of all, I thank **Allah** for the patience, strength, will, and health he has given me to complete this journey.

I also want to thank myself for persevering through years of struggle and overcoming obstacles, responsibilities, and challenges. It was never easy.

A heartfelt thanks to my family, including my mother, my brothers Souhaib and Ahmed, my sister Israa, my grandfather, and my grandmother, with special gratitude to my mother for her sacrifices, unwavering support, and constant encouragement.

I would like to express my heartfelt gratitude to my supervisor for her unwavering moral and material support throughout my PhD journey.

Finally, I extend my gratitude to the president of the committee and the examiners for graciously accepting the invitation to serve as jury members for this thesis defense and for their valuable evaluation and review of my work.

# Table of contents

|                                                                                       |            |
|---------------------------------------------------------------------------------------|------------|
| <b>List of figures</b>                                                                | <b>xii</b> |
| <b>List of tables</b>                                                                 | <b>xiv</b> |
| <b>List of abbreviations</b>                                                          | <b>xvi</b> |
| <b>1 Introduction</b>                                                                 | <b>1</b>   |
| 1.1 Research context and motivation . . . . .                                         | 1          |
| 1.2 Research contributions . . . . .                                                  | 2          |
| 1.2.1 COVID-19 diagnosis based on X-RAY and CT scans . . . . .                        | 2          |
| 1.2.2 Retinal disease diagnosis based on OCT images . . . . .                         | 3          |
| 1.2.3 Skin lesions diagnosis based on dermatoscopic images . . . . .                  | 3          |
| 1.2.4 Development of a novel hybrid approach for medical image segmentation . . . . . | 3          |
| 1.3 Structure of the Thesis . . . . .                                                 | 4          |
| 1.4 Publications . . . . .                                                            | 5          |
| <b>2 Background and literature review</b>                                             | <b>7</b>   |
| 2.1 Introduction . . . . .                                                            | 7          |
| 2.2 Introduction to medical image analysis . . . . .                                  | 7          |
| 2.2.1 Definition and significance . . . . .                                           | 7          |
| 2.2.2 Historical perspective and evolution . . . . .                                  | 8          |
| 2.3 CAD and detection systems . . . . .                                               | 9          |
| 2.3.1 Overview of CAD . . . . .                                                       | 9          |
| 2.3.2 CAD applications . . . . .                                                      | 10         |
| 2.3.3 Key challenges in CAD systems . . . . .                                         | 15         |
| 2.4 Theoretical background . . . . .                                                  | 16         |
| 2.4.1 Data acquisition and quality assurance . . . . .                                | 16         |
| 2.4.2 Machine learning . . . . .                                                      | 19         |

|          |                                                                                                   |           |
|----------|---------------------------------------------------------------------------------------------------|-----------|
| 2.4.3    | Deep learning . . . . .                                                                           | 21        |
| 2.5      | Overall literature review . . . . .                                                               | 32        |
| 2.5.1    | COVID-19 diagnosis . . . . .                                                                      | 32        |
| 2.5.2    | Retinal disease diagnosis and detection . . . . .                                                 | 37        |
| 2.5.3    | Skin lesion diagnosis and detection . . . . .                                                     | 45        |
| 2.6      | Conclusion . . . . .                                                                              | 54        |
| <b>3</b> | <b>COVID-19 diagnosis from chest radiograph and computed tomography scans</b>                     | <b>56</b> |
| 3.1      | Introduction . . . . .                                                                            | 56        |
| 3.2      | Methodology . . . . .                                                                             | 57        |
| 3.2.1    | Baseline CNN . . . . .                                                                            | 57        |
| 3.2.2    | SE-Net . . . . .                                                                                  | 57        |
| 3.2.3    | ResNet model . . . . .                                                                            | 60        |
| 3.2.4    | InceptionNet model . . . . .                                                                      | 60        |
| 3.2.5    | Loss functions . . . . .                                                                          | 60        |
| 3.3      | Experiments and results . . . . .                                                                 | 61        |
| 3.3.1    | Datasets . . . . .                                                                                | 61        |
| 3.3.2    | Preprocessing . . . . .                                                                           | 62        |
| 3.3.3    | Experimental settings . . . . .                                                                   | 62        |
| 3.3.4    | Performance evaluation . . . . .                                                                  | 63        |
| 3.3.5    | Experimental results . . . . .                                                                    | 64        |
| 3.4      | Discussion . . . . .                                                                              | 71        |
| 3.5      | Conclusion and future perspectives . . . . .                                                      | 71        |
| 3.5.1    | Research contributions and implications . . . . .                                                 | 72        |
| 3.5.2    | Possible future improvements . . . . .                                                            | 72        |
| <b>4</b> | <b>Enhanced retinal disease diagnosis and detection using optical coherence tomography images</b> | <b>74</b> |
| 4.1      | Introduction . . . . .                                                                            | 74        |
| 4.2      | Part I: Retinal diseases classification and identification using HTC-retina . .                   | 75        |
| 4.2.1    | Motivation for HTC-Retina development . . . . .                                                   | 75        |
| 4.2.2    | HTC-Retina model contribution . . . . .                                                           | 76        |
| 4.2.3    | Methodology . . . . .                                                                             | 76        |
| 4.2.4    | Experiments and results . . . . .                                                                 | 82        |
| 4.2.5    | Discussion . . . . .                                                                              | 91        |
| 4.3      | Part II: Intraretinal cystoid fluid detection and segmentation using DA-UNet-Plus . . . . .       | 93        |

|          |                                                                                   |            |
|----------|-----------------------------------------------------------------------------------|------------|
| 4.3.1    | Motivation for DA-Unet-Plus development . . . . .                                 | 93         |
| 4.3.2    | Contribution of DA-UNet-Plus . . . . .                                            | 93         |
| 4.3.3    | Methodology . . . . .                                                             | 94         |
| 4.3.4    | Experiments and results . . . . .                                                 | 99         |
| 4.3.5    | Discussion of DA-Unet-Plus results . . . . .                                      | 102        |
| 4.4      | Conclusion and future work perspectives . . . . .                                 | 104        |
| 4.4.1    | HTC-Retina model . . . . .                                                        | 104        |
| 4.4.2    | DA-UNet-Plus model . . . . .                                                      | 106        |
| <b>5</b> | <b>Enhanced skin lesion diagnosis and localization using dermatoscopic images</b> | <b>108</b> |
| 5.1      | Introduction . . . . .                                                            | 108        |
| 5.2      | Part I: Skin lesions classification using JILDYA-Net . . . . .                    | 109        |
| 5.2.1    | Motivation for JILDYA-Net development . . . . .                                   | 109        |
| 5.2.2    | The main contribution of JILDYA-Net . . . . .                                     | 110        |
| 5.2.3    | Methodology . . . . .                                                             | 111        |
| 5.2.4    | Experimental setup . . . . .                                                      | 114        |
| 5.2.5    | Results and discussion of JILDYA-Net . . . . .                                    | 117        |
| 5.2.6    | Summary of JILDYA-Net . . . . .                                                   | 123        |
| 5.3      | Part II: Skin lesion segmentation and localization using DA-UNet-Plus . . .       | 124        |
| 5.3.1    | Motivation for DA-UNet-Plus usage . . . . .                                       | 124        |
| 5.3.2    | DA-UNet-Plus contribution . . . . .                                               | 125        |
| 5.3.3    | Methodology . . . . .                                                             | 125        |
| 5.3.4    | Experimental results of DA-UNet-Plus . . . . .                                    | 126        |
| 5.3.5    | Summary of DA-UNet-Plus usage in skin-related conditions . . . .                  | 130        |
| 5.4      | Conclusion and future work . . . . .                                              | 131        |
| <b>6</b> | <b>General conclusion</b>                                                         | <b>133</b> |
| 6.1      | Learned lessons . . . . .                                                         | 133        |
| 6.2      | Some challenges of CAD systems in medical image analysis . . . . .                | 135        |
|          | <b>References</b>                                                                 | <b>137</b> |
|          | <b>Appendix A Glossary</b>                                                        | <b>152</b> |

# List of figures

|      |                                                                                                                     |     |
|------|---------------------------------------------------------------------------------------------------------------------|-----|
| 2.1  | The intersection of medicine and computer science in CAD . . . . .                                                  | 10  |
| 2.2  | General flowchart of a CAD system based on DL and medical images . . .                                              | 11  |
| 2.3  | A diagram of MLP architecture . . . . .                                                                             | 22  |
| 2.4  | A diagram of CNN architecture . . . . .                                                                             | 25  |
| 2.5  | Vision transformer architecture . . . . .                                                                           | 29  |
| 2.6  | The FNet model architecture . . . . .                                                                               | 32  |
| 3.1  | Baseline CNN architecture . . . . .                                                                                 | 58  |
| 3.2  | Proposed SE-Net architecture . . . . .                                                                              | 59  |
| 3.3  | Confusion matrices of different models for COVID-19 vs. healthy classifica-<br>tion based on CXR scans . . . . .    | 65  |
| 3.4  | Confusion matrices of different models for COVID-19 vs. non-COVID-19<br>classification based on CXR scans . . . . . | 66  |
| 3.5  | Confusion matrices of different models for multi-class classification task<br>based on CXR scans . . . . .          | 67  |
| 3.6  | Confusion matrices of different models for COVID-19 vs. non-COVID-19<br>classification based on CT scans . . . . .  | 68  |
| 4.1  | The overall architecture of the HTC-Retina model . . . . .                                                          | 78  |
| 4.2  | HTC-Retina performance in the training and validation sets . . . . .                                                | 83  |
| 4.3  | Confusion matrix of each dataset generated by the HTC-Retina model . . .                                            | 85  |
| 4.4  | The grad class activation mapping visualization of the HTC-Retina model .                                           | 86  |
| 4.5  | HTC-Retina gains. . . . .                                                                                           | 88  |
| 4.6  | The overall architecture of the DA-UNet-Plus model . . . . .                                                        | 95  |
| 4.7  | The convolutional block diagram . . . . .                                                                           | 95  |
| 4.8  | The residual attention block diagram . . . . .                                                                      | 96  |
| 4.9  | The DANet head bottleneck block diagram . . . . .                                                                   | 97  |
| 4.10 | Correlation between predicted and ground truth sizes of IRC regions . . . .                                         | 101 |



|      |                                                                                  |     |
|------|----------------------------------------------------------------------------------|-----|
| 4.11 | Example cases of segmentation results of the DA-UNet-Plus model . . . . .        | 103 |
| 4.12 | Comparison of DA-UNet-Plus performance and SOTA models . . . . .                 | 104 |
| 5.1  | The layout of the JILDYA-Net model . . . . .                                     | 112 |
| 5.2  | Datasets' information and distribution . . . . .                                 | 115 |
| 5.3  | Flow diagram of the datasets pre-processing, augmentation, and preparation       | 116 |
| 5.4  | Confusion matrices generated by the proposed model . . . . .                     | 119 |
| 5.5  | JILDYA-Net <i>vs.</i> SOTA methods for the HAM10000 dataset . . . . .            | 120 |
| 5.6  | JILDYA-Net <i>vs.</i> SOTA methods for the ISCI 2019 dataset . . . . .           | 120 |
| 5.7  | JILDYA-Net improvement factor <i>w.r.t.</i> SOTA for the HAM10000 dataset . .    | 122 |
| 5.8  | JILDYA-Net improvement factor <i>w.r.t.</i> SOTA for the ISIC 2019 dataset . . . | 122 |
| 5.9  | Correlation between predicted and ground truth sizes of skin lesions . . . .     | 128 |
| 5.10 | Visual segmentation results of the DA-UNet-Plus model . . . . .                  | 129 |
| 5.11 | Comparison of DSC of DA-UNet-Plus with SOTA methods . . . . .                    | 131 |

# List of tables

|      |                                                                                                                             |    |
|------|-----------------------------------------------------------------------------------------------------------------------------|----|
| 2.1  | Summary of related works on COVID-19 diagnosis . . . . .                                                                    | 35 |
| 2.1  | Summary of related works on COVID-19 diagnosis . . . . .                                                                    | 36 |
| 2.1  | Summary of related works on COVID-19 diagnosis . . . . .                                                                    | 37 |
| 2.2  | Summary of related works on retinal disease classification using traditional CNN and transfer learning approaches . . . . . | 39 |
| 2.3  | Summary of related works on retinal disease classification using advanced CNN approaches . . . . .                          | 40 |
| 2.4  | Summary of related works on retinal disease classification using emerging transformer-based approaches . . . . .            | 42 |
| 2.5  | Summary of related works on retinal fluid segmentation . . . . .                                                            | 45 |
| 2.6  | Summary of related works on skin lesion classification using traditional CNN. . . . .                                       | 47 |
| 2.7  | Summary of related works on skin lesion classification using lightweight architectures. . . . .                             | 49 |
| 2.8  | Summary of related works on skin lesion classification using advanced approaches. . . . .                                   | 51 |
| 2.9  | Summary of related works on skin lesion segmentation using traditional CNN and SOTA segmentation architectures. . . . .     | 53 |
| 2.10 | Summary of related works on skin lesion segmentation using advanced architectures and hybrid models. . . . .                | 55 |
| 3.1  | Detailed configuration of the datasets based on the classification task and image modality. . . . .                         | 63 |
| 3.2  | Results of different models for COVID-19 vs. healthy classification based on CXR scans. . . . .                             | 64 |
| 3.3  | Results of different models for COVID-19 vs. non-COVID-19 classification based on CXR scans. . . . .                        | 65 |
| 3.4  | Results of different models for multi-class classification task based on CXR scans. . . . .                                 | 66 |

|     |                                                                                                                        |     |
|-----|------------------------------------------------------------------------------------------------------------------------|-----|
| 3.5 | Results of different models for COVID-19 <i>vs.</i> non-COVID-19 classification based on CT scans. . . . .             | 68  |
| 3.6 | Models' computational complexity and performance comparison . . . . .                                                  | 69  |
| 3.7 | Comparison of COVID-19 diagnosis results based on CXR scans . . . . .                                                  | 70  |
| 3.8 | Comparison of COVID-19 diagnosis results based on CT scans . . . . .                                                   | 70  |
| 4.1 | Best validation results . . . . .                                                                                      | 84  |
| 4.2 | Detailed evaluation results of the HTC-Retina model . . . . .                                                          | 84  |
| 4.3 | Comparison of classification results of different models on the test set. . . .                                        | 87  |
| 4.4 | Ablative studies on CPTE, RDP-ConvNet, and positional embeddings . . . .                                               | 89  |
| 4.5 | Comparison results on the noisy and original test set of each dataset. . . . .                                         | 90  |
| 4.6 | External validation results of the HTC-Retina model. . . . .                                                           | 91  |
| 4.7 | DA-UNet-Plus performance in IRC detection . . . . .                                                                    | 99  |
| 5.1 | Evaluation results of JILDYA-Net on the original datasets. . . . .                                                     | 117 |
| 5.2 | Evaluation results of JILDYA-Net on different classification tasks and datasets.                                       | 117 |
| 5.3 | JILDYA-Net <i>vs.</i> SOTA methods for the HAM10000 dataset . . . . .                                                  | 118 |
| 5.4 | JILDYA-Net <i>vs.</i> SOTA methods for the ISCI 2019 dataset . . . . .                                                 | 119 |
| 5.5 | Computational statistics of JILDYA-Net on HAM10000 and ISIC 2019 datasets.                                             | 121 |
| 5.6 | Maximum and minimum JILDYA-Net improvement factor <i>w.r.t.</i> SOTA for both HAM10000 and ISIC 2019 datasets. . . . . | 123 |
| 5.7 | DA-UNet-Plus performance in skin lesion segmentation and localization. . .                                             | 127 |
| 5.8 | Performance comparison of DA-UNet-Plus with SOTA methods . . . . .                                                     | 130 |

# List of abbreviations

|             |                                         |                 |                                            |
|-------------|-----------------------------------------|-----------------|--------------------------------------------|
| <b>AI</b>   | Artificial Intelligence                 | <b>CSR</b>      | Central Serous Retinopathy                 |
| <b>AM</b>   | Attention Mechanism                     | <b>DL</b>       | Deep Learning                              |
| <b>ANN</b>  | Artificial Neural Network               | <b>DME</b>      | Diabetic Macular Edema                     |
| <b>AK</b>   | Actinic Keratosis                       | <b>DR</b>       | Diabetic Retinopathy                       |
| <b>AUC</b>  | Area Under the Curve                    | <b>DF</b>       | Dermatofibroma                             |
| <b>AMD</b>  | Age-related Macular Degeneration        | <b>DSC</b>      | Dice Similarity Coefficient                |
| <b>BCC</b>  | Basal Cell Carcinoma                    | <b>DA</b>       | Dual Attention                             |
| <b>BKL</b>  | Benign Keratosis                        | <b>EAM</b>      | External Attention Module                  |
| <b>BCE</b>  | Binary Cross-Entropy                    | <b>FT</b>       | Fourier Transform                          |
| <b>CCE</b>  | Categorical Cross-Entropy               | <b>FFT</b>      | Fast Fourier Transform                     |
| <b>CNN</b>  | Convolutional Neural Network            | <b>FFN</b>      | Feed-Forward Network                       |
| <b>CV</b>   | Computer Vision                         | <b>FP</b>       | False Positive                             |
| <b>CAD</b>  | Computer-Aided Diagnosis                | <b>FN</b>       | False Negative                             |
| <b>CPTE</b> | Convolutional Patch and Token Embedding | <b>GELU</b>     | Gaussian Error Linear Unit                 |
| <b>CBAM</b> | Convolutional Block Attention Module    | <b>GPA</b>      | Global Average Pooling                     |
| <b>CAM</b>  | Channel Attention Module                | <b>Grad-CAM</b> | Gradient-weighted Class Activation Mapping |
| <b>CXR</b>  | Chest X-ray                             | <b>ICR</b>      | Intraretinal Cystoid Fluid                 |
| <b>CT</b>   | Computed Tomography                     | <b>ISIC</b>     | International Skin Imaging Collaboration   |
| <b>CNV</b>  | Choroidal NeoVascularization            | <b>IoU</b>      | Intersection over Union                    |

---

|                    |                                                    |             |                                 |
|--------------------|----------------------------------------------------|-------------|---------------------------------|
| <b>ML</b>          | Machine Learning                                   | <b>SOTA</b> | State-Of-The-Art                |
| <b>MLP</b>         | Multi Layer Perceptron                             | <b>SGD</b>  | Stochastic Gradient Descent     |
| <b>MHSA</b>        | Multi-Head Self-Attention                          | <b>SE</b>   | Squeeze-and-Excitation          |
| <b>MRI</b>         | Magnetic Resonance Imaging                         | <b>SDR</b>  | Stochastic Depth Regularization |
| <b>MH</b>          | Macular Hole                                       | <b>SCC</b>  | Squamous Cell Carcinoma         |
| <b>MEL</b>         | Melanoma                                           | <b>TP</b>   | True Positive                   |
| <b>NV</b>          | Melanocytic Nevus                                  | <b>TN</b>   | True Negative                   |
| <b>NLP</b>         | Natural Language Processing                        | <b>TPR</b>  | True Positive Rate              |
| <b>OCT</b>         | Optical Coherence Tomography                       | <b>TPR</b>  | False Positive Rate             |
| <b>PAM</b>         | Position Attention Module                          | <b>ViT</b>  | Vision Transformer              |
| <b>ReLU</b>        | Rectified Linear Unit                              | <b>ViTE</b> | Vision Transformer Encoder      |
| <b>RDP-ConvNet</b> | Residual Depthwise-Pointwise Convolutional Network | <b>VASC</b> | Vascular Lesions                |

# Chapter 1

## Introduction

### 1.1 Research context and motivation

Since the invention of the X-ray cathode tube 130 years ago or so, medical images have played a crucial and indispensable role in diagnosis, analysis, and inpatient treatment. This advancement led to the appearance of various image modalities such as computed tomography (CT), magnetic resonance imaging (MRI), optical coherence tomography (OCT), and other types according to medical sub-field. Indeed, the evolution of medical imaging techniques generates vast amounts of data that must be accurately interpreted. Considering the advancements in medical imaging, several challenges persist in analyzing the collected data within a limited timeframe, especially when relying solely on human intervention. These challenges include the time-consuming nature of interpretation, the significant costs associated with training medical experts, and the susceptibility to human error in the analytical process. Therefore, there is an urgent need for computer-aided diagnosis (CAD) to alleviate the workload of medical professionals. These systems are essential for delivering efficient, accurate, and personalized diagnoses and treatments, ultimately enhancing patient care and reducing the burden on healthcare systems.

Recently, with the rapid advancements in artificial intelligence (AI) and the emergence of machine learning (ML) as a transformative breakthrough, the development of accurate CAD systems has become a tangible reality. This progress is primarily due to ML's ability to autonomously learn and capture informative features that represent the underlying regularities or patterns in data, playing a pivotal role in various medical image analysis tasks. Among the most recent advancements in ML is deep learning (DL), which elevates feature extraction to a more advanced learning stage. Reverse to image processing techniques and handcrafted feature extraction, DL can automatically extract and discover meaningful features from raw data, leading to high-level interpretation of different medical images. The advent of

convolutional neural networks (CNNs) in the last four decades or so, however, introduced a paradigm shift. CNNs, with their ability to automatically learn and extract hierarchical features directly from raw image data, offered a more sophisticated approach to medical image analysis including classification, segmentation, and detection. By leveraging large datasets and advanced architectures, these networks can discern intricate patterns and anomalies within medical images that might be challenging for human observers to detect.

This research is driven by the need to improve diagnostic accuracy, efficiency, and personalization in medical imaging, addressing the limitations of traditional methods that struggle with the increasing volume and complexity of medical data. As diseases like cancer and COVID-19 demand early and precise diagnosis, DL methods offer a powerful solution by automating feature extraction and analysis, reducing the burden on healthcare professionals and minimizing human error. The present thesis aims to bridge the gap between research and clinical practice, developing robust, interpretable models that can be seamlessly integrated into real-world healthcare settings, thereby advancing personalized medicine and enhancing patient outcomes.

## **1.2 Research contributions**

This thesis explores the application of DL techniques to the field of medical image analysis, focusing on the development and evaluation of novel models tailored to specific imaging modalities and diagnostic tasks. The research spans several critical areas within medical imaging, each representing distinct challenges and opportunities for innovation in automated diagnosis and disease management, with a particular emphasis on three medical sub-fields: chest radiology, ophthalmology, and dermatology. Our contributions encompass a comprehensive literature analysis, the development of specific advanced DL models, including the training and pre-training parameters and hyperparameters, the usage of ad-hoc datasets, culminating in the submissions of notable publications of research in peer-reviewed venues.

### **1.2.1 COVID-19 diagnosis based on X-RAY and CT scans**

The first area of focus is the application of advanced DL models for the diagnosis of COVID-19, a global health crisis that has underscored the need for rapid and accurate diagnostic tools. This research section examines how DL models, particularly CNNs and advanced techniques like attention mechanisms, can be applied to X-ray and CT scan data to detect COVID-19 with high accuracy. This part of the thesis explores the development of a CNN model that incorporates attention layers, and evaluates various transfer learning approaches

including ResNet, and InceptionNet, for both binary and multi-class classification tasks of COVID-19 cases and other lung diseases. The findings demonstrate the high performance of the developed model in the COVID-19 diagnosis for both classification tasks.

### **1.2.2 Retinal disease diagnosis based on OCT images**

The second area of focus addresses the diagnosis of seven distinct retinal diseases based on OCT imaging. The OCT scan provides detailed cross-sectional images of the retina, which are critical for diagnosing conditions such as age-related macular degeneration, diabetic retinopathy, and other conditions. This research section explores the application of hybrid DL approach: HTC-Retina as an advanced Hybrid model that integrates the capabilities of Vision Transformers (ViTs) and CNNs to enhance the classification accuracy of seven distinct retinal diseases. By performing comprehensive analysis of OCT images, the model extracts the most salient and disease-specific features, thereby improving diagnostic precision for each condition.

### **1.2.3 Skin lesions diagnosis based on dermatoscopic images**

The third area of focus delves into the diagnosis of skin lesions, including melanoma and other skin cancers and lesions, using dermatoscopic images. This research section introduces the development of a novel lightweight model, JILDYA-Net, specifically designed for the classification of skin lesions. The research examines the inclusion of lightweight advanced architecture to build a robust model that can effectively diagnose different skin lesions and be easily integrated into low-resource devices, which leads to support clinicians in making timely and informed decisions. The findings of the JILDYA-Net model have shown promising results in terms of efficiency and computational complexity.

### **1.2.4 Development of a novel hybrid approach for medical image segmentation**

Additionally, the thesis presents a novel hybrid model named DA-UNet-Plus, which incorporates advanced DL techniques by combining the U-Net structure with enhancements such as residual attention blocks, and dual attention bottleneck. This model is specifically designed for the segmentation of medical images. In the first stage, the model is used to segment, detect, and quantify the intraretinal cystoid fluid from the OCT image; a fluid appearing in some retinal diseases such as age-related macular degeneration and diabetic macular edema. In the second stage, the model is adopted to segment and localize the skin



lesion from dermoscopic images. The aim of developing this hybrid approach is to address the challenges of accurately segmenting medical images within complex patterns. These experiments' findings highlight the notable performance of the DA-UNet-Plus model in the segmentation task on both intraretinal cystoid fluid detection and skin lesion localization.

## 1.3 Structure of the Thesis

The thesis is organized into four chapters, each focusing on a specific aspect of DL for medical image analysis. This thesis organization is also based on our research contributions and scientific publications. The structure is designed to guide the reader through the research's theoretical foundations, methodological innovations, and practical applications. This leads to a comprehensive exploration of how DL can enhance medical diagnostics. The thesis begins with a concise overview of medical image analysis and CAD systems, highlighting the associated challenges. It then introduces several state-of-the-art (SOTA) DL algorithms and reviews recent advancements in medical image analysis using these methods, grounded in a thorough literature review. Furthermore, the thesis details our contributions across three critical medical sub-fields: chest radiology, ophthalmology, and dermatology. Each of these contributions is presented in a dedicated chapter, ensuring that the thesis is structured so that each chapter is self-contained and can be read independently.

### **Chapter 2: Background and literature review**

This chapter provides a foundational overview of medical imaging and DL. It introduces the different medical imaging modalities and highlights the relevant ones. Also, it delves into CAD systems and highlights the imposed challenges. The chapter also presents a global overview of the evolution of ML and DL models. Subsequently, the chapter provides a literature review of the SOTA deep models developed for medical image classification and segmentation. This background sets the stage for the subsequent chapters by contextualizing the challenges and opportunities that DL presents in the medical imaging domain.

### **Chapter 3: COVID-19 diagnosis from chest radiograph and computed tomography scans**

Chapter 3 delves into the application of DL for diagnosing COVID-19 using X-ray and CT scan images. It details the development of a CNN model with the use of advanced techniques such as the attention mechanism tailored for this task and evaluates the performance of various transfer learning models for binary and multi-class classification. The chapter discusses

the methodology, experimental setup, and results, highlighting the effectiveness of DL in detecting COVID-19 from radiographic data. The material in this chapter builds upon two published conference papers “A deep learning model for CXR-based COVID-19 detection” [1], and “COVID-19 detection using CXR images: simple CNN *vs* transfer learning” [2].

#### **Chapter 4: Enhanced retinal disease diagnosis and detection using optical coherence tomography images**

In Chapter 4, the focus shifts to the diagnosis of retinal diseases using OCT imaging. This chapter explores the application of advanced hybrid DL techniques for the segmentation and classification of retinal images. The first part of this chapter presents HTC-Retina, along with a thorough analysis of its performance in identifying these retinal conditions. Part one of this chapter has been published in [3]. Additionally, the second chapter’s part introduces a novel hybrid model named DA-UNet-Plus for intraretinal fluid detection and segmentation. The chapter underscores the potential of DL and hybrid models to improve the early detection and management of retinal diseases.

#### **Chapter 5: Enhanced skin lesion diagnosis and localization using dermatoscopic images**

Chapter 5 addresses the diagnosis of skin lesions, including melanoma and other lesion types, using dermatoscopic images. It introduces the JILDYA-Net model. Additionally, this chapter presents the adaptation of the DA-UNet-Plus model for skin lesion segmentation and localization. The chapter provides a detailed account of the models’ development, training, and evaluation, emphasizing their contributions to improving diagnostic accuracy and efficiency in dermatology.

#### **Conclusion**

The thesis concludes with a summary of the key findings, highlighting the contributions made to the field of medical image analysis through the application of DL. The conclusion also reflects on the implications of the research for future studies and directions.

## **1.4 Publications**

The following papers are integral to the research conducted during the doctoral program. While additional papers were published, they have been excluded from this thesis to maintain coherence, as they are not directly relevant to the content presented here.

**a. International conferences**

- **Ayoub Laouarem**, Chafia Kara-Mohamed, El-Bay Bourenane, and Aboubekour Hamdi-Cherif. A deep learning model for CXR-based COVID-19 detection. In 2021 International Conference on Engineering and Emerging Technologies (ICEET), pages 1–5. IEEE, 2021.
- **Ayoub Laouarem**, Chafia Kara-Mohamed, El-Bay Bourenane, and Aboubekour Hamdi-Cherif. Covid-19 detection using cxr images: simple cnn vs. transfer learning. In 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET), pages 1–6. IEEE, 2021.
- **Ayoub Laouarem**, Chafia Kara-Mohamed, El-Bay Bourenane, A deep convolutional neural network architecture for skin cancer detection. In 2021 International Conference on Computer and Information Sciences (ICCIS), pages 1–6. IEEE, 2021.

**b. International journals**

- **Ayoub Laouarem**, Chafia Kara-Mohamed, El-Bay Bourenane, and Aboubekour Hamdi-Cherif. Htc-retina: A hybrid retinal diseases classification model using transformer-convolutional neural network from optical coherence tomography images. *Computers in Biology and Medicine*, 178:108726, 2024.
- **Ayoub Laouarem**, Chafia Kara-Mohamed, El-Bay Bourenane, and Aboubekour Hamdi-Cherif. JILDYA-Net: An efficient lightweight multi-class classification architecture for skin lesions. Under review, 2024.

# **Chapter 2**

## **Background and literature review**

### **2.1 Introduction**

This chapter provides a concise overview of medical image analysis, encompassing its definition, significance, and historical evolution. It further discusses CAD systems, elucidating their applications and inherent challenges. The objective is to establish a foundational understanding of medical image analysis and CAD systems within a broader context. Subsequently, the chapter presents a brief background on ML and SOTA DL models. A brief exposition of these models' architectures and functionalities is provided. Finally, the chapter reviews related research works that have inspired our research and significantly contributed to the advancement of both medical image analysis and the development of sophisticated CAD systems. The techniques discussed represent a selection of those utilized in our research, acknowledging that numerous other methods have been developed across various medical fields.

### **2.2 Introduction to medical image analysis**

#### **2.2.1 Definition and significance**

The medical field has witnessed significant advances in computerized medical imaging in the last two decades. These advances have led to new multidimensional imaging modalities [4]. Therefore, the processing and analyzing medical images have become essential clinical tools in diagnostic and decision-making processes in radiology, pathology, ophthalmology, dermatology, cardiology, neurology, and other medical sub fields. Medical image analysis is the process of employing advanced computational techniques to examine and interpret various

medical images, such as X-rays, computed tomography (CT) scans, magnetic resonance imaging (MRI) scans, ultrasound images, fundus images, optical coherence tomography (OCT) images, dermoscopic images, mammograms, etc. It combines computer science, mathematics, and medicine to assist healthcare professionals in diagnosing diseases, planning treatments, and monitoring patient progress.

The importance of medical image analysis is self-evident: improved accuracy, and faster diagnoses in patients. Existing ways to interpret medical images typically require a lot of time and are susceptible to human error because they depend on expert analyses from radiologists and other health professionals. While time-consuming, specialist assessment is being largely replaced by efficient automated and semi-automated image analysis platforms capable of swiftly analyzing high numbers of datasets with considerably less variability to provide critical bedside clinician support that may influence patient outcomes [5]. Medical image analysis has numerous applications, including detecting tumors, monitoring disease progression, guiding surgical procedures, and evaluating treatment effectiveness. For example, in X-rays images-based pneumology, image analysis can help identify the size, shape, and location of lesions, assisting in both diagnosis and treatment planning. In ophthalmology, it helps diagnose retinal diseases, while in dermatology, it aids in the early detection of skin cancers.

### 2.2.2 Historical perspective and evolution

The field of medical imaging and image analysis has evolved due to the collective contributions from many areas of medicine, engineering, and basic sciences. It has a rich history that dates back to the discovery of X-rays by Wilhelm Conrad Roentgen in 1895 [6]. This groundbreaking revelation changed the field of medicine by permitting non-invasive visualization of the body's internal structures. Throughout the years, many imaging modalities have been developed, which led to improved medical diagnostics [4]:

- **X-rays:** The first and most widely used imaging technique, primarily for detecting bone fractures, infections, and tumors.
- **Computed Tomography:** Introduced in the 1970s, CT scanning provides detailed cross-sectional images of the body, useful for diagnosing conditions such as cancers, cardiovascular diseases, and infections.
- **Magnetic Resonance Imaging:** Developed in the 1980s, MRI uses magnetic fields and radio waves to produce high-resolution images of soft tissues, making it invaluable for neurological, musculoskeletal, and cardiovascular applications.

- **Positron Emission Tomography** : A functional imaging technique that visualizes metabolic processes in the body, often used in oncology and neurology.
- **Fundus Imaging**: Captures detailed images of the interior surface of the eye, including the retina, optic disc, macula, and posterior pole. It is crucial for diagnosing and monitoring conditions like diabetic retinopathy, glaucoma, and macular degeneration.
- **Optical Coherence Tomography**: A non-invasive imaging technique that provides high-resolution cross-sectional images of the retina, essential for diagnosing and managing retinal diseases such as age-related macular degeneration, diabetic retinopathy, and glaucoma.
- **Dermatoscopic Imaging**: Used for examining skin lesions with a dermatoscope, which provides magnification and illumination. This technique is vital for the early detection and diagnosis of skin cancers, including melanoma.
- **Ultrasound**: Utilizes high-frequency sound waves to create images of internal organs and tissues. It is widely used in obstetrics, cardiology, and emergency medicine.

The evolution of medical imaging analysis has been linked to computer technologies and the advancement of artificial intelligence. In the past, radiologists primarily depended on their visual assessments and expertise to interpret medical images, often resulting in qualitative assessments. Nowadays, the advancement of digital imaging and the development of powerful computers enable the making of computerized image analysis tools. These tools facilitate more objective and quantitative evaluations, thus enhancing diagnostic accuracy and consistency. Recently, deep and ML methods have revolutionized medical image analysis. These methods leverage large datasets and complex algorithms to automatically detect patterns and anomalies in medical images using different classification, and segmentation techniques. These advancements have paved the way for the development of sophisticated CAD systems that can assist clinicians in making more accurate and timely decisions.

## 2.3 CAD and detection systems

### 2.3.1 Overview of CAD

CAD is an emerging field that has shown useful results being a reliable tool for medical professionals dealing with complex diagnostic images, and making precise evaluations in the past few years. With the advancement of digital medical imaging technologies, CAD

systems are now playing an increasingly important role in the medical diagnostic field. The application of these technologies is one in which medical images are analyzed automatically by algorithms, including artificial intelligence, ML, computer vision (CV), and medical image processing (Fig. 2.1), yielding useful information to assist in abnormalities and disease detection [7, 8]. Therefore, the main goal of CAD systems is to assist healthcare professionals in the early detection, diagnosis, and treatment planning of various diseases and abnormal signs, ultimately improving patient outcomes and reducing healthcare costs.

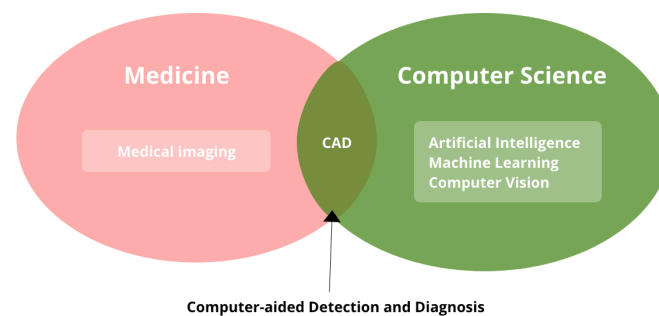


Fig. 2.1 The intersection of medicine and computer science in CAD

There are two principal aspects of CAD research, including “Detection” (CAdE) and “Diagnosis” (CAdx). The CAdE is defined as the technology of lesion region localization in an image. It aims to decrease observational oversights of physicians interpreting medical images by marking the lesion parts or structures visible in the medical images. The CAdx can be defined as identifying the potential diseases that aim to provide additional assistance and support to the clinicians. Fig. 2.2 shows the general flowchart of a CAD system based on DL and medical images.

### 2.3.2 CAD applications

Over the years, CAD has expanded its reach across a wide range of medical fields including radiology, pathology, ophthalmology, dermatology, and more. This expansion has been driven by the integration of advanced artificial intelligence, and computer vision technologies, which have significantly enhanced the capabilities and applications of CAD systems. In our research project, we concentrated on three specific medical fields.

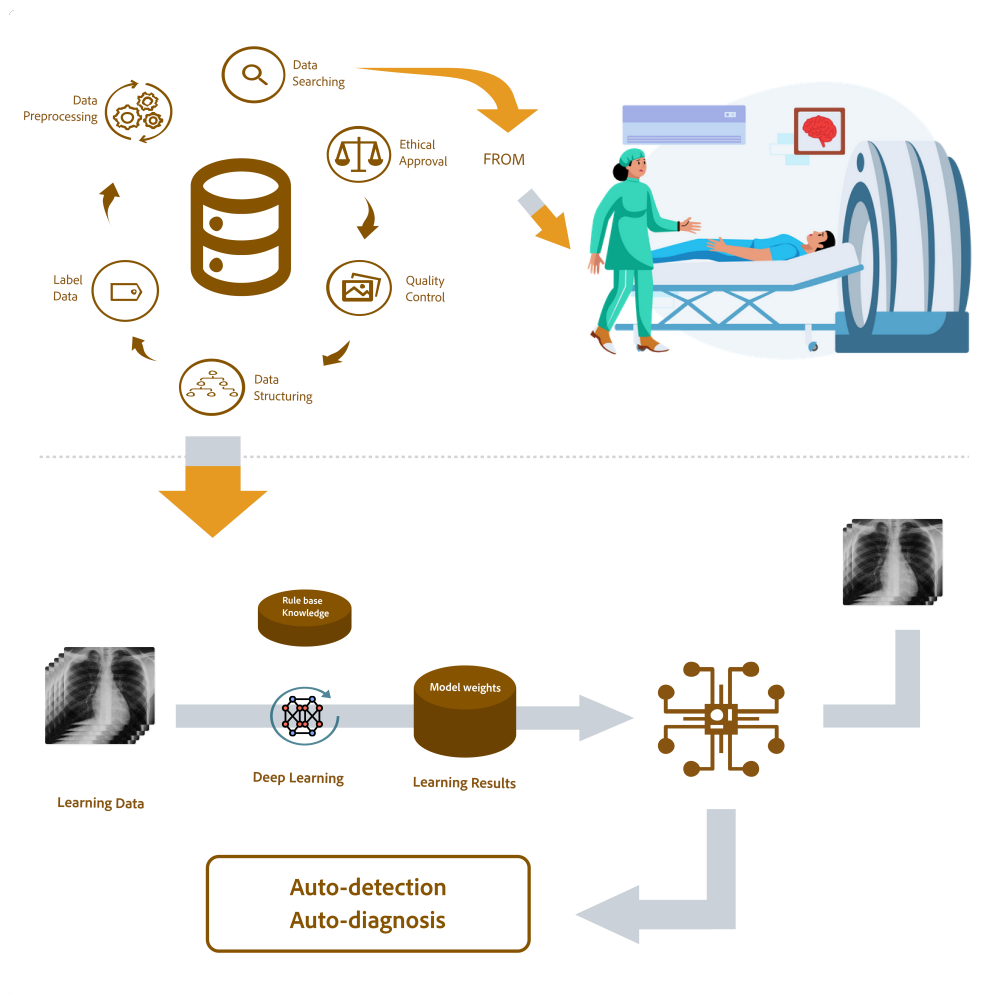


Fig. 2.2 General flowchart of a CAD system based on DL and medical images

### 2.3.2.1 Application in chest radiology

Chest diseases include many conditions that affect the organs and structures within the thoracic cavity, including the lungs, heart, and surrounding tissues. Among the most common and impactful chest diseases are pneumonia, COVID-19, and lung opacity disease.

- **Pneumonia:** is an inflammatory condition of the lung primarily affecting the alveoli. It is typically caused by infections with viruses, bacteria, or other microorganisms. Symptoms can range from mild to severe, including cough, fever, chills, and difficulty breathing. Diagnosis is often confirmed through clinical examination and imaging, with chest radiographs being a crucial tool for identifying the extent and nature of lung involvement.



- **COVID-19:** caused by the SARS-CoV-2 virus, emerged as a global pandemic, significantly affecting respiratory health worldwide. The disease primarily targets the respiratory system, leading to symptoms such as fever, cough, and shortness of breath. Severe cases can result in acute respiratory distress syndrome, requiring intensive medical intervention. Imaging, particularly chest CT scans and radiographs, has proven vital in the diagnosis and management of COVID-19, revealing characteristic patterns such as ground-glass opacities and bilateral infiltrates.
- **Lung opacity:** refers to conditions where areas of the lung appear denser than normal on imaging studies, such as CXR or CT scans. These opacities can be caused by a variety of factors, including infections, inflammations, tumors, or chronic lung diseases like interstitial lung disease. The presence of lung opacities can indicate underlying pathology that requires further investigation to determine the exact cause and appropriate treatment. Accurate identification and characterization of lung opacities are essential for proper diagnosis and management.

CAD technology can be effectively applied in chest radiology to reduce the diagnostic workload for radiologists while enhancing the consistency and reliability of image interpretation. CAD systems employ advanced algorithms, including DL and artificial intelligence, to analyze CXR and CT scans. These systems identify abnormalities that may indicate diseases such as pneumonia, COVID-19, and lung opacity, thereby supporting radiologists in making more accurate diagnoses. Chapter 3 presents our proposed system for the diagnosis of COVID-19 and other lung diseases using DL techniques. [1, 2]

### 2.3.2.2 Application in ophthalmology

Retinal diseases include a range of conditions that affect the retina. These diseases can lead to significant visual impairment or blindness if not diagnosed and treated promptly. Among the most prevalent retinal diseases [9]:

- **Age-related Macular Degeneration (AMD)** is a progressive ocular condition characterized by the gradual deterioration of central vision. This degeneration occurs as a result of age-related damage to the macula, leading to visual impairment primarily in the elderly population.
- **Diabetic Macular Edema (DME)** is an ocular disorder characterized by vascular damage in the retina. If left untreated, DME can result in the accumulation of fluid in the macula, leading to retinal swelling and, ultimately, irreversible vision loss.

- **DRUSEN** is an early manifestation of AMD, which refers to yellow deposits beneath the retina. While individual drusen are not indicative of eye diseases, their accumulation in significant quantities can contribute to the development of AMD and subsequent vision impairment.
- **Choroidal NeoVascularization (CNV)** is a prevalent, sight-threatening condition associated with significant visual impairment. It is characterized by the aberrant proliferation of blood vessels. These blood vessels can break through the layers of the retina, leading to leakage of fluid and blood, and causing damage to the macula.
- **Central Serous Retinopathy (CSR)** is a pathological condition characterized by fluid accumulation in the subretinal space, resulting in retinal detachment and subsequent visual impairment or loss.
- **Diabetic Retinopathy (DR)** is a diabetic complication that impacts ocular health. It arises due to the impairment of blood vessels within the retina, the light-sensitive tissue located at the posterior of the eye.
- **Macular Hole (MH)** is an infrequent ocular condition that affects the central region of the retina responsible for central vision. It is characterized by developing either a complete or partial-thickness discontinuity in the macular tissue. MH can lead to the impairment of central vision accompanied by visual distortion.

CAD systems can be effectively applied in ophthalmology to assist ophthalmologists and enhance the consistency and reliability of image interpretation. These systems use sophisticated algorithms, including DL and artificial intelligence, to analyze retinal images from modalities such as OCT and fundus photography, identifying abnormalities and aiding in diagnosis. Chapter 4 presents our proposed two systems, for the diagnosis of retinal diseases [3] and intraretinal fluid detection using advanced DL techniques.

### 2.3.2.3 Application in dermatology

Skin diseases encompass a wide variety of conditions that affect the skin, ranging from common, benign lesions to serious, life-threatening malignancies. Among the most prevalent and concerning skin diseases are skin lesions, which can include benign growths, inflammatory conditions, and malignant tumors [10]:

- **Melanoma (MEL)**: is a serious form of skin cancer originating in the melanocytes, the cells responsible for skin pigment. Melanoma is known for its ability to spread to

other parts of the body if not detected early. Early signs include changes in the size, shape, or color of moles. Prompt diagnosis and treatment are crucial for improving survival rates.

- **Melanocytic Nevus (NV):** commonly known as a mole, is a benign growth of melanocytes. While most nevi are harmless, they can sometimes develop into melanoma. Regular monitoring for changes in appearance is important to detect any potential malignancy early.
- **Basal Cell Carcinoma (BCC):** is the most common type of skin cancer, originating in the basal cells of the epidermis. It typically appears as a pearly or waxy bump and is slow-growing but can cause significant local damage if untreated. BCC rarely metastasizes but requires treatment to prevent extensive tissue destruction.
- **Actinic Keratosis (AK):** is a precancerous condition caused by long-term sun exposure. It appears as rough, scaly patches on the skin and can potentially develop into squamous cell carcinoma if left untreated. Early treatment is recommended to prevent progression to malignancy.
- **Benign Keratosis (BKL):** including seborrheic keratosis, is a common non-cancerous skin growth. These lesions can vary in appearance but often present as waxy, raised spots. Although benign, they can be mistaken for malignant lesions, necessitating accurate diagnosis.
- **Dermatofibroma (DF):** is a common benign skin nodule, often resulting from minor skin injuries. These firm, raised lesions are usually harmless and do not require treatment unless they cause discomfort or cosmetic concerns.
- **Vascular Lesions (VASC):** such as hemangiomas and port-wine stains, are caused by abnormalities in blood vessels. These lesions can vary in appearance and may require treatment for cosmetic reasons or if they affect function or health.
- **Squamous Cell Carcinoma (SCC):** arises from the squamous cells in the epidermis. It often appears as a firm, red nodule or a flat lesion with a scaly crust. SCC can metastasize if not treated promptly, making early detection and treatment critical for preventing spread.

CAD systems can analyze dermoscopic images to distinguish between benign and malignant skin lesions with high accuracy. By identifying specific features and patterns associated with malignancy, such as asymmetry, irregular borders, and varied coloration.

This technological advancement allows for more precise diagnostics and better treatment strategies for skin conditions. Chapter 4 presents our proposed system, for the diagnosis of different skin lesions [168] using advanced DL techniques.

### **2.3.3 Key challenges in CAD systems**

Despite significant advancements in CAD systems, there remain many challenges in the development, implementation, and adoption phases. The following subsections describe the key challenges that a CAD system can face [8, 11, 12]

#### **2.3.3.1 Data collection and quality assessment**

The first basic step for a CAD system is collecting pertinent data. The high-quality annotated medical imaging datasets are essential for building an accurate decision-making system. However, a challenge can be faced in obtaining a large diverse well-annotated datasets due to many factors such as privacy concerns, variability in imaging protocols, and the labor-intensive nature of manual annotation. In addition to this, there are two main issues related to data collection, namely bias and imbalanced datasets that can affect the effectiveness of the CAD system.

#### **2.3.3.2 Algorithm performance**

The most important challenge in this regard is ensuring the robustness and generalizability of CAD algorithms across different patient populations, imaging devices, and clinical settings. Algorithms trained on limited or biased datasets may falter in real-world scenarios, where variability in imaging modalities, and clinical environments can impact their performance. This challenge is the most important one because the misdiagnoses could potentially impact patient health and care requiring stringent procedures for evaluation and the absence of such would make them unacceptable barriers to implementation.

#### **2.3.3.3 Integration into clinical workflow**

Adopting CAD systems to existing clinical workflow is associated with practical and technical challenges. Healthcare centers are faced with challenges like interoperability between already existing medical imaging systems, security of data, and staff training in using CAD technology properly. Seamless integration can be achieved by working closely with healthcare providers, software developers, and regulatory agencies to define standards for how the CAD system is used in clinical departments.

### **2.3.3.4 Validation and clinical trials**

Strict validation through clinical trials is necessary to demonstrate the effectiveness and safety of CAD systems in diverse clinical settings. Such trials can be time-consuming and costly, but they are essential for regulatory approval and widespread adoption. However, CAD systems should not be a substitute for the decisions of healthcare professionals and clinicians, but as supporting tools to improve diagnostic accuracy and speed. Therefore, careful consideration must be given to the presentation and utilization of CAD results by healthcare providers, ensuring that these tools augment rather than supplant clinical decision-making.

Addressing these challenges is crucial for unlocking CAD's full potential in boosting diagnostic accuracy, improving patient outcomes, and advancing the field of medicine. By overcoming these hurdles, CAD can evolve into an essential tool in healthcare providers' diagnostic staff and offer significant potential for improving medical image analysis, leading to more precise and fast diagnoses, and enhancing patient care delivery.

In conclusion, while CAD systems hold promise for enhancing medical diagnosis, addressing the current challenges through continued research, innovation, and collaboration is crucial for their successful integration into clinical practice.

## **2.4 Theoretical background**

### **2.4.1 Data acquisition and quality assurance**

Data plays a crucial role in the development of CAD systems and the application of DL models for medical imaging analysis. The quality, quantity, and diversity of the data can significantly influence the performance and generalizability of the models. Therefore, the availability of such data is a decisive factor in the success of DL models and CAD systems. It is well-known that the collection of such medical data is inherently difficult due to many challenges encountered during the process. These challenges include privacy and security concerns, quality and consistency of the data, and the difficulty of aggregating data that is dispersed across multiple hospitals.

Medical data can be collected from various sources and through different modalities including, private data from hospitals and clinics, as well as public data from the Internet or other public sources. However, it is crucial to ensure that the collected data is of high quality and relevance to the task at hand to achieve optimal performance from the CAD systems and ML models. Ensuring the quality of medical data is indispensable for developing reliable

and accurate DL models. The following key factors are among the most important factors that must be verified to ensure the data quality.

#### **2.4.1.1 Well-Annotation**

Medical datasets must be well-annotated under the supervision of experienced clinicians. These annotations involve image labeling with relevant information, such as lesion location, disease type and presence, and other helpful information, serving to build an effective model.

#### **2.4.1.2 Preprocessing**

The image preprocessing steps such as normalization, resizing, noise reduction, and artifact removal are essential for standardizing and enhancing image clarity, leading to model performance improvement.

##### **a. Normalization**

Image normalization is a common preprocessing step in image analysis. It adjusts pixel values to a common scale, ensuring consistency across different images. This process enhances the convergence speed and performance of DL models by reducing gradient variance and improving feature learning. Additionally, normalization improves generalization to unseen data and enhances image quality by boosting contrast. These benefits make normalization a standard practice for achieving robust and reliable results in various computer vision and ML applications.

##### **b. Resizing**

Image resizing is an essential preprocessing step in image analysis. It adjusts the dimensions of images to a fixed size required by DL models. This ensures uniform input sizes, facilitating batch processing and reducing computational load. Resizing helps maintain aspect ratio and image quality, ensuring that crucial features are preserved while fitting the model's architecture.

##### **c. Noise reduction**

Image noise reduction is a critical preprocessing technique in image analysis aimed at eliminating unwanted variations or distortions that can hide the true details in an image. By using techniques like Gaussian filtering or median filtering, we can significantly improve the quality of an image. This makes the underlying structures much clearer and more accurately

represented. As a result, DL models can perform better since they are less affected by noise and can concentrate on the important features. This ultimately boosts the accuracy and reliability of image interpretation and analysis.

#### 2.4.1.3 Diversity

The diversity in data contributes to building robust models that can generalize better to different clinical scenarios. This diversity can lie in imaging conditions, image modalities and formats, multiple disease variations, and patient demographics.

#### 2.4.1.4 Imbalance

The most addressed issue in medical data is class imbalance, where some conditions are more prevalent than others. Since some diseases or lesions rarely occur, collecting large data samples can be difficult. Therefore, the lack of data in specific conditions, in turn, leads to the creation of the class imbalance problem. However, techniques like oversampling, undersampling, and synthetic data generation, can help address this issue.

One of the most used techniques to address the class imbalance problem is data augmentation which is a widely employed technique designed to artificially expand the size of a dataset by generating modified versions of existing images. This approach enhances the generalization capabilities of ML and DL models, particularly in computer vision tasks. Common medical image augmentation techniques include [13]:

- **Rotation:** Rotating the image by specific angles to make the model invariant to object orientation.
- **Flipping:** Horizontally or vertically flipping the image to produce mirror images.
- **Scaling:** Zooming in or out to simulate variations in image magnification and improve model robustness.
- **Translation:** Shifting the image along the x or y axis to enhance the model's robustness to object position within the image.
- **Cropping:** Randomly cropping portions of the image to enable the model to focus on different regions.
- **Color Jittering:** Randomly altering the image's brightness, contrast, saturation, and hue.

- **Elastic Transformations:** Applying random elastic distortions to simulate deformations.

## 2.4.2 Machine learning

Machine Learning (ML) is a sub-domain of Artificial Intelligence that involves the development of methods or algorithms that allow software applications to become more accurate in predicting and detecting outcomes without explicit programming [14].

In the context of medical image analysis, the integration of ML involves image processing, image segmentation, feature extraction, and image classification using different algorithms. Nowadays, with the advancement of ML approaches such as DL and neural networks, they have demonstrated remarkable success in handling medical data through many tasks such as detecting tumors, identifying diabetic retinopathy, and diagnosing pneumonia and COVID-19 and other tasks, resulting in significant advancement in the medical imaging analysis field.

### 2.4.2.1 Classification

Classification is one of the most widely utilized techniques in data analysis, enabling the development of decision models that predict the behavior of future instances. The main objective is to predict a class label which is selected from a predefined set of probabilities. This means building a learning system that performs the research of such a procedure according to a model. This learning system can be based on probabilistic assumptions (e.g., Naive Bayes Classifier), proximity concepts (e.g., Nearest Neighbors), or hypothesis space exploration (e.g., Decision Trees, Support Vector Machine, and Neural Networks).

The classification can be divided into binary classification, which is the special case of distinguishing between exactly two classes (e.g., COVID-19 and non-COVID-19 cases), and multiclass classification, which involves distinguishing between more than two classes (e.g., COVID-19, Pneumonia, Tuberculosis, and Healthy cases).

The evaluation of the classification process can be measured by different factors including:

- **Accuracy:** refers to the proportion of correct classifications. Formally, it defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.1)$$

- **Precision:** refers to the proportion of correctly predicted positive instances out of all predicted positive instances. Formally, it defined as:

$$Precision = \frac{TP}{TP + FP} \quad (2.2)$$



- **Sensitivity or Recall:** refers to the proportion of actual positive cases that the model correctly predicts. Formally, it defined as:

$$Recall/Sensitivity = \frac{TP}{TP + FN} \quad (2.3)$$

- **F1-score:** refers to the combination of precision and recall into a single score. Formally, it defined as:

$$F1\_Score = 2 \cdot \frac{Precision * Recall}{Precision + Recall} \quad (2.4)$$

- **Specificity:** refers to the proportion of negative instances which correctly predicted. Formally, it defined as:

$$Specificity = \frac{TN}{TN + FP} \quad (2.5)$$

- **AUC (Area Under the ROC Curve):** is frequently employed to evaluate the effectiveness of binary classification models by calculating the area under the ROC curve. Essentially, a higher AUC score reflects the model's superior ability to differentiate between the two classes. Mathematically, AUC can be expressed as:

$$AUC = \int_0^1 \frac{TP}{TP + FN} d \frac{FP}{TN + FP} \quad (2.6)$$

Here, TP, TN, FP, and FN refer to true positives, true negatives, false positives, and false negatives, respectively

#### 2.4.2.2 Segmentation

Segmentation is a fundamental technique in data analysis, particularly within the realm of image processing. It refers to the process of partitioning an image into several regions or segments that are perceptually similar, where each representing a meaningful part of the image, such as an object or a region of interest. The segmentation procedure aims to simplify the representation of an image into something more meaningful and easier to analyze, which is useful in many computer vision applications.

Segmentation plays a critical role in various applications giving it considerable advantages, in the healthcare sector, image segmentation enables precise identification and analysis of medical images, facilitating the diagnosis and treatment of disease. It enables the precise delineation of organs, tumors, and abnormalities, facilitating the planning and follow-up of surgical operations.

The evaluation of the segmentation process can be measured by different factors including:

- **Dice Similarity Coefficient (DSC):** measures the overlap between the predicted segmentation and the ground truth. It is a harmonic mean of precision and recall, emphasizing how similar the predicted and actual segmentations are. Mathematically, DSC can be expressed as:

$$DSC = \frac{2 \cdot |A \cap B|}{|A \cap B| + |A \cup B|} \quad (2.7)$$

- **Intersection over Union (IoU) or Jaccard index:** measures the overlap between the predicted segmentation and the ground truth divided by their union. It is a widely used metric for evaluating the accuracy of the segmentation model.

$$IoU = \frac{|A \cap B|}{|A \cup B|} \quad (2.8)$$

In addition to **Accuracy**, **Sensitivity**, **Precision**, and **Specificity**. Here, TP, TN, FP, FN, A, and B refer to true positive pixels, the true negative pixels, false positive pixels, false negative pixels, the set of predicted pixels, and the set of ground truth pixels respectively.

### 2.4.3 Deep learning

Deep learning (DL) is a subset of ML that uses artificial neural networks with multiple layers. These networks are designed to model and understand complex patterns in data [15]. In medical image analysis, DL approaches have been the pivotal key to successful CAD systems development. Indeed, handling the high complexity and unstructured medical data is difficult to process using traditional rule-based methods and classic ML algorithms. Therefore, these problems are solved by using advanced DL approaches such as convolutional neural networks and vision transformer models. These approaches have been particularly effective at handling such data, leading to significant advancements in medical image analysis tasks.

#### 2.4.3.1 Feed-forward neural networks

A feed-forward neural network is the core of DL, which is inspired by the oversimplified structure and function of the human brain. It is the simplest artificial neural network (ANN) wherein connections between neurons are feed-forward without any cycles or loops. Whereas, each neuron receives input, processes it through an activation function, and passes the output

to the next layer. This kind of network is usually used in classification and linear filtering problems [16].

The multi-layer perceptron (MLP) is a kind of feedforward neural network that is composed of an input layer, some hidden layers, and an output layer (see Fig. 2.3). Each layer consists of many neurons and each neuron uses a nonlinear activation function to mimic the brain neuron activity. These MLPs are employed in the solution of diverse problems like pattern classification, identification systems, and robotics.

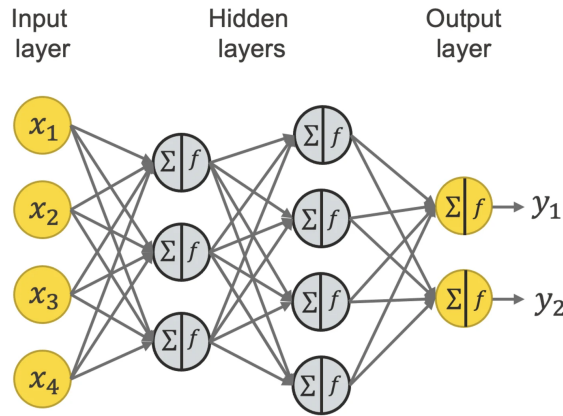


Fig. 2.3 A diagram of MLP architecture consisting of an input layer, hidden layers, and an output layer

#### 2.4.3.2 Activation functions

In neural networks, the activation function introduces non-linearity into the network, enabling it to learn complex patterns. There are a variety of activation functions, the popular and the used ones in our contributions are listed in the following:

- **Sigmoid function:** Produces an output between 0 and 1, useful for binary classification.

$$f(x) = \frac{1}{1 + e^{-x}} \quad (2.9)$$

- **Tanh function:** Produces an output between -1 and 1, often used in hidden layers.

$$f(x) = \tanh(x) \quad (2.10)$$

- **ReLU function (Rectified Linear Unit):** Outputs the input directly if it is positive; otherwise, it outputs zero. ReLU is widely used due to its efficiency in training deep

networks.

$$f(x) = \max(0, x) \quad (2.11)$$

- **GELU function(Gaussian Error Linear Unit):** is a recent activation function that combines both ReLU and Gaussian noise properties.

$$f(x) = x \cdot \Phi(x) \approx x \cdot \frac{1}{2} (1 + \tanh(\sqrt{\frac{2}{\pi}} (x + 0.044715x^3))) \quad (2.12)$$

- **Softmax function:** converts the raw output scores of a network into a probability distribution over predicted output classes. In addition, it ensures that the sum of the output probabilities is equal to 1, which provides an interpretable result for classification tasks.

$$f(x_j) = \frac{e^{x_j}}{\sum_j e^{x_j}} \quad (2.13)$$

### 2.4.3.3 Backpropagation and optimization

The backpropagation is a key algorithm used for training ANNs, which is based on an optimization algorithm, such as Stochastic Gradient Descent (SGD), Adam, Mini-batch Gradient Descent, and others, that attempts to minimize the error of the network by moving down the gradient of the error curve. The first stage is forward propagation, wherein the network inputs are propagated layer by layer until the corresponding outputs are produced. These outputs are then compared to the respective desired responses, resulting in deviations (errors) that are used to adjust the weights and thresholds of all neurons. Consequently, these errors necessitate the application of the second stage, known as backward propagation, during which the adjustments of the synaptic weights and thresholds of all neurons in the network are executed. The successive application of the forward and backward stages allows for the automatic adjustment of the synaptic weights and thresholds of the neurons in each iteration. This iterative process gradually reduces the sum of the errors produced by the network responses to the desired outputs.

#### a. Loss function

The loss function plays the role of the objective or the cost function for the optimization problems in the neural networks. It supervises the training stage of networks and measures how well the model's predictions match the actual outcomes. Thus, the choice of the loss function depends on the task type and the models' specific requirements.

In medical image classification tasks, the cross-entropy loss [17] is the commonly used function that measures the performance of a classification model whose output is a probability distribution over classes. The choice of the applied cross-entropy loss depends on the type of classification task—binary or multi-class. Consequently, two major methods are widely used:

- **Binary Cross-Entropy (BCE) loss:** is utilized for binary classification problems, where the task is to categorize inputs into one of two classes (e.g., healthy vs. not healthy). It measures the performance of a model whose output is a probability value between 0 and 1. The BCE loss is can defined as:

$$L_{BCE} = -\frac{1}{N} \sum_{i=0}^{N-1} y_i \log \hat{y}_i + (1 - y_i) \cdot \log(1 - \hat{y}_i) \quad (2.14)$$

where  $N$  is the number of samples.  $y_i$  represents the label of the  $i^{th}$  sample and  $\hat{y}_i$  denotes the probability of the positive class for the  $i^{th}$  sample. The BCE loss quantifies the dissimilarity between the predicted probabilities  $\hat{y}_i$  and the true labels  $y_i$  for each sample. This loss function penalizes the model for larger discrepancies between the predicted and actual labels, thereby guiding the model to improve its accuracy.

- **Categorical Cross-Entropy (CCE) loss:** is employed for multi-class classification problems where there are more than two possible classes, such as classifying multiple types of diseases. This loss function evaluates the performance of a model whose output is a probability distribution over  $K$  possible classes. The CCE loss can be defined as:

$$L_{CCE} = -\frac{1}{N} \sum_{i=0}^{N-1} \sum_{j=0}^{K-1} y_i^j \log(p_i^j) \quad (2.15)$$

where  $N$  is the number of samples,  $K$  is the number of categories.  $p_i^j$  represents the  $j^{th}$  probability of the  $i^{th}$  sample, and  $y_i^j$  represents the  $j^{th}$  label of the  $i^{th}$  sample.

For the medical image segmentation task, the loss function is based on the dice coefficient [18], which measures the overlap between the predicted segmentation and the ground truth, and is defined as:

$$L_{Dice} = 1 - \frac{2 \sum_{i=1}^N y_i \hat{y}_i}{\sum_{i=1}^N y_i + \sum_{i=1}^N \hat{y}_i} \quad (2.16)$$

where  $y_i$  is the ground truth label for pixel  $i$  and  $\hat{y}_i$  is the predicted label for pixel  $i$ . Dice Loss is well-suited for tasks with highly imbalanced classes, such as detecting small lesions.

#### 2.4.3.4 Convolutional neural networks

Convolutional neural network (CNN) is a type of feed-forward neural network architecture, in which the connection pattern between neurons (nodes) is inspired by the animal's visual cortex. CNNs are designed for processing grid-like data such as images which are the dominant data type in the computer vision community. CNNs are currently the most effective models for image recognition, classification, and segmentation due to the use of convolutional layers which can automatically learn spatial hierarchies of features. As the medical image relies heavily on spatial relationships between pixels, CNNs excel at capturing these relationships through convolutional filters, which learn to detect edges, textures, and patterns crucial for accurate diagnosis.

Several representative CNN architectures are widely recognized for their effectiveness in medical image classification and diagnosis, including ResNet [19], DenseNet [20], VGGNet [21], InceptionNet [22], and EfficientNet [23]. For image segmentation, notable architectures include the Fully Convolutional Network [24], U-Net [25], V-Net [26], and their variants [27, 28]. Figure 2.4 illustrates a sample of CNN architecture structure that commonly consists of a convolutional layer, pooling layer, batch normalization, and skip connection.

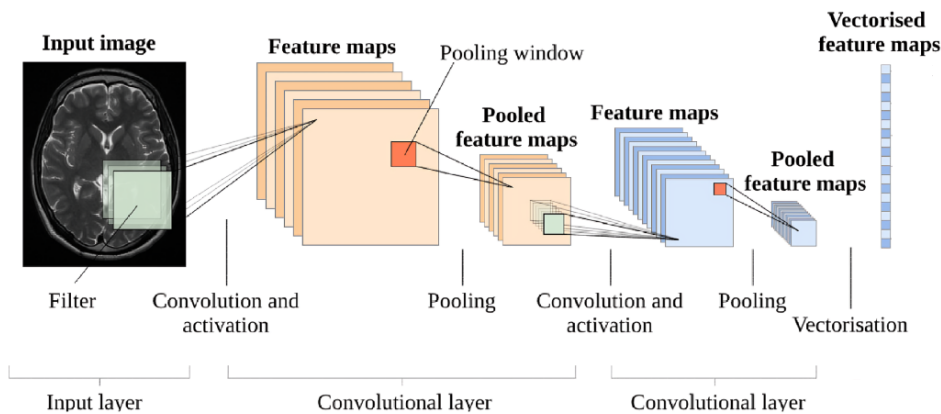


Fig. 2.4 A diagram of CNN architecture mainly consisting of convolutional layers and pooling layers.

##### a. Convolution operation

The convolution operation is the fundamental component of the building block of the CNN. Its role is to detect local features at different positions of feature inputs from the previous layer using learnable kernels  $K_{i,j}^l$ . Essentially, these kernels are the connection weights between the feature maps of the current layer and those of the previous layer. The convolutional operation can be mathematically represented as follows:

$$Y_i^{(l)} = h\left(\sum_{j=1}^{N^{(l-1)}} K_{i,j}^{(l)} * Y_j^{(l)} + B_i^{(l)}\right) \quad (2.17)$$

where  $Y_i^{(l)}$  is the  $i^{th}$  feature map at  $l^{th}$  layer,  $B_i^{(l)}$  is a bias matrix,  $h$  is an activation function.

### b. Pooling operation

The pooling operation involves the downsampling of feature maps, which decreases the network's parameters while expanding the network's receptive field. Pooling works independently on each depth layer of the input, modifying it solely at a superficial level. Typically, pooling operations utilize either the max function or the average function. In essence, the purpose of pooling is to minimize the number of weights, thereby reducing computational time as well as the probability of overfitting.

### c. Batch normalization

Batch normalization is typically positioned post-activation within a neural network layer to standardize the mean and variance of inputs, thereby mitigating the impacts of weight scale and input data manipulation [29]. This procedure functions as a network regularizer, expediting the training phase and reducing sensitivity to parameter initialization.

### d. Skip connection

Skip connection or Residual connection is a significant innovation in deep neural networks. It is a kind of residual learning that addresses the problem of vanishing gradients and enables the training of substantially deeper networks [19, 20]. This connection skips and bypasses one or more layers by adding the input to the output of a subsequent layer. This additive shortcut helps preserve the flow of gradients during backpropagation, thus mitigating the degradation problem. Therefore, the skip connections are foundational elements in modern DL architectures.

#### 2.4.3.5 Attention mechanism

Attention mechanisms have become a powerful approach in DL, significantly improving the performance of models. In medical image analysis, the attention mechanisms can improve classification and segmentation tasks by enabling the model to focus on the most relevant parts of the image, enhancing its ability to identify and delineate different structures or abnormalities accurately. In the beginning, the attention mechanism was developed

for natural language processing (NLP) tasks, and due to the high impact on the models' performance improvements, it has been adopted and applied to other domains and tasks including computer vision. The core idea of the attention mechanism is to assign different weights to different parts of the input, allowing the model to focus more on the important regions while ignoring the less relevant ones.

#### a. Soft attention

In the soft attention mechanism, each part of the input is assigned a weight  $\alpha_i$ , and the output is a weighted sum of these parts [30, 31]. Assuming that the input sequence  $x = (x_1, x_2, x_3, \dots, x_n)$ , soft attention computes a context vector  $c$  that captures the most important features of the input sequence as a weighted sum of the input vectors:

$$c = \sum_{i=1}^n \alpha_i x_i \quad (2.18)$$

where  $\alpha_i$  are the attention weights for each input vector  $x_i$ . The attention weight is calculated using an attention function  $s$ , which typically depends on the hidden state  $h_t$  and some additional context information.

$$e_i = s(h_t, x_i) \quad (2.19)$$

The scores  $e_i$  are then normalized using the softmax function to obtain the attention weights:

$$\alpha_i = \frac{\exp(e_i)}{\sum_{i=1}^n \exp(e_i)} \quad (2.20)$$

Overall, soft attention offers an effective approach to selectively focus on the most pertinent parts of the input sequence, which improves the model's performance in classification or segmentation tasks.

#### b. Self-Attention

The self-attention mechanism [32] represents a critical development, fundamentally transforming how models analyze and comprehend sequential data. It has been particularly successful in transformer architectures. The self-attention mechanism functions by computing a set of attention scores that determine the influence of each element in the sequence on every other element. This process involves three linear transformations query  $Q$ , key  $K$ , and value  $V$ .

Let  $X = (x_1, x_2, x_3, \dots, x_n)$  be the input sequence, the self-attention mechanism computes:



$$Attention(Q, K, V) = softmax(\frac{XW^Q(XW^K)^T}{\sqrt{d_k}})XW^V \quad (2.21)$$

where the  $W^Q$ ,  $W^K$ , and  $W^V$  are weight matrices for the query, key, and value projections, respectively.  $d_k$  is the dimensionality of the key vectors. The softmax function ensures that the attention scores are normalized to form a probability distribution.

Overall, this mechanism offers an effective approach that allows the model to capture dependencies between different parts of the input sequence effectively.

#### 2.4.3.6 Vision Transformer

One of the most significant revolutions in fundamental artificial intelligence and DL is transformer network, developed for NLP tasks by Vaswani et al. [32]. The Transformer introduces a novel learning approach that eliminates all recurrence operations in DL models, replacing them with self-attention layers effectively to model long-range dependencies in data. Recently, transformers have been extended to computer vision tasks. Dosovitskiy et al. [33] introduced the Vision Transformer (ViT) architecture. Unlike CNNs that rely on convolution operations to extract local features, ViTs leverage the Self-Attention mechanism to process image data and capture long-range dependencies. This novel computer vision approach has shown promising results in image classification and segmentation tasks, leading to the development of various transformer-based models like the Swin Transformer [34], Twins [35], and CrossViT [36], serving as general-purpose backbones for computer vision.

ViTs excel at understanding complex medical conditions by capturing long-range dependencies and global context within images. Unlike traditional methods that analyze images locally, ViTs process information holistically, integrating details from different areas to uncover hidden relationships. This ability to recognize complex patterns across the entire image improves diagnostic accuracy, especially for diseases with subtle or widespread features, and opens up possibilities for groundbreaking medical discoveries.

##### a. Patch embedding and linear projection

Patch embedding is the process of transforming images into image patches. Analogous to NLP, where sentences are composed of words or sub-word tokens, images are divided into patches that serve a similar role. In ViT, the image is divided into fixed-size patches through patch embedding. Each patch is flattened and linearly embedded into a vector, creating a sequence of patch embeddings. For an image of dimensions  $H_{image} \times W_{image}$  and patch size  $H_{patch} \times W_{patch}$ , the image is divided into  $(\frac{H_{image}}{H_{patch}} \times \frac{W_{image}}{W_{patch}})$  patches. After constructing the

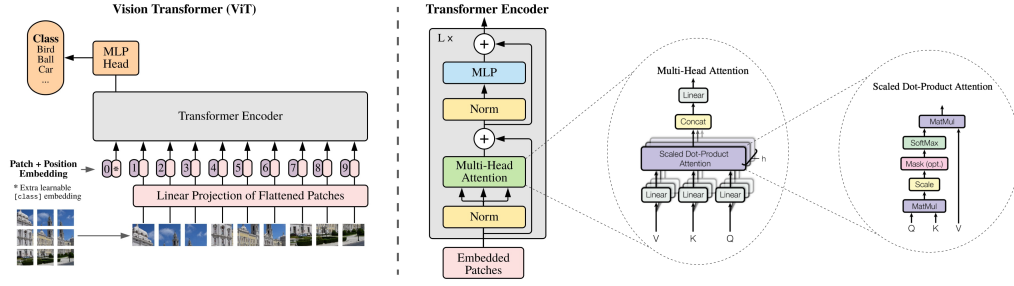


Fig. 2.5 Vision transformer architecture. [33]

image patches, a linear projection layer is employed to map the image patch arrays to patch embedding vectors.

### b. Position embedding

In NLP, the order and position of words in a sentence are crucial, as reordering the words can alter the meaning of the entire sentence. Similarly, in image processing, reordering image patches can render the image meaningless, thereby affecting the model's ability to predict what the complete image represents accurately. Consequently, a mechanism to enable the model to infer the order or position of the patches is necessary. To address this, ViTs incorporate positional embeddings into the patch embeddings. These positional embeddings help the model understand the spatial relationships between patches, thereby preserving the structural integrity of the image.

### c. Transformer encoder

The encoder is one of the fundamental components in the Transformer architecture. In ViT, the combined patches with positional encodings are fed into a standard Transformer Encoder, consisting of multiple layers of Multi-Head Self-Attention (MHSA) and Feed-Forward Networks (FFN). Replacing recurrent and convolutional layers, the MHSA layers allow the model to weigh the importance of different tokens in the input sequence relative to each other. This enables the model to capture global context and dependencies across the entire image. The MHSA layer includes multiple projections of the queries  $Q$ , keys  $K$ , and values  $V$ , corresponding to the number of heads, as illustrated in Fig. 2.5. These projections are learned by training weight matrices that determine the weights of the current patch, the extent of attention that should be paid to other patches in the image, and the vector representations of all the patches in the image.

$$head_i = Attention(Q, K, V) \quad (2.22)$$

$$MultiHead(Q, K, V) = concat(head_1, head_2, \dots, head_h)W^O \quad (2.23)$$

The outputs of all heads are concatenated and linearly projected to combine the learned features from each head.

After the MHSA, the output is passed through a position-wise FFN, which is applied independently to each token, consisting of two fully connected layers with a GELU activation. The FFN allows the model to learn more complex representations by combining and transforming the captured information. Residual connections and layer normalization are applied around both the MHSA and FFN sub-layers to ensure stable, efficient training and performance improvements. Overall, the transformer encoder facilitates efficient processing of input sequences by employing self-attention to model dependencies between all pairs of tokens and feedforward network to introduce non-linearities. Incorporating residual connections and layer normalization helps stabilize the training process and mitigate the vanishing gradient problem.

#### d. Classification head

For image classification tasks, a special classification token is appended to the sequence of patch embeddings before feeding them into the transformer encoder. The output corresponding to this classification token is passed through a fully connected layer to generate the final class predictions.

#### 2.4.3.7 Advanced multi-layer perceptron

The significant advancement in the DL field, in addition to the invention of Transformers, has led to the emergence of novel modern MLP architectures such as MLP-Mixer and FNet. These models, while inspired by the Transformer, adopt an attention-free approach. The following subsection provides a comprehensive overview of each model architecture.

##### a. MLP-Mixer

The MLP-Mixer, introduced by Tolstey et al. [37], presents a novel approach that leverages a purely MLP-based architecture for image recognition. Unlike CNN and ViT, without any convolutional or attention mechanisms, the architecture uses only a series of MLP layers. The MLP-Mixer works in the same manner as the ViT. It processes input images by dividing

them into patches and flattening these patches into a sequence of tokens. Each token is then processed through the Mixer layer which is the core block of MLP-Mixer. The Mixer layer consists of two MLP sub-blocks:

- **Token-Mixing MLP:** This sub-block focuses on individual image patches (tokens) and aims to capture interactions between spatial locations. It operates independently on each channel of the feature map.
- **Channel-Mixing MLP:** This sub-block focuses on information across channels and allows features from different channels to interact. It operates on each token independently.

The combination of these two MLP layers facilitates information exchange across both spatial dimensions and feature channels, achieving competitive performance on image classification tasks.

#### b. FNet

The FNet is an MLP proposed by Lee-Thorp et al. [38]. It presents an innovative approach to transformer encoder architectures as illustrated in Fig 2.6. It replaces the self-attention sublayers with straightforward linear transformations that blend input tokens to reduce the computational cost of the self-attention mechanism. The FNet works in the same manner as a transformer. The input data undergoes patch projection and position embedding procedures, resulting in a sequence of token creation. The architecture comprises two primary sub-layers:

- **Fourier Transform Layer:** This layer applies the 2D Fourier Transform (FT) to the input tokens, using the Fast Fourier Transform (FFT) [39], with 1D FT along the sequence dimension,  $\mathcal{F}_{seq}$ , and 1D FT along the hidden dimension,  $\mathcal{F}_h$ .

$$y = \Re(\mathcal{F}_{seq}(\mathcal{F}_h(x))) \quad (2.24)$$

- **Feed-Forward Network:** Following the Fourier Transform, the tokens are processed through a standard FFN to refine the extracted features.

By leveraging the Fourier Transform, FNet achieves competitive performance with significantly lower computational costs, making it suitable for various applications.

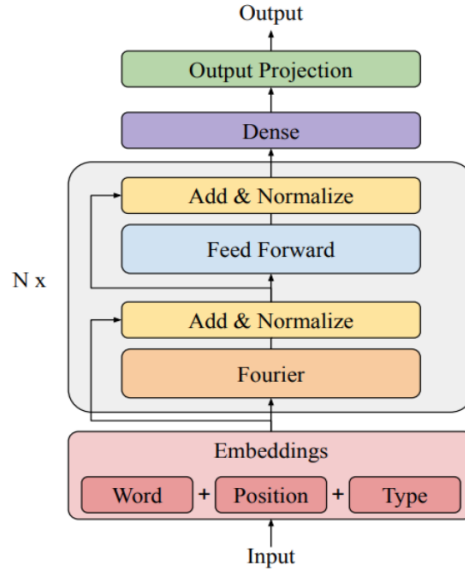


Fig. 2.6 The FNet model architecture. [38]

## 2.5 Overall literature review

DL techniques have gained widespread use and become powerful tools in various areas of medical image analysis, including CAD, image segmentation, and image-guided therapy. In this section, the relevant works from the literature review that are related to our contributions are presented. Each chapter's related works are discussed in a separate section.

### 2.5.1 COVID-19 diagnosis

The outbreak of the COVID-19 pandemic has spurred a significant amount of research into effective diagnostic methods through different DL models. As the disease primarily affects the respiratory system, medical imaging techniques such as chest X-rays (CXR) and CT scans have been crucial in COVID-19 diagnosis and management. Most research studies have used CNNs for this task. Some of these research studies used transfer learning via multiple pre-trained models, including AlexNet [40], ResNet [19], [41], DenseNet [20], VGGNet [21], InceptionNet [22], and EfficientNet [23]. The others proposed custom models. Most of the research studies were based on X-rays and CT scans. Mozaffari et al. [42] presented a survey on DL models for COVID-19 detection, which mentioned the most proposed models and solutions from the beginning of the pandemic.

### 2.5.1.1 Chest X-Ray based diagnosis

Due to the limited number of images corresponding to COVID-19 cases, many studies have employed transfer learning and pre-trained models for classification tasks. For instance, Karar et al. [43] compared several pre-trained models using a dataset containing 306 X-ray images, including both positive COVID-19 and other pneumonia cases, achieving an accuracy of 99.9% with cascaded DL classifiers, including VGG16, DenseNet169, and ResNet50. Similarly, Narin et al. [44] used a 7,496 X-ray images dataset of COVID-19, normal and bacterial disease cases. They conducted a five-fold cross-validation method to train and test a ResNet50 model, resulting in achieving the classification accuracies of 96.1%, 99.5%, and 99.7% for the different classes of COVID-19 vs. normal, COVID-19 vs. viral disease, and COVID-19 vs. bacterial disease, respectively. In another study [45], different pre-trained models were compared using a dataset comprising 6000 X-ray images of non-COVID-19 patients and 500 images of COVID-19 patients. Out of these, 5,300 images were used for training and 1,200 for testing, with MobileNet achieving the highest accuracy of 98.6%. Castiglioni et al. [46] used ten various ResNet models and datasets containing 610 images including COVID-19 and normal cases. These models were trained on this dataset and obtained an ROC-AUC of 0.80 for the classification task. Apostolopoulos et al. [41] conducted a comparative study using VGG19, MobileNet, and Inception models. Their results showed that MobileNet and VGG19 achieved the best performance, with an accuracy of over 98%. The work reported in [47], was a comparative study using different models such as VGG16, VGG19, MobileNet [48], InceptionNet, Xception, InceptionResNetV3 [49], DenseNet20, and NASNetLarge [50]. The authors used a dataset containing CXR images from patients with confirmed COVID-19, common bacterial pneumonia, and normal incidents. The highest achieved accuracy was around 96% with the best performance given by the VGG16 model. Berrimi et al [51] used InceptionV3, DenseNet201, and modified DenseNet architecture. These models were trained on this public CXR dataset and achieved 92.35%, 63%, and 85% respectively. In [52], the authors proposed a model for classifying diseases related to the coronavirus family. To enhance system performance, they employed UNet [25] for image segmentation and compared the classification performance of four DL models: SqueezeNet [53], ResNet18, Inception V3, and DenseNet-201. Ultimately, Inception V3 achieved the best results, with a classification accuracy of 98.2%, sensitivity of 97.7%, and specificity of 97.1%. In a study by [54], the EfficientNet [23] model was employed to design a COVID-19 detection system. The model was trained on a dataset consisting of 500 X-ray images of healthy cases, 504 images of non-COVID-19 pulmonary diseases, and 504 COVID-19 images. To evaluate the model, the data were divided into 10 portions, and the system was executed 10 times with different test samples. Each test was run for 11 epochs.

The system ultimately achieved average accuracies of 99.62% for two-class classification and 96.70% for three-class classification.

On the other hand, some research introduced custom architectures. In [55], the authors proposed DarkCovidNet, a novel deep CNN architecture, based on the YOLO architecture [56]. After collecting data from various sources, the authors classified the data into three categories: COVID-19, healthy, and pneumonia. This model was trained for binary classification as well as multi-class classification. For the binary task, the model achieved an accuracy of 98.08%, a sensitivity of 95.13%, and a precision of 98.03%. For the three-class classification, the model attained an accuracy of 87.02%, a sensitivity of 92.18%, and a precision of 89.96%. In [57], the authors proposed COVID-Net. The architecture heavily utilizes a lightweight residual PEPX (Projection Expansion Projection Extension) design pattern, consisting of a human-machine collaborative design strategy and a machine-driven design exploration stage. The results demonstrated a sensitivity and accuracy of 93.3%. Ismael and Şengür [58] utilized three approaches: deep feature extraction with SVM, an end-to-end 21-layer CNN model, and fine-tuning of a pre-trained CNN model. The author used 180 positive and 200 negative cases. The results showed an accuracy of 90.58%, a sensitivity of 88.00%, a specificity of 93.30%, and an F1-score of 90.70%. In a paper by [59], the authors proposed a parallel-dilated CNN which includes a parallel stack of convolutional layers, an activation layer, and a max-pooling layer. In addition, Grad-CAM [60] and Grad-CAM++ [61] approaches are used to highlight the important regions. They used a total of 2905 CXR images. The proposed approach gave an accuracy, precision, and recall of 96.58%, 96.58%, and 96.59%, respectively.

### 2.5.1.2 Chest CT scans based diagnosis

The CT scans provide detailed chest images and offer a higher sensitivity for detecting COVID-19-related abnormalities. Many studies have leveraged DL techniques to enhance the diagnostic accuracy of CT scans for COVID-19. For instance, Yu et al. [62] designed a detection system based on the GoogleNet model. The authors tested three different GoogleNet architectures, including the original and modified versions by adding or removing layers such as dropout and fully-connected layers. The best performance was achieved with an accuracy of 87.5%, a sensitivity of 90.9%, and a specificity of 84%. In a study by [63], the authors first preprocessed the data using three steps: histogram stretching to normalize contrast, image cropping to eliminate trivial and unimportant information, and image resizing to reduce the volume of information. For the proposed DenseNet-OTLS model, they employed DenseNet and the composite learning factor method. They divided the model's layers into three groups: frozen layers with a learning factor of zero, middle layers with a learning factor of one, and

new layers with a learning factor of ten. The results achieved were an accuracy of 96.3%, a specificity of 96.25%, and a sensitivity of 96.35%. The aforementioned study by Berrimi et al. [51] also trained their models to detect COVID-19 from CT scans. The best result was achieved with the modified DenseNet model, which attained an accuracy of 95.98%. In [64], low-dose and ultra-low-dose scan protocols, along with a two-stage CapsNet [65], were used, achieving an accuracy of 94.1%. Zhang et al. [66, 66] proposed 7-layer and 5-layer CNNs, with the achieved accuracies for the two-class classification task being 94.03% and 93.64%, respectively. In [67], the authors employed a publicly accessible dataset with 1252 COVID-19 and 1230 non-COVID chest CT images, utilizing two pre-trained models (MobileNetV2 and DarkNet19) and a newly designed lightweight model. Using a repeated ten-fold holdout validation method, they achieved the highest classification accuracy of 98.91% with transfer-learned DarkNet19.

Table 2.1 summarizes some of the key research in COVID-19 diagnosis and detection, highlighting the employed architecture, the image type, the number of classes, and the achieved score.

Table 2.1 Summary of related works on COVID-19 diagnosis

| Reference               | Architecture                                                      | Image type | #Class | Score                                                                                                                       |
|-------------------------|-------------------------------------------------------------------|------------|--------|-----------------------------------------------------------------------------------------------------------------------------|
| Karar et al. [43]       | Cascaded deep learning classifiers (VGG16, ResNet50, DenseNet169) | X-rays     | 4      | Accuracy of 99.9%                                                                                                           |
| Narin et al. [44]       | ResNet                                                            | X-rays     | 2      | Accuracies of 96.1, 99.5, and 99.7% for COVID-19 vs. normal, COVID-19 vs. viral disease, and COVID-19 vs. bacterial disease |
| Jabber et al. [45]      | MobileNet                                                         | X-rays     | 2      | Accuracy of 98.6%                                                                                                           |
| Castiglioni et al. [46] | ResNet                                                            | X-rays     | 2      | ROC-AUC of 0.80                                                                                                             |



Table 2.1 Summary of related works on COVID-19 diagnosis

| Reference                  | Architecture         | Image type | #Class | Score                                                                                                                                                                                |
|----------------------------|----------------------|------------|--------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Apostolopoulos et al. [68] | VGG19<br>MobileNet   | X-rays     | 3      | <b>VGG19:</b><br>Accuracies of 98.75%, and 93.48% for 2-class and 3-class respectively<br><b>MobileNet:</b><br>Accuracies of 97.40%, and 93.85% for 2-class and 3-class respectively |
| Berrimi et al [51]         | InceptionV3          | X-rays     | 3      | Accuracy of 92.35% by InceptionV3 in X-Rays                                                                                                                                          |
|                            | Modified DenseNet    | CT scans   | 2      | Accuracy of 95.98% by the modified DenseNet in CT scans                                                                                                                              |
| Marques et al. [54]        | EfficientNet         | X-rays     | 3      | Accuracy of 96.70%                                                                                                                                                                   |
| Ozturk et al. [55]         | DarkCovidNet (CNN)   | X-rays     | 2      | <b>2-class score:</b><br>Accuracy of 98.08%<br>Sensitivity of 95.13%<br>Precision of 98.03%                                                                                          |
|                            |                      |            | 3      | <b>3-class score:</b><br>Accuracy of 87.02%<br>Sensitivity of 92.18%<br>Precision of 89.96%                                                                                          |
| Wang et al. [57]           | COVID-Net (CNN)      | X-rays     | 3      | Accuracy of 93.30%                                                                                                                                                                   |
| Ismael et al. [58]         | SVM + CNN            | X-rays     | 2      | Accuracy of 90.58%<br>Sensitivity of 88.00%<br>Specificity of 93.30%                                                                                                                 |
| Chowdhury et al. [59]      | Parallel-dilated CNN | X-rays     | 4      | Accuracy of 96.58%                                                                                                                                                                   |
| Yu et al. [62]             | GoogleNet            | CT scans   | 2      | Accuracy of 87.5%<br>Sensitivity of 90.9%<br>Specificity of 84%                                                                                                                      |
| Zhang et al. [63]          | DenseNet-OTLS        | CT scans   | 2      | Accuracy of 96.30<br>Sensitivity of 96.35<br>Specificity of 96.25<br>Precision of 96.29                                                                                              |
| Afshar et al. [64]         | CapsNet              | CT scans   | 3      | Accuracy of 94.1%                                                                                                                                                                    |

Table 2.1 Summary of related works on COVID-19 diagnosis

| Reference             | Architecture | Image type | #Class | Score                                                                |
|-----------------------|--------------|------------|--------|----------------------------------------------------------------------|
| Zhang et al. [69, 66] | 7-layer CNN  | CT scans   | 2      | Accuracy of 94.03%                                                   |
|                       | 5-layer CNN  |            |        | Accuracy of 93.64%                                                   |
| Gupta et al. [67]     | DarkNet19    | CT scans   | 2      | Accuracy of 98.91%<br>Sensitivity of 98.96%<br>Specificity of 98.86% |

## 2.5.2 Retinal disease diagnosis and detection

### 2.5.2.1 Retinal disease classification

#### a. Traditional CNN and transfer learning approaches

Many research adopt traditional CNNs and transfer learning techniques to retinal disease diagnosis. These methods leveraged pre-trained models and fine-tuned them on medical imaging datasets to improve classification performance. For instance, Treder et al. [70] investigated using the pre-trained InceptionV3 model [22], to detect AMD in cross-sectional OCT B-scans. The model demonstrated a sensitivity of 100%, specificity of 92.0%, and accuracy of 96.0%.

Rong et al. [71] proposed a surrogate-assisted CNN with image denoising, thresholding, and morphological dilation for mask extraction to generate surrogate images for CNN training. This approach achieved AUROC values of 97.8% and 98.6% for age-related macular degeneration (AMD) detection in internal validation and external testing datasets, respectively.

Shenghua He et al. [72] proposed a deep CNN for image classification, consisting of nine convolutional blocks and two fully connected layers. The CNN was trained and evaluated on a dataset of 269 OCT images and achieved an average prediction accuracy of 86.6%, although the test set accuracy was not reported.

In a study by Kermany et al. [73], the authors demonstrated that DL-based analysis of OCT B-scans could accurately differentiate images with CNV and DME from images with DRUSEN and normal tissue. They used the pre-trained InceptionV3 [14] model, which achieved an AUC value of 99.9%, a sensitivity of 97.8%, a specificity of 97.4%, and an accuracy of 96.6%. Additionally, a limited model trained with only 1,000 cross-sectional OCT images performed similarly to a model trained on a larger dataset of 108,312 images (AUC 0.988, sensitivity 96.6%, specificity 94.0%, and accuracy 93.4%). The researchers also made the OCT dataset with ground truth labels available for further research. Li et al. [74]

improved upon previous work by fine-tuning a VGG-16 network [21], resulting in slightly better performance with an AUC value of 1, a sensitivity of 97.8%, a specificity of 99.4%, and an accuracy of 98.6%.

Motozawa et al. [75] proposed two CNN models to classify Wet AMD, Dry AMD, and healthy tissue, from OCT 2D B-scans. The models achieved AUROC values of 0.995 and 0.991, with sensitivities of 100% and 98.4%, specificities of 98.8% and 91.1%, and accuracies of 99.0% and 93.3%, respectively.

Frank Y. Shih et al. [76] proposed a DNN approach for the four types of retinal disease classification based on the VGG16 pre-trained network model. Given the limited size of the training dataset, the researchers employed transfer learning and data augmentation techniques to analyze the classification results on retinal images. They evaluated various classifiers with varying parameters and different types of data in training the network architecture. They found that the fine-tuned models performed better than CNNs. The highest accuracy achieved during the test phase was 99.48%.

Berrimi et al. [77] proposed and evaluated a deep CNN architecture for a classification task. The proposed CNN consisted of three convolutional layers with dropout and batch normalization layers. The CNN achieved an accuracy of 98.65%, a recall of 98.50%, a precision of 98.75%, and an F1-score of 98.50%.

Table 2.2 summarizes some of the key research that used traditional CNN and transfer learning approaches in retinal disease classification, highlighting the employed architecture, the dataset, the number of diseases, and the achieved score.

## **b. Advanced CNN approaches**

As the field progressed, more sophisticated CNN architectures were developed to address the complexities of retinal disease classification. These advanced models incorporated innovations such as deeper networks, specialized layers, and ensemble methods to enhance accuracy and robustness.

For instance, Karthik et al. [80] introduced a contemporary diagnostic system for classifying OCT images encompassing eight distinct diseases. In their approach, they replaced the residual connection within three ResNet architectures with the EdgeEn block and cross-activation mechanism to enhance the contrast of derivatives and generate more pronounced features. This modification effectively contributed to improved classification accuracy. Notably, the ResNet50 model achieved the most favorable results, exhibiting an accuracy of 90.30%, a recall of 91.00%, a precision of 91.00%, and an F1-score of 91.00%.

In another work [81], the authors proposed an algorithm for classifying retinal diseases in OCT images. The algorithm uses an ensemble of four improved ResNet50 [19] models to

Table 2.2 Summary of related works on retinal disease classification using traditional CNN and transfer learning approaches

| Reference               | Architecture                       | Dataset                | Diseases                       | Score                                                                                    |
|-------------------------|------------------------------------|------------------------|--------------------------------|------------------------------------------------------------------------------------------|
| Rong et al. [71]        | CNN                                | Private Dataset        | AMD<br>Normal                  | AUC of 98.6%                                                                             |
|                         |                                    | Duke OCT Dataset [78]  | DME<br>Normal                  | AUC of 97.8%                                                                             |
| Treder et al. [70]      | InceptionV3                        | Private dataset        | AMD<br>Normal                  | Accuracy of 96.0%<br>Sensitivity of 100%<br>Specificity of 92.0%                         |
| Li et al. [74]          | VGG16                              | Mendeley OCT Data [79] | CNV<br>DME<br>DRUSEN<br>Normal | Accuracy of 98.6%<br>Sensitivity of 97.6%<br>Specificity of 99.4%                        |
| Shenghua He et al. [72] | Deep CNN                           | Private dataset        | CPC<br>Normal                  | Accuracy of 86.6%                                                                        |
| Kermany et al. [73]     | InceptionV3                        | Mendeley OCT Data [79] | CNV<br>DME<br>DRUSEN<br>Normal | Accuracy of 96.6%<br>Sensitivity of 97.8%<br>Specificity of 97.4%                        |
| Motozawa et al. [75]    | 18 layer-CNN<br>(2 configurations) | Private dataset        | Dry AMD<br>Wet AMD<br>Normal   | Accuracy of 99.0%, 93.3%<br>Sensitivity of 100%, 98.4%<br>Specificity of 98.8%, 91.1%    |
| Berrimi et al [77]      | CNN                                | Mendeley OCT Data [79] | CNV<br>DME<br>DRUSEN<br>Normal | Accuracy of 98.65%<br>Sensitivity of 98.50%<br>Precision of 98.75%<br>F1-score of 98.50% |

classify OCT images into four classes: CNV, DME, Drusen, and Normal. The algorithm was evaluated using a patient-level 10-fold cross-validation process on a development dataset, achieving a high accuracy of 97.3%, sensitivity of 96.3%, and specificity of 98.5%. The study also discussed using occlusion testing and integrating medical history data to improve model performance.

Gokhan Altan [82] proposed a lightweight, explainable CNN architecture called DeepOCT for identifying macular edema on OCT images. The model focused on achieving high classification performance while utilizing pre-trained architectures and shallow dense layers. DeepOCT also aimed to visualize feature activation maps' most relevant regions and pathologies. The model incorporated the block-matching and 3D filtering algorithm (BM3D), flattened the retinal layers to mitigate the effects of different macula positions, and excluded non-retinal layers through cropping. DeepOCT achieved accuracy, sensitivity, and specificity rates of 99.20%, 99.6%, and 98.40%, respectively, for identifying OCT images with macular edema.

Table 2.3 summarizes some of the key research that used advanced CNN approaches in retinal disease classification.

Table 2.3 Summary of related works on retinal disease classification using advanced CNN approaches

| Reference           | Architecture                                       | Dataset                | Diseases                                                 | Score                                                                |
|---------------------|----------------------------------------------------|------------------------|----------------------------------------------------------|----------------------------------------------------------------------|
| Karthik et al.[80]  | ResNet + EdgeEn block + cross-activation mechanism | Mendeley OCT Data [79] | CNV<br>DME<br>DRUSEN<br>Normal                           | Accuracy of 98.20%<br>Recall of 98.00%                               |
|                     |                                                    | OCT-C8 Dataset [83]    | AMD<br>CNV<br>CSR<br>DME<br>DR<br>DRUSEN<br>MH<br>Normal | Accuracy of 90.30%<br>Recall of 91.00%                               |
| Feng Li et al. [81] | Ensemble-ResNet50                                  | Mendeley OCT Data [79] | CNV<br>DME<br>DRUSEN<br>Normal                           | Accuracy of 97.3%<br>Sensitivity of 96.3%<br>Specificity of 98.5%    |
| Gokhan Altan. [82]  | Explainable CNN (DeepOCT)                          | Mendeley OCT Data [79] | CNV<br>DME<br>DRUSEN<br>Normal                           | Accuracy of 99.20%<br>Sensitivity of 99.60%<br>Specificity of 98.40% |
|                     |                                                    | SERI Dataset [84]      | DME<br>Normal                                            | Accuracy of 99.12%<br>Sensitivity of 98.58%<br>Specificity of 99.66% |

### c. Emerging transformer-based approaches

The introduction of Transformer architectures in medical image analysis brought a new wave of innovation, particularly in retinal disease diagnosis. Researchers began integrating ViT with CNNs to combine the strengths of both approaches, leading to hybrid models that offer superior performance in OCT image classification. For instance, Junyong Shen et al. [85] proposed a Structure-Oriented Transformer (SoT) framework for retinal disease grading, utilizing a self-attention mechanism to capture the relationship between lesions and the retina. This model outperformed SOTA methods in grading neovascular AMD and demonstrated superior classification and generalizability on a publicly available retinal disease dataset.

Another study by [86] introduced the Hybrid ConvNet-Transformer network (HCTNet), incorporating a low-level feature extraction module based on the residual dense block for initial feature generation. Two parallel branches, utilizing Transformer and Convolution architectures, capture global and local context within OCT images. A feature fusion block, employing an adaptive re-weighting mechanism, combines these global and local features. HCTNet achieved superior performance, surpassing pure ViT and several ConvNet-based

classification methods. Validation on two public retinal OCT datasets demonstrated HCT-Net's overall accuracies of 91.56% and 86.18%, respectively.

P. Dutta et al. [87] proposed Conv-ViT, a hybrid ViT-CNN model. Conv-ViT leverages the capabilities of transfer learning-based CNN models, specifically Inception-V3 and ResNet-50, for extracting texture information, while a ViT captures shape-based features. This hybrid approach facilitates robust learning of shape-based texture features, achieving an accuracy of approximately 94% in classifying retinal diseases into four distinct classes.

In a distinct study, Huajie Wen et al. [88] employed a hybrid model called the Lesion-Localization Convolution Transformer (LLCT) to enhance the classification accuracy, sensitivity, and specificity of OCT images in ophthalmic disease diagnosis. This architecture integrates convolution and self-attention, utilizing custom feature maps from a CNN as input sequences for the self-attention network. The model achieved notable performance metrics, with an overall accuracy of approximately 97.7%, a sensitivity of 97.7%, and a specificity of 99.2%.

In [89], the authors proposed three architectures ViT, T2T (Tokens-To-Token) ViT, and Mobile-ViT, to detect CNV, Drusen, DME, and normal using OCT images from a public dataset. The predictive accuracies of ViT, T2T-ViT, and Mobile-ViT achieved on the dataset for the classification task were 95.14%, 96.07%, and 99.17% respectively.

Table 2.4 summarizes some of the key research in retinal disease classification using emerging Transformer-based approaches.

### 2.5.2.2 Retinal fluid segmentation

Retinal segmentation has been the focus of extensive research, with numerous methodologies proposed to enhance the accuracy and efficiency of detecting retinal diseases at early stages. Among these approaches, CNNs and their various adaptations have emerged as the most widely adopted techniques. Researchers have leveraged the powerful feature extraction capabilities of CNNs to address the complexities of retinal imaging, resulting in significant advancements in the field.

#### a. Supervised approaches

Supervised approaches in retinal fluid segmentation rely on fully labeled datasets, where each pixel of an image is annotated with the correct class. These methods have been pivotal in advancing the accuracy of retinal fluid detection, particularly through the use of CNNs and their various adaptations. For instance, Schlegl et al. [90] developed an encoder-decoder CNN architecture to segment macular fluid. Utilizing a dataset of 1200 labeled OCT images,

Table 2.4 Summary of related works on retinal disease classification using emerging transformer-based approaches

| Reference                | Architecture                   | Dataset                | Diseases                                                 | Score                                                                                       |
|--------------------------|--------------------------------|------------------------|----------------------------------------------------------|---------------------------------------------------------------------------------------------|
| Junyong Shen et al. [85] | Structure-Oriented Transformer | nAMD Private dataset   | AMD type 1<br>AMD type 2                                 | Accuracy of 88.42%<br>Sensitivity of 88.41%<br>Specificity of 88.42%                        |
|                          |                                | Duke OCT Dataset [78]  | AMD<br>DME<br>Normal                                     | Accuracy of 99.35%<br>Sensitivity of 99.25%<br>Specificity of 99.55%                        |
| Ma et al. [86]           | CNN-Transformer (HCTNet)       | Mendeley OCT Data [79] | CNV<br>DME<br>DRUSEN<br>Normal                           | Accuracy of 91.56%<br>Sensitivity of 88.57%<br>Precision of 88.11%                          |
|                          |                                | Duke OCT Dataset [78]  | AMD<br>DME<br>Normal                                     | Accuracy of 86.18%<br>Sensitivity of 85.40%<br>Precision of 88.53%                          |
| P. Dutta et al. [87]     | CNN-Transformer (Conv-ViT)     | Mendeley OCT Data [79] | CNV<br>DME<br>DRUSEN<br>Normal                           | Accuracy of 94.46%<br>Sensitivity of 94.0%<br>Precision of 94.0%                            |
| Huajie Wen et al. [88]   | CNN-Transformer (LLCT)         | Mendeley OCT Data [79] | CNV<br>DME<br>DRUSEN<br>Normal                           | Accuracy of 97.7%<br>Sensitivity of 97.7%<br>Precision of 97.8%<br>F1-Score of 97.75%       |
| Akça et al. [89]         | ViT<br>T2T ViT<br>Mobile-ViT   | Mendeley OCT Data [79] | CNV<br>DME<br>DRUSEN<br>Normal                           | ViT Acc: 95.14%<br>T2T-ViT Acc: 96.07%<br>Mobile-ViT Acc: 99.17%                            |
| Laouarem et al. [3]      | HTC-Retina                     | Duke OCT Dataset [78]  | AMD<br>DME<br>Normal                                     | Accuracy of 99.77%<br>Sensitivity of 99.77%<br>Precision of 99.77%<br>Specificity of 99.80% |
|                          |                                | Mendeley OCT Data [79] | CNV<br>DME<br>DRUSEN<br>Normal                           | Accuracy of 99.4%<br>Sensitivity of 99.41%<br>Precision of 99.39%<br>Specificity of 99.7%   |
|                          |                                | OCT-C8 Dataset [83]    | AMD<br>CNV<br>CSR<br>DME<br>DR<br>DRUSEN<br>MH<br>Normal | Accuracy of 97.00%<br>Sensitivity of 97.00%<br>Precision of 97.04%<br>Specificity of 99.56% |

this architecture achieved a precision of 91% and a recall of 84% for intraretinal fluid segmentation. However, for subretinal fluid segmentation, it attained a precision of 61% and a recall of 81%.

In a study by [91], the authors proposed an encoder-decoder CNN architecture called DeepRetina to segment retinal layers. This network utilizes a modified Xception architecture

as the encoder backbone. DeepRetina achieved a 90% IoU and a 92% sensitivity score on a publicly available dataset.

In another study [92], the authors introduced a modified U-Net [25] architecture-based segmentation method to segment macular fluid in OCT images. Trained on 1,289 annotated OCT scans, this architecture achieved a 91.1% Dice coefficient score.

Chen et al. [93] proposed a deep network embedded with rough fuzzy discretization (RFDDN). They used fuzzy c-means clustering to determine the membership degrees of pixels for each segmentation region and designed an individual fitness function based on rough fuzzy set theory. A genetic algorithm was employed to find the optimal breakpoints for discretizing the features of OCT fundus images. The authors incorporated this feature discretization as a pre-module of the deep neural network and introduced a deep supervised attention mechanism to extract important multi-scale information. RFDDN achieved a 95% Dice coefficient score, 97% sensitivity, and 89% specificity.

In [94], the authors developed a DL-based semantic segmentation method to analyze lesion tissues in vessel OCT images, aiming to ease the workload for cardiovascular specialists. They segmented the region of interest, extracted patches for input, and used dense blocks with skip connections in their model. Their dense-block-SegNet (DBSegNet) was trained and tested on multiple datasets and achieved sensitivities of 91.81% for calcified, 92.81% for fibrous, and 91.78% for lipid plaques.

Venkatesh et al. [95] conducted a retrospective, cross-sectional study to analyze spectral domain OCT (SDOCT) findings in type two macular telangiectasia (MacTel) and correlate them with clinical stages and visual acuity. They examined 212 eyes from 108 patients using SDOCT imaging. Key findings included hyperreflective middle retinal layers, inner retinal cavities, and retinal pigment clumps. The study found specific SD-OCT features associated with poor vision in both non-proliferative and proliferative groups.

Ahmed et al. [96] developed an automated method for detecting and quantifying retinal fluid in OCT scans. They first preprocessed the images to reduce noise and artifacts. Then, they trained a U-net DL model to identify and segment fluid regions and quantify cystoid macular edema and its progression. The method was evaluated against manual segmentation by experts and showed high correlation and superior performance. The model achieved an accuracy of 99.81%, a DSC of 83.34%, a precision of 86.56%, a recall of 80.89%, and an AUC of 99.48%.

The work reported in [97], presented a DL algorithm designed for the automatic segmentation and quantification of intraretinal cystoid fluid in SD-OCT volumes. This proposed algorithm employs a cascade of two FCNNs, each tasked with distinct but complementary objectives. The first FCNN is responsible for delineating the entire retinal area, serving as



a critical preprocessing step. The output of this network is then seamlessly integrated into the optimization process of the second FCNN, which focuses on the precise segmentation of intraretinal cystoid fluid (IRC). This approach achieved an overall dice coefficient of 75.4%. Additionally, the method yielded an intraclass correlation coefficient of 93.6% in the quantification of IRC.

Roy et al. [98] introduced ReLayNet, a fully convolutional deep architecture specifically designed for end-to-end segmentation of retinal layers and fluid masses in OCT images. ReLayNet employs a contracting path of convolutional blocks to extract a hierarchical set of contextual features, which are crucial for accurate segmentation. This is followed by an expansive path of convolutional blocks that perform semantic segmentation with precision. The model's training was guided by a joint loss function, combining weighted logistic regression and dice overlap loss, to enhance segmentation performance. Notably, ReLayNet achieved a Dice coefficient score of 77% in fluid mass segmentation.

## **b. Semi-Supervised approaches**

Other research uses semi-supervised approaches due to the difficulties of obtaining labeled data and ground truth samples. Semi-supervised approaches address the challenge of limited labeled data in retinal segmentation by combining a small amount of labeled data with a larger set of unlabeled images. These methods aim to enhance segmentation performance while reducing the reliance on extensive manual annotation.

Liu et al. [99] introduced a self-ensembling semi-supervised method for retinal fluid segmentation. This approach includes one encoder and three decoders within a modified U-Net architecture. The three decoders generate an output probability map, contour map, and distance map, which are combined to produce the final segmentation map, effectively distinguishing close fluid regions. Additionally, the method incorporates the Selective Kernel Module (SKM) [100], which adjusts receptive fields according to fluid scale. The training process utilizes weighted dice, contour, regression, and consistency loss functions. This method achieved a 76% mean DSC on the RETOUCH dataset.

In [101], the authors introduced a modified U-Net architecture with enlarged kernel size and an expanded receptive field, designed to effectively leverage both labeled and unlabeled data for OCT image segmentation. The segmentation model was trained using a semi-supervised, adversarial approach. This approach named SGNet which is a GAN-based automated classifier was developed and evaluated for its ability to detect progressive retinal disorders and fluid regions. The results demonstrated that the model exhibited exceptional generalizability, particularly in scenarios with limited availability of labeled data.

To summarize the related research on retinal fluid segmentation, Table 2.5 provides a comprehensive overview, highlighting key aspects such as the learning approach, the employed architecture, the dataset utilized, and the performance metrics achieved in these studies.

Table 2.5 Summary of related works on retinal fluid segmentation

| Reference              | Network type     | Architecture                                 | Dataset                           | Score                                                                                                              |
|------------------------|------------------|----------------------------------------------|-----------------------------------|--------------------------------------------------------------------------------------------------------------------|
| Schlegl et al. [90]    | Fully supervised | CNN with encoder-decoder                     | Private Dataset                   | Intraretinal fluid:<br>Precision of 91%<br>Recall of 84%<br>Subretinal fluid:<br>Precision of 61%<br>Recall of 81% |
| Lee et al. [92]        | Fully supervised | Modified U-Net Framework                     | Private Dataset                   | DSC of 91.1%                                                                                                       |
| Chen et al. [93]       | Fully supervised | Deep network with rough fuzzy discretization | Private Dataset                   | DSC of 95.0%<br>Sensitivity of 97.0%<br>Specificity of 89.0%                                                       |
| Ren et al. [94]        | Fully supervised | DL with level-set method                     | Private Dataset                   | Sensitivities of 91.81% for calcified, 92.81% for fibrous, and 91.78% for lipid plaques                            |
| Venkatesh et al. [102] | Fully supervised | CNN with 2 MacTel structure                  | Private Dataset                   | Based on Correlation                                                                                               |
| Liu et al. [99]        | Semi-supervised  | Encoder and 3 decoders with modified Unet    | RETOUCH dataset                   | DSC of 76%                                                                                                         |
| Ahmed et al. [96]      | Fully supervised | U-Net                                        | Extracted from Mendeley Data [79] | Accuracy of 99.8%<br>DSC of 83.3%<br>Precision of 86.6%<br>Recall of 80.9%<br>AUC of 99.5%                         |
| Freerk et al. [97]     | Fully supervised | FCNN                                         | EUGENDA dataset                   | DSC of 75.4%                                                                                                       |
| Roy et al. [98]        | Fully supervised | ReLayNet (CNN)                               | Duke OCT dataset [78]             | DSC of 77%                                                                                                         |
| Liu et al. [101]       | Semi-supervised  | SGNet (GAN approach)                         | Duke OCT dataset [78]             | DSC of 80%                                                                                                         |

### 2.5.3 Skin lesion diagnosis and detection

Skin lesion diagnosis holds significant importance in healthcare. Skin cancer, a prevalent form of malignancy in humans, is primarily diagnosed visually by clinical experts. This process typically commences with a primary clinical screening, followed by dermoscopic assessment, biopsy, and histopathological testing. The advancement of medical imaging led to the development of DL models and CAD systems for automated skin lesion diagnosis and detection. Indeed, numerous researchers have employed conventional networks and DL methods for skin lesion classification and segmentation.

### 2.5.3.1 Skin lesion classification and identification

#### a. Traditional CNN

Traditional CNN approaches have laid the foundation for skin lesion analysis, leveraging their strong feature extraction capabilities to effectively identify and classify skin conditions. For instance, In [103], the authors proposed a Deep CNN model optimized for efficiency and performance. Using dermoscopic images from ISIC databases, The DCNN achieved a precision of 94%, sensitivity of 93%, and specificity of 91% in the ISIC-17 dataset.

Jinnai et al. [103] trained a faster, region-based CNN (FRCNN) [104] and compared its performance with ten dermatologists. FRCNN achieved 86.2% accuracy for 6-class classification, and 91.5% accuracy for binary classification, surpassing dermatologists in accuracy, sensitivity, and specificity.

In [105], the authors introduced a method using Mask and Region-based CNN [106] techniques for automatically cropping the region of interest in dermoscopic images, followed by lesion classification using a ResNet152 structure. The evaluation showed significant improvements in accuracy and balanced accuracy compared to previous models. Notably, the proposed model achieved an accuracy of 90.4%, sensitivity of 82.0%, specificity of 92.5%, and an AUC of 0.872, indicating its capability for accurate discrimination between benign and malignant lesions without bias.

In [107], the authors proposed a CNN model and focused on exploring different nonlinear activation functions. Experimental analysis of PH2 and ISIC datasets demonstrates that a CNN model with parameterized Leaky ReLU [109] function outperforms alternative networks, achieving 97.50% accuracy, 98.00% precision, and 98.00% sensitivity in classifying skin lesions into three classes.

Shetty et al. [108] developed a CNN architecture and employed data augmentation techniques on the HAM10000 [101] dataset to enhance the model's learning capabilities. The performance of the proposed model was benchmarked against traditional ML algorithms, demonstrating superior results. The model achieved an accuracy of 94%, with a precision of 88% and a recall of 85%.

Azeem et al. [109] developed SkinLesNet, a four-layer CNN architecture. The model was trained and evaluated on three datasets: an augmented PAD-UFES-20, HAM10000, and ISIC 2017, and compared to pre-trained ResNet50 and VGG16 models. SkinLesNet achieved accuracies of 96%, 90%, and 92%, and sensitivities of 92%, 87%, and 82% on the respective datasets.

Table 2.6 offers a comprehensive overview of related research that utilized traditional CNN architectures for skin lesion classification.

Table 2.6 Summary of related works on skin lesion classification using traditional CNN.

| Reference           | Architecture                            | Dataset         | #Class | Score                                                                                                        |
|---------------------|-----------------------------------------|-----------------|--------|--------------------------------------------------------------------------------------------------------------|
| Imran et al. [110]  | Deep CNN                                | ISIC 2019       | 8      | Accuracy of 89.6%<br>Precision of 90.7%<br>Sensitivity of 89.6%<br>Specificity of 97.6%<br>AUROC of 0.991    |
| Jinnai et al. [103] | Faster, region-based CNN                | Private dataset | 6      | Binay-accuracy of 91.5%<br>Accuracy of 86.2%<br>Sensitivity of 83.3%<br>Specificity of 94.5%                 |
| Acosta et al. [105] | Mask R-CNN + ResNet152                  | ISIC 2017       | 3      | Accuracy of 90.4%<br>Sensitivity of 82.0%<br>Specificity of 92.5%<br>AUC of 0.872                            |
| Rasel et al. [107]  | CNN model with parameterized Leaky ReLU | PH2             | 3      | Accuracy of 97.5%<br>Sensitivity of 98.0%<br>Specificity of 98.7%<br>Precision of 98.0%<br>F1-Score of 98.0% |
|                     |                                         | ISIC archive    | 3      | Accuracy of 98.6%<br>Sensitivity of 98.3%<br>Specificity of 99.3%<br>Precision of 98.3%<br>F1-Score of 98.0% |
| Shetty et al. [108] | CNN                                     | HAM10000        | 7      | Accuracy of 94%<br>Precision of 88%<br>Recall of 85%<br>F1-Score of 86%                                      |
| Azeem et al. [109]  | Multi-layer CNN                         | HAM10000        | 7      | Accuracy of 90%<br>Precision of 89%<br>Sensitivity of 87%<br>F1-Score of 85%                                 |
|                     |                                         | PAD-UFES-20     | 3      | Accuracy of 96%<br>Precision of 97%<br>Sensitivity of 92%<br>F1-Score of 92%                                 |
|                     |                                         | ISIC 2017       | 3      | Accuracy of 92%<br>Precision of 80%<br>Sensitivity of 82%<br>F1-Score of 75%                                 |

### b. Lightweight approaches

Lightweight architectures have gained traction in skin lesion analysis due to their efficiency and suitability for deployment on resource-constrained devices. These approaches focus on reducing model complexity and size while maintaining high performance, making them ideal for real-time applications and mobile health solutions. As an example, Hoang et al. [111] introduced a novel skin lesion segmentation and classification approach utilizing a

wide-ShuffleNet architecture [112]. In this approach, they employed entropy-based weighted sum first-order cumulative moment (EW-FCM) computation to isolate the lesion region from the background. The segmented lesion images were then passed through a wide ShuffleNet model for the classification task.

Similarly, Srinivasu et al. [113] proposed an alternative approach for lesion classification using the HAM10000 dataset [114], which combined the strengths of MobileNet and Long Short-Term Memory (LSTM) networks. The selection of MobileNet was motivated by its computational efficiency, reduced model size, and suitability for deployment on mobile devices. The LSTM network played a crucial role in capturing temporal information associated with the features extracted by MobileNetV2. The authors reported that the LSTM integration with MobileNetV2 significantly enhanced the model's overall classification performance.

Oztel et al. [115] developed a DL model intended for use in a mobile application. First, they trained various small-sized pre-trained networks for skin lesion image classification. Then, the most suitable network in terms of both performance and mobile compatibility was selected and transformed into the TensorFlow Lite format. The proposed system achieved a classification accuracy of 74.27% for seven classes on a combined dataset.

In [116], the authors introduced an end-to-end framework for classifying seven types of skin lesions. They developed two algorithms for image segmentation and feature extraction using modified Gaussian Filtering and Segmentation-based Fractal and designed a lightweight deep CNN called S-MobileNet, experimenting with different activation functions (ReLU, Mish) and compression techniques to optimize performance. They trained and evaluated their model on the HAM10000 dataset, achieving improved results when using their proposed preprocessing techniques, Mish activation, and a compressed S-MobileNet architecture.

Table 2.7 offers a comprehensive overview of related research that utilized lightweight architectures for skin lesion classification.

### **c. Advanced approaches**

Advanced CNN techniques represent a significant evolution in the field of skin lesion analysis, incorporating sophisticated methods to enhance accuracy, robustness, and generalization capabilities. These techniques often involve complex architectures, novel training strategies, and the integration of additional computational mechanisms such as attention modules or multi-scale processing. For instance, Xiaoyu et al. [118] introduced deep metric attention learning CNN (DeMAL-CNN) that utilized a triplet-based network for deep metric learning to optimize embeddings, expanding the training sample size and enhancing robustness, in addition to the use of mixed attention mechanism to improve skin lesion localization ability.

Table 2.7 Summary of related works on skin lesion classification using lightweight architectures.

| Reference              | Architecture                           | Dataset            | #Class | Score                                                                                |
|------------------------|----------------------------------------|--------------------|--------|--------------------------------------------------------------------------------------|
| Hoang et al. [111]     | EW-FCM + wide-ShuffleNet               | HAM10000           | 7      | Accuracy of 86.33%<br>Sensitivity of 86.33%<br>Specificity of 97.72%                 |
|                        |                                        | ISIC 2019          | 8      | Accuracy of 82.56%<br>Sensitivity of 82.56%<br>Specificity of 97.51%                 |
| Srinivasu et al. [113] | MobileNet + LSTM                       | HAM10000           | 7      | Accuracy of 85.34%<br>Sensitivity of 92%<br>Specificity of 88.24%                    |
| Oztel et al. [115]     | ResNet-18                              | PAD-UFES-20        | 6      | Accuracy of 74.6%<br>Sensitivity of 62.8%<br>Precision of 75.2%<br>F1-Score of 65.9% |
|                        |                                        | MSLD + PAD-UFES-20 | 7      | Accuracy of 74.3%<br>Sensitivity of 71.4%<br>Precision of 76.8%<br>F1-Score of 73.6% |
| Razia et al. [116]     | S-Mobile Net                           | HAM10000           | 7      | Accuracy of 98.15%<br>Precision of 96.23%<br>F1-Score of 94.59%                      |
| Laouarem et al. [117]  | JILDYA-Net<br>(FNet + Attention + CNN) | HAM10000           | 7      | Accuracy of 98.80%<br>Sensitivity of 95.87%<br>Specificity of 99.30%                 |
|                        |                                        | ISIC 2019          | 8      | Accuracy of 98.72%<br>Sensitivity of 94.92%<br>Specificity of 99.27%                 |

By combining triplet loss with classification loss during training, DeMAL-CNN achieved an accuracy of 87.3%, sensitivity of 65.3%, and specificity of 94.7%.

In [119], the authors introduced G-DMN, a two-stage framework for skin lesion classification. G-DMN combines CycleGAN [120] for dataset expansion and Dense-MobileNetV2 (DMN) for classification. The first stage utilizes the CycleGAN to augment the dataset using a novel image pairing strategy. The second stage utilizes DMN, a lightweight model derived from MobileNetV2, enhancing feature reuse and focal area focus. Tested on the HAM10000 dataset [101], G-DMN achieved 87.07% accuracy, 80.13% precision, 75.28% sensitivity, 96.19% specificity, 77.26% F1-Score, and 0.952 AUC, with only 5.33M parameters, demonstrating superior performance with fewer parameters compared to classical models.

In another study by Wang et al. [121], a Double Branch Network (DBN), was developed using a two-branch network model with a common backbone structure. Their approach incorporated the CFEBlock for common feature extraction and FusionBlock for combining features between branches. They introduced the CSLI dataset, combining publicly available data with newly collected images, containing 3361 clinical dermatology images across six disease categories. The comprehensive evaluations showed strong performance across various diseases on the test data.

Alsahafi et al. [122] introduced a novel method by using the power of ResNet architecture. Instead of employing horizontal sliding filters, they used multiple convolution filters for multi-layer feature extraction and incorporated a unique approach for cross-channel correlation through sliding dot product filters. To mitigate the imbalanced dataset challenges, the authors implemented a novel strategy: transforming the dataset into a vector encompassing both image data and corresponding weights. The evaluation was conducted on the ISIC-2019 [123], and ISIC-2020 [124] datasets, demonstrating superior performance to existing deep CNNs in the skin lesions multiclass classification.

Qasim et al. [125] explored the application of deep spiking neural networks [126] for skin lesion classification, specifically melanoma detection. Leveraging the surrogate gradient descent method, the researchers trained a spiking VGG-13 model on a subset of the ISIC 2019 dataset [123], comprising 3,670 melanoma and 3,323 non-melanoma images. Their proposed architecture achieved an accuracy of 89.57% and an F1-score of 90.07%, demonstrating superior performance compared to VGG-13 and AlexNet, while notably utilizing fewer trainable parameters.

Table 2.8 offers a comprehensive overview of related research that utilized advanced CNN architectures for skin lesion classification.

### **2.5.3.2 Skin lesion segmentation and localization**

Segmentation and localization are critical in skin lesion analysis, as they enable the precise identification and delineation of lesion boundaries from surrounding healthy skin. Accurate segmentation is essential for reliable diagnosis, as it isolates the area of interest, allowing for detailed examination and classification. Effective localization aids in focusing attention on the most relevant regions, reducing the influence of irrelevant features. Many studies have been conducted to improve segmentation and localization techniques, with some researchers employing traditional CNN and SOTA segmentation models, while others explore advanced hybrid models incorporating attention mechanisms and transformers.

Table 2.8 Summary of related works on skin lesion classification using advanced approaches.

| Ref                   | Architecture                                                                                       | Dataset             | # Class | Score                                                                                                                          |
|-----------------------|----------------------------------------------------------------------------------------------------|---------------------|---------|--------------------------------------------------------------------------------------------------------------------------------|
| Xiaoyu et al. [118]   | Deep metric attention CNN                                                                          | ISIC 2016           | 2       | Accuracy of 87.3%<br>Sensitivity of 65.3%<br>Specificity of 94.7%<br>AUC of 0.858                                              |
|                       |                                                                                                    | PH2                 | 3       | Accuracy of 93.5%<br>Sensitivity of 95.0%<br>Specificity of 93.1%<br>AUC of 0.974                                              |
| Wang et al. [119]     | CycleGAN + Improved MobileNetV2                                                                    | HAM10000            | 7       | Accuracy of 87.07%<br>Precision of 80.13%<br>Sensitivity 75.28%<br>Specificity of 96.19%<br>F1-Score of 77.26%<br>AUC of 0.952 |
| Wang et al. [121]     | Double branch network                                                                              | HAM10000            | 7       | Accuracy of 91.6%<br>Precision of 85.9%<br>Sensitivity of 84.2%<br>Specificity of 97.6%<br>F1-Score of 85.0%<br>AUC of 0.985   |
|                       |                                                                                                    | Private data (CSLI) | 6       | Accuracy of 79.5%<br>Precision of 77.8%<br>Sensitivity of 75.6%<br>Specificity of 95.6%<br>F1-Score of 76.2%<br>AUC of 0.934   |
| Alsahafi et al. [122] | Deep residual network using multilevel feature extraction and cross-channel correlation (Skin-Net) | ISIC 2019           | 8       | Accuracy of 94.7%<br>Sensitivity of 70.8%<br>Specificity of 96.8%                                                              |
|                       |                                                                                                    | ISIC 2020           | 2       | Accuracy of 99.0%<br>Sensitivity of 96.6%<br>Specificity of 99.4%                                                              |
| Qasim et al. [125]    | Deep spiking neural networks                                                                       | ISIC 2019           | 8       | Accuracy of 89.57%<br>F1-Score of 90.07%                                                                                       |

### a. Traditional CNN and SOTA segmentation architectures

Maron et al. [127] demonstrated that supplementing NNs with both segmented and un-processed skin images, rather than solely relying on raw images, does not hinder overall performance while effectively mitigating the impact of irrelevant skin features. This finding signifies a crucial step towards robust and reliable segmentation in dermatology. Similarly, Zhao et al. [128] enhanced the U-Net architecture, developing a novel model for skin lesion



segmentation and introducing a new loss function. This resulted in exceptional performance, achieving an IoU of 84.73%.

In [129] The authors proposed a two-step system for automatic skin lesion segmentation. First step they remove the hairline artifacts from the image using morphological operators. The cleaned images are then processed by an encoder-decoder CNN, which performs downsampling and upsampling to generate high-resolution segmentation maps. The model employs a Tversky loss function [130] to minimize errors between predicted and target outputs. Evaluated on four publicly available datasets, including ISIC 2016, ISIC 2017, ISIC 2018, and ph2, the proposed model achieved mean IoU scores of 87.1%, 77.8%, 85.2%, and 88.8%, respectively.

In [131], an end-to-end deep edge CNN (DE-Net) was developed, which is based on an encoder-decoder structure. The DE-Net includes an edge information-guided module (EIGM) to integrate boundary details with contextual information. An entirety-center-edge (ECE) loss function was also introduced to enhance boundary accuracy without compromising overall performance. The model was evaluated on the ISIC 2017 dataset and achieved a DSC of 87.92% and an IoU of 80.53%.

Yanagisawa et al. [132] trained a CNN based on the DeepLabv3+ [133] network for skin lesion area segmentation. Their approach achieved results comparable to manual cropping, demonstrating its potential for drastically reducing the time and effort required for data annotation by clinicians.

Wang et al. [134] sought to streamline the annotation process by employing a collaborative learning deep CNN (CL-DCNN) framework. They used lesion segmentation masks to provide explicit contour information, effectively guiding the network's learning process toward refined segmentation. This method yielded promising results, achieving an IoU of 79.1%.

Table 2.9 summarizes related research in skin lesion segmentation using CNNs and SOTA segmentation models, highlighting key aspects such as the learning approach, the employed architecture, the dataset utilized, and the performance metrics achieved in these studies.

## **b. Advanced architectures and hybrid models**

Advanced architectures and hybrid models have further pushed the boundaries of skin lesion analysis by integrating sophisticated techniques to enhance segmentation and localization. These approaches often combine traditional CNNs with novel mechanisms such as attention modules and transformers, aiming to improve feature extraction, contextual understanding, and overall model performance. The subsequent works highlight the innovations and advancements achieved through these cutting-edge methods in skin lesion analysis.

Table 2.9 Summary of related works on skin lesion segmentation using traditional CNN and SOTA segmentation architectures.

| Reference          | Network type     | Architecture                                       | Dataset   | Score                 |
|--------------------|------------------|----------------------------------------------------|-----------|-----------------------|
| Maron et al. [127] | Fully supervised | U-Net                                              | ISIC 2018 | IoU of 75%            |
| Zhao et al. [128]  | Fully supervised | Enhanced U-Net++                                   | ISIC 2018 | IoU of 84.73%         |
| Kaur et al. [129]  | Fully supervised | Morphological operators + CNN with encoder-decoder | ISIC 2016 | IoU of 87.1%          |
|                    |                  |                                                    | ISIC 2017 | IoU of 77.8%          |
|                    |                  |                                                    | ISIC 2018 | IoU of 85.2%          |
|                    |                  |                                                    | PH2       | IoU of 88.8%          |
| Gu et al. [131]    | Fully supervised | End-to-end deep edge CNN                           | ISIC 2017 | Accuracy of 94.50%    |
|                    |                  |                                                    |           | DSC of 87.92%         |
|                    |                  |                                                    |           | IoU of 80.53%         |
|                    |                  |                                                    |           | Sensitivity of 88.04% |
| Wang et al. [134]  | Fully supervised | Collaborative learning deep CNN                    | ISIC 2017 | Specificity of 96.59% |
|                    |                  |                                                    |           | DSC of 86.7%          |
|                    |                  |                                                    |           | IoU of 79.1%          |
|                    |                  |                                                    |           | Sensitivity of 86.5%  |
|                    |                  |                                                    |           | Specificity of 95.9%  |

In a work by Chen et al. [135], a recurrent attention CNN (O-Net) was proposed for the segmentation of skin lesions. O-Net integrates an attention class feature module with a recurrent O-shape structure, enhancing iterative refinement of segmentation results. This design leverages attention feature maps to improve coarse-to-fine feature representation and detailed tissue extraction. The model was validated on ISIC-2017 and PH2 datasets, achieving DSC of 87.04% and 92.12%, and IoU of 80.36% and 86.15%, respectively.

Zhang et al. [136] introduced a novel Channel Attention Module (CMD) designed to fully use high- and low-level feature information, effectively bridging the semantic gap often observed between encoder and decoder components in segmentation models. This approach facilitates richer feature representations and enhances the robustness of the overall segmentation process. Their method achieved an accuracy of 95.7% and a DSC of 86.4%. This underscores the potential of incorporating attention mechanisms for enhanced feature utilization in challenging segmentation tasks.

Alahmadi and Alghamdi [137], developed a semi-supervised method to reduce the workload of dermatologists and aid in early cancer prediction. This approach addresses the challenge of limited annotated datasets and the risk of overfitting in neural networks. The proposed method incorporates unlabelled samples during training, combining supervised and unsupervised strategies to enrich the encoder module. The authors designed a surrogate task using convolutional and Transformer representations to learn data-driven features from the

images themselves. The algorithm demonstrated effectiveness on three datasets: ISIC 2017 (DSC of 89.8%), ISIC 2018 (DSC of 90.5%), and PH2 (DSC of 94.0%).

In another work [138], a model named APT-Net was presented, which is a new CNN-transformer hybrid network, based on an encoder-decoder architecture. APT-Net introduces an adaptive position encoding module to enhance positional information and a dual-path parallel decoder to efficiently process multiscale feature maps. APT-Net tested on two major skin image datasets ISIC 2017 and PH2, achieved scores of 78.3% IoU and 86.3% DSC on ISIC 2017, and even better scores of 89.9% IoU and 94.6% DSC on PH2.

In [139], the authors developed the Dynamic Spatial Group Enhanced Network (DSEUNet) to address the need for efficient and accurate skin cancer segmentation in resource-limited settings, such as mobile healthcare. DSEUNet significantly reduces computational load and parameter quantity while maintaining competitive segmentation accuracy. Extensive trials on the ISIC 2017 and ISIC 2018 datasets demonstrated that DSEUNet outperforms traditional models like UNet in both accuracy and efficiency. This research advances the development of accessible and efficient tools for medical image analysis, particularly in resource-constrained environments.

Leo et al. [140] proposed DBNet-SI, a U-shaped network, which incorporates a dual-branch module (MSI) combining shift window attention and inception structures. This module captures multiscale global contextual information and long-range pixel dependencies. The cross-branch bidirectional interaction within the MSI module enhances information complementarity. Additionally, a progressive feature enhancement and information compensation module (PFEIC) was designed to progressively enhance fine-grained features. The network was tested on the ISIC 2017 and ISIC 2018 datasets. It achieved a DSC of 92.5% and 93% and IoU of 87.2% and 88.2% for both datasets respectively.

Table 2.10 summarizes related research in skin lesion segmentation using advanced architectures and hybrid models, highlighting key aspects such as the learning approach, the employed architecture, the dataset utilized, and the performance metrics achieved in these studies.

## 2.6 Conclusion

In conclusion, in this chapter, we delved into the critical role and the historical development of medical image analysis. We explored CAD systems and highlighted the potential applications and the significant challenges inherent in these systems. By presenting an overview of ML and SOTA DL models, we have set the stage for appreciating the sophisticated techniques driving current advancements. Furthermore, we reviewed the related research works that

Table 2.10 Summary of related works on skin lesion segmentation using advanced architectures and hybrid models.

| Reference             | Network type     | Architecture                                                                       | Dataset   | Score                                                                                             |
|-----------------------|------------------|------------------------------------------------------------------------------------|-----------|---------------------------------------------------------------------------------------------------|
| Chen et al. [135]     | Fully supervised | Recurrent attention CNN                                                            | ISIC 2017 | DSC of 87.04%<br>IoU of 80.36%                                                                    |
|                       |                  |                                                                                    | PH2       | DSC of 92.12%<br>IoU of 86.15%                                                                    |
| Zhang et al. [136]    | Fully supervised | Multi-scale aggregation network using Channel-attention dilated convolution module | ISIC 2018 | Accuracy of 95.7%<br>DSC of 86.4%<br>IoU of 81.6%<br>Sensitivity of 91.5%<br>Specificity of 96.7% |
| Alahmadi et al. [137] | Semi-supervised  | CNN + Transformer                                                                  | ISIC 2017 | DSC of 90.58%                                                                                     |
|                       |                  |                                                                                    | ISIC 2018 | DSC of 89.8%                                                                                      |
|                       |                  |                                                                                    | PH2       | DSC of 94.0%                                                                                      |
| Zhang et al. [138]    | Fully supervised | CNN-transformer hybrid network                                                     | ISIC 2017 | Accuracy of 93.7%<br>DSC of 86.3%<br>IoU of 78.3%<br>Sensitivity of 86.6%<br>Specificity of 96.4% |
| Li et al. [139]       | Fully supervised | Dynamic Spatial Group Enhanced UNet                                                | ISIC 2017 | IoU of 79.2%<br>DSC of 88.4%                                                                      |
|                       |                  |                                                                                    | ISIC 2018 | IoU of 80.5%<br>DSC of 89.2%                                                                      |
| Leo et al. [140]      | Fully supervised | U-shaped network (DBNet-SI)                                                        | ISIC 2017 | Accuracy of 95.9%<br>DSC of 92.5%<br>IoU of 86.1%<br>Sensitivity of 90.9%<br>Specificity of 98.0% |
|                       |                  |                                                                                    | ISIC 2018 | Accuracy of 96.4%<br>DSC of 93.0%<br>IoU of 87%<br>Sensitivity of 92.0%<br>Specificity of 98.4%   |

underscore the collaborative nature of progress in this domain, showcasing the innovations that have informed and inspired our research. This foundational knowledge equips readers with a comprehensive understanding of the principles and methodologies that underpin our subsequent discussions and contributions to the medical image analysis field and CAD systems.

## **Chapter 3**

# **COVID-19 diagnosis from chest radiograph and computed tomography scans**

### **3.1 Introduction**

The COVID-19 pandemic has posed unprecedented challenges to global healthcare systems. Early and accurate diagnosis of COVID-19 is critical for controlling the spread of the virus and providing timely treatment to affected individuals. Although RT-PCR is one of the most popular and effective tools used in SARS-CoV-2 detection, its sensitivity is insufficient for the low viral load in test specimens. Additionally, the RT-PCR tool is only affordable in some countries due to its high cost and scarce availability [141]. Medical imaging, specifically CXR and CT scans, has emerged as a critical adjunct in COVID-19 diagnosis and management. These imaging modalities provide crucial insights into the extent and severity of pulmonary involvement, facilitating the differentiation of COVID-19 from other respiratory pathologies. In practice, these imaging modalities deployments have proven effective for detecting COVID-19. However, the substantial increase in COVID-19 cases, particularly during peak periods, has significantly strained medical personnel and radiologists, leading to considerable fatigue and exhaustion. Therefore, an automated diagnostic system could significantly alleviate this burden, reducing costs, time, and effort while enhancing overall efficiency. DL techniques, particularly CNNs [142], are widely recognized for their effectiveness and efficiency in image processing. They have achieved SOTA results in various domains of medical imaging analysis [143–145]. However, one of the significant challenges in developing an automated system based on DL is the need for a substantial volume of data,

which is crucial for the good performance of DL models. To address this challenge, many researchers have opted to utilize publicly available datasets and employ transfer learning techniques to overcome the scarcity of COVID-19 data as discussed in the literature review section in the previous chapter.

As a contribution to COVID-19 diagnosis, we initially introduce an 11-layer CNN model as a baseline framework for further developments. Building upon this foundation, we integrate an attention mechanism into the 11-layer CNN, which enhances its performance by improving focus on relevant features. Additionally, we thoroughly evaluate several transfer learning models, including ResNet, and InceptionNet. Our experiments assess each model's effectiveness in both binary classification tasks, distinguishing between COVID-19 and non-COVID-19 cases, and multi-class classification tasks, where additional conditions such as pneumonia and lung opacity are included. This approach aims to refine the models' ability to detect COVID-19-specific features within medical scans, thereby advancing diagnostic accuracy.

## 3.2 Methodology

### 3.2.1 Baseline CNN

The baseline CNN model [1] comprises two primary components: a convolutional block and a feed-forward network. The convolutional block includes four convolutional layers with 32, 64, 128, and 96 filters, respectively, each with a kernel size of 3x3. Each convolutional layer employs the ReLU activation function to enhance the non-linearity of the feature maps, followed by downsampling using Max-Pooling with a size of 2x2 and a stride of 1. Additionally, Dropout layers are incorporated within the convolutional block to prevent overfitting and enhance generalization. The feed-forward network consists of two hidden layers with 512 and 128 neurons followed by an output layer, which is tailored to the specific classification task and the number of classes involved. This layer utilizes the softmax activation function for multi-class classification tasks, while for binary classification tasks, it employs the sigmoid activation function. Fig. 3.1 illustrates the proposed baseline model.

### 3.2.2 SE-Net

Attention mechanisms are becoming increasingly prevalent and are among the most popular techniques in recent DL architectures. Initially introduced for machine translation tasks, the concept of attention in neural networks is inspired by the selective focus observed in human

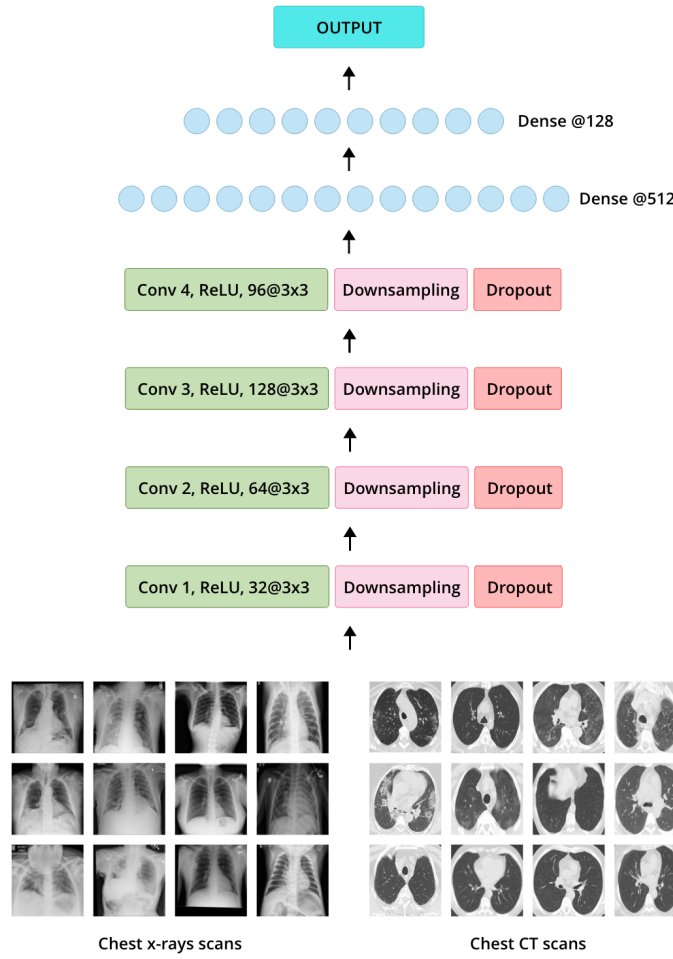


Fig. 3.1 **Baseline CNN architecture.**

visual processing. Therefore, attention mechanisms enable models to dynamically allocate focus to specific portions of input data that are crucial for task performance, while ignoring less relevant parts. This approach not only improves the accuracy of the models but also enhances processing efficiency.

In this approach, we incorporate a Squeeze-and-Excitation (SE) attention mechanism [151] into the baseline CNN model. Specifically, the output feature map of each convolutional layer is passed through an SE-Attention block. This block enhances the network's representational power by enabling dynamic channel-wise feature recalibration. Consequently, the network can focus on the most informative features while suppressing less useful ones, potentially improving overall model performance.

The process of the SE-attention block can be summarized in three steps:

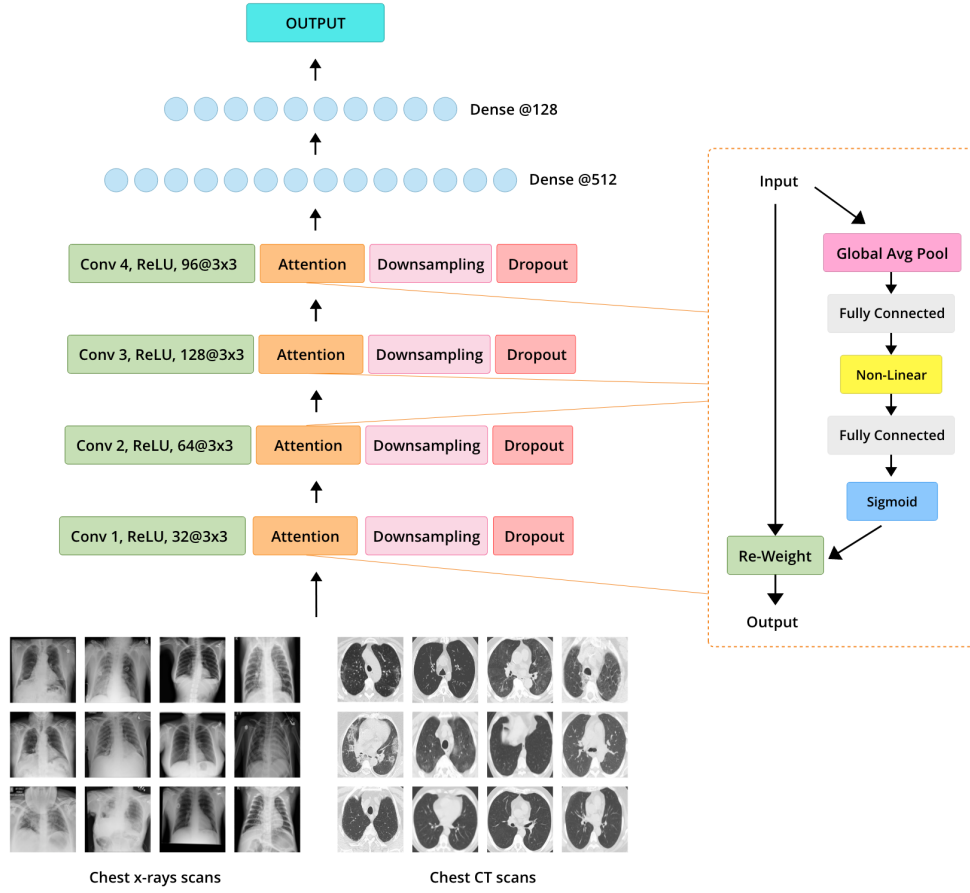


Fig. 3.2 Proposed SE-Net architecture.

- **Squeeze step**

The SE block first applies global average pooling to each channel of the feature map. This operation reduces each  $H \times W \times C$  feature map to a  $1 \times 1 \times C$  vector. This vector summarizes the global spatial information of each channel.

- **Excitation step**

The resulting  $1 \times 1 \times C$  vector is then passed through two FC layers with a ReLU activation in between. The first FC layer reduces the dimensionality of the vector, and the second FC layer restores it to the original number of channels  $C$ . This results in a  $1 \times 1 \times C$  vector of learned channel-wise weights.

- **Scale step**

The learned weights are then used to scale the original feature maps channel-wise.



This means that each channel of the feature map is multiplied by its corresponding weight from the  $1 \times 1 \times C$  vector.

Fig. 3.2 illustrates the detailed block diagram of the SE-Attention layer and its integration into the baseline model.

### 3.2.3 ResNet model

The ResNet model [19], or deep residual neural network, was first introduced in 2015 to address challenges associated with training deeper CNNs, such as training difficulty, accuracy saturation, and degradation. The primary innovation of ResNet is the skip connection, which propagates information across layers, enabling the construction of deeper networks. These skip connections facilitate the network's ability to capture global features effectively.

### 3.2.4 InceptionNet model

The Inception model [22], also known as GoogLeNet, was first introduced in 2014 and has undergone multiple variations and improvements since then. Inception V2 represents a significant advancement, enhancing accuracy while simultaneously simplifying the model's architecture. This version incorporates refinements that optimize performance and efficiency, making it a robust choice for various image classification tasks. Furthermore, Inception V3 introduces additional enhancements, including refined Batch Normalization and additional factorization, building upon the improvements made in Inception V2. These advancements further increase the model's accuracy and efficiency, solidifying Inception V3 as a highly effective architecture for image classification tasks.

### 3.2.5 Loss functions

The loss functions employed in the models are binary cross-entropy and categorical cross-entropy. Binary cross-entropy, efficient for two-class problems (binary classification tasks), ensures a straightforward and effective optimization path. In contrast, categorical cross-entropy, capable of handling multiple classes (multi-class classification tasks), provides a more nuanced approach for complex classification scenarios. These functions quantify the discrepancy between the predicted and actual outcomes, aiming to minimize this discrepancy during the training process.

### 3.3 Experiments and results

This section outlines the used datasets, the preprocessing steps, the core experiments, results, and key findings of our contributions to COVID-19 diagnosis using CXR and CT scans. The experiments were conducted for both binary and multi-class classification tasks utilizing the datasets mentioned above. Initially, we present the results achieved by the baseline CNN and the SE-Net, followed by a comparative analysis to demonstrate the impact of the attention mechanism. Subsequently, we display the results obtained from the application of transfer learning models, such as ResNet and InceptionNet, and compare these with our proposed model across the datasets. This comparison considers critical aspects such as training time and parameter count to evaluate model performance comprehensively

#### 3.3.1 Datasets

##### a. COVID Chestxray Dataset

The COVID chestxray dataset [146] is an image data collection, totaling 679 frontal chest X-ray images from 412 people from 26 countries. This data collection is largely compiled from public databases on websites.

##### b. COVID-19 Radiography Database

This is a database of chest X-ray images [147, 148] for COVID-19-positive cases along with Normal and Viral Pneumonia scans. This database was created by a team of researchers from Qatar University and the University of Dhaka, Bangladesh along with their collaborators from Pakistan and Malaysia in collaboration with medical doctors. The dataset consists of 3,616 COVID-19-positive cases along with 10,192 Normal, 6,012 Lung Opacity (Non-COVID lung infection), and 1,345 Viral Pneumonia images and corresponding lung masks.

##### c. COVIDx CXR-2

This is an open-access benchmark dataset [57] constructed from other publicly available data sources like the COVID Chestxray Dataset [146], and COVID-19 Radiography Database [147, 148], comprising 16,400 CXR images across 15,000 patient cases. The dataset consists of 2,300 COVID-19 cases along with 14,100 non-COVID-19 cases. This dataset is structured into training and testing directories and includes a text file that contains the image name with the corresponding class i.e., positive *vs.* negative.

#### **d. COVID CXR Dataset**

This is a publicly available, balanced Chest X-ray dataset [149] encompassing four classes: COVID, Normal, Lung Opacity, and Viral Pneumonia, with 8,769, 7,662, 8,192, and 5,410 images respectively. This dataset was curated from an initially imbalanced Chest X-ray dataset (COVID-19 Radiography Database [147, 148]). To achieve a balanced version, both under-sampling and over-sampling techniques, leveraging image processing methods were employed.

#### **e. SARS-COV-2 CT Scan Dataset**

This is a publicly available collection of chest CT scans [150], comprising a total of 2,482 scans. Specifically, it includes 1,252 CT scans of COVID-19-positive cases and 1,230 CT scans of non-infected patients. These data have been gathered from real patients in São Paulo, Brazil hospitals.

### **3.3.2 Preprocessing**

To ensure consistency across the training data, we employ two main procedures. First, a normalization procedure that resizes all images to 224x224 pixels and applies min-max normalization. This step ensures that each pixel value is scaled to a range between 0 and 1, which helps in stabilizing and speeding up the training process by providing uniform input data to the model.

Second, we use data augmentation techniques, including Rotation, Zoom, Horizontal Flip, and Shift to enrich the dataset and enhance the model's learning capabilities. These techniques not only increase the diversity of the training data but also help prevent overfitting by exposing the model to a wider range of scenarios. By implementing these procedures, we ensure that our DL models are well-prepared to handle real-world variations and complexities in image data, leading to more robust and accurate performance.

### **3.3.3 Experimental settings**

In our study, we conducted experiments using a TESLA T4 GPU on Google Colab <sup>1</sup> and Kaggle <sup>2</sup>. During the training phase, the models employed the Adam optimizer, a variant of gradient descent algorithm, and were trained for 50 epochs. The learning rate was adjusted throughout the epochs using a plateau reduction strategy to enhance model performance. The

---

<sup>1</sup>Google Colab: [colab.research.google.com](https://colab.research.google.com)

<sup>2</sup>Kaggle: [www.kaggle.com](https://www.kaggle.com)

experiments encompassed several tasks. Initially, we trained and evaluated the models for a binary classification task to assess their performance in differentiating between normal and COVID-19 cases. Subsequently, we utilized the COVIDx CXR-2 dataset for another binary classification task, training the models to distinguish between COVID-19 positive and negative cases, where negatives include both normal and other pneumonia cases.

For multi-class classification, we employed the COVID CXR Dataset, training the models to classify four categories: COVID-19, Normal, Lung Opacity, and Viral Pneumonia based on X-ray scans. Additionally, we trained and evaluated the models using the SARS-COV-2 CT Scan Dataset for CT scans, focusing on classifying COVID-19 and non-COVID-19 cases. Table 3.1 presents the detailed configurations of the datasets used for each task.

Table 3.1 Detailed configuration of the datasets based on the classification task and image modality.

| Task ref   | Classes                                                | Task type   | Image modality | Training set | Validation set | Test set |
|------------|--------------------------------------------------------|-------------|----------------|--------------|----------------|----------|
| CXR Task 1 | COVID-19<br>Healthy                                    | Binary      | X-ray scan     | 11,871       | 3,391          | 3,391    |
| CXR Task 2 | COVID-19<br>Non-COVID-19                               | Binary      | X-ray scan     | 21,328       | 4,473          | 4,482    |
| CXR Task 3 | COVID-19<br>Healthy<br>Lung Opacity<br>Viral Pneumonia | Multi-class | X-ray scan     | 18,020       | 6,002          | 5,999    |
| CT Task    | COVID-19<br>Healthy                                    | Binary      | CT scan        | 1,240        | 496            | 496      |

Throughout the experiments, all the hyperparameters were fixed. All the networks used the same settings and hyperparameters during the training, validation, and testing phases

### 3.3.4 Performance evaluation

To assess the classification performance, we utilized the sigmoid method for the binary task and the softmax method for the multi-class task, transforming logits into class probabilities and selecting the highest probability value as the predicted category. In this study, we employed accuracy, precision, sensitivity, F1-score, and specificity as evaluation metrics. Additionally, we used AUC in the binary task to visualize the trade-off between the TPR and the FPR.

### 3.3.5 Experimental results

In this section, we present the results of our experiments for the baseline CNN, SE-Net, ResNet50V2, and InceptionV3. The first subsection details the experimental findings for CXR scan classification, while the second subsection addresses the results for CT scan classification.

#### 3.3.5.1 X-ray scan classification

Three primary sets of experiments were conducted based on the classification of CXR scans. The first experiment focused on distinguishing between COVID-19 and healthy cases (CXR Task 1), the second on differentiating between COVID-19 and non-COVID-19 cases (CXR Task 2), and the last on classifying normal cases and three diseases: COVID-19, lung opacity, and viral pneumonia (CXR Task 3).

In the first experiment (CXR Task 1), the baseline model achieved an accuracy of 97.73% and a sensitivity of 96.74%. The SE-Net model demonstrated superior performance with an accuracy of 99.53% and a sensitivity of 99.89%. The InceptionV3 model achieved an accuracy of 99.41% and a sensitivity of 98.86, while the ResNet50 model attained an accuracy of 96.69% and a sensitivity of 96.74%. Table 3.2 presents the comprehensive classification report for each model.

Table 3.2 Results of different models for COVID-19 vs. healthy classification based on CXR scans.

| Model        | Accuracy     | Sensitivity  | Precision  | F1-Score     | Specificity | ROC-AUC      |
|--------------|--------------|--------------|------------|--------------|-------------|--------------|
| Baseline CNN | 97.73        | 96.74        | 98.81      | 97.76        | 98.78       | 97.79        |
| InceptionV3  | 99.41        | 98.86        | <b>100</b> | 99.43        | <b>100</b>  | 99.42        |
| ResNet50V2   | 96.69        | 96.74        | 96.74      | 96.74        | 96.63       | 96.68        |
| SE-Net       | <b>99.53</b> | <b>99.89</b> | 99.21      | <b>99.54</b> | 99.15       | <b>99.51</b> |

Fig. 3.3 presents the confusion matrices for each model, illustrating the prediction percentages in a visual format.

Based on the results presented in Table 3.2 and Fig. 3.3, it is evident that the SE-Net demonstrated superior performance in classifying COVID-19 cases, achieving a 99.89% (1749 sample) correct classification rate, highlighting the effectiveness of the attention mechanism in helping the model to capture the most relevant features. The SE-Net model misclassified only 0.85% (14 samples) of COVID-19 cases as healthy. In contrast, the InceptionV3 model accurately predicted all healthy cases without any misclassifications and misclassified 1.14% (20 samples) of the COVID-19 cases. Overall, the best model was the

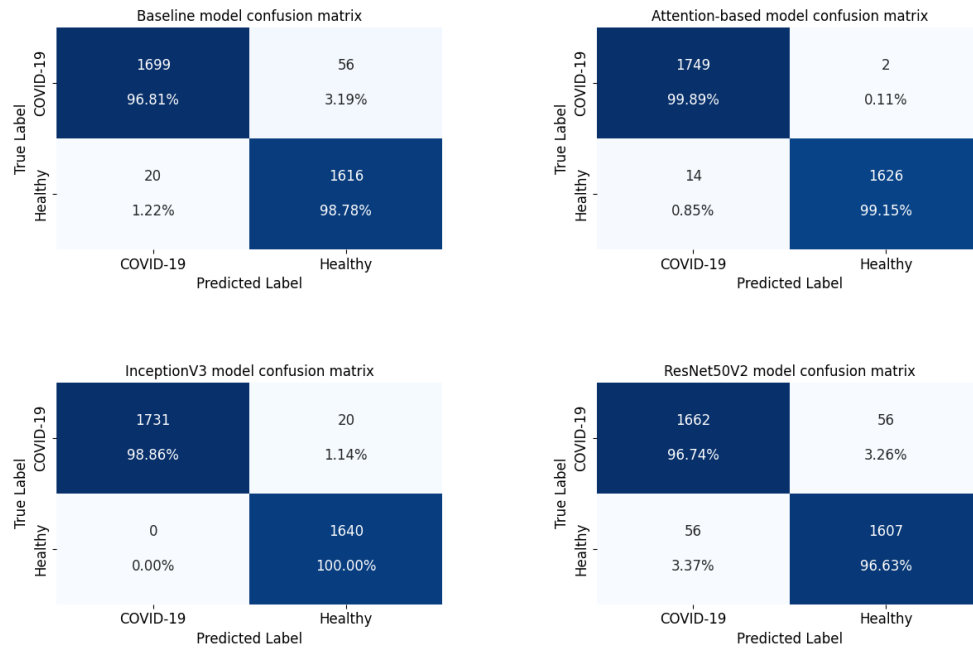


Fig. 3.3 Confusion matrices of different models for COVID-19 vs. healthy classification based on CXR scans.

SE-Net particularly due to its high sensitivity in detecting COVID-19 cases. This is crucial as accurate classification of COVID-19 cases has a significant impact on public healthcare.

In the second experiment (CXR Task 2), which involved classifying COVID-19 vs. non-COVID-19 cases, the complexity of the task increased compared to the first experiment. This heightened complexity arose because the non-COVID-19 category included other lung diseases, posing a significant challenge for the models to generate optimal weights. Based on the highlighted results in Table 3.3 and the confusion matrices in 3.4, we observe that the performance of all models decreased compared to the first task. The SE-Net achieved the best results across all the metrics with an almost accuracy of 94% followed by InceptionV3, baseline model, and ResNet50V2.

Table 3.3 Results of different models for COVID-19 vs. non-COVID-19 classification based on CXR scans.

| Model        | Accuracy     | Sensitivity  | Precision    | F1-Score     | Specificity  | ROC-AUC      |
|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Baseline CNN | 91.27        | 90.78        | 91.91        | 91.34        | 91.78        | 91.27        |
| InceptionV3  | 92.55        | 91.00        | 94.47        | 92.69        | 94.25        | 92.61        |
| ResNet50V2   | 90.18        | 89.55        | 91.00        | 90.26        | 92.63        | 90.84        |
| SE-Net       | <b>93.93</b> | <b>93.24</b> | <b>94.73</b> | <b>93.98</b> | <b>94.65</b> | <b>93.94</b> |

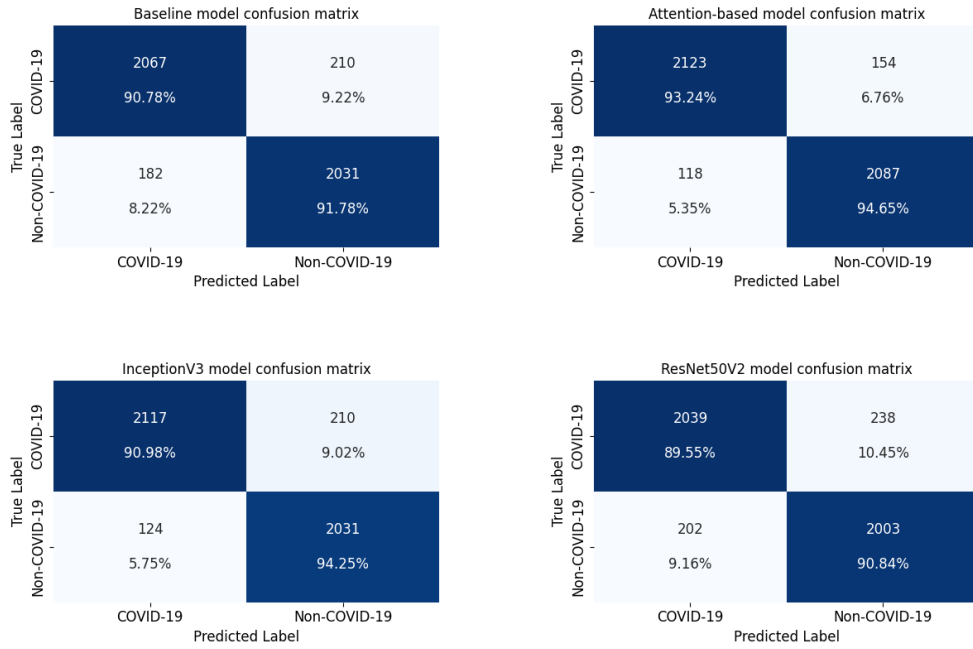


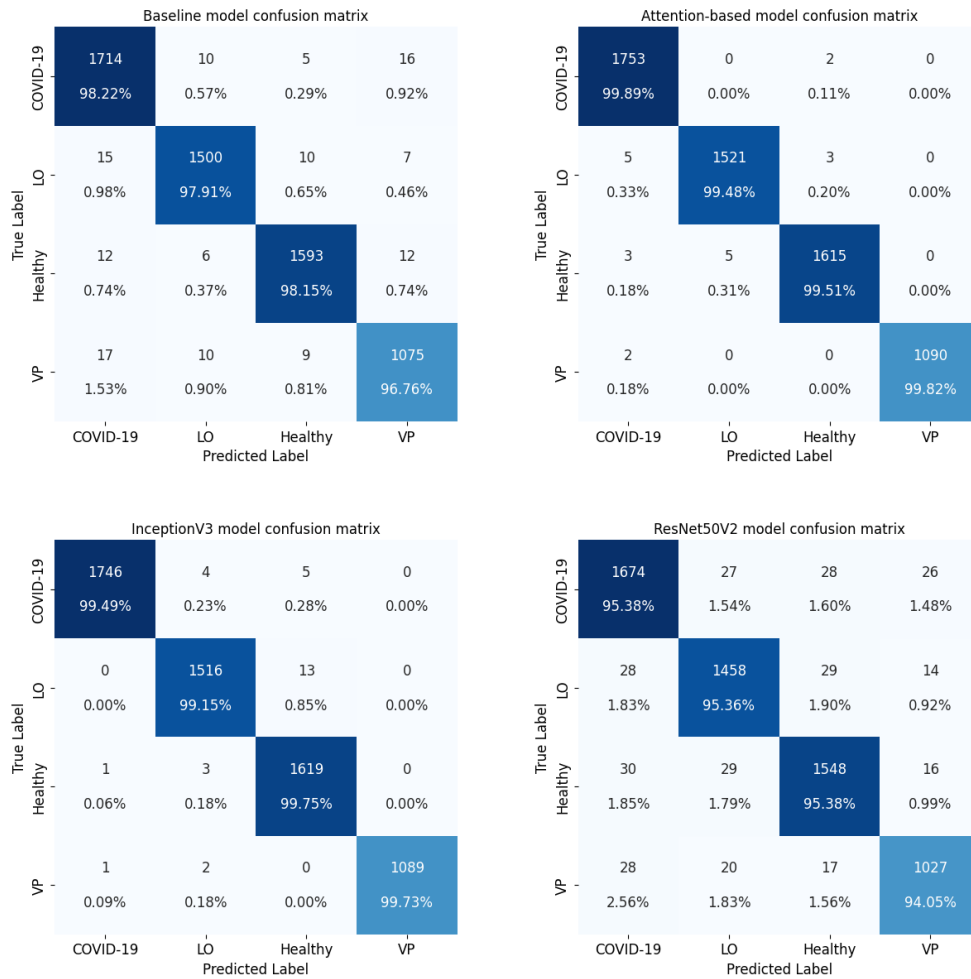
Fig. 3.4 Confusion matrices of different models for COVID-19 vs. non-COVID-19 classification based on CXR scans.

As shown in Fig. 3.4, the SE-Net model misclassified 5.35% (118 samples) of COVID-19 cases as non-COVID-19 and 6.76% (154 samples) of non-COVID-19 cases as COVID-19. This misclassification is attributed to the increased class distribution and overlap between categories. Similarly, while InceptionV3 accurately predicted all the negative cases in Task 1, it misclassified 5.75% (124 samples) of non-COVID-19 cases in this task. This decline in performance affected all the trained models, highlighting the challenges posed by the more complex class distribution.

In the third experiment (CXR Task 3), which involved a multi-class classification task, the models were not forced to group dissimilar conditions (like Pneumonia and Healthy) into a single category (non-COVID), reducing ambiguity. Consequently, the performance across all models improved. Table 3.4 and Fig. 3.5 illustrate the experiment results.

Table 3.4 Results of different models for multi-class classification task based on CXR scans.

| Model        | Accuracy     | Sensitivity  | Precision    | F1-Score     | Specificity  |
|--------------|--------------|--------------|--------------|--------------|--------------|
| Baseline CNN | 97.85        | 97.85        | 97.87        | 97.86        | 99.28        |
| InceptionV3  | 99.52        | 99.53        | 99.55        | 99.54        | 99.83        |
| ResNet50V2   | 95.15        | 95.13        | 95.15        | 95.14        | 98.37        |
| SE-Net       | <b>99.67</b> | <b>99.67</b> | <b>99.70</b> | <b>99.69</b> | <b>99.89</b> |



**Fig. 3.5 Confusion matrices of different models for multi-class classification task based on CXR scans.**

As these results show, the SE-Net model achieved superior performance for the third consecutive time, with an accuracy and sensitivity of 99.67%. InceptionV3's accuracy increased by nearly 9% compared to CXR Task 2, with similar improvements observed in the other models. This performance enhancement is attributed to the models' ability to learn distinct patterns for each condition or class.

### 3.3.5.2 CT scan classification

In the CT scan classification task, the models were trained to classify COVID-19 vs. non-COVID-19 (healthy) cases. The achieved results are shown in Table 3.5. Fig. 3.6 visualizes the confusion matrices of each model that were generated during the test phase.



Table 3.5 Results of different models for COVID-19 vs. non-COVID-19 classification based on CT scans.

| Model        | Accuracy     | Sensitivity | Precision    | F1-Score     | Specificity  | ROC-AUC      |
|--------------|--------------|-------------|--------------|--------------|--------------|--------------|
| Baseline CNN | 97.00        | 97.30       | 96.92        | 97.11        | 96.64        | 96.97        |
| InceptionV3  | 99.19        | 99.22       | <b>99.22</b> | 99.22        | <b>99.16</b> | 99.20        |
| ResNet50V2   | 99.00        | 98.84       | 99.22        | 99.03        | <b>99.16</b> | 99.03        |
| SE-Net       | <b>99.40</b> | <b>100</b>  | 98.85        | <b>99.42</b> | 98.74        | <b>99.38</b> |

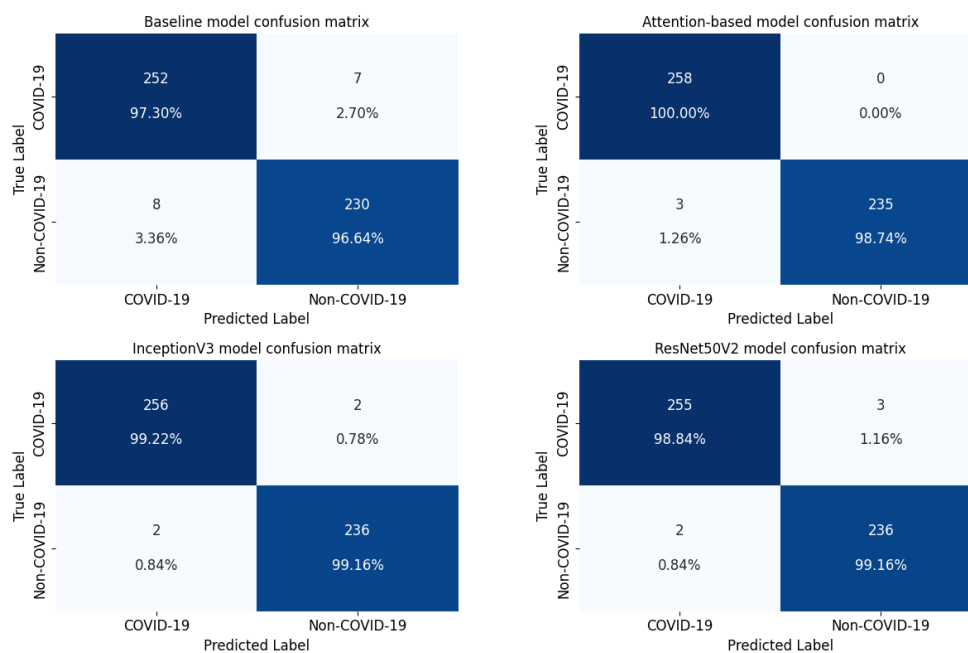


Fig. 3.6 Confusion matrices of different models for COVID-19 vs. non-COVID-19 classification based on CT scans.

Based on the performance metrics, the SE-Net achieved the highest overall accuracy (99.40%), with perfect sensitivity (100%) and a high F1-score (99.42%), highlighting its exceptional ability to correctly identify COVID-19 cases. This suggests that attention mechanisms are highly effective in distinguishing between COVID-19 and non-COVID-19 cases. InceptionV3 and ResNet50V2 also performed well, with accuracies of 99.19% and 99.00% respectively, and both models demonstrated strong precision (99.22%) and specificity (99.16%). The baseline model, while effective, lagged behind with an accuracy of 97.00%, showing the advantages of more advanced architectures in improving classification performance.

### 3.3.5.3 Computational complexity

In DL, computational complexity plays a crucial role in evaluating the performance and efficiency of models. Table 3.6 provides insights into the number of parameters, training time, and inference time for each model across each task.

Table 3.6 Models' computational complexity and performance comparison

| Model       | # Parameters | Task       | Train time (s) | Inference time (s) |
|-------------|--------------|------------|----------------|--------------------|
| InceptionV3 | 24.1M        | CXR Task 1 | 2800           | 6.5                |
|             |              | CXR Task 2 | 6200           | 7.9                |
|             |              | CXR Task 3 | 4800           | 8.2                |
|             |              | CT Task    | 600            | 1.2                |
| ResNet50V2  | 23.5M        | CXR Task 1 | 1400           | 7                  |
|             |              | CXR Task 2 | 3300           | 8.8                |
|             |              | CXR Task 3 | 2500           | 9.6                |
|             |              | CT Task    | 550            | 2                  |
| SE-Net      | <b>7.3M</b>  | CXR Task 1 | <b>800</b>     | <b>2.5</b>         |
|             |              | CXR Task 2 | <b>2200</b>    | <b>6.3</b>         |
|             |              | CXR Task 3 | <b>1600</b>    | <b>7.2</b>         |
|             |              | CT Task    | <b>250</b>     | <b>1</b>           |

The table highlights the efficiency of the SE-Net, demonstrating faster training and inference times across all tasks compared to larger models such as InceptionV3 and ResNet50V2, while also achieving superior accuracy and sensitivity. This suggests that attention mechanisms can effectively handle complex tasks by selectively focusing on relevant features. Overall, model selection should consider the trade-off between accuracy, speed, and computational resources. Based on these results, the SE-Net model emerges as the optimal choice for the various classification tasks conducted in our experiments.

### 3.3.5.4 Comparison with SOTA methods

Tables 2.7 and 2.8 present a comparison of the results achieved by the proposed SE-Net model with SOTA models in terms of accuracy, sensitivity, and precision. These tables highlight the superior performance of our model relative to previous works.

The results underscore that the developed SE-Net model outperforms the SOTA, showcasing the effectiveness of the attention mechanism in enhancing the overall performance of DL models. Furthermore, our model was tested on a larger test set and yielded the best results, demonstrating its robustness and generalizability.

Table 3.7 Comparison of COVID-19 diagnosis results based on CXR scans. N/A: value is not available.

| Authors                    | Model                     | Classes                                          | Test set | Accuracy (%) | Sensitivity (%) | Precision (%) |
|----------------------------|---------------------------|--------------------------------------------------|----------|--------------|-----------------|---------------|
| Narin et al. [44]          | ResNet50                  | COVID-19<br>Healthy                              | 630      | 96.1         | 91.8            | 76.5          |
| Apostolopoulos et al. [68] | VGG19                     | COVID-19<br>Healthy<br>Pneumonia                 | 1,428    | 92.85        | 99.10           | N/A           |
| Berrimi et al [51]         | InceptionV3               | COVID-19<br>Healthy<br>Pneumonia                 | 226      | 92.35        | N/A             | N/A           |
| Ozturk et al. [55]         | DarkCovidNet              | COVID-19<br>Healthy                              | 125      | 98.08        | 95.13           | 98.03         |
|                            |                           | COVID-19<br>Healthy<br>Pneumonia                 | 225      | 87.02        | 92.18           | 89.96         |
| Wang et al. [57]           | COVID-Net                 | COVID-19<br>Healthy<br>Pneumonia                 | 600      | 93.3         | 93.3            | N/A           |
| Chowdhury et al. [59]      | Parallel-dilated CovidNet | COVID-19<br>Healthy<br>Pneumonia                 | 239      | 96.58        | 96.59           | 96.58         |
| This study                 | SE-Net                    | COVID-19<br>Healthy                              | 3,391    | 99.53        | 99.89           | 99.21         |
|                            |                           | COVID-19<br>Non-COVID-19                         | 4,482    | 93.93        | 93.24           | 94.73         |
|                            |                           | COVID-19<br>Healthy<br>Lung Opacity<br>Pneumonia | 5,999    | 99.67        | 99.67           | 99.70         |

Table 3.8 Comparison of COVID-19 diagnosis results based on CT scans. N/A: value is not available.

| Authors            | Model             | Classes             | Test set | Accuracy (%) | Sensitivity (%) | Precision (%) |
|--------------------|-------------------|---------------------|----------|--------------|-----------------|---------------|
| Berrimi et al [51] | Modified DenseNet | COVID-19<br>Healthy | 496      | 95.98        | N/A             | N/A           |
| Yu et al. [62]     | GoogLeNet-COD     | COVID-19<br>Healthy | 88       | 87.5         | 90.9            | N/A           |
| Zhang et al. [63]  | DenseNet-OTLS     | COVID-19<br>Healthy | 192      | 96.3         | 96.35           | 96.29         |
| Zhang et al. [69]  | 7-layer CNN       | COVID-19<br>Healthy | 640      | 94.03        | 94.44           | N/A           |
| Zhang et al. [66]  | 5-layer CNN       | COVID-19<br>Healthy | 640      | 93.64        | 93.28           | N/A           |
| Gupta et al. [67]  | DarkNet19         | COVID-19<br>Healthy | 496      | 98.91        | 98.96           | 98.88         |
| This study         | SE-Net            | COVID-19<br>Healthy | 496      | 99.40        | 100             | 98.85         |

## 3.4 Discussion

In this part of the thesis, we explore the effectiveness of the attention mechanism in enhancing the overall performance of DL models. The experimental results demonstrate that incorporating attention layers into the baseline model significantly improved performance, with increases in accuracy and sensitivity of up to 3% across different tasks. This underscores the high impact of attention mechanisms in enhancing the model's focus on the most relevant features and regions of interest in CXR and CT scans.

In DL, higher parameter counts generally imply more intricate architectures capable of capturing finer details and nuances in the data. For this study, we utilized large pre-trained models, such as Inception and ResNet, which were trained on the extensive ImageNet dataset. Our findings indicate that deeper and pre-trained architectures do not always outperform smaller models, particularly in the domain of medical image analysis, including COVID-19 diagnosis from CXR and CT scans. The baseline model yielded commendable results, not significantly trailing behind the pre-trained models. The introduction of attention layers, –which did not increase the model size– enabled the model to weigh the importance of different features dynamically. This resulted in richer and more meaningful feature representations, allowing the SE-Net model to outperform the pre-trained models across all experimental tasks. The bottom line is that large pre-trained models, such as ResNet or InceptionNet, often include extensive and sometimes redundant layers that may not contribute significantly to the specific task. In contrast, attention-based models can achieve similar or superior results with fewer layers by effectively highlighting and amplifying crucial information. This demonstrates the efficiency of attention mechanisms in enhancing model performance while maintaining a more streamlined architecture.

## 3.5 Conclusion and future perspectives

This chapter discussed the COVID-19 diagnosis and provided an in-depth examination of utilizing CXR and CT scans for diagnosing COVID-19 and other lung diseases, highlighting the limitation of RT-PCR availability and emphasizing the potential of medical imaging in enhancing diagnostic accuracy. It also proposed a novel attention-based model named SE-Net and presented other classification models. The results demonstrated that SE-Net showed notable and promising performance in identifying and classifying COVID-19 and other lung cases from CXR and CT scans.

To expand the research, multiple classification tasks were investigated to assess the performance across different binary and multi-class tasks and different scan modalities.

The findings demonstrated the superior performance of the SE-Net model, which consistently outperformed the other models in terms of accuracy, sensitivity, precision, and specificity. Specifically, in binary classification tasks distinguishing between COVID-19 and non-COVID-19 cases, as well as in multi-class classification tasks involving additional conditions such as pneumonia and lung opacity, the SE-Net achieved remarkable accuracy and robustness. Indeed, the inclusion of the attention mechanism significantly contributed to the model's ability to focus on relevant features, thereby enhancing its diagnostic accuracy.

### 3.5.1 Research contributions and implications

- *Timely relevance:* During the COVID-19 pandemic, the focus on disease diagnosis was highly relevant given the global implications, addressing an urgent need for efficient and accurate diagnostic methods.
- *Comprehensive overview:* The chapter explained the importance of early and accurate COVID-19 diagnosis, the limitation of RT-PCR availability, and the role of medical imaging. It covered various imaging modalities and their contributions to COVID-19 detection.
- *Methodological rigor:* The description of datasets, including the COVID Chestxray Dataset, COVID-19 Radiography Database, COVIDx CXR-2, and SARS-COV-2 CT Scan Dataset, provided a strong foundation for reproducibility and validation of results.
- *Model development and improvement:* Beginning with constructing a baseline 11-layer CNN model, the research advanced methodically towards more sophisticated models. The integration of an attention mechanism demonstrated a progressive approach to model improvement. Additionally, evaluating transfer learning models, including ResNet and InceptionNet further highlights the depth of analysis.
- *Attention mechanism:* The inclusion of the Squeeze-and-Excitation attention mechanism reflects the innovative application of advanced techniques to amplify the performance of DL models. By selectively emphasizing relevant features within the images, this approach significantly enhanced the model's diagnostic capability.

### 3.5.2 Possible future improvements

- *Relative novelty:* While the use of CNNs and attention mechanisms is well-executed, these approaches are relatively standard in the current landscape of medical imaging

research. The chapter could be improved to include other novel or more recent unique methodological contributions.

- *Generalizability and data diversity*: Ensure that the datasets used for training the models are diverse and representative of different demographics to enhance the generalizability of the models.
- *Scalability*: The chapter mentions the use of public datasets and transfers learning to mitigate data scarcity. However, the scalability of these models to real-world clinical settings, particularly in low-resource environments, remains an area that could be addressed more thoroughly.
- *Robustness*: Evaluate the models' robustness by testing them under varying data conditions, including different imaging qualities and noisy environments
- *Interpretability*: Prioritize interpretability by offering clear explanations of model predictions to foster trust and confidence among healthcare professionals.
- *Integration into clinical workflows*: The transition from research to practical implementation in healthcare settings can be expanded. Discussions on integration into existing clinical workflows, potential barriers, addressing potential challenges, solutions, and steps for implementation would add practical value.

Overall, the chapter is well-structured and informative, offering valuable insights into the application of DL techniques for COVID-19 diagnosis through medical imaging, with promising results that pave the way for further advancements in this field, particularly in systems leveraging attention mechanisms. By addressing potential improvements, the chapter's impact and utility could be significantly enhanced, further benefiting both researchers and practitioners. Strengthening its practical implementations would increase its comprehensiveness, readability, and relevance, making it a more valuable resource in the fields of medical imaging and AI.

## **Chapter 4**

# **Enhanced retinal disease diagnosis and detection using optical coherence tomography images**

This chapter addresses the diagnosis and detection of various retinal diseases and highlights the application of advanced DL techniques for the analysis of OCT images, focusing on both classification and segmentation tasks. It begins by introducing the HTC-Retina model, designed for the diagnosis of seven distinct retinal diseases. Following this, the chapter delves into the model's refinement through the integration of the advanced DA-UNet-Plus segmentation architecture, further enhancing its diagnostic accuracy and performance.

### **4.1 Introduction**

The prevalence of retinal diseases has witnessed a notable surge in recent years, causing significant visual impairments and abnormalities. These conditions encompass a range of disorders that inflict damage upon various components of the retina, consequently leading to temporary or permanent vision loss. The advancement of medical imaging techniques, particularly OCT, has significantly enhanced the ability of physicians and experts to examine retinal diseases. This technology facilitates the quantification, analysis, and treatment planning for various retinal disorders, making early diagnosis and detection of early-stage pathologies crucial to preventing irreversible vision loss. Therefore an automated diagnosis and detection system could significantly contribute to the initial diagnosis and detection of retinal diseases and help physicians in fast analysis, and treatment planning.

In this chapter, we first address the diagnosis of seven retinal diseases, namely AMD, DME, DRUSEN, CNV, CSR, DR, and MH, utilizing a hybrid classification model. We present the HTC-Retina model, a novel approach developed to diagnose seven retinal diseases using OCT datasets. The concept of this hybrid model is inspired by extensive literature reviews, addressing the limitations observed in existing methods.

Following this, the second part details a novel hybrid segmentation model named DA-UNet-Plus which is inspired by insights from the existing works. DA-UNet-Plus model is destined to detect and quantify intraretinal fluid which appears in AMD and DME cases.

By employing the hybrid approach in both HTC-Retina and DA-UNet-Plus, we aim to achieve high performance by leveraging the strengths of each technique to mitigate the limitations of other methods

## **4.2 Part I: Retinal diseases classification and identification using HTC-retina**

### **4.2.1 Motivation for HTC-Retina development**

Among the various ML approaches, CNNs have demonstrated significant success in medical image classification by efficiently extracting local features from the entire image and providing well-informed inductive biases. This success is notably evident in the classification of OCT images for retinal disease diagnosis, where researchers have achieved considerable breakthroughs. However, in the analysis of retinal diseases, relying on more local lesion features may be required to achieve improved performance. Therefore, global retinal information along with local lesion information is needed to enhance classification performance.

On the other hand, the emergence of ViTs as a dominant architecture in computer vision offers new avenues for medical image analysis. With the leveraging of the self-attention mechanisms to capture long-range dependencies among image pixels, thus exhibiting promising results in image classification tasks. However, the application of transformer architectures alone poses certain challenges, particularly with small-scale datasets. Transformers, lacking the local inductive bias inherent in CNNs, typically require large amounts of training data. This limitation becomes pronounced in the context of retinal OCT image classification, where the availability of training images is relatively constrained.

To address these challenges, combining the strengths of CNNs and ViTs presents a promising solution. CNNs excel in efficient parameter utilization and the extraction of local features and multi-scale feature maps through convolutional operations. Meanwhile, ViTs' self-attention mechanisms facilitate the processing of long-range and global dependencies



within an image. By integrating these complementary capabilities, a hybrid model can be developed to robustly and efficiently process image information, enhancing overall classification performance. In recent studies [152–154], researchers have explored the combination of CNN and ViT in image analysis tasks. The objective was to leverage the respective strengths of each algorithm and create a hybrid model capable of achieving superior performance across various data sizes.

## 4.2.2 HTC-Retina model contribution

The HTC-Retina model is designed to classify seven different types of retinal diseases from OCT images. The model is a hybrid architecture that combines the strengths of transformers in capturing long-range dependencies and CNNs in extracting hierarchical local features. This proposed model incorporates three modules: the convolutional patch and token embedding (CPTE) module, the vision transformer encoder (ViTE) module, and the residual depthwise-pointwise ConvNet (RDP-ConvNet) module. The detailed architecture of the HTC-Retina is described in Section 4.2.3.1. The HTC-Retina effectiveness evaluation was passed through several experiments across three datasets, each comprising different types of retinal diseases and a diverse range of retinal conditions. This work is published in the Journal of Computers in Biology and Medicine [3].

## 4.2.3 Methodology

### 4.2.3.1 HTC-Retina structure

ViTs [33] are widely recognized for their utilization of self-attention [32] to conduct mutual modeling for all input patches, a process that can be computationally costly and data-intensive. Incorporating domain-specific knowledge into the model architecture can mitigate this problem. In the context of retinal disease classification, the OCT scans may contain irrelevant information, obscuring important features. Moreover, some retinal layer features may not be readily apparent. ViT models excel at capturing global dependencies, but they may struggle with local feature extraction [155]. In contrast, the convolutional kernel is a learnable filter that can selectively attenuate irrelevant information and enhance local features, complementing ViT's attention mechanism. By integrating these two components, the model enables meticulous examination of retinal layers and facilitates learning intricate feature correlations within the retinal images.

The HTC-Retina model comprises four main components: the CPTE, ViTE, RDP-ConvNet, and the classifier. Fig. 4.1 illustrates the detailed representation of the HTC-Retina

architecture. The initial step involves feeding the OCT image into the CPTE module, facilitating the extraction of pertinent features while concurrently capturing low-level local dependencies and patterns essential for the formation of overlapping tokens and patches. Concurrently, the sequence of patches and tokens is subjected to processing in two parallel modules: the ViTE module and the RDP-ConvNet module. This parallel processing scheme serves the purpose of encoding global and high-level local features of the retinal image, as well as modeling the intricate relationship between specific lesion regions and the entire retinal layer. The ViTE module includes self-attention and a multilayer perceptron, while the RDP-ConvNet module comprises a simple fully convolutional block. Each module's output then undergoes global average pooling to produce fixed-length representations, followed by a dropout layer. Subsequently, a feature fusion module is employed to concatenate the extracted global and local contextual features effectively. Finally, a fully connected layer with a softmax activation function generates classification outcomes.

#### a. Convolutional patch and token embedding module

The convolutional patch and token embedding module is a simple convolutional network that collects low-level local characteristics from the input image and creates patches and overlapping tokens by running it through a convolutional block. This block is a standard architecture that includes a convolution, ReLU activation [156], and a max pool. Given an image or feature map  $f \in \mathbb{R}^{W \times H \times C}$

$$x = \text{MaxPool}(\text{ReLU}(\text{CONV}(f))) \quad (4.1)$$

where W is the width, H is the height, C is the number of channels of the of the feature map and the convolution operation (CONV) has d filters, the same number as the embedding dimension of the ViTE and the RDP-ConvNet backbone. As illustrated in Fig. 4.1(a), the CPTE module is made up of three hierarchical convolutional layers of 64, 128, and 256 filters, respectively. Each convolutional layer employs 3x3 filters, followed by a zero-padding layer that pads the input image by one pixel on each side. The max-pooling layer then decreases the feature map's spatial dimensions by a factor of two. After the process, the CPTE yields a set of spatially localized patches, which are subsequently utilized as input for the RDP-ConvNet module. Simultaneously, it also generates a sequential arrangement of tokens (reshaping the patches to a series of tokens), serving as input for the ViTE module.

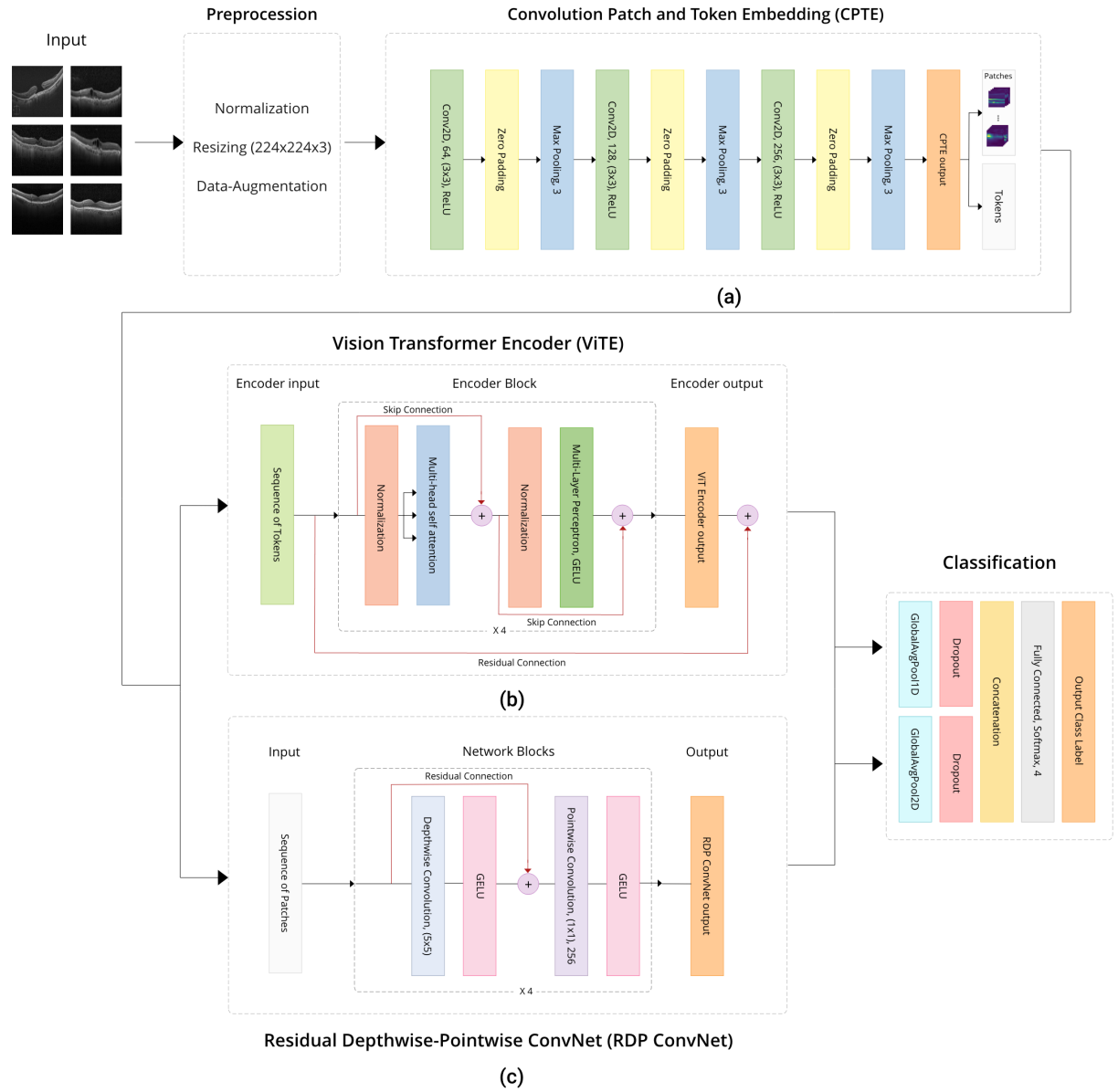


Fig. 4.1 **The overall architecture of the HTC-Retina model.** First, the inputs are fed to a CPTE to obtain patches and tokens. Second, the sequence of patches and tokens is subjected to parallel processing in two modules, namely ViTE and RPD-ConvNet, each composed of multiple blocks. Following this, a global average pooling block is applied to the output of each module, followed by dropout regularization. Lastly, a fusion module is employed to effectively concatenate the extracted global and local features, which are fed into a fully connected softmax layer to generate classification predictions.

## b. Vision transformer encoder module

The encoder component comprises a sequence of four identical encoders, as illustrated in Fig. 4.1(b). Each encoder unit is composed of several layers, including a normalization

layer (NL) [157], an MHSA layer, an MLP, and a stochastic depth regularization (SDR) mechanism [158]. The MHSA layer first applies three linear transformations to the input embeddings. Each embedding vector  $x_i$  is first transformed into a set of query  $Q_i$ , key  $K_i$ , and value  $V_i$  vectors. These vectors are then used to compute the similarity between all tokens by taking the dot product between the query and key vectors of each pair of tokens. This similarity matrix is then scaled by the square root of the dimension of the tokens to improve optimization and decrease variance. The softmax function is applied to the scaled similarity matrix to obtain an attention matrix, which is multiplied by the corresponding value vectors to obtain enhanced tokens. To ensure robust feature extraction, the transformer employs ensemble learning by performing these operations  $h$  times, corresponding to the number of attention heads ( $h$ ), and concatenates the output tokens. This process enables the model to capture correlations between different tokens and learn relevant features at multiple levels of abstraction. The detailed process is further described as follows:

$$head_i = Attention(Q, K, V) \quad (4.2)$$

$$Attention(Q, K, V) = softmax(\frac{xW^Q(xW^K)^T}{\sqrt{d_k}})xW^V \quad (4.3)$$

$$output = concat(head_1, head_2, \dots, head_h)W^{output} \quad (4.4)$$

where  $X$  is the sequence of tokens.  $W^Q$ ,  $W^K$ , and  $W^V$ ,  $W^{output}$  are trainable parameter matrices.  $d_k$  is the dimensionality of the key vectors.  $h$  is the number of parallel attention heads.

Subsequently, an MLP is incorporated into the MHSA layer, consisting of several non-linear layers activated by the GELU [159] function. The MHSA and MLP layers are equipped with skip connections analogous to those present in residual networks, and they undergo layer normalization. Furthermore, the training process involves the application of stochastic depth regularization, a technique that randomly drops a set of layers during training, whereas during inference, all layers are retained. This technique is similar to Dropout, but instead of operating on individual nodes within a layer, it operates on a block of layers. Specifically, assuming that the previous layer's representation is denoted as  $x_{l-1}$ ,  $LN$  and  $SDR$  represent linear normalization and stochastic depth regularization, respectively, the output of the current layer can be expressed as

$$\hat{x}_l = SDR(MHSA(LN(x_{l-1}))) + x_{l-1} \quad (4.5)$$

$$x_l = SDR(MLP(LN(x_l))) + \hat{x}_l \quad (4.6)$$

Finally, at the end of the encoder, a residual connection [19] is added between the output of the last layer and the original input embeddings, which helps to preserve the information.

Overall, the transformer encoder allows for efficient processing of input sequences by using self-attention to model dependencies between all pairs of tokens, and feedforward networks to introduce non-linearities between tokens. The use of residual connections and layer normalization helps to stabilize the training process and prevent the vanishing gradient problem.

### c. Residual depthwise-pointwise ConvNet module

While the ViTE excels at collecting global and long-range contextual information, it ignores local features and the underlying two-dimensional neighborhood structure in images. Based on recent advances in the area, a parallel RDP-ConvNet module is designed to capture high-level local features [160–162]. The RDP-ConvNet module as illustrated in Fig. 4.1(c) is built as a fully convolutional block, with two vital convolutional operations: depthwise convolution with a subsequent residual connection, followed by pointwise convolution. The depthwise convolution involves grouped convolution with the number of groups equal to the number of channels.

Subsequently, the pointwise convolution, characterized by a kernel size of 1, is applied. Each convolution is followed by a GELU activation function layer [159]. Specifically, assuming that the previous layer's representation is denoted as  $x_{l-1}$  and  $\sigma$  is the non-linear activation of GELU, the output of the current layer can be expressed as:

$$\hat{x}_l = \sigma(\text{ConvDepthwise}(x_{l-1})) + x_{l-1} \quad (4.7)$$

$$x_{l+1} = \sigma(\text{ConvPointwise}(\hat{x}_l)) \quad (4.8)$$

### d. Concatenation and classification module

The global features extracted by the ViTE module and the high-level local features extracted by the RDP-ConvNet are passed through global average pooling to transform the output features into 1-D representation. After that, a concatenation approach is applied to outputs, which creates a combined feature representation that incorporates both global and local information. By concatenating the features, the fused representation captures both the overall context provided by the global features and the fine-grained details captured by the local

features. Assuming that  $O_{ViTE}$ ,  $O_{RDPCovNet}$  produced as the ViTE and the RDP-ConvNet outputs respectively, and  $GAP$  given as the Global Average Pooling operation, the final output Output<sub>f</sub> of the fusion module can be expressed as

$$Output_f = \text{concat}(GAP_{1D}(O_{ViTE}), GAP_{2D}(O_{RDPCovNet})) \quad (4.9)$$

Following the concatenation, the output is passed through a fully connected neural network comprising a single layer having four neurons and a softmax activation function, which facilitates the assignment of each vector to its corresponding class. The output of this layer takes the form of a probability distribution, with a cumulative sum of 1. Accordingly, the neuron exhibiting the highest probability value denotes the class label.

#### 4.2.3.2 Datasets

To develop effective DL models for retinal disease classification and segmentation, a substantial amount of well-annotated data is essential. However, acquiring and annotating such data in medical imaging presents significant challenges due to various factors previously discussed in Chapter 2. Therefore, we chose to utilize existing OCT image datasets.

##### a. Duke OCT Dataset

The duke OCT dataset [78] is a collection of OCT images <sup>1</sup> from the VIP laboratory, at Duke University. It consists of SD-OCT volumetric scans acquired from 45 patients with three kinds of retinal diseases: 15 normal patients, 15 patients with early age-related macular degeneration, and 15 patients with diabetic macular edema. It contains 3231 images divided into three categories: NORMAL, AMD, and DME. We refer to this data as the OCT-2014.

##### b. Mendeley OCT Dataset

The mendeley OCT dataset [79] is a collection of OCT images <sup>2</sup>, which contains 84,495 images divided into four categories: NORMAL, CNV, DME, and DRUSEN. The dataset is organized into training, testing, and validation folders containing subfolders for each image category. Since the dataset is imbalanced, we applied various modifications and augmentation techniques to create a balanced version. The resulting balanced dataset comprises a total of 48,574 images, evenly distributed across the four disease categories. We refer to this data as the OCT-2017.

<sup>1</sup>Duke OCT Dataset: [people.duke.edu/~sf59/Srinivasan\\_BOE\\_2014\\_dataset.htm](http://people.duke.edu/~sf59/Srinivasan_BOE_2014_dataset.htm)

<sup>2</sup>Mendeley OCT Dataset: [data.mendeley.com/datasets/rscbjbr9sj/2](http://data.mendeley.com/datasets/rscbjbr9sj/2)

### c. OCT-C8 Dataset

This dataset [83] is an open-access benchmark dataset<sup>3</sup> containing 24,000 images divided into eight categories: AMD, CNV, CSR, DME, DR, DRUSEN, MH, and NORMAL where each category contains 3000 images. The dataset is also organized into training, testing, and validation folders containing subfolders for each image category. We refer to this data as the OCT-C8.

#### 4.2.3.3 Preprocessing and augmentations

To ensure consistency across the training data, all images were resized to  $224 \times 224$  pixels and normalized using max-min normalization as follows:

$$x_{normalized} = (x - x_{min}) \frac{x_{normalized-max} - x_{normalized-min}}{x_{max} - x_{min}} \quad (4.10)$$

where  $x$  refers to the original intensity of the input image,  $x_{min}$  and  $x_{max}$  refers to the minimum and maximum pixel values in the original image,  $x_{normalized-min}$  and  $x_{normalized-max}$  are the required minimum and maximum values for the normalized image, respectively.

Data augmentation techniques, including Crop, RandomGamma, Rescaling, and HorizontalFlip were used to improve the dataset and enhance the model's performance. The augmentation and transformation techniques applied to the input OCT scans were chosen on purpose and thoughtfully to avoid the random selection of parameters based on subjective intuition. Such uninformed parameter choices can result in undesirable transformations being applied to the OCT scans that are sensible, leading to feature deformation and loss. The aforementioned techniques offer benefits such as increased spatial robustness, improved handling of image variations, adaptability to different lighting conditions, standardization of image sizes, and orientation invariance. These advantages contribute to the enhanced performance and generalization capabilities of the model in the task of OCT image classification.

#### 4.2.4 Experiments and results

This section delineates the used ad hoc datasets, the experimental settings, results, and key findings of the HTC-Retina model. The efficacy of HTC-Retina has been validated through many experiments and evaluation steps. The subsequent subsections offer a comprehensive overview of these experiments and the evaluation methodology and results.

---

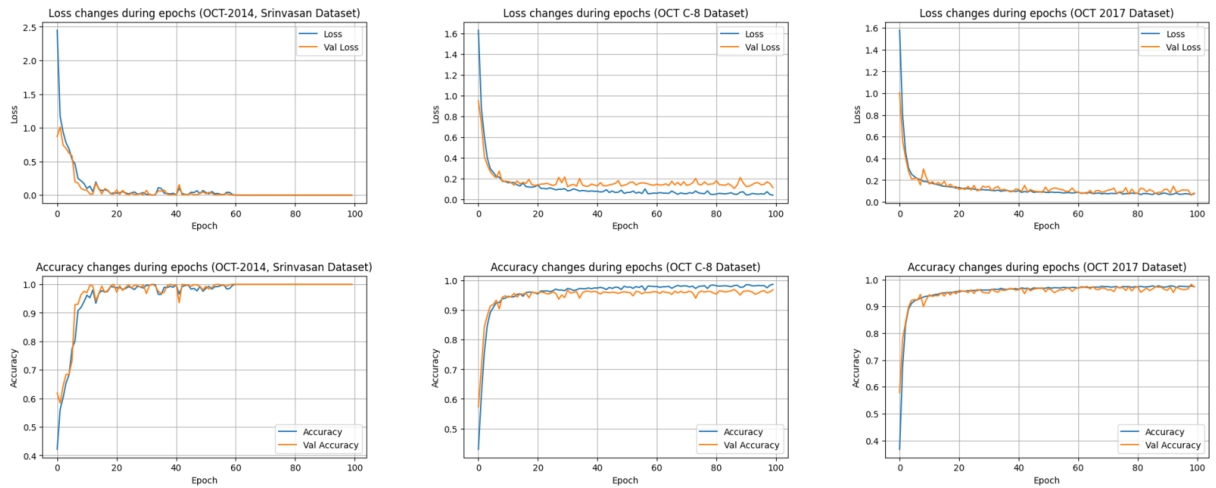
<sup>3</sup>OCT-C8 Dataset: [kaggle.com/datasets/obulisainaren/retinal-oct-c8](https://kaggle.com/datasets/obulisainaren/retinal-oct-c8)

#### 4.2.4.1 Experimental settings

The HTC-Retina was built using the TensorFlow framework and the experiments were conducted using an NVIDIA Tesla P100 GPU. The model was trained to optimize the cross-entropy loss [17]. AdamW (Stochastic Gradient Descent) [163] was used as the optimizer, with a weight decay of  $1e^{-4}$ . Furthermore, the model was trained for 100 epochs with a batch size of 128. The initial learning rate was set to  $1e^{-3}$ , with a plateau reduction strategy used across the epochs. In our experiments, we used three datasets: Duke OCT Dataset, Mendeley OCT Dataset, and OCT-C8 for the classification and identification of different retinal diseases. Each dataset's details are described above in Section 4.2.3.2.

#### 4.2.4.2 Experimental results

During the experiment, we systematically recorded the error and accuracy variations of the training and validation sets, hence that the validation set contains OCT scans that were not used in the training phase. These recorded values were the foundation for evaluating the model's classification performance. To visually illustrate the observed changes, we presented in Fig. 4.2 the error and the accuracy trends of the training set and the validation set of the three used datasets.



**Fig. 4.2 HTC-Retina performance in the training and validation sets.** The blue line presents the accuracy and loss of the training set, while the orange one presents the accuracy and loss of the validation set.

The curves indicate a rapid decrease in both training and validation errors of each dataset during the initial epochs, followed by a stabilization phase after the 60th epoch. Similarly, the accuracy rate rapidly increases, leveling off after the 60th epoch. Based on the graphs,



we observe that the model is doing well since the highest difference between training and validation accuracy is between 0.05% and 0.15%. Therefore, there are no signs of overfitting.

Table 4.1 shows the highest validation accuracies and the lowest validation losses of the HTC-Retina model.

Table 4.1 Best validation results

| Dataset  | Validation accuracy | Validation loss |
|----------|---------------------|-----------------|
| OCT-2017 | 98.30%              | 0.0664          |
| OCT-2014 | 99.80%              | 0.0001          |
| OCT-C8   | 96.99%              | 0.0472          |

#### 4.2.4.3 Fine-grained results

To examine the fine-grained performance of the model, we present a comprehensive classification report on the test datasets in Table 4.2. This report provides detailed metrics for each disease category's accuracy, sensitivity, precision, F1-score, and specificity.

Table 4.2 Detailed evaluation results of the HTC-Retina model

| Dataset  | Disease Class | Evaluation metrics (%) |             |           |          |             |
|----------|---------------|------------------------|-------------|-----------|----------|-------------|
|          |               | Accuracy               | Sensitivity | Precision | F1-Score | Specificity |
| OCT-2017 | CNV           | 98.80                  | 99.70       | 98.36     | 99.02    | 99.03       |
|          | DME           | 99.60                  | 98.76       | 99.58     | 99.18    | 99.86       |
|          | DRUSEN        | 99.60                  | 99.60       | 99.60     | 99.60    | 99.86       |
|          | NORMAL        | 99.60                  | 99.60       | 100.00    | 99.80    | 100         |
|          | Average       | 99.40                  | 99.41       | 99.39     | 99.40    | 99.70       |
| OCT-2014 | AMD           | 100                    | 100         | 99.32     | 99.65    | 99.66       |
|          | DME           | 99.32                  | 99.32       | 100       | 99.65    | 100         |
|          | NORMAL        | 100                    | 100         | 100       | 100      | 100         |
|          | Average       | 99.77                  | 99.77       | 99.77     | 99.76    | 99.80       |
| OCT-C8   | AMD           | 100                    | 100         | 100       | 100      | 100         |
|          | CNV           | 92.85                  | 92.86       | 95.60     | 94.20    | 98.77       |
|          | CSR           | 100                    | 100         | 98.87     | 99.43    | 99.38       |
|          | DME           | 92                     | 92          | 97.28     | 94.57    | 99.83       |
|          | DR            | 99.15                  | 99.14       | 99.43     | 99.28    | 99.63       |
|          | DRUSEN        | 95.70                  | 95.71       | 91.78     | 93.71    | 99.91       |
|          | MH            | 99.15                  | 99.14       | 100       | 99.57    | 100         |
|          | NORMAL        | 97.15                  | 97.14       | 93.41     | 95.24    | 99.02       |
|          | Average       | 97.00                  | 97.00       | 97.04     | 97.01    | 99.56       |

To facilitate a comprehensive analysis of the classification results, we present in Fig. 4.3 the confusion matrices that were acquired after testing the model on a separate test set.

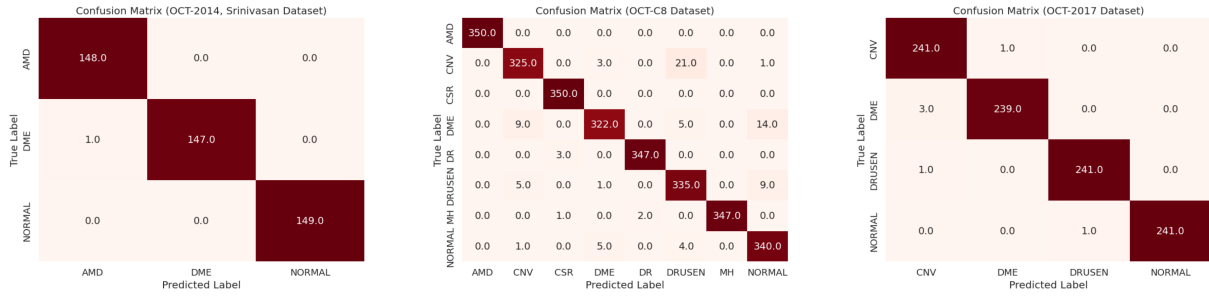


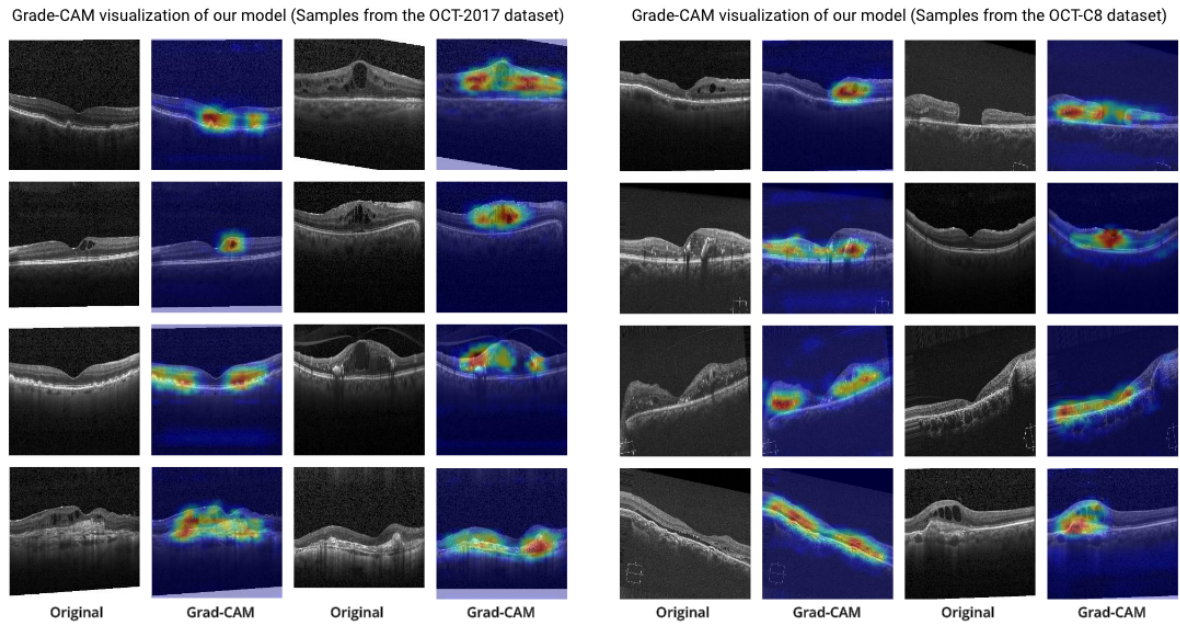
Fig. 4.3 Confusion matrix of each dataset generated by the HTC-Retina model.

Based on these results, the proposed model demonstrated a notable classification performance in terms of accuracy and sensitivity. More specifically, in the OCT-2014 dataset, the model exhibited remarkable performance in the classification of AMD, DME, and Normal cases with an accuracy and sensitivity of 99.77%. As we observed in Fig. 4.3, the proposed model misclassifies one sample from the DME class. In the OCT-C8 dataset, the proposed HRC-Retina model achieves a notable overall accuracy rate of 97.00% for the classification of eight distinct cases. The analysis reveals that the proposed model demonstrates a high level of accuracy in correctly predicting the categories, particularly in cases related to AMD, CSR, DR, and MH, where the predictions are accompanied by a high confidence score. However, there are instances of misclassification observed in the DME vs. CNV cases, as both of these conditions involve the accumulation of fluid, leading to potential confusion for the HRC-Retina model. Similarly, misclassifications are observed in cases of CNV vs. DRUSEN, which represent the two primary forms of age-related macular degeneration. The complexity and similarities between these conditions may contribute to the model's confusion in accurately classifying them. There are also instances of misclassification observed in the normal and other classes, particularly when dealing with cases that exhibit subtle lesions. Identifying such cases poses a challenge even for experienced medical experts. Regarding the OCT-2017 dataset, the model successfully classified 980 out of 986 cases, resulting in an accuracy and sensitivity of 99.40%. This achievement highlights the model's remarkable predictive performance. However, a few misclassifications were observed in cases of CNV and DME, due to the previously mentioned reason.

#### 4.2.4.4 Features map visualization

In medical imaging analysis, the explainability of DL models is important due to ethics and life health involvement. Indeed, it brings transparency, interpretability, and validation to the DL models, empowering clinicians, researchers, and patients to understand, trust, and effectively utilize artificial intelligence technologies in healthcare. In this section, we present

the Class Activation Mapping (CAM) visualizations using the Grad-CAM method [60] as a way to visualize the evidence supporting the model’s predictions. The CAM technique generates a heat map highlighting regions within the image strongly associated with the target class. Intensified red coloration indicates a higher correlation with the predicted category. In Fig. 4.4, the Grad-CAM visualization distinctly delineates the regions of interest. Notably, in the OCT images of diseased conditions, the lesion regions exhibit intensified red coloration, indicating the model’s attention to the areas that are diagnostically significant and aligning with the focus of ophthalmologists during decision-making processes.



**Fig. 4.4 The grad class activation mapping visualization of the HTC-Retina model.** The regions characterized by reddish hues exhibit the highest significance in the classification process, followed by regions containing yellowish pixels, whereas the regions with bluish hues demonstrate the least contribution to the classification outcome.

#### 4.2.4.5 Validation and comparative analysis

To validate the effectiveness of the HTC-Retina model, we compared its performance with widely used classification approaches: VGG19 [21] and ResNet50 [19], pure ViT, and other SOTA methods. Table 4.3 presents a comparative analysis of different models’ performances on the three used datasets for OCT image classification.

In a comparative analysis of various models, the HTC-Retina consistently demonstrated superior performance, surpassing SOT methods and other selected architectures across all three datasets. As depicted in Table 4.3, HTC-Retina achieved remarkable accuracy

Table 4.3 Comparison of classification results of different models on the test set.

| Datasets | Models                 | Evaluation matrices (%) |              |              |              |              |
|----------|------------------------|-------------------------|--------------|--------------|--------------|--------------|
|          |                        | Accuracy                | Sensitivity  | Precision    | F1-Score     | Specificity  |
| OCT-2017 | VGG-19                 | 94.01                   | 94.01        | 94.01        | 94.01        | 96.08        |
|          | ResNet50               | 94.83                   | 94.63        | 95.12        | 94.87        | 96.35        |
|          | Pure ViT               | 93.90                   | 93.85        | 94.05        | 93.95        | 95.20        |
|          | InceptionV3 [73]       | 96.6                    | 97.80        | N/A          | N/A          | 97.40        |
|          | Deep CNN [77]          | 98.65                   | 98.50        | 98.75        | 98.50        | N/A          |
|          | Multi-ResNet50 [81]    | 90.2                    | 86.9         | 86.5         | N/A          | 96.60        |
|          | ConViT [87]            | 94.46                   | 94.00        | 94.00        | 94.00        | N/A          |
|          | CoAtNet [153]          | 97.82                   | 97.82        | 97.86        | 97.84        | 99.38        |
|          | LLCT [88]              | 97.70                   | 97.70        | 97.80        | 97.75        | 99.20        |
|          | HCTNet [86]            | 91.56                   | 88.57        | 88.11        | N/A          | N/A          |
|          | HTC-Retina (Ours)      | <b>99.40</b>            | <b>99.41</b> | <b>99.39</b> | <b>99.40</b> | <b>99.70</b> |
| OCT-2014 | VGG-19                 | 99.32                   | 99.32        | 99.62        | 99.47        | 99.75        |
|          | ResNet50               | 99.20                   | 99.15        | 99.50        | 99.32        | 99.66        |
|          | Pure ViT               | 99.52                   | 99.45        | 99.64        | 99.55        | <b>99.87</b> |
|          | CNN                    | 92.64                   | 90.21        | N/A          | N/A          | 95.88        |
|          | Multi-ResNet50 [81]    | 95.53                   | N/A          | N/A          | N/A          | N/A          |
|          | SoT [85]               | 99.35                   | 99.25        | 99.25        | 99.25        | 99.55        |
|          | CoAtNet [153]          | 98.85                   | 98.83        | 98.80        | 98.81        | 99.38        |
|          | HCTNet [86]            | 86.18                   | 85.40        | 88.53        | N/A          | N/A          |
|          | HTC-Retina (Ours)      | <b>99.77</b>            | <b>99.77</b> | <b>99.77</b> | <b>99.76</b> | 99.80        |
| OCT-C8   | VGG-19                 | 90                      | 89.98        | 90.03        | 90           | 94.32        |
|          | ResNet50               | 91.50                   | 91.43        | 92.12        | 91.77        | 94.79        |
|          | Pure ViT               | 92.65                   | 92.52        | 93.03        | 92.77        | 94.98        |
|          | ResNet34 [80]          | 92.40                   | 92.00        | 92.00        | 92.00        | N/A          |
|          | ResNet50 [80]          | 90.30                   | 91.00        | 91.00        | 91.00        | N/A          |
|          | ResNet101 [80]         | 84.50                   | 86.00        | 86.00        | 86.00        | N/A          |
|          | CoAtNet (Hybrid) [153] | 94.92                   | 94.92        | 95.02        | 94.93        | 97.26        |
|          | HTC-Retina (Ours)      | <b>97.00</b>            | <b>97.00</b> | <b>97.04</b> | <b>97.01</b> | <b>99.56</b> |

percentages of 99.40%, 99.77%, and 97%, along with sensitivities of 99.41%, 99.77% and 97.00% on the OCT-2017, OCT-2014, and OCT-C8 datasets. These results surpassed CNN-based models and baseline approaches such as VGG-19, ResNet50, InceptionV3, ViT, and hybrid-based CNNViT models like ConViT, CoAtNet, LLCT, and HCTNet. Notably, in the OCT-2017 dataset, HTC-Retina outperformed ConViT, CoAtNet, LLCT, and HCTNet by 4.94, 1.58, 1.7, and 7.84 percentage points (pp.) in accuracy, respectively. In the OCT-2014 dataset, HTC-Retina surpassed SoT, CoAtNet, and HCTNet by 0.42, 0.92, and 13.59 pp., respectively. These experimental results show that the proposed HTC-Retina can achieve the desired classification performance on the OCT-2017, OCT2014, and OCT-C8 datasets, further validating the rationality of our model design.

The improvement brought about by HTC-Retina in comparison to a specific method is quantified by the ratio of the metric value of HTC-Retina to the corresponding value of that method. This improvement is denoted as “gain”, as all the evaluation metrics aim for

maximization. Using Table 4.3, we promptly compute this gain, for each metric. For instance, if the metric is accuracy and the SOTA method is ConViT, then the HTC-Retina accuracy gain with respect to ConViT is computed as  $99.40/94.46 = 1.0522$ . Note that any gain must exceed unity strictly; otherwise, it is regarded as a loss. This calculation is performed for overall accuracy, sensitivity, specificity, and all SOTA methods across all datasets. These gain values are visualized in charts in Fig. 4.5 . The findings are evident: All gains are higher than unity and therefore, HTC-Retina exhibits no loss compared to any of the SOTA methods, across all metrics and all datasets; one exception being the specificity of Pure ViT for the OCT-2014 dataset, being 99.87 compared to our method, yielding 99.80. However, even in this exceptional adversarial case, the gain remains very close to 1 (exactly 0.999).

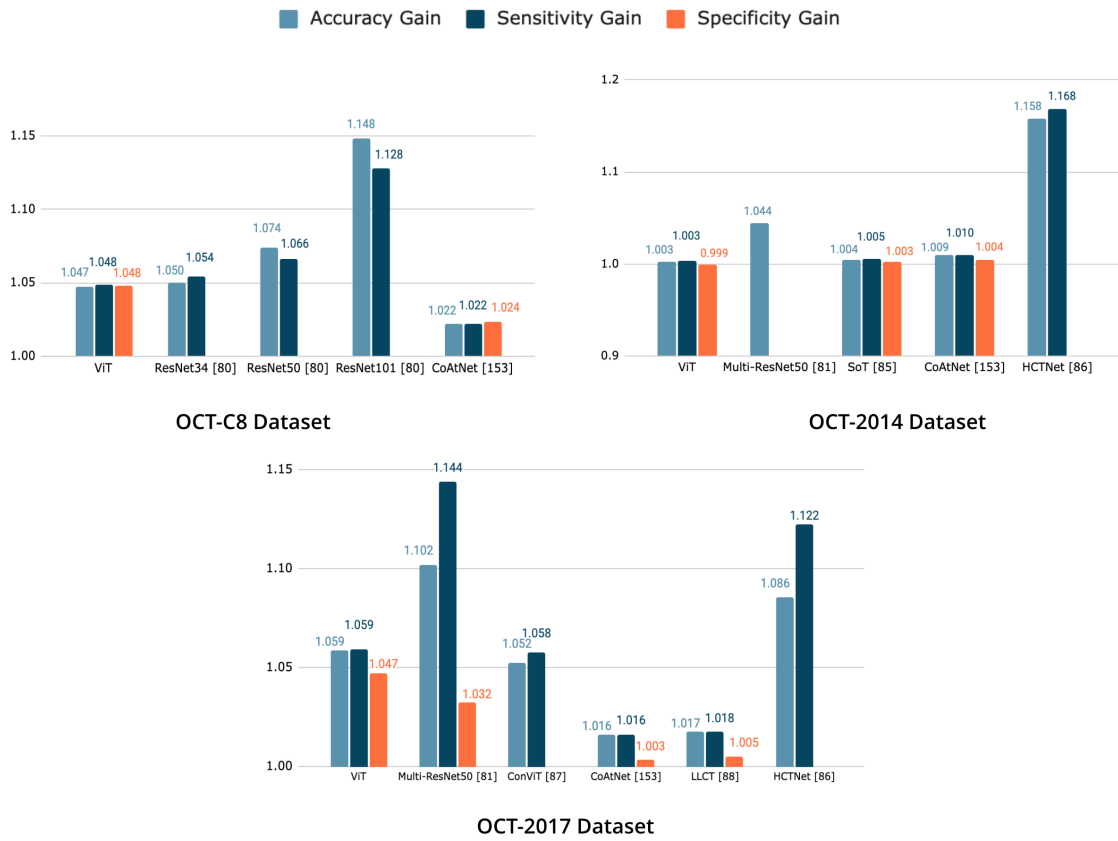


Fig. 4.5 HTC-Retina gains.

#### 4.2.4.6 Ablation studies

Ablation studies involve the deliberate removal or modification of specific components within a model to rigorously evaluate their impact to its overall performance. These studies are

important in ML for isolating and quantifying the contribution of each individual element on the model's effectiveness. Four ablation studies were conducted to validate the effectiveness of the proposed HTC-Retina model. These studies aimed to assess the impact of the CPTE and RDP-ConvNet modules on the performance of HTC-Retina in OCT scans. Furthermore, the influence of removing positional embeddings from the model was also examined. The effects of each modification on the proposed HTC-Retina model are presented in Table 4.4.

Table 4.4 Ablative studies on CPTE, RDP-ConvNet, and positional embeddings

| Model      | CPTE | PE  | ViTE | RDP-ConvNet | Accuracy (%) |       |          |       |        |       |
|------------|------|-----|------|-------------|--------------|-------|----------|-------|--------|-------|
|            |      |     |      |             | OCT-2017     |       | OCT-2014 |       | OCT-C8 |       |
|            |      |     |      |             | Val.         | Test. | Val.     | Test. | Val.   | Test. |
| HTC-Retina | No   | Yes | Yes  | No          | 92.75        | 93.90 | 99.49    | 99.52 | 92.58  | 92.65 |
|            | Yes  | Yes | Yes  | No          | 95.10        | 98.14 | 99.60    | 99.62 | 94.36  | 94.35 |
|            | Yes  | Yes | Yes  | Yes         | 97.92        | 99.03 | 99.71    | 99.72 | 96.65  | 96.70 |
|            | Yes  | No  | Yes  | Yes         | 98.30        | 99.40 | 99.80    | 99.77 | 96.99  | 97.00 |

The results presented in Table 4.4 provide valuable insights into the effects of these modifications on the performance of HTC-Retina across different datasets.

In the first two ablation studies, the HTC-Retina architecture was evaluated by excluding the RDP-ConvNet module while retaining the positional embeddings (PE) and ViTE modules. These studies aimed to examine the influence of the CPTE module on the overall performance of the HTC-Retina model. The results indicated a decline in accuracy across the validation and test sets. The test accuracy of the model exhibited notable gains, with improvements of 4.24% on the OCT-2017 dataset, 0.13% on OCT-2014, and 1.7% on OCT-C8. These findings indicate that the CPTE module holds significant importance in enhancing the model's performance by effectively capturing pertinent local features, mainly when dealing with small datasets.

The third ablation study was conducted to explore the influence of the RDP-ConvNet module on the HTC-Retina architecture. The experimental results consistently improved the model's accuracy across all the validation and test sets. The employment of the depthwise and pointwise convolution operations boosted the model's performance, resulting in test accuracy gains of 0.89% on OCT-2017, 0.10% on OCT-2014, and 2.35% on OCT-C8. This finding underscores the significance of the RDP-ConvNet module in effectively capturing fine-grained details and high-level local features, thereby enhancing the model's discriminative performance.

To validate the study as reported in [152] and further investigate the incorporation of enhanced local information within the HTC-Retina model, a final ablation study was conducted to assess the necessity of position embedding. The findings indicate that removing

position embeddings does not adversely impact the performance of the proposed model. On the contrary, their omission leads to an improvement in accuracy, with enhancements of 0.37% observed in the OCT-2017 dataset, 0.05% in OCT-2014, and 0.30% in the OCT-C8 dataset, as reported in [152]. Hence, the final architecture of the HTC-Retina model adopts the exclusion of position embedding.

#### 4.2.4.7 Noise robustness evaluation

To assess the robustness of the HTC-Retina model to Gaussian noise, we introduced noise to the test set of each dataset. The average Peak Signal-to-Noise Ratio (PSNR) between the original and the noisy versions was 28.4, indicating a moderate noise level. We also computed the Structural Similarity Index (SSIM) further to assess the similarity between the original and noisy images, obtaining a value of 57.9.

Without retraining the model on the noisy datasets, we directly evaluated its performance using the pre-trained model on the original datasets. The comprehensive evaluation metrics are presented in Table 4.5, illustrating the comparison between the model’s performance on the original and noisy datasets. The results indicate that the HTC-Retina model exhibits commendable performance in the presence of noise, showcasing its robustness compared to its performance on the original dataset.

Table 4.5 Comparison results on the noisy and original test set of each dataset.

| Datasets | Type     | Evaluation matrices (%) |                    |           |          |             |
|----------|----------|-------------------------|--------------------|-----------|----------|-------------|
|          |          | Accuracy                | Recall/Sensitivity | Precision | F1-Score | Specificity |
| OCT-2017 | Original | 99.40                   | 99.41              | 99.39     | 99.40    | 99.70       |
|          | Noisy    | 99.07                   | 99.08              | 99.06     | 99.07    | 99.62       |
| OCT-2014 | Original | 99.77                   | 99.77              | 99.77     | 99.76    | 99.80       |
|          | Noisy    | 98.13                   | 98.13              | 98.22     | 98.17    | 99.60       |
| OCT-C8   | Original | 97.00                   | 97.00              | 97.04     | 97.01    | 99.56       |
|          | Noisy    | 94.70                   | 94.71              | 94.79     | 94.70    | 99.23       |

#### 4.2.4.8 External validation

To evaluate the performance and generalizability of the HTC-Retina model, an external validation was conducted, encompassing the assessment of the model on independent datasets to ascertain its reliability, robustness, and suitability. Specifically, the pre-trained HTC-Retina model, originally trained on the OCT-C8 dataset, was evaluated on additional datasets. The performance of the pre-trained model on the OCT-2014 and OCT-2017 datasets is presented in Table 4.6.

Table 4.6 External validation results of the HTC-Retina model.

| Model                               | Datasets | Evaluation matrices (%) |             |           |          |             |
|-------------------------------------|----------|-------------------------|-------------|-----------|----------|-------------|
|                                     |          | Accuracy                | Sensitivity | Precision | F1-Score | Specificity |
| HTC-Retina<br>Pre-trained on OCT-C8 | OCT-2017 | 95.23                   | 95.30       | 95.23     | 95.23    | 95.22       |
|                                     | OCT-2014 | 94.37                   | 94.45       | 94.37     | 94.31    | 96.88       |

Based on the aforementioned results, the performance of the HTC-Retina model that pre-trained on OCT-C8 in disease prediction was notable when applied to two distinct OCT datasets: OCT-2017, comprising cases of CNV, DME, DRUSEN, and Normal, and OCT-2014, consisting of cases of AMD, DME, and Normal. Therefore, it is evident that the model demonstrates accurate classification capabilities, yielding reliable predictions for OCT scan classification tasks. This evaluation proves that our model exhibits independence from specific data samples, demonstrating its robustness and generalizability to diverse datasets. The results indicate the model's capacity to effectively accommodate variations in image characteristics, acquisition settings, and potential temporal changes between different datasets.

#### 4.2.5 Discussion

The main objective of our study was to develop an innovative and precise hybrid model, named HTC-Retina, for the classification and identification of retinal diseases using OCT scans. Inspired by many recent research [152, 153, 86, 88, 160], the proposed HTC-Retina model was constructed by integrating ViT and CNN architectures. This synergistic approach effectively addressed the limitations inherent in each architecture and capitalized on their respective strengths, resulting in a highly proficient model for the classification task of OCT scans. In contrast to SOTA hybrid models [85, 86, 88], which incorporate CNN blocks preceding the ViT architecture, our proposed HTC-Retina model directly utilizes the output sequence of tokens of the CPTE module as input to the ViT encoder, eliminating patch splitting, projection, and positional embedding. This novel approach enhances model flexibility, accommodating input resolutions beyond strict divisibility by the pre-set patch size. The CPTE module efficiently embeds the image, generating more nuanced tokens for the transformer. Additionally, the CPTE module is a straightforward CNN network, characterized by its simplicity and reduced complexity, in contrast to the module employed in [86, 87], which integrated dense blocks and residual blocks. Regarding the ViT component of our proposed model, the final configuration comprises merely four encoder blocks, in contrast to HCTNet



[86] and LLCT [88], both employing 12 blocks. This design choice renders our model less complex. As for the remaining components the proposed model employs the same structure of the HCTNet but using different techniques. While HCTNet incorporates a parallel convolutional branch with residual dense blocks and a fully connected layer, HCT-Retina adopts a lightweight architecture, based on depthwise and pointwise convolution operation (RDP-ConvNet module), akin to ViT. Operating directly on patches as input, RDP-ConvNet aims to capture relevant local features through a less expressive operation. For the combination module, HTC-Retina utilizes global average pooling which reduces dimensionality, enhances translation invariance, improves generalization, and reduces sensitivity instead of a re-weighting mechanism. The network further simplifies and reduces complexity by directly concatenating output features, facilitating access to multi-scale features.

To evaluate the effectiveness of the proposed HTC-Retina, a comparative analysis was conducted across three distinct datasets, benchmarking against SOTA methods, including hybrid ViT-CNN models, CNN models, and baseline approaches. The results, as detailed in Section 4.2.4, underscore the efficacy and feasibility of HTC-Retina in classifying OCT diseases, surpassing other SOTA methods. Additionally, to validate its remarkable performance, the model underwent testing on noisy versions of the datasets and external validation. The notable performance in these scenarios underscores the robustness and generalizability of HTC-Retina across diverse datasets.

Despite the successes demonstrated by our proposed hybrid model, HTC-Retina, it is essential to acknowledge certain limitations inherent in our study. The primary constraint lies in the lack of diversity and abundance in real sample datasets, necessitating the integration of augmented data to compensate. This potential source of bias may influence the model's robustness in real-world scenarios. Additionally, the model's performance, particularly in diseases with shared characteristics (CNV, DME, DRUSEN) or diseases with tiny lesions that require very great scrutiny, even for an expert, encounters difficulties, particularly in the presence of numerous other diseases. The fine-tuning process for the hybrid architecture poses another challenge, demanding meticulous attention to hyperparameter optimization. Achieving an optimal balance between transformer and CNN components may necessitate extensive experimentation, impacting the scalability of the model.

## 4.3 Part II: Intraretinal cystoid fluid detection and segmentation using DA-UNet-Plus

### 4.3.1 Motivation for DA-Unet-Plus development

Most available analysis software tools for OCT are limited to measuring retinal layer thickness, despite prior studies demonstrating the limited value of this biomarker for visual prognosis and disease management [164]. In practice, most physicians rely on qualitative OCT biomarkers, such as the presence of Intraretinal Cystoid Fluid (IRC), to inform retreatment decisions in individualized therapy for macular diseases [165]. These biomarkers are also sophisticated prognostic factors for visual outcomes, with IRC being one of the most significant variables associated with vision loss. The IRC is a critical marker for various retinal diseases, including DME and AMD. Therefore, accurate detection and segmentation of IRC are essential for effective diagnosis, treatment planning, and monitoring of these conditions. Traditional manual segmentation methods are time-consuming and prone to inter-observer variability, underscoring the need for automated solutions.

Traditional methods often struggle with the complex and variable nature of retinal structures, resulting in suboptimal segmentation outcomes. However, with advancements in DL approaches, particularly CNNs and attention mechanisms, the segmentation process can be automated and more effective. The U-Net framework, a commonly used CNN architecture for segmentation in the medical imaging field, has proven to be robust and efficient. The U-Net architecture excels by capturing low-level features, which provide information about the shape of each class, and high-level features, which utilize this information to determine class assignments for each position. This combination allows the model to restore the spatial dimensions of the original input and accurately classify each location [25].

Building on this framework as both the baseline and backbone, integrating advanced attention mechanisms such as dual attention (DA) and squeeze-and-excitation (SE) attention can result in a powerful and efficient model. We show that this integration addresses the need for precise and effective detection of fluid regions within retinal images. The main objective is to leverage these advanced attention mechanisms to enhance the U-Net performance and improve the model's robustness.

### 4.3.2 Contribution of DA-UNet-Plus

The second contribution of this chapter is the DA-UNet-Plus model, which is designed to detect and identify IRC from OCT images. The DA-UNet-Plus model is a hybrid architecture

that combines the strengths of the traditional U-Net architecture with the innovative dual attention network (DANet) head, incorporating position and channel attention mechanisms and squeeze-and-excitation residual blocks based-SE attention. The detailed architecture of the DA-UNet-Plus is described in Section 4.3.3.1.

### 4.3.3 Methodology

#### 4.3.3.1 DA-UNet-Plus structure

The U-Net [25] is a well-established fully convolutional network frequently utilized for medical image segmentation. Its design features a contracting path and an expanding path, facilitating precise localization of features and efficient network training. This architecture forms the backbone of the proposed DA-UNet-Plus model. In the DA-UNet-Plus model, rather than employing the standard convolution block as a bottleneck in the U-Net architecture, the integration of a DANet head [166] is preferred. This substitution enhances the model by incorporating the powerful representation capabilities of the position attention module (PAM) and channel attention module (CAM). The U-Net's contracting and expansive paths effectively address the challenge of combining low-level and high-level features, thereby overcoming the limitations inherent in the standalone DANet architecture. This integrated approach ensures improved segmentation performance, particularly in medical imaging tasks. The DA-UNet-Plus model consists of an encoder, a bottleneck, and a decoder as illustrated in Fig. 4.6.

##### a. Encoder

The encoder is designed to progressively reduce spatial dimensions while capturing intricate features. It consists of three stages, each stage comprises convolutional blocks with ReLU activation and residual connections integrated with SE attention blocks, which adaptively recalibrate channel-wise feature responses. The convolutional blocks apply convolutional filters of size 64, 128, and 256 with kernel sizes of 3 respectively, followed by batch normalization and ReLU activation layer. Fig 4.7 displays the detailed convolutional block architecture.

As for the residual attention blocks, they use the skip connection as in the ResNet architecture [19] with the inclusion of an SE attention layer. The residual attention blocks apply the same filter sizes as the previous convolutional block. Fig. 4.8 illustrates the detailed architecture of the residual attention block. The inclusion of the residual-based-SE-attention blocks aims to enhance feature recalibration and improve gradient flow.

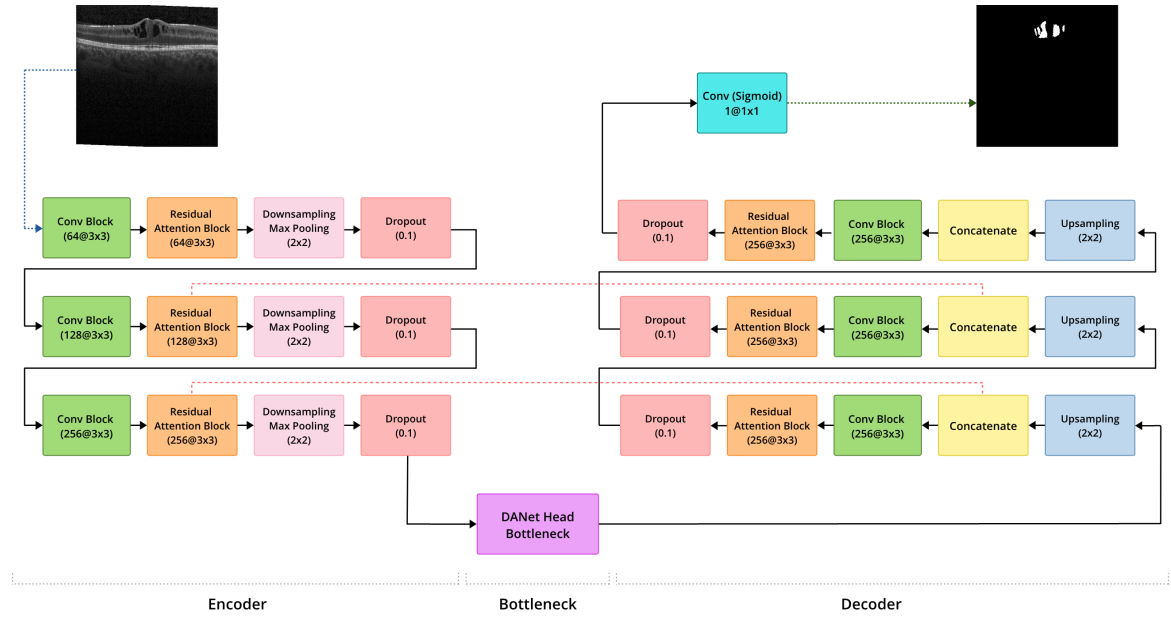


Fig. 4.6 **The overall architecture of the DA-UNet-Plus model.** First, the input is fed to the encoder which consists of three convolutional blocks, each followed by a residual attention block, max-pooling layers, and dropouts. Next, the output of the encoder is passed through a DANet head bottleneck. After, the bottleneck's output is fed to a decoder whose path mirrors the encoder but in reverse. Lastly, the final convolutional layer reduces the number of filters to match the number of target classes.

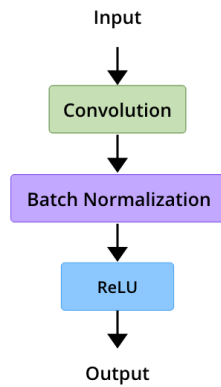


Fig. 4.7 **The convolutional block diagram.**

The encoder also includes max-pooling layers with a stride size of 2 for the downsampling process, and dropout layers to prevent overfitting in each step.

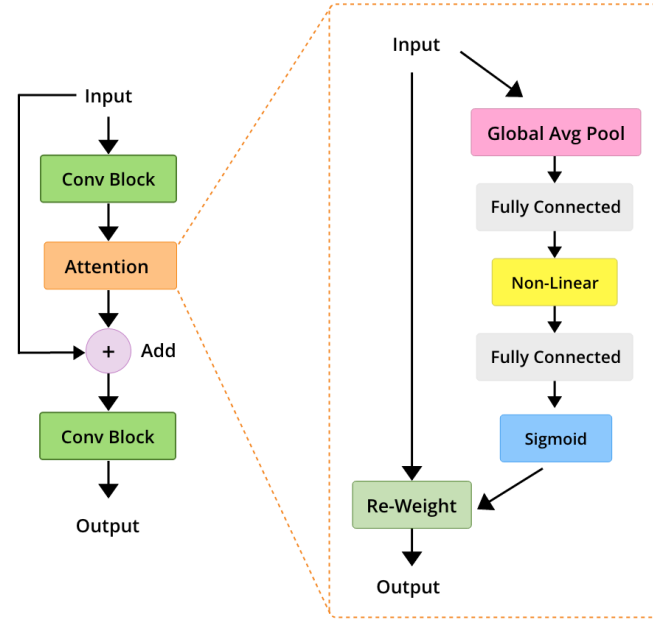


Fig. 4.8 The residual attention block diagram.

Overall, the encoder path progressively reduces the spatial dimensions of the input while increasing the depth of feature maps, enabling the model to capture complex features at various scales.

#### b. Bottleneck based-DANet head

The model's bottleneck consists of a convolutional layer followed by a DANet head, which consists of position and channel attention modules. These modules refine feature maps by focusing on relevant spatial regions and channels, enhancing the model's ability to capture intricate details in the feature maps. Fig. 4.9 presents the detailed architecture of the bottleneck.

The DANet is a correlation-based attention network designed to integrate local features with their global dependencies, thereby capturing rich contextual relationships to address the intra-class variability problem.

- The PAM focuses on spatial attention by considering the weighted sum of all features at each position. These weights are determined by the correlation between the current position and every other position, with higher weights assigned to positions with similar features. This mechanism ensures that any two positions with similar features can mutually enhance each other, regardless of their spatial distance.

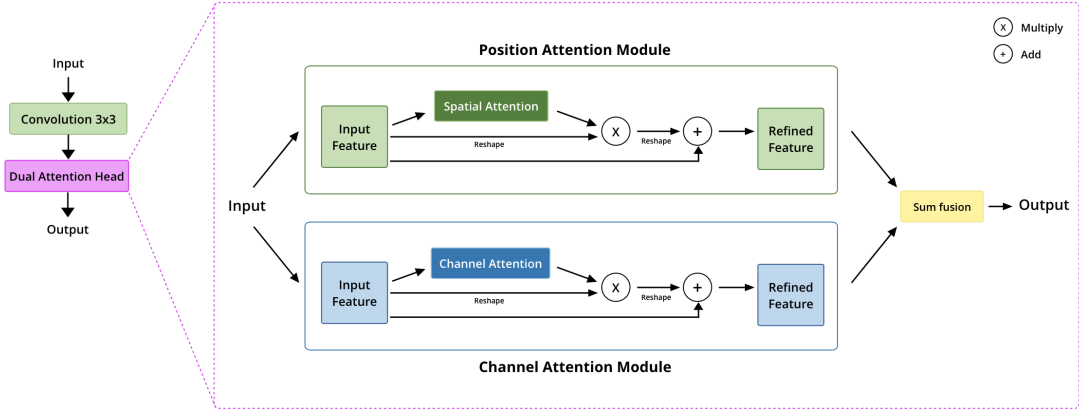


Fig. 4.9 The DANet head bottleneck block diagram.

- The CAM captures the dependencies between different channel maps. It updates each channel map with a weighted sum of all channel maps, using a similar correlation-based approach to that of the PAM.

This dual attention mechanism enables the DANet to effectively integrate spatial and channel-wise contextual information, leading to improved performance in tasks that require detailed feature analysis, such as retinal fluid segmentation.

### c. Decoder

The decoder path mirrors the encoder but in reverse, gradually restoring the original spatial dimensions of the input while reducing the depth of feature maps. It employs upsampling layers and concatenates the corresponding encoder features to preserve spatial information. The decoder includes three stages, each with upsampling, concatenation with encoder features, convolutional layers, residual attention blocks, and dropout layers. The final output layer uses a 1x1 convolution with a sigmoid activation function to produce a pixel-wise probability map for the target segmentation classes.

#### 4.3.3.2 Dataset

The **Intraretinal Cystoid Fluid Dataset** [96] is a publicly available collection<sup>4</sup> specifically designed for image segmentation tasks, focusing on cystoid macular edema ocular disease. This dataset was derived from the Mendeley OCT Dataset, containing 1000 images extracted from DME subjects. Additionally, it includes 200 images sourced from the Institute of

<sup>4</sup>IRC Dataset: [kaggle.com/datasets/zeeshanahmed13/intraretinal-cystoid-fluid](https://kaggle.com/datasets/zeeshanahmed13/intraretinal-cystoid-fluid)

Ophthalmology, Jamshoro. This dataset contains the image with the associated mask which presents the ground truth. We refer to this data as the OCT-IRC-Seg.

#### **4.3.3.3 Preprocessing**

Retinal images obtained through OCT devices rely on the interferometric detection of optical coherent beams, characterized by high spatial coherence and low temporal coherence. This results in significant speckle noise within the acquired OCT images. To mitigate these noise effects and enhance image quality, several image preprocessing steps are applied to the OCT-IRC-Seg dataset sequentially to ensure a well-formatted structure suitable for accurate segmentation by the DA-UNet-Plus model.

##### **a. Normalization**

The intensity values of the images are normalized to a standard range, using the max-min normalization. This step ensures that the pixel values are scaled consistently, improving the convergence and stability of the model during training.

##### **b. Resizing**

The original OCT images are resized to a consistent dimension to standardize the input size for the DA-UNet-Plus model. This step ensures uniformity in image dimensions, facilitating efficient processing and model training.

##### **c. Morphological operation**

A minimum filter is applied to the resized images to reduce speckle noise. This morphological operation helps in smoothing the image by replacing each pixel value with the minimum value in its local neighborhood, effectively diminishing noise while preserving essential structural details.

##### **d. Image contrast**

Finally, the contrast of the preprocessed images is adjusted to ensure that the important features are well highlighted. Enhancing image contrast improves the differentiation of cystoid fluid regions from the surrounding retinal tissue, aiding the DA-UNet-Plus model to accurate segmentation.

### 4.3.4 Experiments and results

This section delineates the experimental settings, results, and key findings of the DA-UNet-Plus model. The subsequent subsections offer a comprehensive overview of the experiment and results on the OCT-IRC-Seg dataset.

#### 4.3.4.1 Experimental settings

The DA-UNet-Plus was built using the TensorFlow framework and the experiments were conducted using an NVIDIA Tesla P100 GPU. The model was trained to optimize the hybrid dice coefficient and binary cross-entropy loss. This loss is defined as:

$$Loss = L_{Dice} + L_{BCE} \quad (4.11)$$

The Adam algorithm was employed as the optimizer for training the model. The training process spanned 50 epochs with a batch size of 20. The OCT-IRC-Seg dataset was divided into 80% for the training set and 20% for the validation set. The test phase was conducted on a test set comprising approximately 200 images.

The performance of the DA-UNet-Plus model is evaluated using accuracy, precision, recall, dice similarity coefficient (DSC), and intersection over union (IoU) (*aka* Jaccard index, as used in statistics and set theory).

#### 4.3.4.2 Evaluation results

To examine the fine-grained performance of the model, we present a comprehensive evaluation report of the DA-UNet-Plus model in Table 4.7. This table shows the performance of DA-UNet-Plus in detecting and segmenting IRC.

Table 4.7 DA-UNet-Plus performance in IRC detection

| Dataset     | Evaluation metrics (%) |        |           |       |       |
|-------------|------------------------|--------|-----------|-------|-------|
|             | Accuracy               | Recall | Precision | DSC   | IoU   |
| OCT-IRC-Seg | 99.84                  | 88.20  | 86.37     | 87.30 | 77.46 |

Based on Table 4.7, the reported results demonstrate the good performance of the DA-UNet-Plus across various key metrics, indicating the effectiveness of our approach in intraretinal cystoid fluid detection and segmentation.



#### **a. Accuracy**

In terms of accuracy, the model achieved an exceptional accuracy of 99.84%. This highlights the model's ability to correctly classify the majority of the pixels in the OCT images. This high accuracy underscores the robustness of the model in handling the inherent noise and variability present in OCT images, ensuring reliable detection of cystoid fluid regions.

#### **b. Recall**

In terms of recall, the model achieved a score of 88.20%. This exhibits a strong sensitivity to true positives, effectively identifying the majority of the cystoid fluid regions. The high recall ensures that most of the fluid regions are detected, contributing to comprehensive diagnostic insights.

#### **c. Precision**

In terms of precision, the model achieved a score of 86.37%. This highlights the model's ability to accurately identify cystoid fluid regions without a high rate of false positives.

#### **d. Dice similarity coefficient**

In terms of DSC, a score of 87.30% illustrates the model's proficiency in accurately overlapping the predicted segmentation maps with the ground truth. The high dice coefficient value indicates that the model's segmentation maps are closely aligned with the actual fluid regions, demonstrating effective boundary delineation and region identification.

#### **e. IoU or Jaccard index**

An IoU of 77.46% further confirms the model's capability in overlapping predicted and actual segmentation regions. The IoU metric, which measures the overlap between the predicted and true segmentations relative to their union, showcases the model's overall performance in capturing the true extent of the fluid regions.

Overall, these metrics collectively validate the efficacy of the DA-UNet-Plus model in intraretinal cystoid fluid detection and segmentation.

#### **4.3.4.3 Correlation plot**

Fig. 4.10 presents the correlation plot which illustrates the relationship between the predicted sizes of intraretinal cystoid fluid regions and the corresponding ground truth sizes in the

OCT images. Each point on the scatter plot represents an individual prediction, with the x-axis denoting the ground truth size and the y-axis denoting the predicted size. The diagonal dashed line represents the ideal scenario where the predicted size perfectly matches the ground truth size.

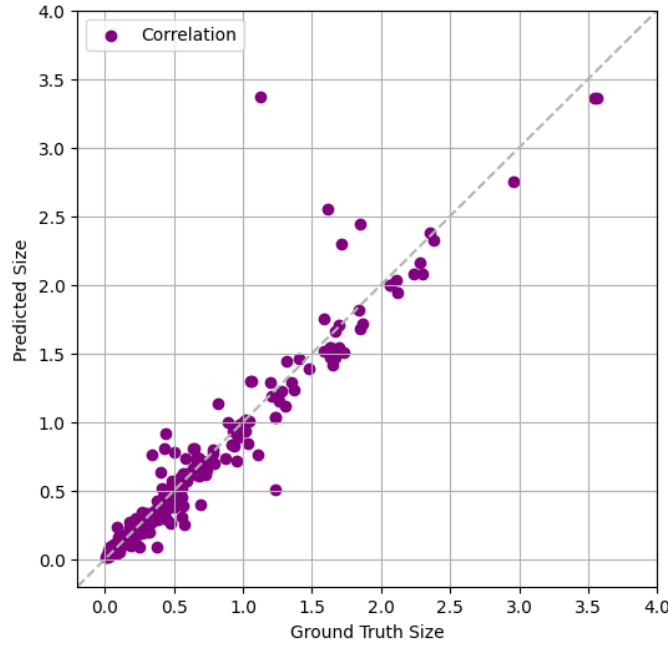


Fig. 4.10 Correlation between predicted and ground truth sizes of IRC regions.

#### Interpretation of the graph

- *Proximity to the diagonal line:* The data points that lie close to the diagonal line indicate a strong positive correlation between the predicted and actual sizes. Based on the figure the majority of the data points are close to the diagonal line, suggesting that the model is highly accurate in predicting the size of cystoid fluid regions, as the predictions closely match the ground truth.
- *Scatter distribution:* Some points exhibit deviation from the line. These deviations highlight instances where the model either overestimates or underestimates the size of the cystoid fluid regions. However, the overall trend demonstrates a consistent and reliable predictive performance.

Overall, the correlation plot signifies a good performance in the model's predictions, as evidenced by the dense clustering of points along the diagonal. The presence of some outliers is expected in complex cases, but their relative scarcity underscores the model's robustness.

This strong correlation between predicted and actual sizes validates the model's efficacy in accurately detecting and quantifying intraretinal cystoid fluid regions, which is crucial for effective clinical assessment and treatment planning

#### 4.3.4.4 Results visualization

Ensuring clinicians, researchers, and patients understand, trust, and effectively utilize artificial intelligence technologies in healthcare hinges on the interpretability and validation of DL models. Fig. 4.11 visualizes the segmentation results of the DA-UNet-Plus model. This figure presents the OCT image in the first row, the ground truth in the second row—created by overlaying the mask image on the OCT image—and the model prediction in the third row, generated by overlaying the predicted mask on the OCT image. The highlighted regions represent the cystoid fluid detected within the retinal layers. As observed in the figure, the predicted masks closely match the ground truth masks, indicating the model's effectiveness in detecting and segmenting the fluid regions. However, some discrepancies between the predicted and actual labels can be observed, which suggests areas for potential improvement in the model's segmentation accuracy.

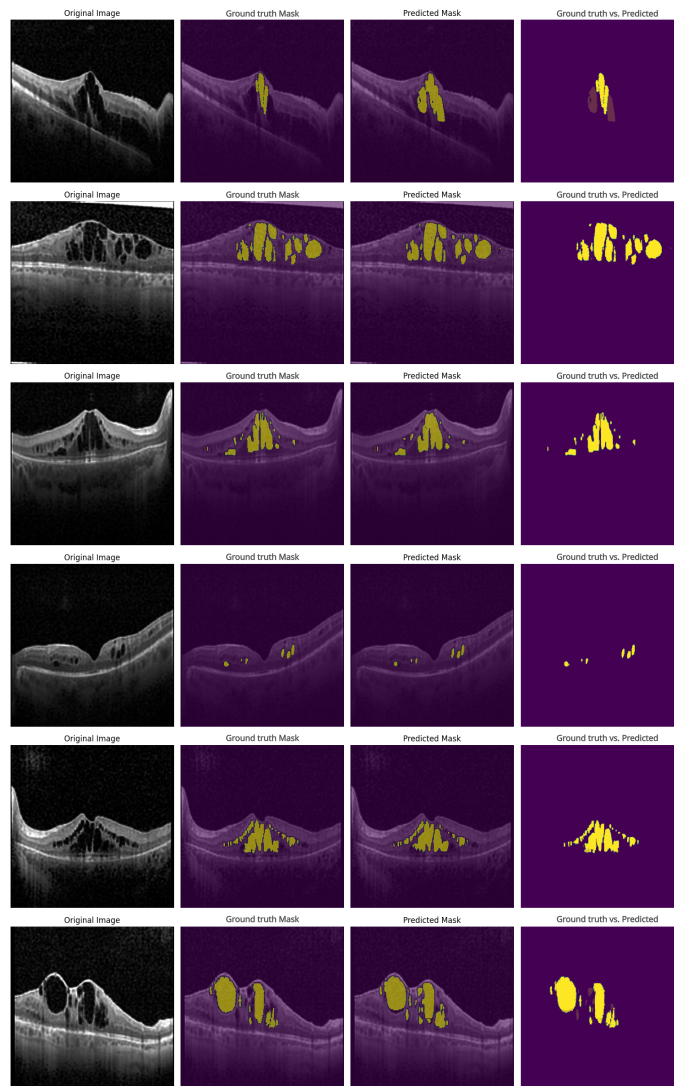
#### 4.3.4.5 DA-UNet-Plus performance vs. SOTA models

This section provides a comparison between the DA-UNet-Plus and other SOTA models used for the same task. Fig. 4.12 shows the performance of different SOTA models in the detection of the cystoid fluid regions and compared to the proposed DA-UNet-Plus based on the dice coefficient score.

The figure shows that the DA-UNet-Plus model surpasses the SOTA models, achieving a higher Dice score of 87.3%. The model outperforms the SOTA models, sometimes by a significant margin. Specifically, an improvement of 4% has been recorded compared to the closest competitor.

### 4.3.5 Discussion of DA-Unet-Plus results

In this study, we assessed the performance of a novel hybrid DL model for the automated detection, segmentation, and quantification of the IRC. The novelty of the proposed model lies in the hybridization of the baseline UNet model with other models such as DANet and the incorporation of advanced techniques such as residual connections and attention mechanisms. This hybridization led to accurate segmentation of the cystoid fluid regions in the OCT image. The proposed model was trained and validated on 1200 OCT images and tested on 200



**Fig. 4.11 Example cases of segmentation results of the DA-UNet-Plus model.** The first column of subfigures shows the OCT images, the yellow region in the second column shows the ground truth mask, and the yellow region in the third column shows the DA-UNet-Plus segmentation results.

samples. The evaluation of the model was based on different metrics including the dice coefficient score, precision, recall, accuracy, and IoU.

In terms of performance, the achieved results described and detailed in the 4.4.2 section demonstrate the DA-UNet-Plus's efficacy, with high accuracy, competitive precision, and recall rates. The dice score and IoU which represent crucial metrics in the medical images segmentation, further corroborate the model's robustness in segmenting cystoid fluid regions accurately. Moreover, the visual results (Figure 3.9) also provide compelling evidence of the model's capability. The comparison between ground truth and predicted masks reveals that

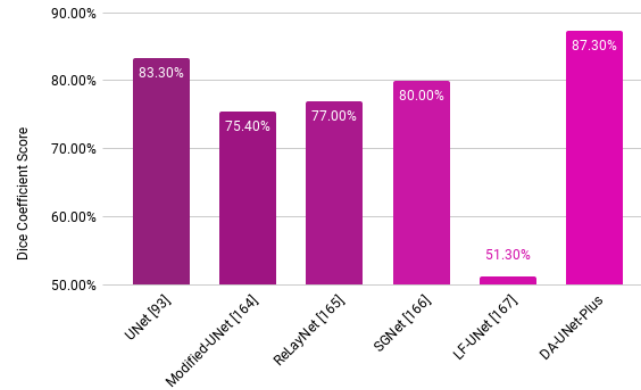


Fig. 4.12 Comparison of DA-UNet-Plus performance and SOTA models.

DA-UNet-Plus can effectively segment fluid regions, closely aligning with the actual. This indicates that the model is adept at learning complex patterns and textures inherent in OCT images.

In summary, DA-UNet-Plus represents a significant step forward in medical image analysis, particularly for retinal OCT images. However, the model has some inconsistencies and differences between predictions and ground truth, which opens the door to further improvements and extensive experiments.

## 4.4 Conclusion and future work perspectives

This chapter presented two main contributions toward retinal disease diagnosis and detection using advanced innovative DL techniques. We describe the strengths of both models and some potential future improvements.

### 4.4.1 HTC-Retina model

#### 4.4.1.1 HTC-Retina strengths

HTC-Retina is a hybrid ViT-CNN model combining transformer and CNN architectures, specifically designed for retinal OCT image classification on small and medium datasets. The objective is to accurately identify retinal diseases. It was proposed and tested for the diagnosis of retinal diseases using different public datasets. This model leverages the strength of ViTs and CNNs to highly classify retinal diseases by deeply analyzing the OCT image and capturing the most relevant features of each disease. Our study highlighted the superior performance of this model compared to the baselines and SOTA methods. The

HTC-Retina underwent an extensive validation process, including external validation and robustness evaluations. During these assessments, it demonstrated exceptional competence in diagnosing retinal diseases and showcased its generalization ability on new data.

Through various experiments, the CPTE and RDP-ConvNet modules were identified as critical components that significantly enhance the performance of the HTC-Retina model. Our model performance was evaluated using the OCT-2017, OCT-C8, and OCT-2014 datasets, achieving remarkable results. Notably, HTC-Retina outperformed existing models while maintaining computational efficiency. The model no longer relies on position embedding, as the CPTE and RDP-ConvNet combined with the ViT capture more local features, thus improving classification accuracy.

Further validation and robustness evaluations demonstrated the model's generalization ability and performance under varying conditions. The HTC-Retina model exhibited excellent performance on new data and maintained robustness when exposed to external noise, underscoring its practical applicability.

#### 4.4.1.2 HTC-Retina's potential future improvements

The work identifies several key areas for potential HTC-Retina future enhancements, described below.

- *Complexity*: The hybrid architecture presents significant challenges in both fine-tuning and practical application. Due to its intricate structure, optimizing hyperparameters requires meticulous attention, making the fine-tuning process particularly demanding. This complexity can affect the model's scalability and ease of use. Additionally, balancing the transformer and CNN components within the architecture often necessitates extensive experimentation, particularly when applied to complex datasets. This may hinder the model's practical deployment in real-world scenarios, where efficiency and adaptability are crucial.
- *Data diversity and bias*: There is a limitation in the diversity and abundance of real sample datasets, leading to reliance on augmented data. This could introduce bias and affect the model's robustness in real-world scenarios.
- *Misclassification issues*: The model struggles with diseases that share characteristics (e.g., CNV, DME, DRUSEN) and those with tiny lesions, especially in cases involving other retinal pathologies, which are challenging to detect even for experts. These limitations highlight the model's difficulties in handling complex cases, underscoring the need for additional training data to enhance its performance and improve its ability to differentiate between similar pathologies.

Achieving an optimal balance between the transformer and CNN components may necessitate extensive experimentation, impacting the scalability of the model, and thus requiring a careful tradeoff between these conflictual objectives.

## 4.4.2 DA-UNet-Plus model

### 4.4.2.1 DA-UNet-Plus strengths

In the second part of this chapter, a hybrid model called DA-UNet-Plus was introduced for segmenting intraretinal cystoid fluid from OCT images. It enhances the baseline UNet by incorporating a DANet head, residual attention blocks, and advanced techniques like residual squeeze-and-excitation (SE) blocks and dual attention mechanisms. These improvements improve the model's feature extraction capabilities and segmentation performance, as shown in the experiments.

The DA-UNet-Plus performance was indeed evaluated using the OCT-IRC-Seg dataset, achieving notable results and overcoming the SOTA methods. The qualitative results underline the model's capability to produce segmentation outputs that closely align with ground truth annotations. The correlation analysis between predicted and ground truth fluid sizes underscores the model's reliability in volumetric assessment. These findings illustrate the robustness and effectiveness of DA-UNet-Plus in handling the complexities of OCT images.

Our study demonstrates the good performance of the DA-UNet-Plus, which outperformed all other SOTA methods and approaches with a dice score of 87.3%. The results illustrate that the hybridization of different models and techniques can yield interesting results and significant advancements in medical image segmentation, especially in our case study of intraretinal cystoid fluid detection and quantification.

### 4.4.2.2 DA-UNet-Plus's potential future improvements

The use of DA-UNet-Plus unveiled several points for potential future improvements, briefly described below.

- *Complexity and computational overhead:* The integration of advanced attention mechanisms, such as dual attention network (DANet) and squeeze-and-excitation (SE) blocks, adds complexity to the model. While this enhances performance, it also increases the computational burden, resulting in longer training times. Though this limitation could impact real-time or resource-constrained environments, advancements in cloud-based solutions offer a potential solution to offset this complexity. However, such cloud

resources can be expensive, making it a trade-off between computational efficiency and cost.

- *Model Interpretability*: While DA-UNet-Plus achieves high segmentation accuracy, the inherent complexity of DL models can pose challenges to interpretability. In clinical settings, this lack of transparency can hinder adoption, as healthcare professionals require a clear understanding of the decision-making process behind the model's outputs to ensure trust and responsible use.
- *Generalization and dataset diversity*: Despite the strong performance demonstrated on specific datasets, there is still room to improve the model's generalization capabilities across different types of retinal diseases and ophthalmic imaging modalities. To enhance its robustness and broaden its clinical applicability, DA-UNet-Plus must be evaluated on larger, more diverse datasets. Expanding the scope of evaluation will help improve its performance across a wider range of clinical cases, including other diseases involving cystoid fluid conditions, such as age-related macular degeneration and retinal vein occlusion.

These limitations highlight the necessity for continued refinement and rigorous testing to reinforce the DA-UNet-Plus generalization capabilities, robustness, and practical applicability in more general clinical environments. On a more applied level, future research should aim at enhancing its generalizability and exploring its application across various ophthalmic imaging modalities and other diseases containing cystoid fluid conditions such as age-related macular degeneration and retinal vein occlusion.



# Chapter 5

## Enhanced skin lesion diagnosis and localization using dermatoscopic images

This chapter focuses on the study of skin lesions—areas of the skin characterized by abnormal growth or appearance relative to surrounding tissue. Similar to Chapter 3, this chapter is divided into two parts. The first part introduces the novel JILDYA-Net model, designed for the localization and diagnosis of skin lesions from dermatoscopic images. The second part explores the adaptation of the DA-UNet-Plus model, introduced in the previous chapter, for enhanced segmentation of these lesions.

### 5.1 Introduction

Some skin lesions can become malignant and develop into cancer. Indeed, skin cancer is a disease in which malignant cells are randomly spread in the skin tissues. Among all cancers, skin cancers are the most common, with incidence rising faster than that of any other cancer; with melanoma being the most aggressive form of skin cancer. Melanoma occurs when there is a dysfunction in the melanin-producing cells. This disease can grow in many body areas such as the back, legs, arms, and face. According to the American Cancer Society, over 60,000 new cases of melanoma are reported annually, making it the sixth most common cancer in men and the seventh in women [167]. Typical signs of melanoma include changes in existing moles or the development of new pigmented or unusual growths on the skin, although it can also appear on normal-looking skin [168]. Given the increasing incidence of melanoma, particularly among individuals under 40, it is imperative to develop advanced methods for early detection and diagnosis. Therefore, monitoring skin lesions through

periodic medical imaging can be crucial for early intervention, treatment, and prevention of the spread of this disease.

Detecting melanoma among benign skin lesions and other skin cancers is still a significant challenge in dermatology as a variety of types of skin lesions can exhibit the visual characteristics of melanoma. Thus, relying solely on visual examination by the naked eye presents significant challenges and requires the expertise of a highly skilled dermatologist. The advancement of medical imaging techniques, particularly the use of dermoscopy, has improved melanoma diagnosis compared to direct visual inspection. However, the dermatologists' accuracy in identifying melanoma using dermoscopic images in typical clinical settings remains below 80% [169]. Therefore, the development of an automated system for the detection and identification of skin lesions, particularly melanoma, is highly recommended. Such a system can significantly aid dermatologists in diagnosis and treatment planning, enhancing both accuracy and efficiency in clinical practice.

In this chapter, we address the diagnosis of eight skin lesions, including melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis (AK), benign keratosis lesion (BKL), dermatofibroma (DF), vascular lesion (VASC), squamous cell carcinoma (SCC), using a novel efficient lightweight hybrid oriented architecture.

The first part introduces our proposed JILDYA-Net model, a novel approach developed to diagnose eight skin diseases. The concept of this model is inspired by extensive literature study. Following this, the second part describes our work on determining the location of the skin lesion spread using a hybrid segmentation model, namely the proposed DA-UNet-Plus segmentation model which was introduced in the previous chapter.

## **5.2 Part I: Skin lesions classification using JILDYA-Net**

### **5.2.1 Motivation for JILDYA-Net development**

Skin lesion classification has become increasingly important yet remains challenging due to the extensive time required for physicians to manually analyze very similar lesions. It is inherently complex due to factors like artifacts, image resolution disparities, and subtle differences among various cancer forms. However, advancements in ML algorithms and computer vision techniques offer promising solutions. These technologies leverage powerful feature extraction capabilities and robust image analysis methods. Indeed, CNNs, Attention Mechanisms, and Transformer-based approaches have shown significant promise in various medical image classification tasks. Notably, MLPs inspired by the Transformer principle,

such as MLP-Mixer, have emerged as SOTA models, demonstrating rapid learning and reduced complexity, with models like FNet achieving noteworthy outcomes.

Integrating these approaches forms a robust framework for lightweight DL methods, which are increasingly gaining attention for their performance in classifying skin lesions. The development of diagnostic systems focuses on several critical factors, including model performance, complexity, and prediction speed. Consequently, these architectures are designed to be computationally efficient, ensuring rapid inference times and minimal memory usage. This efficiency makes them well-suited for deployment on devices with limited processing capabilities, such as mobile phones and edge devices. Despite their lightweight design, these architectures exhibit remarkable versatility, handling diverse datasets with varying lesion complexity. They adapt to different input image sizes and effectively classify a wide range of skin lesions. Furthermore, their real-time processing capabilities enable swift decision-making, crucial for applications like telemedicine and point-of-care diagnostics.

Regarding deployment, lightweight architectures offer portability across various devices, including smartphones and Internet of Things devices. This accessibility enhances the reach of skin lesion classification tools, facilitating broader use in clinical and remote settings. However, challenges remain, particularly in balancing model size and classification accuracy and adapting to new lesion types not present in training data. Despite these challenges, the benefits of lightweight architectures in terms of efficiency, versatility, and accessibility make them a promising solution for advancing dermatology and healthcare practices. Therefore, the objective is to address the scarcity of lightweight architectures, by using advanced lightweight and effective DL methods.

### 5.2.2 The main contribution of JILDYA-Net

We introduce JILDYA-Net, a novel, lightweight, and effective model that integrates the attention mechanism and the FNet as an alternative architecture to traditional transformer models. JILDYA-Net leverages FNet due to its reduced computational complexity, particularly for long sequences. Unlike transformers, which rely on self-attention mechanisms and suffer from quadratic computation scaling with sequence length, FNet utilizes the Fourier Transform to process sequences.

The proposed approach comprises two pivotal components: an attention convolutional block designed to enhance structural features in input images, particularly in dermoscopic images known for their diverse features. Specifically, we introduce a novel convolutional attention component called attention-based structural feature enhancement (ASFE) to enhance skin lesion features, ensuring the capture of corresponding features in each patch. Following this, an encoder-based FNet [38] facilitates faster processing and reduces memory usage,

which is especially beneficial for longer input sequences. Additionally, an external attention module (EAM) [173] is integrated to refine learned representations and emphasize relevant features, thereby accelerating convergence and enhancing model stability during training.

### 5.2.3 Methodology

#### 5.2.3.1 JILDYA-Net overall structure

The JILDYA-Net model is composed of two main parts: an ASFE component, and an encoder-based FNet with an EAM. We first utilize an attention guidance component to enhance the skin lesion features. Then, the output is passed through patch projection and position embedding to form overlapping tokens used as input to the encoder-based FNet. The encoder-based FNet models the relationship between all tokens, thereby obtaining the semantics of the positional relationship between specific lesion regions and the entire skin. After the encoder, we adopt an EAM to refine the learned representations by attending to salient features, enhancing the model's ability to capture intricate spatial dependencies. The attention layer's output is then passed to the global average pooling layer followed by a dropout, which, in turn, passes it to the fully connected layer that assigns each sequence to its corresponding class. The overview of our proposed approach is shown in Fig. 5.1

##### a. Attention-based structural feature enhancement

The convolutional kernel functions as a trainable filter, capable of isolating relevant information while preserving local features. We employ attention guidance to effectively harness structural features, utilizing multiple convolutional layers along with a convolutional attention mechanism module as a trainable image preprocessing stage. As illustrated in Fig. 5.1, we utilize a fundamental convolution unit with a 3x3 kernel and a ReLU activation function. The initial segment of the ASFE component (see Fig. 5.1-a) comprises an ensemble block where an equal number of basic convolution units as input channels process the image concurrently. The outputs of these basic convolution units are then concatenated, and the resultant concatenation undergoes further processing through another convolution unit to aggregate the outcomes. This step is followed by an element-wise addition operation that emphasizes the structural features of the original image. Finally, a Convolutional Block Attention Module (CBAM) [174] is applied to refine the feature representation of the processed image. CBAM integrates channel and spatial attention mechanisms, enabling adaptive adjustment of focus on crucial structural features while dampening irrelevant information. This comprehensive attention mechanism enhances feature representation, emphasizing the scale of the lesion's target region. The detailed process is further described as follows:

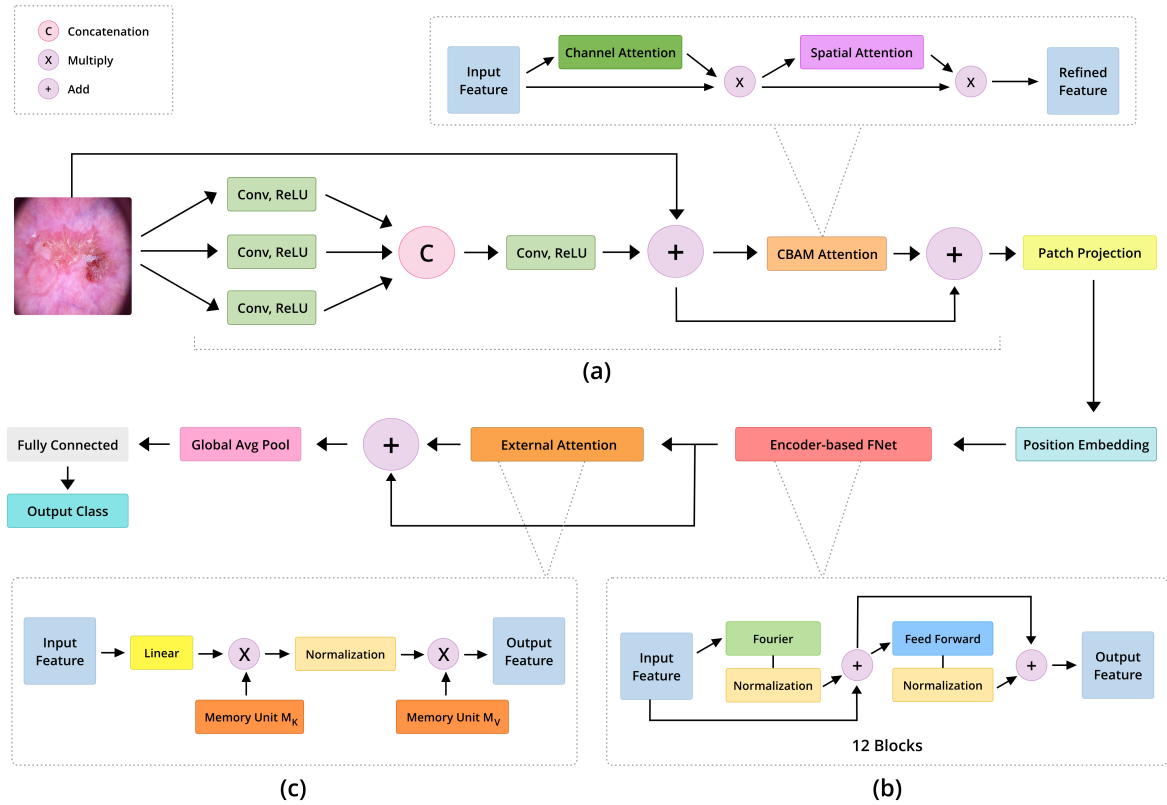


Fig. 5.1 The layout of the JILDYA-Net model. (a) is the ASFE structure diagram, (b) is the detailed network structure of the encoder-based FNet, and (c) is the EAM structure diagram.

$$ConvUnit_i = RELU(Conv(x)) \quad (5.1)$$

$$x = concat(ConvUnit_1, \dots, ConvUnit_n) \quad (5.2)$$

$$x = x \oplus RELU(Conv(x)) \quad (5.3)$$

$$output = CBAM(x) \quad (5.4)$$

where  $x$  donates the input image,  $n$  represents the number of channels, and  $\oplus$  is the element add operation.

### b. Encoder based-FNet

The FNet [38] is a Multi-Layer Perceptron (MLP) that presents an innovative approach to transformer encoder architectures [32]. It replaces the self-attention sublayers with straightforward linear transformations that blend input tokens. The FNet model is based on the Fourier Transforms which replaces a part of the standard BERT model on NLP tasks. Drawing inspiration from the ViT [33] paradigm, we extend the utilization of the FNet to image classification tasks. As shown in Fig. 5.1, the output-enhanced image from the ASFE component undergoes patch projection and position embedding procedures, resulting in a sequence of token generation. These tokens represent the input of the encoder-based FNet (see Fig. 5.1-b) which consists of 12 identical blocks. Each block is composed of two primary sub-layers: a 2D Fourier transformation sub-layer employing the Fast Fourier Transform (FFT) [39], with 1D Fourier Transform applied across patches and channels where this sub-layer facilitates a linear transformation of the input while integrating token mixing, and a feed-forward sub-layer. Both layers incorporate skip connections similar to those found in residual networks, followed by layer normalization. The detailed process is elaborated further as follows:

$$x_{fourier} = \Re(\mathcal{F}_{P,C}(x)) \quad (5.5)$$

$$x_{norm1} = Norm(x_{fourier} + x) \quad (5.6)$$

$$x_{ff} = FF(x_{norm1}) \quad (5.7)$$

$$x_{norm2} = Norm(x_{ff} + x_{norm1}) \quad (5.8)$$

where  $(\mathcal{F}_{P,C})$  is the 2D Fourier transformation operation along the patches and channels,  $Norm$  is the normalization, and  $FF$  stands for the feed-forward layer.

### c. External attention module

The EAM [173], represents a lightweight approach to attention mechanisms, relying on two external, compact, trainable, shared memory components. Initially, the External Attention Module computes the attention map by evaluating the affinities between self-query vectors and a learnable key memory external to the model. Subsequently, it generates a refined feature map by applying this attention map to another learnable value memory external to the model through element-wise multiplication.

As depicted in Fig. 5.1, the output of the encoder-based FNet is subjected to the EAM to enhance learned representations by focusing on pertinent features. This module comprises two successive linear layers and two normalization layers (see Fig. 5.1-c). The linear layers correspond to the two memory units, denoted as  $M_k$  and  $M_v$ . These external memories are trained to discern discriminative features and capture informative components while filtering out irrelevant information. The EAM calculations are concluded as follows:

$$x_q = \text{QueryLinear}(x) \quad (5.9)$$

$$x_{attn} = M_k(x_q) \quad (5.10)$$

$$x_{attn} = \text{Norm}(x_{attn}) \quad (5.11)$$

$$x_{output} = M_v(x_{attn}) \quad (5.12)$$

## 5.2.4 Experimental setup

This section delineates the used datasets, the preprocessing steps, the data augmentation, and the experimental settings of the JILDYA-Net model.

### 5.2.4.1 Datasets

Data serves as the foundation for developing DL solutions and computer-aided diagnosis systems. In these studies, we utilized datasets from online International Skin Imaging Collaboration (ISIC) challenges<sup>1</sup>, focusing on skin lesion classification tasks.

#### a. HAM10000

The HAM10000 (Human Against Machine) dataset [114] is a large collection of multi-source dermatoscopic images<sup>2</sup> of common pigmented skin lesions, where histopathology validates each skin lesion diagnosis. It consists of 10,015 dermoscopic images of seven skin lesion types: actinic keratoses and intraepithelial carcinoma, basal cell carcinoma, benign keratosis-like lesions, dermatofibroma, melanoma, melanocytic nevi, and vascular lesions.

<sup>1</sup>ISIC Challenges: [challenge.isic-archive.com](http://challenge.isic-archive.com)

<sup>2</sup>HAM10000: [dataverse.harvard.edu/dataset.xhtml](http://dataverse.harvard.edu/dataset.xhtml)

### b. ISIC 2019

ISIC 2019 [123] is a public dataset<sup>3</sup> comprised of HAM10000 [114], BCN [170], and MSK [172] datasets. It contains a total of 25,331 labeled images that are publicly available for classification tasks with eight labels: melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, vascular lesion, squamous cell carcinoma.

#### 5.2.4.2 Preprocessing and augmentation

In this study, we used the HAM10000 and ISIC 2019 datasets for the classification of different skin lesions. Fig. 5.2 illustrates the datasets' information and distribution.

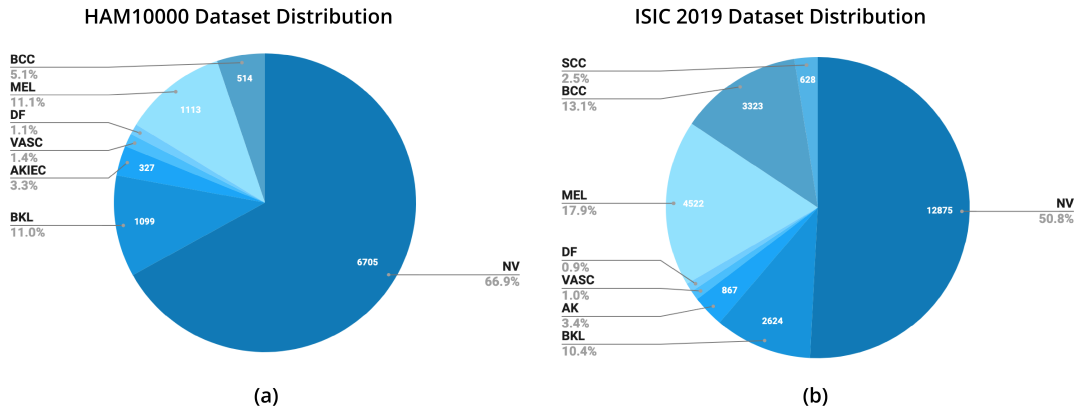


Fig. 5.2 Datasets' information and distribution.

Fig. 5.2 clearly illustrates a significant issue with class imbalance across the datasets, which can substantially impact the performance of DL methods. Specifically, the BCC, DF, and VASC classes are notably underrepresented, posing challenges such as overfitting and increased training complexity. Conversely, the NV lesion class dominates, constituting 50% or more of the total dataset. This imbalance risks introducing bias into the model, skewing its predictions toward the overrepresented NV class and potentially diminishing the accuracy and generalizability of the model across other classes.

To address this problem, we employ data augmentation techniques to expand the dataset by generating additional samples for underrepresented classes; while applying minimal or no augmentation to overrepresented classes to ensure balanced sample distribution within the training set. This strategy aims to mitigate issues associated with imbalanced class distributions, overfitting, and limited availability of training images.

We utilized rotation, translation, and flipping techniques for data augmentation. The rotation involves adjusting the angle of an image by degrees ranging from 45 to 180. The



translation involves shifting the image by a percentage of 12.5 in both horizontal and vertical directions. Flipping involves mirroring the image horizontally and vertically. These augmentation procedures are applied only to the training and validation sets.

Fig. 5.3 displays the four derived datasets from the HAM10000 and ISIC 2019 datasets. The initial version of each dataset was allocated for multi-class classification tasks, while the second version was designated for binary-class classification tasks.

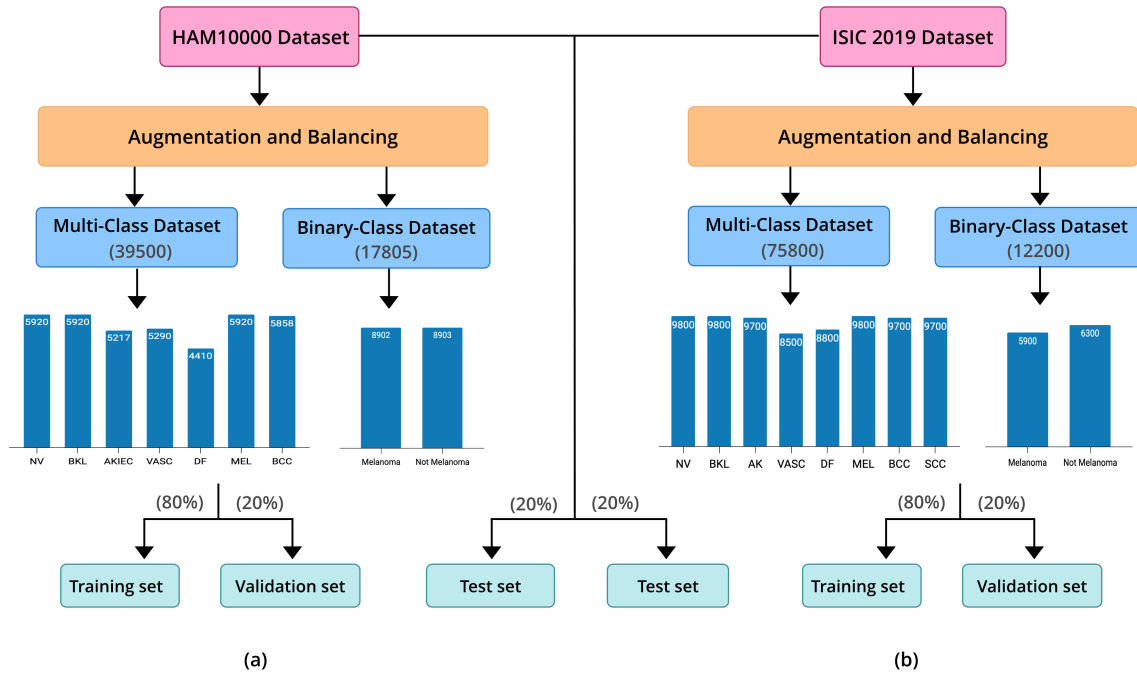


Fig. 5.3 **Flow diagram of the datasets pre-processing, augmentation, and preparation.** (a) The HAM10000 dataset flow diagram. (b) The ISIC 2019 dataset flow diagram.

In the pre-processing step, all images are resized to 224×224 pixels and normalized using max-min normalization. Each augmented dataset is split into 80% for training and 20% for validation while the test set only contains the original images.

### 5.2.4.3 Software tools and model settings

The JILDYA-Net was built using the TensorFlow framework and the experiments were conducted using an NVIDIA Tesla P100 GPU. The model was trained to optimize the cross-entropy loss [17]. The optimizer, weight decay, momentum, epoch, and batch size are AdamW [163],  $1e^{-4}$ , 0.6, 50, and 128, respectively. The initial learning rate is  $5e^{-4}$  and decreased after a few epochs using the plateau reduction strategy.

### 5.2.5 Results and discussion of JILDYA-Net

This section presents the results and key findings of the JILDYA-Net model. The efficacy of JILDYA-Net has been validated through many experiments and tasks. The subsequent subsections offer a comprehensive overview of these experiments and the evaluation methodology and results.

#### 5.2.5.1 Evaluation results

We conducted many experiments to examine the fine-grained performance of the JILDYA-Net model. Initially, we trained and tested the model on the original datasets (before balancing). Table 5.1 shows the evaluation results of the pre-trained model for multi-class skin lesions.

Table 5.1 Evaluation results of JILDYA-Net on the original datasets.

| Dataset   | Test Split | Accuracy (%) | Sensitivity (%) | Specificity (%) |
|-----------|------------|--------------|-----------------|-----------------|
| HAM10000  | 20%        | 92.80        | 86.60           | 98.30           |
| ISIC 2019 |            | 86.10        | 84.20           | 97.70           |

The results demonstrate that class imbalance significantly impacts the model's overall performance in differentiating between various lesion categories. The evaluation indicates that the model exhibits a pronounced bias toward the NV class, which constitutes the majority of the dataset.

Next, we trained and tested the model using the balanced versions of the two datasets, as described in Section 5.2.4.2 to avoid biased model training and skewed performance metrics. To assess the good performance of the JILDYA-Net, we evaluate it on both binary classification (i.e., distinguishing between melanoma and not melanoma lesions) and multi-class classification tasks. Table 5.2 presents a comprehensive classification report of the pre-trained model on the augmented datasets from the testing phase across all datasets.

Table 5.2 Evaluation results of JILDYA-Net on different classification tasks and datasets.

| Classification task | Dataset   | Test Split | Accuracy (%) | Sensitivity (%) | Specificity (%) | AUC   |
|---------------------|-----------|------------|--------------|-----------------|-----------------|-------|
| Binary-Class        | HAM10000  | 20%        | 94.30        | 94.30           | 94.30           | 0.943 |
|                     | ISIC 2019 | from       | 96.40        | 96.40           | 96.40           | 0.964 |
| Multi-Class         | HAM10000  | original   | 98.80        | 95.87           | 99.30           | 0.986 |
|                     | ISIC 2019 | dataset    | 98.72        | 94.92           | 99.27           | 0.984 |

As shown in Table 5.2, the results indicate that the proposed model achieves remarkable performance in the two classification tasks. Specifically, the model demonstrated outstanding ability in classifying malignant cases (i.e., melanoma vs. non-melanoma), achieving an AUC,

accuracy, and sensitivity of 94.30%, 94.30%, and 88.71% for the HAM10000 and 96.40%, 96.40%, 94.66% ISIC 2019 datasets.

In the multi-classification task on the HAM10000 dataset, the model attained an accuracy of 98.80%, a sensitivity of 95.87%, a specificity of 99.30%, and an overall AUC of 98.6%. For the ISIC 2019 dataset, the model achieved an accuracy of 98.72%, a sensitivity of 94.92%, a specificity of 99.27%, and an overall AUC of 98.4%.

To facilitate a comprehensive analysis of the classification results, we provide a graphical representation of the confusion matrices generated by the proposed method, as depicted in Fig. 5.4. These matrices are invaluable for understanding the frequency of correct and incorrect classifications made by the model across different classes within the test set.

From the illustrated confusion matrices in Fig. 5.4, the proposed model demonstrates a high level of accuracy in correctly predicting skin lesions. However, there are misclassifications in BCC vs BKL, MEL vs BKL, and MEL vs NV, and this is due to intraclass variation, and the high degree of similarity between skin lesions especially in the early stages.

### 5.2.5.2 Comparison of JILDYA-Net model with SOTA methods

We present a comprehensive comparative analysis with SOTA methods utilizing the HAM10000 and ISIC 2019 datasets. Tables 5.3 and 5.4 summarize the comparison of our model's performance metrics, including accuracy, sensitivity, specificity, and AUC, against those of previous studies employing the same datasets. Fig. 5.5 and Fig. 5.6 provide visual representations of the results summarized in Tables 5.3 and 5.4, respectively. The reported results of the proposed model were derived using a pre-trained model on augmented and balanced datasets to ensure a fair and rigorous evaluation process.

Table 5.3 JILDYA-Net vs. SOTA methods for the HAM10000 dataset. N/A: value is not available.

| Authors                | Method                         | Accuracy (%) | Sensitivity (%) | Specificity (%) | AUC          |
|------------------------|--------------------------------|--------------|-----------------|-----------------|--------------|
| Hoang et al. [111]     | EW-FCM + wide-ShuffleNet       | 86.33        | 86.33           | 97.72           | N/A          |
| Srinivasu et al. [113] | MobileNet V2-LSTM              | 85.34        | 92.00           | 88.24           | N/A          |
| Shetty et al. [108]    | CNN                            | 95.18        | 85              | N/A             | N/A          |
| Wang et al. [119]      | G-DMN                          | 87.07        | 75.28           | 96.19           | 0.952        |
| Wang et al. [121]      | DBN                            | 91.58        | 84.20           | 97.58           | 0.985        |
| Azeem et al. [109]     | SkinLesNet (DCNN)              | 90.00        | 87.00           | N/A             | N/A          |
| Sarmad et al. [175]    | Deep CNN + MC-SVM              | 98.57        | 93.89           | 96.37           | <b>0.988</b> |
| Ours                   | JILDYA-Net (ASFE + FNet + EAM) | <b>98.80</b> | <b>95.87</b>    | <b>99.30</b>    | 0.986        |

After examination of Table 5.3 and Figure 5.5, we can see that the JILDYA-Net outperforms the SOTA models. The model achieves better classification results in the multi-classification task for the HAM10000 dataset compared to previous work.

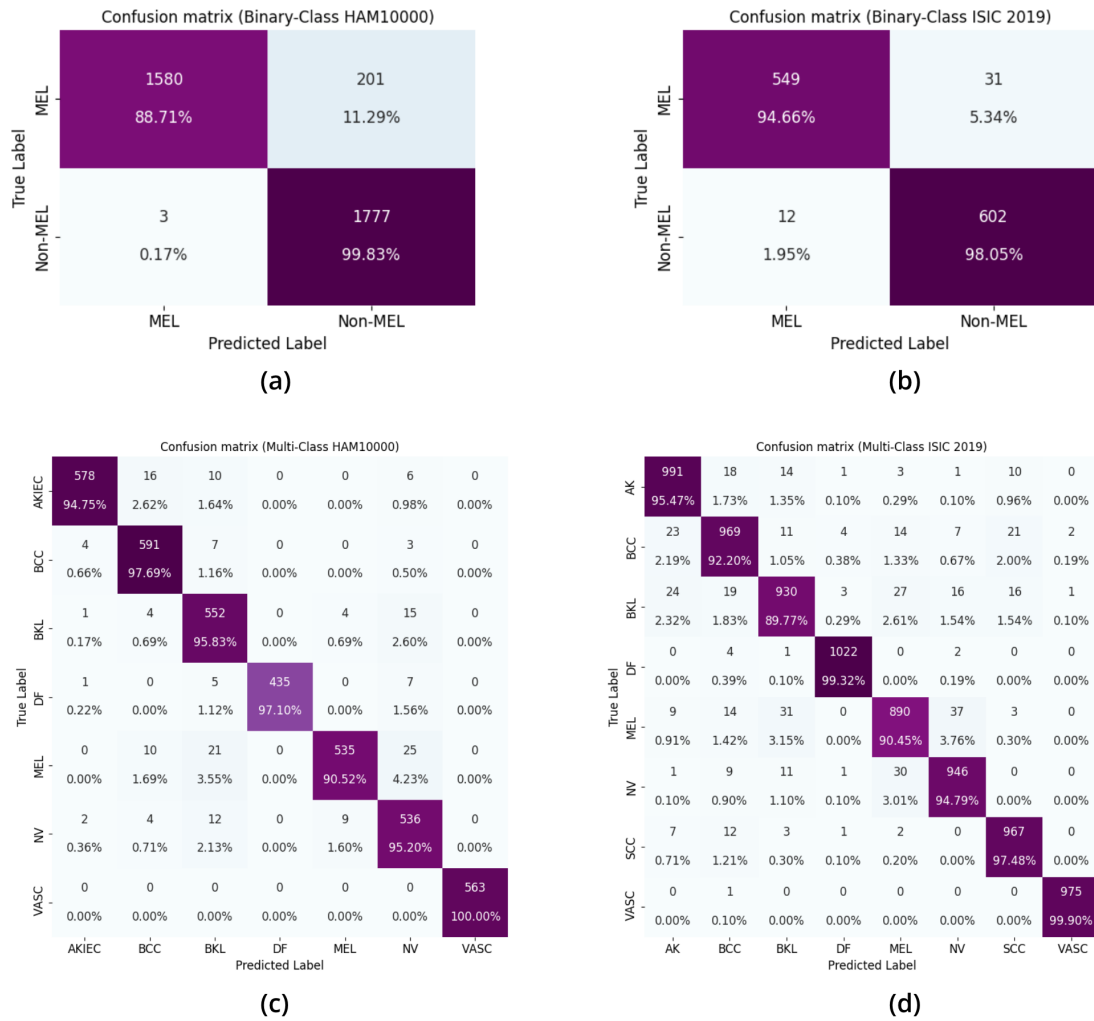


Fig. 5.4 **Confusion matrices generated by the proposed model.** (a) Binary-class HAM10000 confusion matrix. (b) Binary-class ISIC 2019 confusion matrix. (c) Multi-class HAM10000 confusion matrix (d) Multi-class ISIC 2019 confusion matrix.

Table 5.4 JILDYA-Net *vs.* SOTA methods for the ISCI 2019 dataset. N/A: value is not available.

| Authors               | Method                         | Accuracy (%) | Sensitivity (%) | Specificity (%) | AUC          |
|-----------------------|--------------------------------|--------------|-----------------|-----------------|--------------|
| Hoang et al. [111]    | EW-FCM + wide-ShuffleNet       | 82.56        | 82.56           | 97.51           | N/A          |
| Imran et al. [110]    | DCNN                           | 89.58        | 89.58           | 97.57           | 0.954        |
| Alsahafi et al. [122] | Skin-Net (RCNN)                | 94.65        | 70.78           | 96.78           | N/A          |
| Qasim et al. [125]    | Spiking VGG-13                 | 89.57        | N/A             | N/A             | N/A          |
| Sarmad et al. [175]   | Deep CNN + MC-SVM              | 93.47        | 84.34           | 87.53           | <b>0.985</b> |
| Ours                  | JILDYA-Net (ASFE + FNet + EAM) | <b>98.72</b> | <b>94.92</b>    | <b>99.27</b>    | 0.984        |

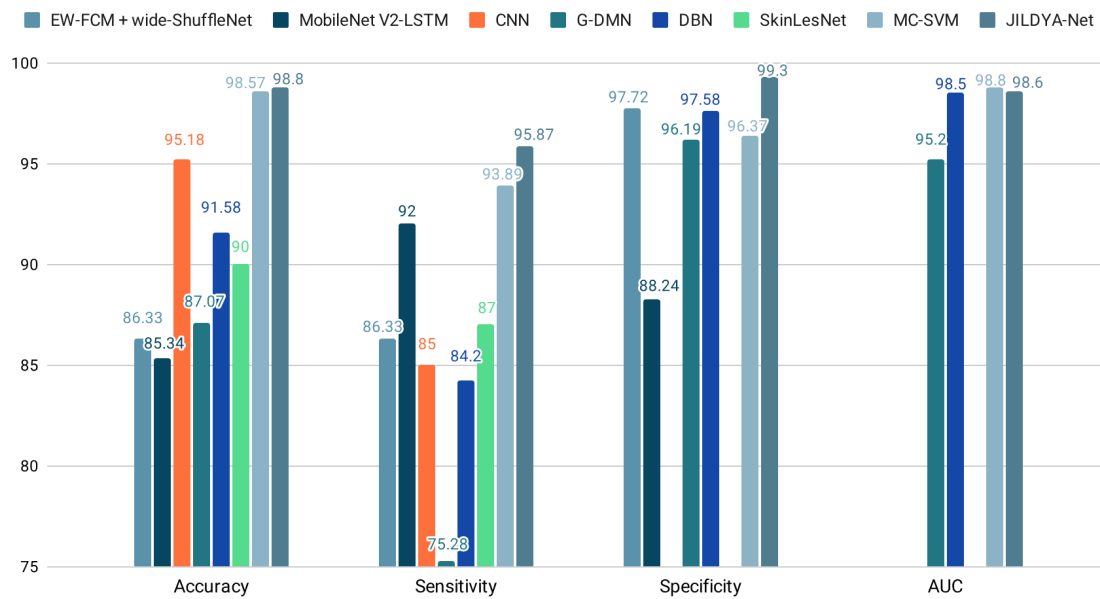


Fig. 5.5 JILDYA-Net vs. SOTA methods for the HAM10000 dataset.

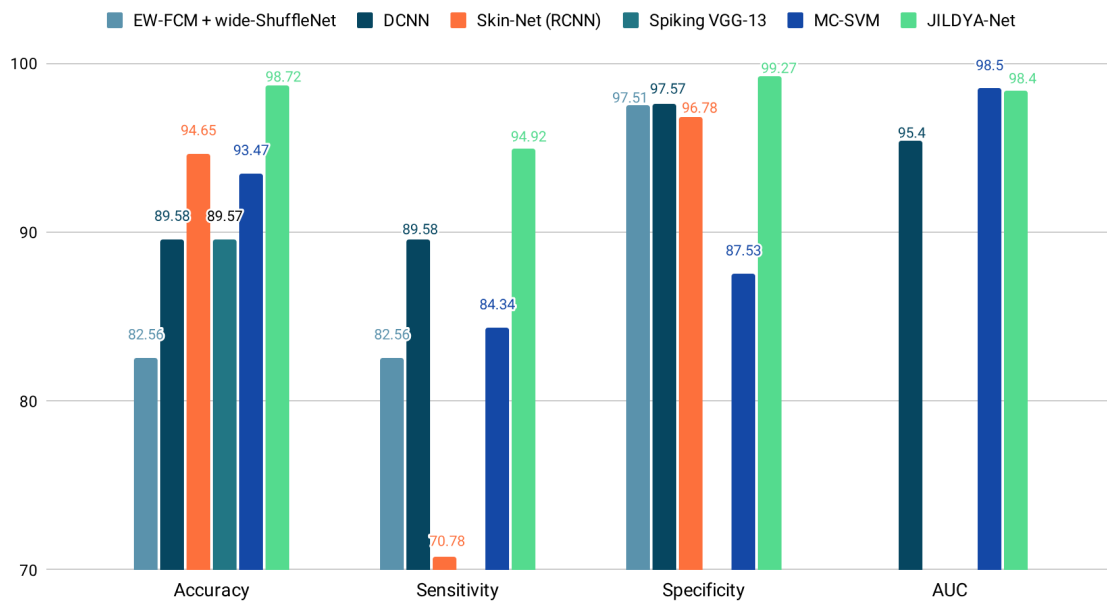


Fig. 5.6 JILDYA-Net vs. SOTA methods for the ISCI 2019 dataset.

Additionally, based on Table 5.4 and Figure 5.6, the results highlight that JILDYA-Net outperforms SOTA methods sometimes by a large margin in terms of sensitivity.

### 5.2.5.3 JILDYA-Net computational complexity

The computational complexity of any method plays a crucial role, especially in domains where efficiency is of utmost importance. As the aim was to develop a fast, accurate, and less complex method, we meticulously examined several factors such as time complexity, memory utilization, and model size in terms of parameters to make an overview of the computational complexity of the JILDYA-Net model.

All experiments were performed using an NVIDIA P100 GPU. Table 5.5 depicts the computational statistics of the proposed method, including training and test times, total parameters, and peak memory usage, for both the HAM10000 and ISIC 2019 datasets.

Table 5.5 Computational statistics of JILDYA-Net on HAM10000 and ISIC 2019 datasets.

| Factor                 | Value    |           |
|------------------------|----------|-----------|
|                        | HAM10000 | ISIC 2019 |
| Training Time (s)      | 6,150    | 13,250    |
| Test Time (s)          | 8        | 19        |
| Total Parameters       | 2.3M     |           |
| Peak Memory Usage (GB) | 2.7      |           |

### 5.2.5.4 Discussion of JILDYA-Net results

To assess the effectiveness of the proposed approach, we compared the model's performance against SOTA methods analysis across two distinct datasets. Fig. 5.7 and Fig. 5.8 offer insights into the performance improvements of JILDYA-Net compared to several SOTA methods across different metrics, including accuracy, sensitivity, specificity, and AUC. The following calculation quantifies these improvements:

$$Improvement\ factor = \left( \frac{JILDYA - Net_{performance}}{SOTA - method_{performance}} - 1 \right) * 100 \quad (5.13)$$

The obtained results (Fig. 5.7 and 5.8) show that the calculated improvements of JILDYA-Net over SOTA methods reveal significant enhancements across accuracy, sensitivity, specificity, and AUC metrics. Notably, the JILDYA-Net model shows substantial improvements in accuracy, sensitivity, and specificity compared to the best-performing SOTA models such as [113, 110, 121, 175].

For both ISIC 2019 and HAM10000 datasets, it is clear that JILDYA-Net outperforms all other SOTA methods on all metrics. The AUC metric is the only metric where JILDYA-Net loses 0.1% (on ISIC 2019 dataset) and 0.2% (on HAM1000 dataset) *w.r.t.* to [175]. The minimum and maximum improvement factor are reported in Table 5.6.

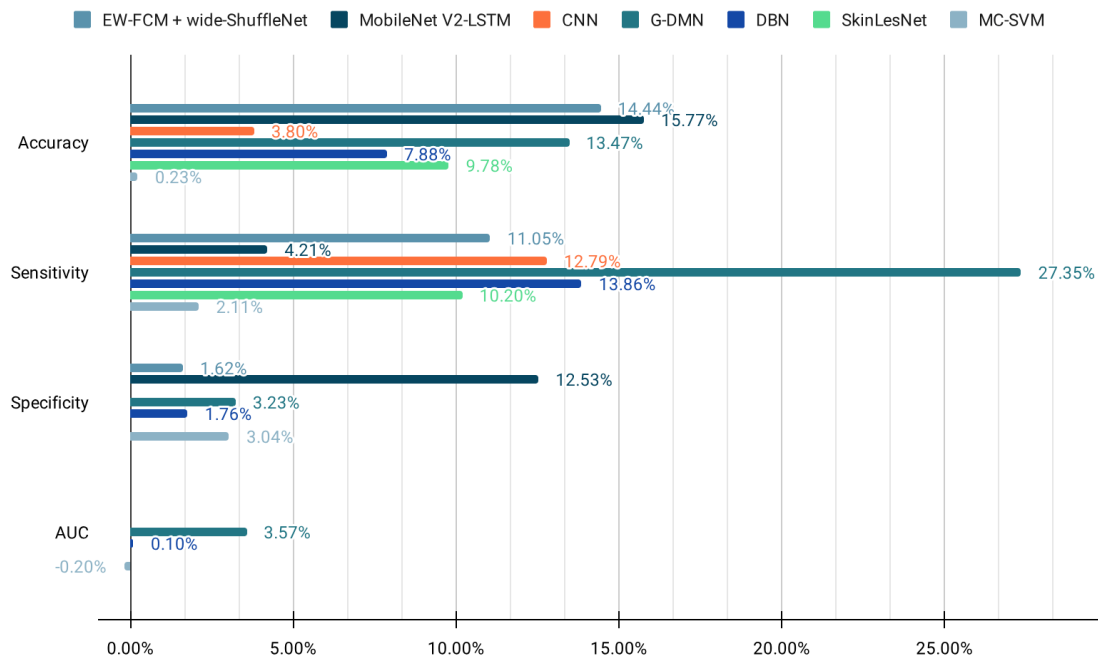


Fig. 5.7 JILDYA-Net improvement factor *w.r.t.* SOTA for the HAM10000 dataset.

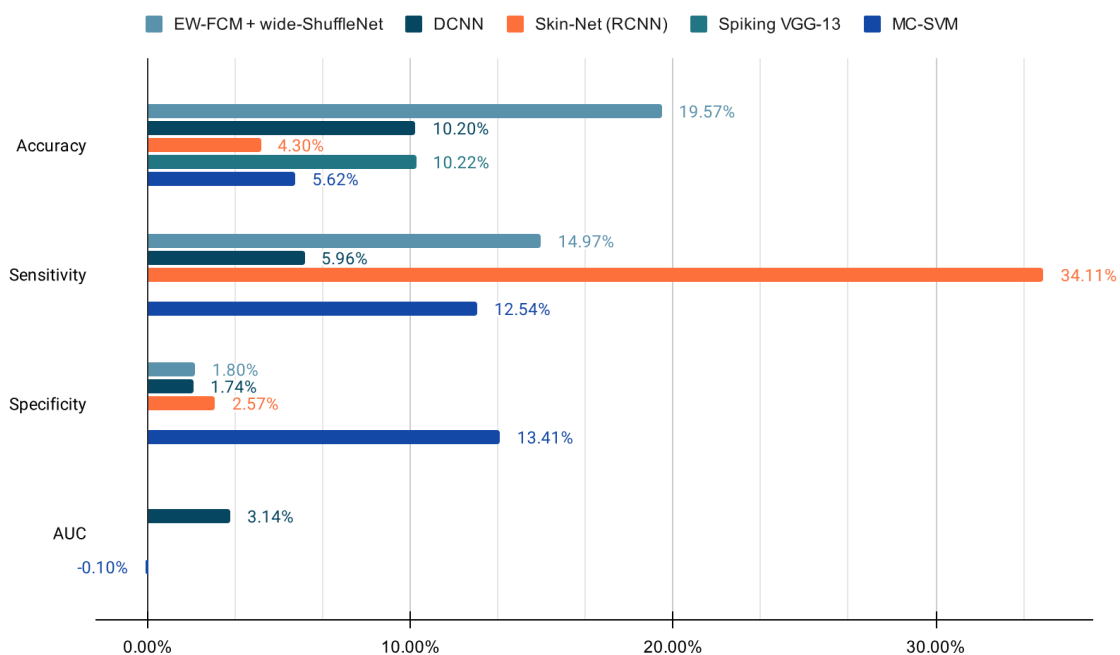


Fig. 5.8 JILDYA-Net improvement factor *w.r.t.* SOTA for the ISIC 2019 dataset.

On the other hand, given the significance of cancerous lesion detection, particularly melanoma, JILDYA-Net demonstrates noteworthy performance in classifying melanoma

Table 5.6 Maximum and minimum JILDYA-Net improvement factor *w.r.t.* SOTA for both HAM10000 and ISIC 2019 datasets.

| Dataset   | Accuracy       | Sensitivity    | Specificity    | AUC            |
|-----------|----------------|----------------|----------------|----------------|
| HAM10000  | [0.23 ; 15.77] | [2.11 ; 27.35] | [1.76 ; 12.53] | [-0.20 ; 3.57] |
| ISIC 2019 | [4.30 ; 19.57] | [5.96 ; 34.11] | [1.80 ; 13.41] | [-0.10 ; 3.14] |

cases, achieving a classification rate of 90%. This achievement is notable considering the challenges associated with detecting cancerous lesions, even for dermatologists, especially in the early stages.

Although our proposed approach has shown success, it is crucial to recognize its inherent limitations. The model's sensitivity to class imbalance and small datasets may result in diminished performance. Additionally, due to its lightweight architecture, the model may encounter challenges when dealing with high-level complex scenarios.

### 5.2.6 Summary of JILDYA-Net

The first part of this chapter introduced a novel ad-hoc approach named JILDYA-Net for skin lesion classification from dermoscopic images. In this approach, advancements in DL architectures are leveraged to design a lightweight framework capable of achieving both high accuracy and computational efficiency. By replacing complex components with simpler, yet effective modules, this approach strikes a balance between model performance and resource utilization, making the solution suitable for deployment on mobile devices and embedded systems without compromising classification accuracy.

To evaluate the effectiveness of JILDYA-Net, extensive experiments were conducted comparing its performance against SOTA models across two commonly used datasets: HAM10000 and ISIC 2019. Numerically, JILDYA-Net outperforms SOTA methods in accuracy, sensitivity, and specificity, with improvements of up to 19.57%, 34.11%, and 13.41%, respectively, on augmented and balanced versions of the HAM10000 and ISIC 2019 datasets. Notably, the lightweight architecture of JILDYA-Net facilitates easy integration into real-time applications, enabling healthcare professionals to leverage its capabilities for rapid and accurate diagnosis of skin cancer.

While the JILDYA-Net achieved satisfactory classification results, there is room for improvement, especially with access to additional well-annotated datasets. Further investigation is required to assess its performance on diverse datasets and lesion types. Additionally, integrating semantic segmentation models alongside image classification could enhance skin lesion localization and detection accuracy.



## 5.3 Part II: Skin lesion segmentation and localization using DA-UNet-Plus

The second part of this chapter explores the adaptation of the DA-UNet-Plus model, initially developed for detecting intraretinal fluid as detailed in the previous chapter, to the task of skin lesion segmentation and localization.

### 5.3.1 Motivation for DA-UNet-Plus usage

Skin lesion segmentation and spread localization are critical components in the accurate diagnosis. Effective automatic segmentation of skin lesions enables the precise delineation of lesion boundaries, which is an essential prerequisite for CAD systems. The spread localization process further identifies the extent of lesion proliferation, providing vital information about the progression of the disease. Accurately segmenting skin lesions poses a formidable challenge due to the diverse visual characteristics they exhibit, such as significant variations in color, texture, shape, and size. The task is further complicated by the often indistinct boundaries between skin lesions and normal skin, making precise localization and differentiation in images difficult. These complexities are compounded by the artifacts' presence like hair, shadows, and reflections, which can obscure lesion boundaries and further complicate the segmentation process. Traditional image processing techniques frequently fall short in addressing these challenges, underscoring the need for the development of advanced methods capable of achieving robust and precise segmentation.

With the continuous progress of CNNs in medical imaging and advancements in DL methods, effective segmentation of skin lesions has become feasible. This progress is largely attributed to the power of recent methods and architectures such as U-Net [25], V-Net [26], and their variants [27, 28]. These advanced architectures have significantly improved the accuracy and reliability of skin lesion segmentation, addressing the complex challenges posed by diverse lesion characteristics and artifacts in medical images. Recently, the hybridization of DL methods has demonstrated a huge impact on the enhancement of many model performances. Indeed, hybrid architectures that combine CNNs with other DL modules, such as attention mechanisms, have shown significant promise in improving segmentation outcomes [134, 138, 140]. These hybrid models enhance the ability to accurately delineate lesion boundaries and localize the spread, providing comprehensive insights into the lesion's characteristics.

### 5.3.2 DA-UNet-Plus contribution

The primary objective of this study is to evaluate the efficacy and versatility of the DA-UNet-Plus model in various medical imaging tasks beyond its original scope, i.e. that of detecting intraretinal fluid. By extending the use of DA-UNet-Plus to skin lesion segmentation, this research aims to demonstrate the model's adaptability and robustness in handling different types of medical images with varying complexities. The study involves a comprehensive analysis of the model's performance in accurately segmenting and localizing skin lesions, which present unique challenges due to their diverse visual characteristics and the presence of artifacts.

### 5.3.3 Methodology

#### 5.3.3.1 Adaptation of DA-UNet-Plus

DA-UNet-Plus, detailed in Chapter 3, is a hybrid DL model designed for medical image segmentation. It integrates several innovative components to enhance its performance, including SE-attention residual blocks for efficient feature extraction and improved gradient flow, and dual attention mechanisms (both spatial and channel attention) to capture long-range dependencies and refine feature representations. Initially, the model developed for intraretinal fluid detection in OCT images, DA-UNet-Plus has demonstrated good performance.

#### 5.3.3.2 Datasets

##### a. ISIC 2017 dataset

The ISIC 2017 dataset [170] is a publicly available resource<sup>3</sup> comprising 2,600 images, with 2,000 designated for training and 600 for testing. For both the training and test sets, the dataset includes corresponding ground truth labels and patient metadata, providing a comprehensive foundation for analysis. The lesions within the dataset are categorized into four distinct class groups: melanoma; nevus or seborrheic keratosis; seborrheic keratosis; and melanoma or nevus. Additionally, the dataset includes metadata on the patient's approximate age and gender, further enriching the context for model training and evaluation.

##### b. ISIC 2018 dataset

The ISIC 2018 [171] is a public dataset<sup>3</sup> utilized for instance segmentation, semantic segmentation, and object detection tasks in medical research. This dataset comprises 15,414 images

---

<sup>3</sup>ISIC Data: [challenge.isic-archive.com/data](https://challenge.isic-archive.com/data)

with 18,055 labeled objects across six different classes, including skin cancer, pigment network, globule, and others like milia-like cysts, negative network, and streaks. The dataset includes 11,720 unlabeled images without annotations. Additionally, the dataset can be divided into three tasks: disease classification with 11,720 images, lesion segmentation with 3,694 images, and attribute detection with 3,694 images.

### **5.3.3.3 Data preprocessing**

In the segmentation task of skin lesions, we utilized the ISIC 2017 and ISIC 2018 datasets, which contain dermoscopic images accompanied by corresponding masks that represent the ground truth. To mitigate noise effects and enhance image quality, several image preprocessing steps were applied, ensuring a well-formatted structure suitable for accurate segmentation by the DA-UNet-Plus model. These preprocessing steps included normalization, resizing, and morphological operations.

#### **a. Normalization**

The normalization step is employed to scale pixel values to a standard range. This step reduces the impact of varying lighting conditions and skin tones, ensuring uniformity across the dataset.

#### **b. Resizing**

Resizing the images to a consistent dimension is crucial for efficient processing by the neural network. This step ensures that all input images have the same resolution, facilitating batch processing and reducing computational complexity.

#### **c. Morphological operations**

Morphological operations such as dilation and erosion, are applied to refine the boundaries of the skin lesions. These operations help in reducing noise and small artifacts in the image, thereby improving the accuracy of the segmentation model.

### **5.3.4 Experimental results of DA-UNet-Plus**

This section delineates the experimental settings, results, and key findings of the DA-UNet-Plus model. The subsequent sub-sections offer a comprehensive overview of the experiment and results on the ISIC 2017 and ISIC 2018 datasets

#### 5.3.4.1 DA-UNet-Plus settings

The DA-UNet-Plus was built using the TensorFlow framework and the experiments were conducted using an NVIDIA Tesla P100 GPU. The model was trained to optimize the hybrid dice coefficient and binary cross-entropy loss. This loss is defined as:

$$Loss = L_{Dice} + L_{BCE} \quad (5.14)$$

The Adam algorithm was utilized as the optimizer for training the model. The training process spanned 50 epochs with a batch size of 80. The ISIC datasets were divided into 70% for the training set and 30% for validation and testing. The performance of the DA-UNet-Plus model was rigorously evaluated using multiple metrics, including accuracy, recall, precision, dice similarity coefficient (DSC), and IoU.

#### 5.3.4.2 DA-UNet-Plus evaluation results

To examine the fine-grained performance of the model, we present a comprehensive evaluation report of the DA-UNet-Plus model in Table 5.7. This table shows the performance of DA-UNet-Plus in skin lesion segmentation and localization.

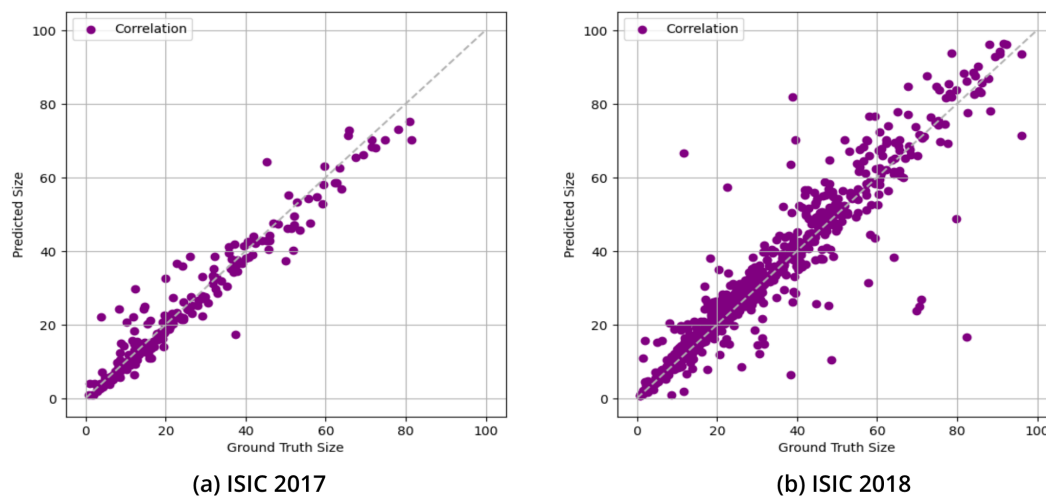
Table 5.7 DA-UNet-Plus performance in skin lesion segmentation and localization.

| Dataset   | Evaluation metrics (%) |        |           |      |      |
|-----------|------------------------|--------|-----------|------|------|
|           | Accuracy               | Recall | Precision | DSC  | IoU  |
| ISIC 2017 | 97.0                   | 93.0   | 92.30     | 92.6 | 86.2 |
| ISIC 2018 | 96.15                  | 91.70  | 94.25     | 93.1 | 87.1 |

Based on the aforementioned results, the DA-UNet-Plus demonstrated a notable performance in segmenting and localizing the skin lesions, across various metrics, indicating its effectiveness and strong performance in this task.

The DA-UNet-Plus achieved a recall of 93.0% and 91.7% for the ISIC datasets which exhibit a strong sensitivity to true positives, effectively identifying the majority of the skin lesions. Moreover, the dice scores of 92.6% and 93.10% mean that the model is achieving a close alignment between its predicted lesion boundaries and the actual lesion boundaries defined in the ground truth data. Additionally, the IoU values indicate the model's accuracy in delineating skin lesions from the background skin.

The second set of results, as shown in the correlation plots (Fig. 5.9), displays the comparison of the predicted lesion sizes to the ground truth sizes for both datasets.



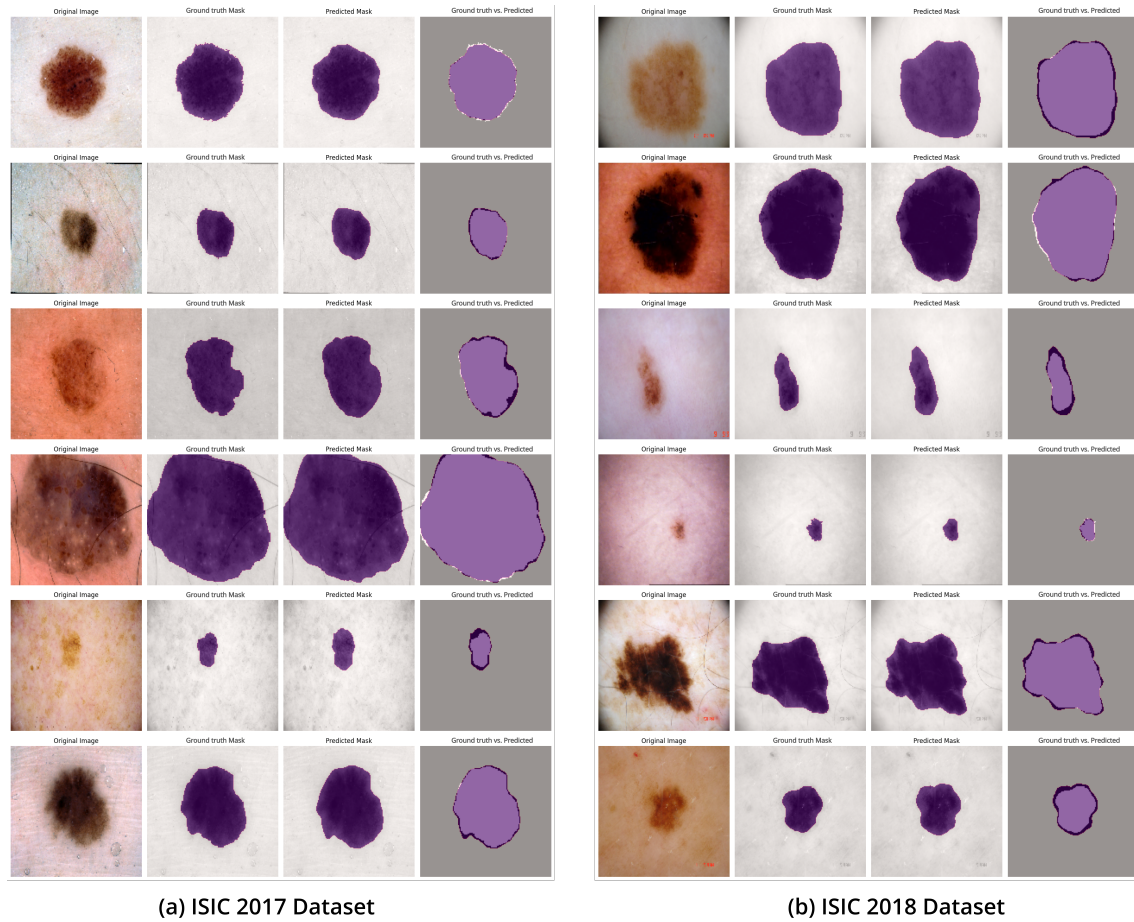
**Fig. 5.9 Correlation between predicted and ground truth sizes of skin lesions.** (a) Correlation plot of the ISIC 2017 dataset, (b) Correlation plot of the ISIC 2018 dataset.

For both datasets, the scatter plots reveal that the majority of data points are clustered near the diagonal line, which represents perfect correlation. This proximity signifies a strong positive correlation between the predicted sizes and the ground truth sizes of the lesions. While some outlier data points are evident due to the presence of complex cases, indicating instances where the model either overestimates or underestimates the size of the skin lesions, the overall trend suggests good performance of the DA-UNet-Plus model in accurately estimating lesion sizes.

In summary, the strong correlation between predicted and actual sizes validates the model's efficacy in accurately segmenting and localizing the lesions, which indicates that the DA-UNet-Plus model is well-suited for skin lesion segmentation tasks.

### 5.3.4.3 Visualization of DA-UNet-Plus results

The interpretability and validation of DL models play a crucial role in ensuring that clinicians, researchers, and patients can understand, trust, and effectively integrate these systems into their daily workflows. Fig. 5.10 visualizes the segmentation results of the DA-UNet-Plus model. This figure presents the skin lesion image in the first row, the ground truth in the second row—created by overlaying the mask image on the original image—and the model prediction in the third row, generated by overlaying the predicted mask on the original image. The last row represents a combined view of the actual versus predicted masks. The highlighted regions represent the extent of the skin lesions.



**Fig. 5.10 Visual segmentation results of the DA-UNet-Plus model.** (a) Examples from the ISIC 2017 dataset, (b) Examples from the ISIC 2018 dataset. The first column displays the skin lesions, the second column shows the ground truth masks in purple, the third column presents the segmentation results from DA-UNet-Plus in purple, and the final column provides a combined view of the actual versus predicted masks.

As observed in the figure, the predicted masks exhibit high similarity to the actual masks, with clear boundaries and accurate lesion shapes. While there are instances where the predicted masks slightly deviate from the ground truth, these deviations are relatively minor in the displayed examples. Overall, the visual representation demonstrates the effectiveness of the DA-UNet-Plus model in accurately segmenting skin lesions, reinforcing its potential utility in clinical practice.

#### 5.3.4.4 Comparative analysis of DA-UNet-Plus with SOTA methods

To comprehensively assess the performance of the DA-UNet-Plus model in skin lesion segmentation, we conducted a comparative analysis against SOTA approaches. Table 5.8

provides a detailed comparison of various SOTA models' performance on the ISIC 2017 and ISIC 2018 datasets.

Table 5.8 Performance comparison of DA-UNet-Plus with SOTA methods. N/A: value is not available. (-): not Applied

| Method                                    | ISIC 2017   |             |             | ISIC 2018   |             |             |
|-------------------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                                           | DSC (%)     | IoU (%)     | Recall (%)  | DSC (%)     | IoU (%)     | Recall (%)  |
| U-Net [127]                               | -           | -           | -           | 85.7        | 75.0        | N/A         |
| Recurrent attention CNN [135]             | 87.04       | 80.36       | 89.7        | -           | -           | -           |
| Enhanced U-Net++ [128]                    | -           | -           | -           | 91.7        | 84.73       | N/A         |
| CNN (encoder-decoder) [129]               | 87.5        | 77.8        | N/A         | 92.0        | 85.20       | N/A         |
| Multi-scale aggregation network [136]     | -           | -           | -           | 86.4        | 81.6        | 91.5        |
| CNN + Transformer [137]                   | 90.58       | 81.5        | <b>94.8</b> | 89.88       | 81.62       | 88.20       |
| End-to-end deep edge CNN [131]            | 87.92       | 80.53       | 88.04       | -           | -           | -           |
| Collaborative learning deep CNN [134]     | 86.7        | 79.1        | 86.5        | -           | -           | -           |
| CNN-transformer hybrid network [138]      | 86.3        | 78.3        | 86.6        | -           | -           | -           |
| Dynamic Spatial Group Enhanced UNet [139] | 88.4        | 79.2        | N/A         | 89.2        | 80.5        | N/A         |
| U-shaped network (DBNet-SI) [140]         | 92.5        | <b>87.2</b> | 90.9        | 93.0        | 86.9        | <b>92.0</b> |
| DA-UNet-Plus                              | <b>92.6</b> | 86.1        | 93.0        | <b>93.1</b> | <b>87.1</b> | 91.7        |

According to the results, both the DA-UNet-Plus and DBNet-SI [140] models achieved superior outcomes in the skin lesion segmentation task, outperforming other methods in terms of the DSC and IoU metrics. The performance of DA-UNet-Plus and DBNet-SI was remarkably close, with only a 0.2% to 1% difference, underscoring the efficacy of both models in accurately segmenting and localizing skin lesions.

Given that the DSC is an excellent overall indicator of segmentation accuracy, Fig. 5.11 graphically illustrates the DSC achieved by each model on the ISIC 2017 and ISIC 2018 datasets.

The results clearly demonstrate that DA-UNet-Plus delivers exceptional performance in skin lesion segmentation, highlighting its robustness and versatility in handling various medical image segmentation tasks across different imaging modalities.

### 5.3.5 Summary of DA-UNet-Plus usage in skin-related conditions

In the second part of this chapter, we explored the application of the DA-UNet-Plus hybrid model, initially developed for intraretinal detection, to the segmentation and localization of skin lesions in dermoscopic images. The model's performance was rigorously evaluated using the ISIC 2017 and ISIC 2018 datasets, where it demonstrated remarkable accuracy in producing segmentation outputs closely aligned with ground truth annotations.

The results underscore the robustness and effectiveness of DA-UNet-Plus in addressing the inherent challenges of medical image analysis. Its ability to handle the complexities of diverse image modalities and produce high-quality segmentation highlights its potential for

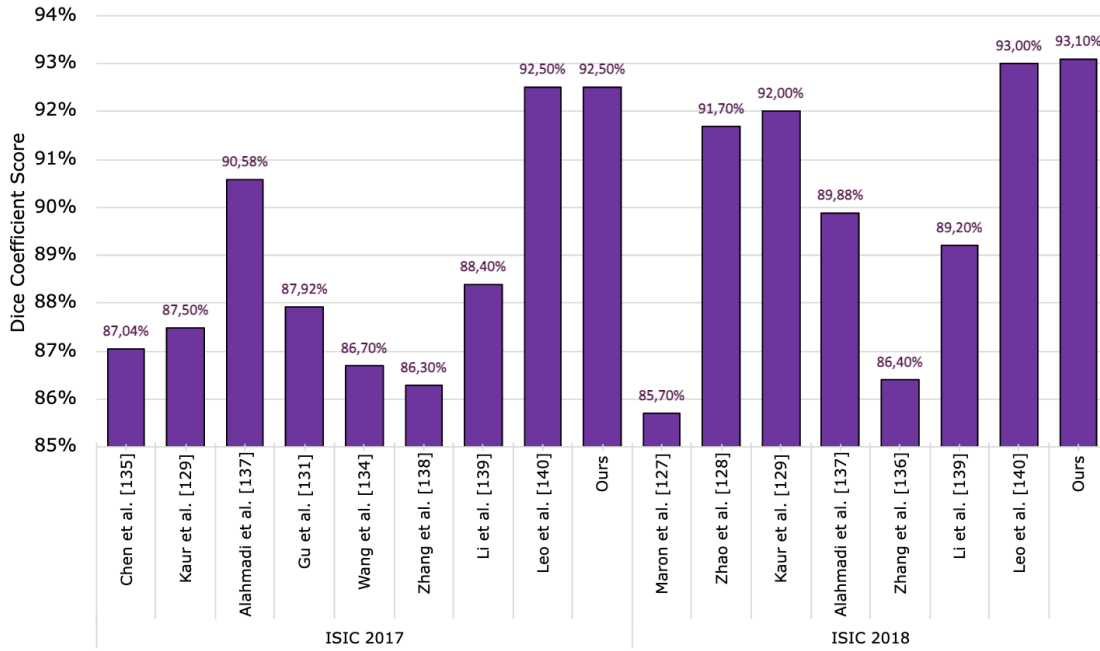


Fig. 5.11 Comparison of DSC of DA-UNet-Plus with SOTA methods.

broader applications in medical imaging tasks. These findings reinforce the model's utility as a reliable tool in the early detection and treatment planning of skin-related conditions.

## 5.4 Conclusion and future work

This chapter presented two main contributions toward skin lesion diagnosis and localization using advanced innovative DL techniques. Firstly, the chapter introduced a novel lightweight model, JILDYA-Net, designed for the diagnosis and classification of skin lesions using two prominent public challenge datasets. This model capitalized on the compact architecture of FNet combined with lightweight attention mechanisms, enabling it to effectively classify skin lesions by thoroughly analyzing dermoscopic images and capturing critical features pertinent to each lesion type. Our study highlighted the performance of JILDYA-Net, which significantly surpasses SOTA methods. Validation of unseen data further affirms the model's robust capability in accurately diagnosing skin lesions.

Secondly, this chapter explored the adaptation of the DA-UNet-Plus model for the segmentation and localization of skin lesions. Originally developed for intraretinal detection, this model integrates the DA-Net head and residual attention blocks into the traditional U-Net architecture. The results of this adaptation demonstrated the DA-UNet-Plus's commendable performance in skin lesion analysis. The findings clearly showed the effectiveness of com-



binning various models and techniques, revealing substantial advancements in medical image segmentation. This hybrid approach not only enhanced the precision of lesion segmentation but also offered significant potential for future developments in medical imaging technologies. For future work, we plan to further refine both JILDYA-Net and DA-UNet-Plus by testing them on additional datasets and experimenting with various architectural configurations. The main potential future contributions fall under the following headings: class imbalance issues, the trade-off between model complexity *vs.* performance, more generalization to unseen data, more context on segmentation challenges, measured impact of variations in image quality, (such as differences in lighting or resolution), on the models' effectiveness, a more detailed comparison with other SOTA methods.

Our ultimate goal is to integrate these two models into a robust CAD system for skin lesions. In this envisioned system, DA-UNet-Plus will be employed for the precise segmentation and localization of lesions, followed by the application of JILDYA-Net to classify the segmented lesions. This integration will potentially result in a comprehensive and powerful model that may significantly enhance the accuracy and efficiency of skin lesion diagnosis, paving the way for more advanced and reliable medical imaging solutions.

# Chapter 6

## General conclusion

### 6.1 Learned lessons

This thesis has explored the transformative potential of DL in medical image analysis by demonstrating how novel models and techniques can enhance diagnostic accuracy, efficiency, and personalization across various medical domains. This research has contributed to the growing body of knowledge on automated medical diagnostics and provided practical advancements that bridge the gap between DL theory and clinical application by addressing key challenges associated with COVID-19 diagnosis, retinal disease detection, and skin lesion classification and segmentation.

Through the thesis and these contributions, three medical subfields were addressed: chest radiology, ophthalmology, and dermatology. These contributions demonstrate that advanced DL methods can deliver precise and reliable diagnostic outputs. The thesis was divided into four chapters. The first chapter presented the background on medical image analysis and CAD, along with a brief overview of the SOTA DL models and related works relevant to our contributions. The subsequent chapters delved into our contributions to medical imaging in the aforementioned medical subfields.

The first contribution of the thesis focused on the diagnosis of COVID-19 and other lung diseases from CXR and CT scans. We thoroughly examined multiple classification tasks and different models. Our methodology suggests that using attention mechanisms like squeeze-and-excitation attention tends to achieve remarkable performance with an accuracy of up to 99% in the diagnosis of COVID-19 and other lung diseases like pneumonia and lung opacity.

The second contribution concentrated on retinal disease diagnosis and detection from OCT images. In this regard, we introduced HTC-Retina, a novel architectural design using the hybridization of ViT and CNN to benefit the strength and mitigate the limitations of

each approach. This developed method's performance, generalization, and robustness were validated through many experiments using different datasets. The HTC-Retina achieved remarkable results with high accuracy in the classification and identification of different retinal cases. This research highlights the use of hybrid models and advanced computer vision methods like transformers to enable more accurate and robust models and demonstrates significant improvements in the automatic detection of retinal diseases as well as contributes to early-stage diagnosis.

The third contribution of this thesis delved into the diagnosis of melanoma, other skin cancers, and various skin lesions. In this regard, we developed JILDYA-Net, a novel lightweight model explicitly designed for efficiency, making it ideal for deployment in clinical settings with limited computational resources. JILDYA-Net leverages advanced DL techniques, such as the attention mechanism and the FNet architecture, a lightweight transformer design, to enhance performance. The model demonstrated remarkable classification accuracy on challenging online skin lesion datasets, underscoring its effectiveness in diagnostic tasks while maintaining low computational demands.

The final contribution introduced DA-UNet-Plus, an advanced hybrid architecture for medical image segmentation. This model integrates a residual attention U-Net structure with a DANet head, specifically optimized for segmentation tasks. Initially, DA-UNet-Plus was applied to detect, quantify, and segment intraretinal fluid in OCT images, yielding notable results and addressing a critical challenge in retinal fluid diagnostics. Subsequently, the model was employed for the segmentation and localization of skin lesions in dermatoscopic images, demonstrating robust performance in accurately delineating skin lesion patterns. This contribution is particularly impactful, showcasing the model's capacity to segment complex medical images and localize abnormalities effectively, thereby offering significant potential for improving diagnosis and enabling more precise, targeted treatments.

A central theme throughout this thesis has been the pursuit of models that are accurate, robust, and interpretable. This research has made significant strides in ensuring that DL models can be seamlessly integrated into healthcare workflows by leveraging techniques like attention mechanisms, hybrid architectural designs, and advanced optimization methods. These efforts align with the broader vision of personalized medicine, where automated tools enhance clinicians' capabilities and enable more tailored patient care.

While the research presented here has yielded promising results, several areas for further exploration remain open. Future work could focus on refining these models to achieve broader generalization across diverse patient populations, enhancing interpretability and transparency, and investigating additional imaging modalities alongside complementary data types. This may lead to the development of multimodal architectures that could further

advance diagnostic accuracy. Moreover, integrating DL models into clinical practice will require ongoing collaboration with healthcare professionals to ensure these tools effectively meet the practical demands of clinical environments.

Overall, the thesis underscores the significant impact that DL can have on medical image analysis and diagnostics, contributes to the ongoing evolution of precision medicine, and highlights the potential for AI-driven CAD tools to revolutionize patient care. This work aspires to inspire future research in the field, encouraging the development of more efficient and effective models for medical image analysis, classification, and segmentation. These advancements will likely pave the way for more accurate diagnostic systems capable of being deployed in real-world clinical settings and utilized in real time for improved patient outcomes.

## 6.2 Some challenges of CAD systems in medical image analysis

This thesis identified some key challenges in medical image analysis, focusing on data issues, limitations of CAD systems, and complexities in deploying DL models.

### a. Data-related issues

- **Variability in quality** : Noise, artifacts, and device differences affect image quality.
- **Annotation**: Large, annotated datasets are difficult to obtain, requiring expert knowledge.
- **Feature complexity**: Subtle differences make distinguishing between healthy and diseased tissues challenging.
- **Multi-modality integration**: Combining data from various imaging types is technically complex.
- **Sensitivity to variations**: Patient-specific factors and imaging variations complicate maintaining system robustness.

### b. CAD system limitations

- **Generalization**: Systems often perform poorly across diverse populations or imaging settings.

- *Clinical integration*: Implementing CAD systems can disrupt workflows and require clinician training.
- *Real-time processing*: Ensuring accurate, fast results is crucial during diagnostic procedures.
- *Error rates*: Reducing false positives and negatives is critical for reliable diagnoses.
- *Uncertainty*: Many systems lack the ability to quantify prediction confidence, affecting decision-making.

**c. DL model challenges**

- *Overfitting*: Limited training data in medical imaging leads to overfitting.
- *Class imbalance*: Datasets often have few diseased samples, making learning difficult.
- *Interpretability*: DL models are difficult to explain and understand for clinicians.
- *Computational demands*: Large models require significant computational resources.
- *Uncertainty estimation*: Quantifying prediction uncertainty remains a challenge.

These challenges highlight the complexity and limitations of deploying basic CAD systems or DL models in medical image analysis, especially in clinical settings where reliability, robustness, and accuracy are paramount; issues worth considering for future research.

# References

- [1] Ayoub Laouarem, Chafia Kara-Mohamed, El-Bay Bourenane, and Aboubekour Hamdi-Cherif. A deep learning model for cxr-based covid-19 detection. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–5. IEEE, 2021.
- [2] Ayoub Laouarem, Chafia Kara-Mohamed, El-Bay Bourenane, and Aboubekour Hamdi-Cherif. Covid-19 detection using cxr images: simple cnn vs. transfer learning. In *2021 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, pages 1–6. IEEE, 2021.
- [3] Ayoub Laouarem, Chafia Kara-Mohamed, El-Bay Bourenane, and Aboubekour Hamdi-Cherif. Htc-retina: A hybrid retinal diseases classification model using transformer-convolutional neural network from optical coherence tomography images. *Computers in Biology and Medicine*, 178:108726, 2024.
- [4] Mathias Unberath and Andreas Maier. Chapter 1 - medical imaging modalities. In Alejandro F. Frangi, Jerry L. Prince, and Milan Sonka, editors, *Medical Image Analysis*, The MICCAI Society book Series, pages 3–19. Academic Press, 2024.
- [5] Felix Ritter, Tobias Boskamp, A. Homeyer, Hendrik Laue, Michael Schwier, Florian Link, and H.-O. Peitgen. Medical image analysis. *IEEE Pulse*, 2(6):60–70, 2011.
- [6] On a new kind of rays. *Nature*, 53(1369):274–276, January 1896.
- [7] Kunio Doi. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31(4–5):198–211, June 2007.
- [8] Di Wu, Jiachun Ni, Wenbin Fan, Qiong Jiang, Ling Wang, Li Sun, and Zengjin Cai. Opportunities and challenges of computer aided diagnosis in new millennium: A bibliometric analysis from 2000 to 2023. *Medicine*, 102(51):e36703, December 2023.
- [9] Glenn J Jaffe and Joseph Caprioli. Optical coherence tomography to detect and manage retinal disease and glaucoma. *American journal of ophthalmology*, 137(1):156–169, 2004.
- [10] Dan Lipsker. *Clinical examination and differential diagnosis of skin lesions*. Springer, 2013.
- [11] Juri Yanase and Evangelos Triantaphyllou. The seven key challenges for the future of computer-aided diagnosis in medicine. *International Journal of Medical Informatics*, 129:413–422, September 2019.

- [12] Matthew J. Leming, Esther E. Bron, Rose Bruffaerts, Yangming Ou, Juan Eugenio Iglesias, Randy L. Gollub, and Hyungsoon Im. Challenges of implementing computer-aided diagnostic models for neuroimages in a clinical setting. *npj Digital Medicine*, 6(1), July 2023.
- [13] Teerath Kumar, Alessandra Mileo, Rob Brennan, and Malika Bendeche. Image data augmentation approaches: A comprehensive survey and future directions. *arXiv preprint arXiv:2301.02830*, 2023.
- [14] Andreas C Müller and Sarah Guido. *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc.", 2016.
- [15] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
- [16] I Nunes Silva, D Hernane Spatti, R Andrade Flauzino, L Liboni, and S Alves. Artificial neural networks: a practical course. *Springer*, 2016.
- [17] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.
- [18] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017.
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [20] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [22] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [23] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

- [24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [25] Olaf Ronneberger, Philipp Fischer, and Thomas 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*, pages 234–241. Springer, 2015.
- [26] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.
- [27] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.
- [28] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.
- [29] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.
- [30] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [31] Derya Soydaner. Attention mechanism in neural networks: where it comes and where it goes. *Neural Computing and Applications*, 34(16):13371–13385, 2022.
- [32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [33] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- [35] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in neural information processing systems*, 34:9355–9366, 2021.



- [36] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021.
- [37] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.
- [38] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4296–4313, 2022.
- [39] E Oran Brigham and RE Morrow. The fast fourier transform. *IEEE spectrum*, 4(12):63–70, 1967.
- [40] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.
- [41] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.
- [42] Javad Mozaffari, Abdollah Amirkhani, and Shahriar B Shokouhi. A survey on deep learning models for detection of covid-19. *Neural Computing and Applications*, 35(23):16945–16973, 2023.
- [43] Mohamed Esmail Karar, Ezz El-Din Hemdan, and Marwa A Shouman. Cascaded deep learning classifiers for computer-aided diagnosis of covid-19 and pneumonia diseases in x-ray scans. *Complex & Intelligent Systems*, 7(1):235–247, 2021.
- [44] Ali Narin, Ceren Kaya, and Ziynet Pamuk. Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. *Pattern Analysis and Applications*, 24:1207–1220, 2021.
- [45] Bhukya Jabber, Jeevan Lingampalli, Cmak Zeelan Basha, and Azmira Krishna. Detection of covid-19 patients using chest x-ray images with convolution neural network and mobile net. In *2020 3rd International conference on intelligent sustainable systems (ICISS)*, pages 1032–1035. IEEE, 2020.
- [46] Isabella Castiglioni, Davide Ippolito, Matteo Interlenghi, Caterina Beatrice Monti, Christian Salvatore, Simone Schiaffino, Annalisa Polidori, Davide Gandola, Cristina Messa, and Francesco Sardanelli. Artificial intelligence applied on chest x-ray can aid in the diagnosis of covid-19 infection: a first experience from lombardy, italy. *MedRxiv*, pages 2020–04, 2020.
- [47] Antonios Makris, Ioannis Kontopoulos, and Konstantinos Tserpes. Covid-19 detection from chest x-ray images using deep learning and convolutional neural networks. In *11th hellenic conference on artificial intelligence*, pages 60–66, 2020.

- [48] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [49] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.
- [50] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.
- [51] Mohamed Berrimi, Skander Hamdi, Raoudha Yahia Cherif, Abdelouahab Moussaoui, Mourad Oussalah, and Mafaza Chabane. Covid-19 detection from xray and ct scans using transfer learning. In *2021 International Conference of Women in Data Science at Taif University (WiDSTaif)*, pages 1–6. IEEE, 2021.
- [52] A Tahir, Y Qiblawey, A Khandakar, T Rahman, U Khurshid, F Musharavati, M Islam, S Kiranyaz, and M Chowdhury. Deep learning for reliable classification of covid-19. *MERS, and SARS from Chest X-Ray Images*. *arXiv*, 2005, 2021.
- [53] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.
- [54] Gonçalo Marques, Deevyankar Agarwal, and Isabel De la Torre Díez. Automated medical diagnosis of covid-19 through efficientnet convolutional neural network. *Applied soft computing*, 96:106691, 2020.
- [55] Tulin Ozturk, Muhammed Talo, Eylul Azra Yildirim, Ulas Baran Baloglu, Ozal Yildirim, and U Rajendra Acharya. Automated detection of covid-19 cases using deep neural networks with x-ray images. *Computers in biology and medicine*, 121:103792, 2020.
- [56] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
- [57] Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific reports*, 10(1):19549, 2020.
- [58] Aras M Ismael and Abdulkadir Şengür. Deep learning approaches for covid-19 detection based on chest x-ray images. *Expert Systems with Applications*, 164:114054, 2021.
- [59] Nihad K Chowdhury, Md Muhtadir Rahman, and Muhammad Ashad Kabir. Pdcovid-net: a parallel-dilated convolutional neural network architecture for detecting covid-19 from chest x-ray images. *Health information science and systems*, 8(1):27, 2020.

- [60] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [61] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.
- [62] Xiang Yu, Shui-Hua Wang, Xin Zhang, and Yu-Dong Zhang. Detection of covid-19 by googlenet-cod. In *Intelligent Computing Theories and Application: 16th International Conference, ICIC 2020, Bari, Italy, October 2–5, 2020, Proceedings, Part I 16*, pages 499–509. Springer, 2020.
- [63] Yu-Dong Zhang, Suresh Chandra Satapathy, Xin Zhang, and Shui-Hua Wang. Covid-19 diagnosis via densenet and optimization of transfer learning setting. *Cognitive computation*, pages 1–17, 2021.
- [64] Parnian Afshar, Moezedin Javad Rafiee, Farnoosh Naderkhani, Shahin Heidarian, Nastaran Enshaei, Anastasia Oikonomou, Faranak Babaki Fard, Reut Anconina, Keyvan Farahani, Konstantinos N Plataniotis, et al. Human-level covid-19 diagnosis from low-dose ct scans using a two-stage time-distributed capsule network. *Scientific Reports*, 12(1):4827, 2022.
- [65] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. *Advances in neural information processing systems*, 30, 2017.
- [66] Yu-Dong Zhang, Suresh Chandra Satapathy, Shuaiqi Liu, and Guang-Run Li. A five-layer deep convolutional neural network with stochastic pooling for chest ct-based covid-19 diagnosis. *Machine vision and applications*, 32:1–13, 2021.
- [67] Kapil Gupta and Varun Bajaj. Deep learning models-based ct-scan image classification for automated screening of covid-19. *Biomedical Signal Processing and Control*, 80:104268, 2023.
- [68] Ioannis D Apostolopoulos and Tzani A Mpesiana. Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. *Physical and engineering sciences in medicine*, 43:635–640, 2020.
- [69] Yudong Zhang, Suresh Chandra Satapathy, Li-Yao Zhu, Juan Manuel Górriz, and Shuihua Wang. A seven-layer convolutional neural network for chest ct-based covid-19 diagnosis using stochastic pooling. *IEEE Sensors Journal*, 22(18):17573–17582, 2020.
- [70] Maximilian Treder, Jost Lennart Lauermann, and Nicole Eter. Automated detection of exudative age-related macular degeneration in spectral domain optical coherence tomography using deep learning. *Graefes Archive for Clinical and Experimental Ophthalmology*, 256:259–265, 2018.

- [71] Yibiao Rong, Dehui Xiang, Weifang Zhu, Kai Yu, Fei Shi, Zhun Fan, and Xinjian Chen. Surrogate-assisted retinal oct image classification based on convolutional neural networks. *IEEE journal of biomedical and health informatics*, 23(1):253–263, 2018.
- [72] Shenghua He, Jie Zheng, Akiko Maehara, Gary Mintz, Dalin Tang, Mark Anastasio, and Hua Li. Convolutional neural network based automatic plaque characterization for intracoronary optical coherence tomography images. In *Medical Imaging 2018: Image Processing*, volume 10574, pages 800–806. SPIE, 2018.
- [73] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.
- [74] Feng Li, Hua Chen, Zheng Liu, Xuedian Zhang, and Zhizheng Wu. Fully automated detection of retinal disorders by image-based deep learning. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 257:495–505, 2019.
- [75] Naohiro Motozawa, Guangzhou An, Seiji Takagi, Shohei Kitahata, Michiko Mandai, Yasuhiko Hiram, Hideo Yokota, Masahiro Akiba, Akitaka Tsujikawa, Masayo Takahashi, et al. Optical coherence tomography-based deep-learning models for classifying normal and age-related macular degeneration and exudative and non-exudative age-related macular degeneration changes. *Ophthalmology and therapy*, 8:527–539, 2019.
- [76] Frank Y Shih and Himanshu Patel. Deep learning classification on optical coherence tomography retina images. *International Journal of Pattern Recognition and Artificial Intelligence*, 34(08):2052002, 2020.
- [77] Mohamed Berrimi and Abdelouaheb Moussaoui. Deep learning for identifying and classifying retinal diseases. In *2020 2nd International Conference on computer and information sciences (ICIS)*, pages 1–6. IEEE, 2020.
- [78] Pratul P. Srinivasan, Leo A. Kim, Priyatham S. Mettu, Scott W. Cousins, Grant M. Comer, Joseph A. Izatt, and Sina Farsiu. Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images. *Biomedical Optics Express*, 5(10):3568, September 2014.
- [79] Daniel Kermany. Labeled optical coherence tomography (oct) and chest x-ray images for classification, 2018.
- [80] Karri Karthik and Manjunatha Mahadevappa. Convolution neural networks for optical coherence tomography (oct) image classification. *Biomedical Signal Processing and Control*, 79:104176, 2023.
- [81] Feng Li, Hua Chen, Zheng Liu, Xue-dian Zhang, Min-shan Jiang, Zhi-zheng Wu, and Kai-qian Zhou. Deep learning-based automated detection of retinal diseases using optical coherence tomography images. *Biomedical optics express*, 10(12):6204–6226, 2019.
- [82] Gokhan Altan. Deepoct: An explainable deep learning architecture to analyze macular edema on oct images. *Engineering Science and Technology, an International Journal*, 34:101091, 2022.

- [83] O.S. Naren. Retinal oct - c8, 2021, version 2, 2022.
- [84] M. Rastgoo F. Meriaudeau G. Lemaitre, J. Massich. Seri dataset, 2019.
- [85] Junyong Shen, Yan Hu, Xiaoqing Zhang, Yan Gong, Ryo Kawasaki, and Jiang Liu. Structure-oriented transformer for retinal diseases grading from oct images. *Computers in Biology and Medicine*, 152:106445, 2023.
- [86] Zongqing Ma, Qiaoxue Xie, Pinxue Xie, Fan Fan, Xinxiao Gao, and Jiang Zhu. Hctnet: a hybrid convnet-transformer network for retinal optical coherence tomography image classification. *Biosensors*, 12(7):542, 2022.
- [87] Pramit Dutta, Khaleda Akther Sathi, Md. Azad Hossain, and M. Ali Akber Dewan. Conv-vit: A convolution and vision transformer-based hybrid feature extraction method for retinal disease detection. *Journal of Imaging*, 9(7):140, July 2023.
- [88] Huajie Wen, Jian Zhao, Shaohua Xiang, Lin Lin, Chengjian Liu, Tao Wang, Lin An, Lixin Liang, and Bingding Huang. Towards more efficient ophthalmic disease classification and lesion location via convolution transformer. *Computer Methods and Programs in Biomedicine*, 220:106832, June 2022.
- [89] Said Akça, Zeynep Garip, Ekin Ekin, and Furkan Atban. Automated classification of choroidal neovascularization, diabetic macular edema, and drusen from retinal oct images using vision transformers: a comparative study. *Lasers in Medical Science*, 39(1), May 2024.
- [90] Thomas Schlegl, Sebastian M. Waldstein, Hrvoje Bogunovic, Franz Endstraßer, Amir Sadeghipour, Ana-Maria Philip, Dominika Podkowinski, Bianca S. Gerendas, Georg Langs, and Ursula Schmidt-Erfurth. Fully automated detection and quantification of macular fluid in oct using deep learning. *Ophthalmology*, 125(4):549–558, April 2018.
- [91] Qiaoliang Li, Shiyu Li, Zhuoying He, Huimin Guan, Runmin Chen, Ying Xu, Tao Wang, Suwen Qi, Jun Mei, and Wei Wang. Deepretina: layer segmentation of retina in oct images using deep learning. *Translational vision science & technology*, 9(2):61–61, 2020.
- [92] Cecilia S. Lee, Ariel J. Tying, Nicolaas P. Deruyter, Yue Wu, Ariel Rokem, and Aaron Y. Lee. Deep-learning based, automated segmentation of macular edema in optical coherence tomography. *Biomedical Optics Express*, 8(7):3440, June 2017.
- [93] Qiong Chen, Lirong Zeng, and Cong Lin. A deep network embedded with rough fuzzy discretization for oct fundus image segmentation. *Scientific Reports*, 13(1), January 2023.
- [94] Xinbo Ren, Yuxia Zhao, Jingbo Fan, Haiyuan Wu, Qian Chen, and Takashi Kubo. Semantic segmentation of superficial layer in intracoronary optical coherence tomography based on cropping-merging and deep learning. *Infrared Physics & Technology*, 129:104542, 2023.

- [95] Xin Yang, Zhiqiang Li, Yingqing Guo, and Dake Zhou. Dcu-net: a deformable convolutional neural network based on cascade u-net for retinal vessel segmentation. *Multimedia Tools and Applications*, 81(11):15593–15607, February 2022.
- [96] Zeeshan Ahmed, Shahbaz Qamar Panhwar, Attiya Baqai, Fahim Aziz Umrani, Munawar Ahmed, and Arbaaz Khan. Deep learning based automated detection of intraretinal cystoid fluid. *International Journal of Imaging Systems and Technology*, 32(3):902–917, October 2021.
- [97] Freerk G Venhuizen, Bram Van Ginneken, Bart Liefers, Freekje Van Asten, Vivian Schreur, Sascha Fauser, Carel Hoyng, Thomas Theelen, and Clara I Sánchez. Deep learning approach for the detection and quantification of intraretinal cystoid fluid in multivendor optical coherence tomography. *Biomedical optics express*, 9(4):1545–1569, 2018.
- [98] Abhijit Guha Roy, Sailesh Conjeti, Sri Phani Krishna Karri, Debdoot Sheet, Amin Katouzian, Christian Wachinger, and Nassir Navab. Relaynet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. *Biomedical optics express*, 8(8):3627–3642, 2017.
- [99] Xiaoming Liu, Shaocheng Wang, Jun Cao, Ying Zhang, and Man Wang. Uncertainty-guided self-ensembling model for semi-supervised segmentation of multiclass retinal fluid in optical coherence tomography images. *International Journal of Imaging Systems and Technology*, 32(1):369–386, September 2021.
- [100] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [101] Xiaoming Liu, Jun Cao, Tianyu Fu, Zhifang Pan, Wei Hu, Kai Zhang, and Jun Liu. Semi-supervised automatic segmentation of layer and fluid region in retinal optical coherence tomography images using adversarial learning. *IEEE Access*, 7:3046–3061, 2018.
- [102] Ramesh Venkatesh, Nikitha Gurram Reddy, Pranjal Mishra, Sameeksha Agrawal, Deepashri Mutalik, Naresh Kumar Yadav, and Jay Chhablani. Spectral domain oct features in type 2 macular telangiectasia (type 2 mactel): its relevance with clinical staging and visual acuity. *International Journal of Retina and Vitreous*, 8(1), April 2022.
- [103] Shunichi Jinnai, Naoya Yamazaki, Yuichiro Hirano, Yohei Sugawara, Yuichiro Ohe, and Ryuji Hamamoto. The development of a skin cancer classification system for pigmented skin lesions using deep learning. *Biomolecules*, 10(8):1123, 2020.
- [104] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.
- [105] Mario Fernando Jojoa Acosta, Liesle Yail Caballero Tovar, Maria Begonya Garcia-Zapirain, and Winston Spencer Percybrooks. Melanoma diagnosis using deep learning techniques on dermatoscopic images. *BMC Medical Imaging*, 21:1–11, 2021.

- [106] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- [107] MA Rasel, Unaizah H Obaidallah, and Sameem Abdul Kareem. convolutional neural network-based skin lesion classification with variable nonlinear activation functions. *IEEE Access*, 10:83398–83414, 2022.
- [108] Bhuvaneshwari Shetty, Roshan Fernandes, Anisha P Rodrigues, Rajeswari Chengoden, Sweta Bhattacharya, and Kuruva Lakshmana. Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. *Scientific Reports*, 12(1):18134, 2022.
- [109] Muhammad Azeem, Kaveh Kiani, Taha Mansouri, and Nathan Topping. Skinlesnet: Classification of skin lesions and detection of melanoma cancer using a novel multi-layer deep convolutional neural network. *Cancers*, 16(1):108, 2023.
- [110] Imran Iqbal, Muhammad Younus, Khuram Walayat, Mohib Ullah Kakar, and Jinwen Ma. Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images. *Computerized medical imaging and graphics*, 88:101843, 2021.
- [111] Long Hoang, Suk-Hwan Lee, Eung-Joo Lee, and Ki-Ryong Kwon. Multiclass skin lesion classification using a novel lightweight deep learning framework for smart healthcare. *Applied Sciences*, 12(5):2677, March 2022.
- [112] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.
- [113] Parvathaneni Naga Srinivasu, Jalluri Gnana SivaSai, Muhammad Fazal Ijaz, Akash Kumar Bhoi, Wonjoon Kim, and James Jin Kang. Classification of skin disease using deep learning neural networks with mobilenet v2 and lstm. *Sensors*, 21(8):2852, 2021.
- [114] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.
- [115] Ismail Oztel, Gozde Yolcu Oztel, and Veysel Harun Sahin. Deep learning-based skin diseases classification using smartphones. *Advanced Intelligent Systems*, 5(12):2300211, 2023.
- [116] Razia Sulthana A, Vinay Chamola, Zain Hussain, Faisal Albalwy, and Amir Hussain. A novel end-to-end deep convolutional neural network based skin lesion classification framework. *Expert Systems with Applications*, 246:123056, July 2024.
- [117] Ayoub Laouarem, Chafia Kara-Mohamed, El-Bay Bourennane, and Aboubekur Hamdi-Cherif. Jildya-net: An efficient lightweight multi-class classification architecture for skin lesions. Under review, 2024.

- [118] Xiaoyu He, Yong Wang, Shuang Zhao, and Chunli Yao. Deep metric attention learning for skin lesion classification in dermoscopy images. *Complex & Intelligent Systems*, 8(2):1487–1504, 2022.
- [119] Hui Wang, Qianqian Qi, Weijia Sun, Xue Li, Boxin Dong, and Chunli Yao. Classification of skin lesions with generative adversarial networks and improved mobilenetv2. *International Journal of Imaging Systems and Technology*, 33(5):1561–1576, 2023.
- [120] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.
- [121] Hui Wang, Qianqian Qi, Weijia Sun, Xue Li, and Chunli Yao. Classification of clinical skin lesions with double-branch networks. *Frontiers in Medicine*, 10:1114362, 2023.
- [122] Yousef S. Alsahafi, Mohamed A. Kassem, and Khalid M. Hosny. Skin-net: a novel deep residual network for skin lesions classification using multilevel feature extraction and cross-channel correlation with detection of outlier. *Journal of Big Data*, 10(1), June 2023.
- [123] Halpern A. Codella N.C.F. Celebi M.E. Combalia M. Gutman D. Helba B. Kittler H. Rotemberg V. Tschandl P. Malveyh, J. Skin lesion analysis towards melanoma detection (isic 2019), 2019.
- [124] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, Allan Halpern, Brian Helba, Harald Kittler, Kivanc Kose, Steve Langer, Konstantinos Liopyris, Josep Malveyh, Shenara Musthaq, Jabpani Nanda, Ofer Reiter, George Shih, Alexander Stratigos, Philipp Tschandl, Jochen Weber, and H. Peter Soyer. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data*, 8(1), January 2021.
- [125] Syed Qasim Gilani, Tehreem Syed, Muhammad Umair, and Oge Marques. Skin cancer classification using deep spiking neural network. *Journal of Digital Imaging*, 36(3):1137–1147, January 2023.
- [126] Amirhossein Tavanaei, Masoud Ghodrati, Saeed Reza Kheradpisheh, Timothée Masquelier, and Anthony Maida. Deep learning in spiking neural networks. *Neural Networks*, 111:47–63, March 2019.
- [127] Roman C Maron, Achim Hekler, Eva Krieghoff-Henning, Max Schmitt, Justin G Schlager, Jochen S Utikal, and Titus J Brinker. Reducing the impact of confounding factors on skin cancer classification via image segmentation: technical model study. *Journal of Medical Internet Research*, 23(3):e21695, 2021.
- [128] Chen Zhao, Renjun Shuai, Li Ma, Wenjia Liu, and Menglin Wu. Segmentation of skin lesions image based on u-net++. *Multimedia Tools and Applications*, 81(6):8691–8717, 2022.
- [129] Ranpreet Kaur, Hamid GholamHosseini, and Roopak Sinha. Skin lesion segmentation using an improved framework of encoder-decoder based convolutional neural network. *International Journal of Imaging Systems and Technology*, 32(4):1143–1158, 2022.



- [130] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International workshop on machine learning in medical imaging*, pages 379–387. Springer, 2017.
- [131] Rui Gu, Lituan Wang, and Lei Zhang. De-net: A deep edge network with boundary information for automatic skin lesion segmentation. *Neurocomputing*, 468:71–84, 2022.
- [132] Yuta Yanagisawa, Kosuke Shido, Kaname Kojima, and Kenshi Yamasaki. Convolutional neural network-based skin image segmentation model to improve classification of skin diseases in conventional and non-standardized picture images. *Journal of Dermatological Science*, 109(1):30–36, 2023.
- [133] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.
- [134] Ying Wang, Jie Su, Qiuyu Xu, and Yixin Zhong. A collaborative learning model for skin lesion segmentation and classification. *Diagnostics*, 13(5):912, 2023.
- [135] Peng Chen, Sa Huang, and Qing Yue. Skin lesion segmentation using recurrent attentional convolutional networks. *IEEE Access*, 10:94007–94018, 2022.
- [136] Ju Zhang, Weidong Pan, Beng Wang, Qing Chen, and Yun Cheng. Multi-scale aggregation networks with flexible receptive fields for melanoma segmentation. *Biomedical Signal Processing and Control*, 78:103950, 2022.
- [137] Mohammad D Alahmadi and Wajdi Alghamdi. Semi-supervised skin lesion segmentation with coupling cnn and transformer features. *IEEE Access*, 10:122560–122569, 2022.
- [138] Ning Zhang, Long Yu, Dezhi Zhang, Weidong Wu, Shengwei Tian, and Xiaojing Kang. Apt-net: Adaptive encoding and parallel decoding transformer for medical image segmentation. *Computers in Biology and Medicine*, 151:106292, 2022.
- [139] Jian Li, Jiawei Wang, Fengwu Lin, Wenqi Wu, Zhao-Min Chen, Ali Asghar Heidari, and Huiling Chen. Dseunet: A lightweight unet for dynamic space grouping enhancement for skin lesion segmentation. *Expert Systems with Applications*, page 124544, 2024.
- [140] Xuqiong Luo, Hao Zhang, Xiaofei Huang, Hongfang Gong, and Jin Zhang. Dbnet-si: Dual branch network of shift window attention and inception structure for skin lesion segmentation. *Computers in Biology and Medicine*, 170:108090, 2024.
- [141] Shailendra K Saxena. *Coronavirus disease 2019 (COVID-19): epidemiology, pathogenesis, diagnosis, and therapeutics*. Springer nature, 2020.
- [142] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.

- [143] Samir S Yadav and Shivajirao M Jadhav. Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big data*, 6(1):1–18, 2019.
- [144] Qing Li, Weidong Cai, Xiaogang Wang, Yun Zhou, David Dagan Feng, and Mei Chen. Medical image classification with convolutional neural network. In *2014 13th international conference on control automation robotics & vision (ICARCV)*, pages 844–848. IEEE, 2014.
- [145] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.
- [146] Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q Duong, and Marzyeh Ghassemi. Covid-19 image data collection: Prospective predictions are the future. *arXiv preprint arXiv:2006.11988*, 2020.
- [147] Muhammad EH Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, et al. Can ai help in screening viral and covid-19 pneumonia? *Ieee Access*, 8:132665–132676, 2020.
- [148] Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M Zughaier, Muhammad Salman Khan, et al. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in biology and medicine*, 132:104319, 2021.
- [149] Santanu Roy, Mrinal Tyagi, Vibhuti Bansal, and Vikas Jain. Svd-clahe boosting and balanced loss function for covid-19 detection from an imbalanced chest x-ray dataset. *Computers in Biology and Medicine*, 150:106092, 2022.
- [150] Eduardo Soares, Plamen Angelov, Sarah Biaso, Michele Higa Froes, and Daniel Kanda Abe. Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification. *MedRxiv*, pages 2020–04, 2020.
- [151] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.
- [152] Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, and Humphrey Shi. Escaping the big data paradigm with compact transformers. *arXiv preprint arXiv:2104.05704*, 2021.
- [153] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in neural information processing systems*, 34:3965–3977, 2021.

- [154] Stéphane d’Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International conference on machine learning*, pages 2286–2296. PMLR, 2021.
- [155] Xiaozheng Xie, Jianwei Niu, Xuefeng Liu, Zhengsu Chen, Shaojie Tang, and Shui Yu. A survey on incorporating domain knowledge into deep learning for medical image analysis. *Medical Image Analysis*, 69:101985, 2021.
- [156] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.
- [157] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [158] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 646–661. Springer, 2016.
- [159] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.
- [160] Asher Trockman and J Zico Kolter. Patches are all you need? *arXiv preprint arXiv:2201.09792*, 2022.
- [161] Bin Sun, Jun Li, Ming Shao, and Yun Fu. Lprnet: Lightweight deep network by low-rank pointwise residual convolution. *arXiv preprint arXiv:1910.11853*, 2019.
- [162] Bin Sun, Jun Li, Ming Shao, and Yun Fu. Lrprnet: Lightweight deep network by low-rank pointwise residual convolution. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8):4440–4450, 2021.
- [163] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [164] Ursula Schmidt-Erfurth and Sebastian M Waldstein. A paradigm shift in imaging biomarkers in neovascular age-related macular degeneration. *Progress in Retinal and eye Research*, 50:1–24, 2016.
- [165] Sebastian M Waldstein, Christian Simader, Giovanni Staurengi, N Victor Chong, Paul Mitchell, Glenn J Jaffe, Chengxing Lu, Todd A Katz, and Ursula Schmidt-Erfurth. Morphology and visual acuity in aflibercept and ranibizumab therapy for neovascular age-related macular degeneration in the view trials. *Ophthalmology*, 123(7):1521–1529, 2016.
- [166] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019.
- [167] Alfonso Baldi, Paola Pasquali, and Enrico P Spugnini. *Skin cancer: a practical approach*. Springer Science & Business Media, 2013.

- [168] Howard L Kaufman and Janice M Mehnert, editors. *Melanoma*. Cancer treatment and research. Springer International Publishing, Cham, Switzerland, 1 edition, nov 2015.
- [169] ME Vestergaard, PHPM Macaskill, PE Holt, and SW Menzies. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. *British Journal of Dermatology*, 159(3):669–676, 2008.
- [170] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.
- [171] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.
- [172] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019.
- [173] Meng-Hao Guo, Zheng-Ning Liu, Tai-Jiang Mu, and Shi-Min Hu. Beyond self-attention: External attention using two linear layers for visual tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5436–5447, 2022.
- [174] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.
- [175] Sarmad Maqsood and Robertas Damaševičius. Multiclass skin lesion localization and classification using deep learning based features fusion and selection framework for smart healthcare. *Neural networks*, 160:238–258, 2023.

# Appendix A

## Glossary

| Lang.               | Concept                                  | Brief description                                                                                                                                                                                                                                                                                       |
|---------------------|------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| General concepts    |                                          |                                                                                                                                                                                                                                                                                                         |
| AR                  | آلية الانتباه                            | The attention mechanism dynamically weights input features based on their relevance to a given context, enabling models to selectively focus on relevant data. This mechanism enhances performance by allowing models to capture long-range dependencies and contextual relationships more effectively. |
| EN                  | Attention mechanism                      |                                                                                                                                                                                                                                                                                                         |
| FR                  | Mécanisme d'attention                    |                                                                                                                                                                                                                                                                                                         |
|                     |                                          |                                                                                                                                                                                                                                                                                                         |
| AR                  | التصنيف باستخدام التعلم العميق           | Classification with deep learning (DL) involves training neural networks to sort data into predefined categories. Models like CNNs and recurrent neural networks (RNNs) learn complex patterns from large datasets through multiple layers, enabling effective tasks such as medical image recognition. |
| EN                  | Classification using deep learning       |                                                                                                                                                                                                                                                                                                         |
| FR                  | Classification par apprentissage profond |                                                                                                                                                                                                                                                                                                         |
|                     |                                          |                                                                                                                                                                                                                                                                                                         |
| AR                  | شبكة عصبونية إلتقافية                    | A convolutional neural network (CNN) is a deep learning model, attempting to mimic the human visual system by using layers of filters to detect patterns. It is effective for tasks like classification, detection, and segmentation.                                                                   |
| EN                  | Convolutional neural network             |                                                                                                                                                                                                                                                                                                         |
| FR                  | Réseau de neurones convolutifs           |                                                                                                                                                                                                                                                                                                         |
|                     |                                          |                                                                                                                                                                                                                                                                                                         |
| AR                  | البنية الخفيفة                           | Lightweight architectures are designed to be computationally efficient and require fewer resources, often using fewer parameters and operations. They aim to achieve faster inference and lower memory usage while maintaining reasonable performance. Examples include JILDYA-Net.                     |
| EN                  | Lightweight architecture                 |                                                                                                                                                                                                                                                                                                         |
| FR                  | Architecture légère                      |                                                                                                                                                                                                                                                                                                         |
|                     |                                          |                                                                                                                                                                                                                                                                                                         |
| HTC-Retina concepts |                                          |                                                                                                                                                                                                                                                                                                         |

| Lang.               | Concept                                                                   | Brief description                                                                                                                                                                                                                                                                                                                                                                                                                    |
|---------------------|---------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| AR                  | النموذج الهجين للتعرف على أمراض الشبكية                                   | HTC-retina is an original hybrid model integrating the strengths of the ViTs and CNNs to enhance the classification of seven retinal diseases by performing in-depth analysis of OCT images and extracting the most pertinent features associated with each condition.                                                                                                                                                               |
| EN                  | HTC-Retina                                                                |                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| FR                  | HTC-Retina                                                                |                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|                     |                                                                           |                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| AR                  | وحدة الرُقْعِ الْإِلْتِفَافِيَّةِ وتضمين التوكينات                        | The convolutional patch and token embedding (CPTe) module is a simple convolutional network that collects low-level local characteristics from the input image and creates patches and overlapping tokens by running it through a convolutional block. This block is a standard architecture that includes a convolution, ReLU activation, and a max pool.                                                                           |
| EN                  | Convolutional patch and token embedding module                            |                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| FR                  | Module d’encodage par patches convolutifs et par tokens                   |                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|                     |                                                                           |                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| AR                  | مَحْوَلٌ بصري                                                             | Visual transformers (ViTs) are a type of DL model designed for image processing tasks, inspired by transformer architecture used in natural language processing. They operate by dividing an image into patches, embedding these patches, and then applying self-attention mechanisms to capture complex dependencies between them. This allows ViTs to effectively handle long-range interactions and global context within images. |
| EN                  | Visual transformer                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| FR                  | Transformateur visuel                                                     |                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|                     |                                                                           |                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| JILDYA-Net concepts |                                                                           |                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| AR                  | جلدية نت                                                                  | JILDYA-Net is an original lightweight model destined for skin lesions classification. It uses an attention-based structural feature enhancement (ASFE) block to improve feature extraction in dermoscopic images and an encoder-based FNet for efficient processing. It also integrates an external attention module (EAM) to refine representations and enhance model stability and convergence.                                    |
| EN                  | JILDYA-Net                                                                |                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| FR                  | JILDYA-Net                                                                |                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|                     |                                                                           |                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| AR                  | كتلة تعزيز الميزات الهيكلية المعتمدة على الانتباه                         | The attention-based structural feature enhancement (ASFE) block is a novel convolutional attention component used to enhance skin lesion features, ensuring the capture of corresponding features in each patch.                                                                                                                                                                                                                     |
| EN                  | Attention-based structural feature enhancement block                      |                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| FR                  | Bloc d’amélioration de caractéristiques structurales basé sur l’attention |                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|                     |                                                                           |                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| AR                  | وحدة الانتباه الخارجي                                                     | The external attention module (EAM) is used in JILDYA-Net to refine learned representations and emphasize relevant features, thereby accelerating convergence and enhancing model stability during training.                                                                                                                                                                                                                         |
| EN                  | External attention module                                                 |                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| FR                  | Module d’attention externe                                                |                                                                                                                                                                                                                                                                                                                                                                                                                                      |

| Lang.                 | Concept                                | Brief description                                                                                                                                                                                                                                                                                                                                                        |
|-----------------------|----------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| AR                    | شبكة مبنية على تحويلات فورييه          | FNet is a neural Network architecture that substitutes self-attention mechanism with a fast Fourier transform (FFT) approach to process sequences. This method reduces computational complexity and memory consumption, enabling efficient handling of long sequences. By leveraging the FFT, FNet maintains performance while accelerating both training and inference. |
| EN                    | FNet                                   |                                                                                                                                                                                                                                                                                                                                                                          |
| FR                    | FNet                                   |                                                                                                                                                                                                                                                                                                                                                                          |
|                       |                                        |                                                                                                                                                                                                                                                                                                                                                                          |
| AR                    | وحدة الانتباه بالكتلة الالتفافية       | The convolutional block attention module (CBAM) is used to refine image features by applying both channel and spatial attention mechanisms, focusing on important structural features and reducing irrelevant information. This enhances the representation of key regions, like lesion targets.                                                                         |
| EN                    | Convolutional block attention module   |                                                                                                                                                                                                                                                                                                                                                                          |
| FR                    | Module d’attention par bloc convolutif |                                                                                                                                                                                                                                                                                                                                                                          |
|                       |                                        |                                                                                                                                                                                                                                                                                                                                                                          |
| DA-UNet-Plus concepts |                                        |                                                                                                                                                                                                                                                                                                                                                                          |
| AR                    |                                        | The DA-UNet-Plus model is a hybrid architecture that combines the strengths of the traditional U-Net architecture with the innovative dual attention network (DANet) head, incorporating position and channel attention mechanisms and squeeze-and-excitation residual blocks based on SE attention.                                                                     |
| EN                    | DA-UNet-Plus                           |                                                                                                                                                                                                                                                                                                                                                                          |
| FR                    | DA-UNet-Plus                           |                                                                                                                                                                                                                                                                                                                                                                          |
|                       |                                        |                                                                                                                                                                                                                                                                                                                                                                          |
| AR                    |                                        | The U-Net is a fully convolutional network used in medical image segmentation. It features contracting and expanding paths that integrate low-level and high-level features, enhancing feature localization, training efficiency, and segmentation performance, particularly in overcoming the limitations of DANet.                                                     |
| EN                    | U-Net                                  |                                                                                                                                                                                                                                                                                                                                                                          |
| FR                    | U-Net                                  |                                                                                                                                                                                                                                                                                                                                                                          |
|                       |                                        |                                                                                                                                                                                                                                                                                                                                                                          |
| AR                    | وحدة الانتباه إلى الموضع               | The PAM is responsible for enhancing spatial focus by emphasizing important regions based on their location. It is destined to improve feature detection and segmentation accuracy in complex images.                                                                                                                                                                    |
| EN                    | Position attention module              |                                                                                                                                                                                                                                                                                                                                                                          |
| FR                    | Module d’attention de position         |                                                                                                                                                                                                                                                                                                                                                                          |
|                       |                                        |                                                                                                                                                                                                                                                                                                                                                                          |
| AR                    | وحدة الانتباه إلى القناة               | The CAM identifies the relationships between various channel maps. It updates each channel map by incorporating a weighted sum of all channel maps, employing a correlation-based method similar to the one used in the PAM.                                                                                                                                             |
| EN                    | Channel attention module               |                                                                                                                                                                                                                                                                                                                                                                          |
| FR                    | Module d’attention de canal            |                                                                                                                                                                                                                                                                                                                                                                          |
|                       |                                        |                                                                                                                                                                                                                                                                                                                                                                          |
| AR                    | شبكة الانتباه المزدوج                  | The DANet architecture uses a correlation-based attention mechanism to combine local features with global dependencies, addressing intra-class variability. Its dual attention system integrates spatial (via PAM) and channel-wise (via CAM) contextual information, enhancing performance in detailed feature analysis tasks like retinal fluid segmentation.          |
| EN                    | DANet architecture                     |                                                                                                                                                                                                                                                                                                                                                                          |
| FR                    | Architecture DANet                     |                                                                                                                                                                                                                                                                                                                                                                          |