République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université A.MIRA-BEJAIA
Faculté Sciences Exactes
Département Informatique
Laboratory of Medical Informatics and Intelligent and Dynamic Environments (LIMED)

THÈSE

Présentée par

AIT KACI AZZOU Samira
(Epouse BOUKERRAM)

Pour l'obtention du grade de

DOCTEUR EN SCIENCES

Filière : Informatique
Option : Cloud Computing

Thème

Analyse d'Images Médicales pour la Rétinopathie Diabétique à l'aide de Techniques Avancées d'Intelligence Artificielle : Optimisation Bayésienne et Architectures Hybrides CNN–Vision Transformer

Soutenue le 08 Janvier 2026 Devant le Jury composé de :

Nom et Prénom	Grade		
Mr K. AMROUN	Professeur	Univ. de Bejaia	Président
Mme D. BOUKREDERA	M.C.A.	Univ. de Bejaia	Rapporteur
Mme H. CHERROUN	Professeure	Univ. de Laghouat	Examinateur
Mme N. KAMEL	Professeure	Univ. Setif 1.	Examinateur
Mr A. ACHROUFENE	M.C.A.	Univ. de Bejaia	Invité

Année Universitaire : 2025/2026

Remerciements

Je tiens à exprimer ma profonde gratitude à toutes celles et ceux qui, de près ou de loin, ont contribué à la réalisation de cette thèse.

En tout premier lieu, j'adresse mes remerciements les plus respectueux à Madame la Docteure **Djamila BOUKREDERA**, qui a accepté d'encadrer ce travail. Je lui suis infiniment reconnaissante pour sa disponibilité, ses conseils avisés, sa confiance et ses encouragements constants.

J'exprime également toute ma gratitude aux membres du jury pour l'honneur qu'ils me font en évaluant ce travail. Je remercie chaleureusement le **Professeur Kamel AMROUN**, Président du jury, ainsi que **Madame la Professeure Nadjet KAMEL**, **Madame la Professeure Hadda CHERROUN** et **Monsieur le Docteur Achour ACHROUFENE**, pour l'intérêt porté à ce manuscrit et pour la bienveillance et la rigueur de leur évaluation.

J'aimerais également exprimer ma gratitude à mon amie et collègue Adel Karima, ses encouragements constants ont grandement contribué à la réalisation de cette thèse.

Mes sincères remerciements vont aussi à mes étudiants de Master en informatique, dont l'aide précieuse a été déterminante pour la réalisation des expérimentations nécessaires à la validation des modèles développés.

Je souhaite également remercier mes collègues enseignants du Département d'Informatique pour leur soutien moral, ainsi que l'ensemble des membres du laboratoire LIMED, qui ont su créer un cadre de travail à la fois stimulant et convivial.

Ma gratitude s'étend aussi aux responsables et au personnel de la Faculté des Sciences Exactes, et plus particulièrement à ceux du Département d'Informatique, pour leur disponibilité et leur appui constant.

Mes remerciements les plus affectueux vont à ma Mère, à mes frères et sœurs, ainsi qu'à mes beaux-frères et belles-sœurs, pour leur affection, leur présence et leur soutien indéfectible.

Enfin, une pensée toute particulière à mon fils **Yanis**, dont les encouragements et l'amour m'ont donné la force et la motivation nécessaires pour mener à bien ce travail.

À la mémoire de mon cher père Saadi et de mon cher mari Abdellah Boukerram, dont l'absence demeure une douleur, mais dont l'amour et les valeurs continuent d'éclairer mon chemin.

Mes Publications

- **Samira AIT KACI AZZOU,** Djamila BOUKREDERA, Yacine ABICHE, Akram AMOKRANE, and Achour ACHROUFENE. Early detection and severity grading of diabetic retinopathy using fine-tuned deep learning models with automated hyperparameter optimization. *SN Computer Science*, 6(6) :1–19, 2025.

- **Samira AIT KACI AZZOU,** Djamila BOUKREDERA, and Sifeddine BAOUZ. A hybrid cnn- transformer approach for precise three-class diabetic retinopathy classification. *Jor- danian Journal of Computers and Information Technology (JJCIT)*, 11(03) :279–299, Sept 2025.

Mes Communications

- **Samira AIT KACI AZZOU,** Djamila BOUKREDERA, Imen BENADJAOUD, « Proposition d'une Architecture GAN pour le Prétraitement et la Classification des Images de la Rétinopathie Diabétique **»,** Colloque International MOAD'2022 (Méthodes et Outils d'Aide à la Décision) Université de Béjaia, 15 - 17 Novembre 2022.
- Yacine ABICHE, Akram AMOKRANE, **Samira AIT KACI AZZOU**, Djamila BOUKREDERA, «Enhacing Diabetic Retinopathy Detection using Transfert Learning**»** 4 eme edition du colloque international sur les objets et les systems connectés (COC'2023) Tunisie, 22-23 juin 2023
- Ferroudja BELLAL,Tinhinane BESSAA ,**Samira AIT KACI AZZOU**, Djamila BOUKREDERA, « Analyse de l'Influence des Caractéristiques sur la Performance du Modèle Prédictif de la Rétinopathie Diabétique ». 4 ème édition du colloque international sur les objets et les systèmes connectés (COC'2023) Tunisie, 22-23 juin 2023
- **Samira AIT KACI AZZOU** Diamila BOUKREDERA; Siffedine BAOUZ. » Early Diabetic Retinopathy Detection with Vision Transformers and Optimized Data Preprocessing. 17th African Conference on Research in Computer Science and Applied Mathematics - Digital Sciences in Africa (CARI'2024), 23 to 26 November 2024 Bejaia-Algeria,

Table des matières

Bibliographie **142**

Table des figures

Liste des tableaux

Liste des acronymes

AMIR Anomalies Microvasculaires Intrarétiniennes

APTOS Asia Pacific Tele-Ophthalmology Society (jeu de données APTOS 2019)

GPU Graphics Processing Unit

KNN K-Nearest Neighbors

AtRD Auto-tuned Retino Detection

AtR3C Auto-tuned Retino 3-Class

AtR5C Auto-tuned Retino 5-Class

CLAHEC Contrast Limited Adaptive Histogram Equalization

CNN Convolutional Neural Networks

DDR Diabetic Retinopathy Debrecen

DL Deep Learning

EXs Exsudats

FGADR Fine-Grained Annotated Diabetic Retinopathy

FO Examen du Fond dóeil

GP Gaussian Process (processus gaussien)

HM Hémorragies

IA Intelligence Artificielle

MA Microanévrismes

ML Machine Learning

NPDR Rétinopathie Diabétique non Proliférative

OB Optimisation Bayesienne

OCT Tomographie à Cohérence Optique

OCTA Optical Coherence Tomography Angiographye

OMD Oedème Maculaire Diabétique

PDR Rétinopathie Diabétique Proliférative

RD Rétinopathie diabétique

ReVi-RD Resnet-Vit Retino Detection

ReVi-3C Resnet-Vit Retino 3 Classes

ReVi-5C Resnet-Vit Retino 5 Classes

TL Transfert Learning

ViT Vision Transformer

ViRD ViT Retino-Detection

ViR3C ViT Retino 3 Classes

ViR5C ViT Retino 5 Classes

MLP Perceptron multicouche (Multi-Layer Perceptron)

MSA Multi-Head Self-Attention

SVM Support Vector Machines

XAI Intelligence artificielle explicable (Explainable Artificial Intelligence)

Introduction Générale

Contexte

Au cours des dernières décennies, une augmentation significative de la prévalence des maladies liées au diabète a été observée à l'échelle mondiale. Le nombre d'adultes atteints de diabète s'élève désormais à environ 537 millions, un chiffre qui devrait atteindre 643 millions d'ici 2030 et 783 millions d'ici 2045. L'Algérie, comme de nombreux pays, n'échappe pas à cette tendance : le diabète affecte actuellement 14,4% de la population âgée de 18 à 69 ans, avec une projection de 3,4 millions d'adultes touchés d'ici 2045, soit une hausse de 8,1% par rapport à aujourd'hui [57].

Un diabète mal traité et non contrôlé, peut mener à une complication grave appelé RD. Cette pathologie entraîne fréquemment la cécité chez les patients âgés de 20 à 74 ans [119].

La nature asymptomatique de la Rétinopathie Diabétique (RD) à un stade précoce pose un défi diagnostique significatif, car la plupart des patients ignorent les dommages rétiniens jusqu'à ce que des complications menaçant la vision se développent.

Le diagnostic clinique de la RD repose sur la détection de lésions rétiniennes, tels que les microanévrismes (MA), les hémorragies (HM), les exsudats durs, l'œdème macualien et la néovascularisation, qui sont classés en cinq stades de gravité allant de l'absence de RD à la RD proliférative [163]. Sur la base de la présence et de la gravité de ces lésions, la RD est classée en deux stades principaux : la rétinopathie diabétique non proliférative (NPDR) qui est ensuite divisée en légère, modérée et sévère et la rétinopathie diabétique prolifique (PDR) [67], où de nouveaux vaisseaux anormaux se forment et présentent un risque élevé de perte de vision.

La pratique clinique actuelle pour le dépistage de la RD repose fortement sur l'interprétation manuelle des images du fond d'œil par les ophtalmologistes, créant une charge de travail importante pour les programmes de dépistage à grande échelle.

Cette démarche présente trois limites critiques :

— une forte variabilité entre les observateurs, même parmi les cliniciens expérimentés.

Le diagnostic peut varier d'un medecin à un autre, surtout dans les stades précoce où les signes (microanévrismes, hémorragies ponctuelles) sont difficiles à détecter ;

— une analyse qui prend beaucoup de temps et ne peut pas s'adapter aux demandes croissantes en matière de dépistagee ;

— et une disponibilité limitée de spécialistes, particulièrement dans les régions mal desservies où la prévalence du diabète est la plus élevée. Ces contraintes entraînent un diagnostic retardé et un risque accru de perte de vision irréversible.

Face à ces contraintes, les techniques manuelles s'avèrent inadaptées à la demande croissante de dépistage. Les progrès récents en intelligence artificielle (IA) offrent des solutions prometteuses pour un diagnostic précoce, qui permettra :

— La Réduction de la charge de travail des ophtalmologistes ;

— La Minimisation des erreurs humaines ;

— et la Détection plus efficace des lésions que les méthodes traditionnelles.

Des avancées récentes en apprentissage profond ont permis des progrès remarquables dans la détection automatisée de la RD.

Les systèmes automatisés exploitent généralement l'apprentissage automatique (ML : Machine Learning) [132, 15] ou l'apprentissage profond (DL : Deep Learning) [106]. Si les approches ML traditionel reposent sur des caractéristiques généralement extraites manuellement, les méthodes Deep Learning (DL) extraient automatiquement des primitives discriminantes à partir d'images du fond d'œil, surmontant ainsi ces limitations [140]. Néanmoins, le DL nécessite des volumes de données importants, une puissance de calcul élevée et des jeux de données équilibrés [41], des défis majeurs en imagerie médicale, **où l'acquisition et l'annotation d'images restent coûteuses** [168].

Problématique

Pour pallier au manque de données, l'apprentissage par transfert (Transfer Learning : TL) s'impose comme une solution prometteuse. Cette approche consiste à adapter des modèles pré-entraînés à des tâches similaires, puis à les affiner sur des données spécifiques. Plusieurs architectures ont été proposées, des modèles basés sur les réseaux de neurones convolutif comme par exemple VGG16 ou Resnet50 qui permettent l'extraction des caracteristiques locales [173, 158] L'apparition récente des Vision Transformers (ViTs) a révélé de nouvelles possibilités : leur capacité à utiliser un mécanisme d'auto-attention globale leur permet de modéliser les corrélations à longue distance entre les lésions rétiniennes éparpillées, offrant ainsi un avantage considérable pour l'analyse prédictive de la RD. Toutefois, leur utilisation en ophtalmologie reste encore largement exploratoire [69].

Bien que les algorithmes de Transfert Learning (TL) aient montré des résultats prometteurs ces dernières années pour la détection automatique de la RD [173, 113], **atteindre une**

précision parfaite en classification de la RD reste un défi persistant.

Pour améliorer les performances des réseaux d'apprentissage par transfert dans la détection de la RD, l'ajustement des hyperparamètres est essentiel.

La plupart des travaux existants utilisent un ajustement des poids (fine-tuning), en dégelant certaines couches lors de l'entraînement pour les adapter à la tâche spécifique du diagnostic de la RD [113, 189]. **Par contre l'ajustement des hyperparamètres est dans la majorité des cas, manuel, nécessitant des itérations longues et fastidieuses.**

Pour dépasser ces limites, et réduire l'intervention humaine dans un système d'apprentissage automatique, **nous proposons l'intégration d'un mécanisme d'auto-tuning des hyperparamètres, automatisant la sélection optimale de ces derniers**.

Contibutions

L'objectif principal de cette thèse est d'apporter des solutions aux problèmes posés en vue d'améliorer la précision de la détection et de la classification précoce de la RD.

De mannière spécifique, :

— nous proposons, trois architectures basées sur des CNN pré-entraînés, InceptionV3, VGG-16, ResNet-50 et Xception, pour l'extraction des caractéristiques discriminantes. Ces architectures sont composées, outre le backbone CNN pré-entraîné, d'un module de prétraitement des images ainsi que d'un module de classification, différencié selon les tâches de détection et de classification en trois ou cinq classes de gravité.

— Afin d'optimiser les performances de ces modèles, nous intégrons un module d'ajustement des hyperparamètres (fine-tuning) . Ce dernier permet de déterminer, dans une première phase, les meilleures configurations d'entraînement. À cet effet, une optimisation bayésienne automatisée est mise en œuvre pour explorer efficacement l'espace des hyperparamètres (tels que le taux d'apprentissage, la taille des batches, la profondeur du classifieur final) et identifier, la configuration qui maximise la performance des architectures.

  Une fois les hyperparamètres définis, les différentes architectures basées sur InceptionV3, VGG-16, ResNet-50 et Xception sont comparées selon les valeurs des metriques d'évaluation obtenues. Les meilleurs modèles seront alors sélectionnés et désignés sous les appellations Auto-tuned Retino Detection (AtRD), Auto-tuned Retino 3-Class (AtR3C) et Auto-tuned Retino 5-Class (AtR5C), en fonction de leur domaine d'application (détection ou classification en 3 ou 5 stades de gravité).

— nous évaluons et comparons l'ensemble des architectures développées sur deux bases de données APTOS 2019 et EyePACS, dans le but d'analyser l'influence de la provenance des images sur les performances des modèles.

Dans cette theèse nous étudions aussi l'apport des Vits à la classification de la RD

prenant en compte le contexte global de l'image, grâce aux mécanismes d'attention multi-têtes qui permettent de modéliser explicitement les dépendances à longue portée entre les différentes régions. Pour cela :

— nous proposons égalements 3 architectures (ViT Retino-Detection (ViRD), ViT Retino 3 Classes (ViR3C) et ViT Retino 5 Classes (ViR5C)) basées sur les ViTs pour la classification de la RD en 2, 3 et 5 stades de gravité respectivement. Les modèles ViT ont été entraînés exclusivement sur la base APTOS, qui avait préalablement montré les meilleures performances lors des expérimentations sur CNN.

— il convient de préciser que, dans le cadre de ce travail, la classification en trois classes en utilisant les ViTs n'a pas été abordée.

Enfin, une derniere contribution pour améliorer la precision de la detection et la classification de la RD consiste à tirer parti des avantages complémentaires des CNNs et des Vits, pour cela

— nous proposons 3 architectures hybrides (Resnet-Vit Retino Detection (ReVi-RD), Resnet-Vit Retino 3 Classes (ReVi-3C) et Resnet-Vit Retino 5 Classes (ReVi-5C)) qui sont obtenus en combinant les modèles (AtRD, AtR3C et AtR5C), qui assurent l'extraction locale des caractéristiques, avec les modèles (ViRD, ViR3C et ViR5C), qui introduisent un mécanisme d'attention globale. L'approche proposée consiste à alimenter les modèles basés Vit avec les cartes de caractéristiques extraites par les modèles basés CNN, permettant ainsi une modélisation conjointe des détails locaux et du contexte global.

— Les modèles hybrides sont entraînés et évalués sur la base APTOS, afin de vérifier si la synergie entre CNN et Transformer permet d'améliorer significativement les performances de détection et de classification, comparativement aux approches utilisant uniquement l'une ou l'autre de ces modèles.

Organisation

Ce document est organisé comme suit :

— Le chapitre 1 est consacré à la présentation de la rétinopathie diabétique (RD), en abordant sa définition, ses symptômes, sa classification ainsi que les principaux mécanismes de détection.

— Le chapitre 2 introduit les outils d'aide au diagnostic, en particulier les réseaux de neurones convolutionnels (CNN), le transfert d'apprentissage et les Vision Transformers (ViT).

— Le chapitre 3 propose une revue de littérature des travaux récents portant sur la classification de la RD à l'aide des CNN et des ViT. Ce chapitre discute également des

techniques d'optimisation des hyperparamètres ainsi que des métriques d'évaluation utilisées pour le calcul des performances de ces architectures.

— Le chapitre 4 détaille la méthodologie de notre approche basée sur les CNN, intégrant un auto-ajustement des hyperparamètres. Il présente aussi une étude comparative entre différents réseaux employés pour l'extraction de caractéristiques et ce, sur deux jeux de données distincts.

— Dans le chapitre 5, nous présentons la méthodologie que nous avons appliquée pour les architectures basées sur les ViT, ainsi que le processus de conception de nos modèles Hybrides. Ce chapitre contient également une comparaison rigoureuse des performances de nos modèles entre eux et par rapport aux approches existantes de l'état de l'art.

— Enfin, la thèse se conclut par une Conclusion Générale qui résume nos contributions et met en lumière les principales perspectives de recherche ouvertes par ce travail.

CHAPITRE 1

Rétinopathie Diabétique

1.1 Introduction

La rétinopathie diabétique (RD) est une complication grave du diabète, caractérisée par des altérations des vaisseaux sanguins rétiniens, et constitue l'une des principales causes de cécité [119, 68]. Le nombre de personnes touchées par le diabète et la RD ne cesse d'augmenter. Selon la Fédération Internationale du Diabète [57], environ 537 millions de personnes dans le monde souffrent de diabète, un chiffre qui devrait atteindre 643 millions d'ici 2030 et 783 millions d'ici 2045. La RD, complication microvasculaire du diabète la plus fréquente, touche environ 25-30% des patients diabétiques et 60% des patients diabétiques de type 2 depuis plus de dix ans. A l'échelle mondiale, environ 93 millions de personnes sont atteintes de RD. De plus, la majorité des personnes diabétiques ne sont pas diagnostiquées pour la RD, car cette maladie reste souvent asymptomatique jusqu'à un stade avancé [140].

Pour bien comprendre la necessité de la detection automatique de la RD, il faut d'abord comprendre c'est quoi la RD. Dans ce qui suit nous explorons la physiopathologie de la RD, ses symptômes et son diagnostic. Nous soulignerons également l'importance de la détection précoce et de la gestion du diabète pour prévenir la progression de la RD.

1.2 Anatomie de l'Oeil

L'œil est un organe sensoriel complexe chargé de capter les images visuelles et de les transmettre au cerveau. Il ajuste en permanence la quantité de lumière qu'il reçoit et s'adapte pour focaliser les objets avec précision. Le globe oculaire est logé dans une cavité osseuse appelée l'orbite, qui contient également les muscles, les nerfs et les vaisseaux sanguins assurant son bon fonctionnement.

Comme le montre la Figure 1.1, l'enveloppe externe du globe oculaire est constituée de

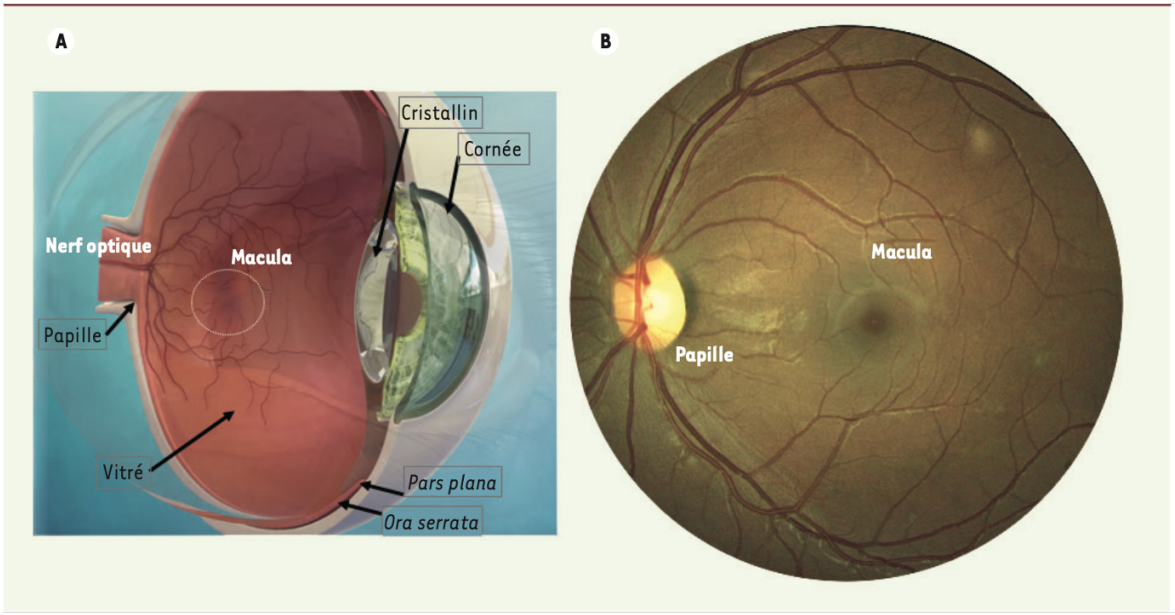


FIGURE 1.1 – A. Représentation schématique d'un globe oculaire. B. Rétinophotographie couleur d'un fond d'œil humain [24]

la sclère et de la cornée. La sclère est une membrane fibreuse, blanche et opaque, conférant à l'œil sa rigidité et sa protection. À l'avant de l'œil, la cornée est une structure transparente, fine et en forme de dôme, jouant un rôle essentiel dans la réfraction de la lumière vers la rétine.

La couche intermédiaire, appelée choroïde, contient l'iris et la pupille. L'iris, qui constitue la partie colorée de l'œil, contrôle l'entrée de la lumière en modulant le diamètre de la pupille. Placé derrière l'iris, le cristallin est une lentille biconvexe qui assure l'accommodation et focalise la lumière de manière précise sur la rétine, permettant ainsi une vision nette [24].

La rétine constitue la couche interne de l'œil et contient les photorécepteurs ainsi que les vaisseaux sanguins assurant son irrigation. La région la plus sensible de la rétine est la macula, responsable de la vision détaillée et centrale. Les photorécepteurs rétiniens sont connectés à des fibres nerveuses qui convergent pour former le nerf optique, lequel transmet les influx nerveux au cerveau pour interprétation des images perçues.

L'intérieur de l'œil comprend plusieurs cavités remplies de milieux liquidiens : La chambre antérieure, située entre la cornée et l'iris, et la chambre postérieure, entre l'iris et le cristallin, toutes deux remplies d'humeur aqueuse, un liquide transparent qui nourrit les structures avasculaires et maintient la pression intraoculaire. La chambre vitréenne, localisée entre le cristallin et le fond de l'œil, contient l'humeur vitrée, une substance gélatineuse qui assure le maintien de la forme du globe oculaire et amortit les chocs [24].

1.3 Physiopathologie de la Rétinopathie Diabétique

La rétinopathie diabétique (RD) est reconnue depuis longtemps comme une maladie microvasculaire. Le diagnostic de la RD repose sur la détection de lésions microvasculaires [171]. Les changements physiologiques dans l'œil voir figure 1.2 causés par le diabète sont à l'origine de la RD.

1.3.1 Définition

La rétinopathie diabétique (RD) est une pathologie microvasculaire induite par un diabète sucré prolongé, entraînant des lésions rétiniennes pouvant compromettre la vision [59, 171]. L'hyperglycémie constitue un facteur clé dans le développement de la RD, déclenchant plusieurs mécanismes pathologiques, notamment l'épaississement de la membrane basale, la perte des péricytes et la non-perfusion des capillaires rétiniens [18].

L'élévation prolongée du glucose sanguin altère la structure et la fonction des capillaires rétiniens, les rendant incompétents sur le plan anatomique et fonctionnel. L'excès de glucose est redirigé vers la voie de l'aldose réductase, transformant les sucres en alcools et affectant les péricytes intramuraux, dont la fonction principale est l'autorégulation des capillaires rétiniens. Cette altération compromet leur rôle dans l'autorégulation des capillaire, conduisant à un affaiblissement pariétal et à la formation de *microanévrysmes*, qui constituent les premières manifestations visibles de la RD. La rupture de ces microanévrysmes engendre des *hémorragies* rétiniennes, qui peuvent être superficielles (en flammes) ou plus profondes (en taches et en points) voir Figure 1.2. Par ailleurs, l'augmentation de la perméabilité vasculaire favorise l'exsudation de fluides et de protéines dans la rétine, entraînant un épaississement rétinien et la formation d'exsudats lipidiques. Lorsque ces anomalies touchent la macula, elles compromettent la vision centrale et aggravent la déficience visuelle [19].

La Rétinopathie Diabétique non Proliférente : constitue le stade initial de la maladie, caractérisé par une augmentation de la perméabilité vasculaire et une occlusion capillaire. Elle se manifeste par diverses lésions rétiniennes, notamment des microanévrysmes, des hémorragies punctiformes et en taches, des exsudats lipidiques (dits "durs"), des nodules cotonneux, une dilatation veineuse irrégulière et des anomalies microvasculaires intrarétiniennes (IRMAs)[59, 171].

La Rétinopathie Diabétique Proliférente représente un stade avancé, marqué par l'apparition de néovaisseaux pathologiques. Ces néovaisseaux, fragiles et anarchiques, sont sujets à des hémorragies intra-vitréennes et peuvent induire un décollement rétinien tractionnel, aboutissant à une altération sévère de la vision.

L'oedème maculaire diabétique (OMD) est une complication qui peut survenir à n'importe quel stade de la RD. Il est causé par l'accumulation de liquide dans la macula, entraînant un épaississement rétinien et une atteinte de la vision centrale [18].

Les spécialistes diagnostiquent la RD en identifiant des indicateurs spécifiques [21]. Les

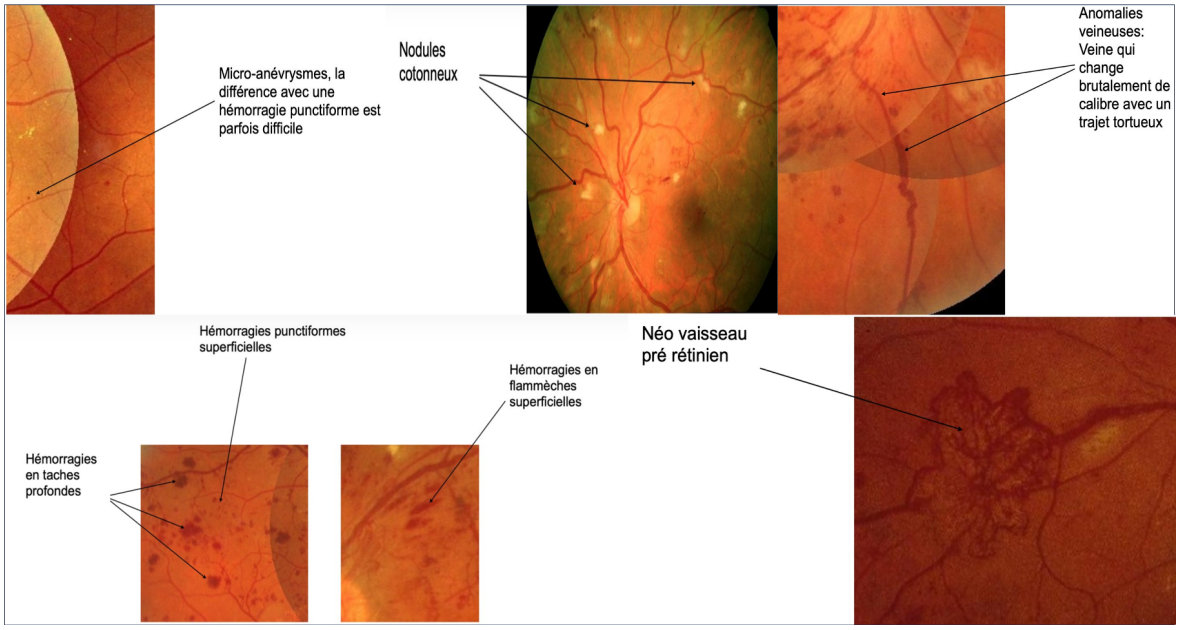


FIGURE 1.2 – Principaux signes de la RD

manifestations de la RD incluent les microanévrismes, les hémorragies intrarétiniennes, les exsudats, l'œdème maculaire, l'ischémie maculaire, la néovascularisation, les hémorragies intravitréennes et le décollement rétinien tractionnel. Les symptômes peuvent ne pas être perceptibles avant que la maladie n'atteigne un stade avancé [18]. La présence et la gravité de ces anomalies déterminent le stade de la maladie [171].

1.3.2 Signes de la Rétinopathie Diabétique

La rétinopathie diabétique (RD) évolue d'anomalies légères (caractérisées par une hyperperméabilité vasculaire) vers une RDNP modérée et sévère (caractérisée par une fuite ou une perte progressive des capillaires rétiniens entraînant une ischémie rétinienne), puis vers une RDP (caractérisée par le développement de néovaisseaux au niveau de la papille optique et de la rétine) (voir Figure 1.3).

Diverses symptômes sont utilisées pour classer la RD à un stade précis de son évolution. La présence de ces caractéristiques permet d'identifier le stade de la RD, facilitant ainsi le diagnostic et le traitement. Dans ce qui suit nous présentons les signes les plus importants dans le cas de la RD.

— **Disque Optique** Le disque optique, également appelé tête du nerf optique (voir Figure 1.4), est la région de la rétine où le nerf optique quitte l'œil. Cette zone, de forme circulaire ou ovale, joue un rôle crucial dans le transfert des informations visuelles de la rétine au cerveau. L'évaluation de l'apparence du disque optique est essentielle pour identifier des pathologies ou d'autres troubles oculaires, car des anomalies dans cette région peuvent révéler des problèmes de santé visuelle importants.

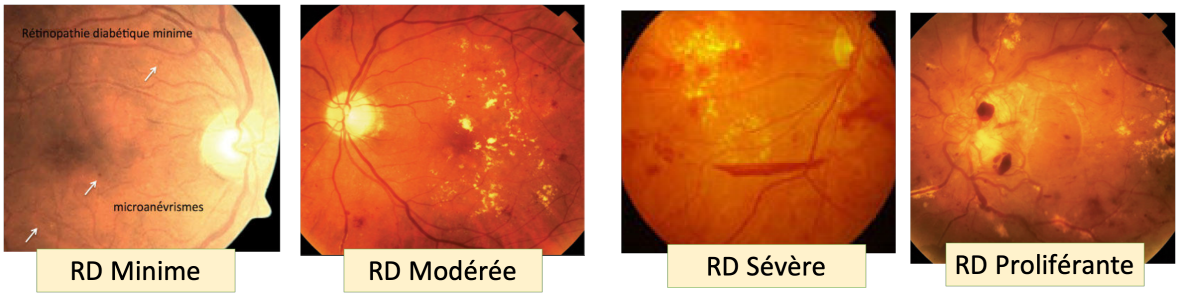


FIGURE 1.3 – Evolution de la RD

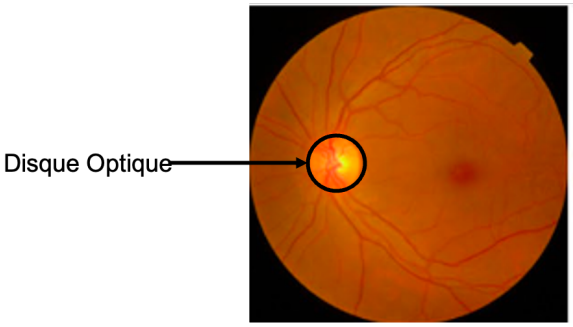


FIGURE 1.4 – Disque Optique

— **Hémorragies** Les Hémorragies (HM) désignent les fuites de sang des vaisseaux sanguins rétiniens, visibles sous forme de taches rouges anormales sur les images rétiniennes. Les HM intrarétiniennes se présentent sous diverses formes des points ou de taches et avec des contrastes variables. Leur diamètre varie généralement de 3 à 10 pixels. Parfois, les HM et les Microanévrismes (MA) apparaissent simultanément et sont désignés sous le terme de "lésions rouges", en raison de leur forme et de leur similarité. Les HM peuvent être aussi petites que les MA ou aussi grandes que le disque optique. L'apparition des HM n'affecte pas significativement la vision. Cependant, la présence de nombreuses HM en tache peut indiquer une ischémie importante, une caractéristique notable de la rétinopathie préproliférative. La figure 1.5 illustre des HM en point et en tache dans une rétine atteinte de RD. Les HM représentent le signe suivant de la RD après les MA.

— Exsudats

Les Exsudats (EXs), dépôts jaunâtres ou blanchâtres qui se forment dans la rétine en raison de fuites lipidiques et protéiques des vaisseaux sanguins endommagés (voir Figure 1.6). Leur taille varie de petites taches à de grandes plaques évoluant en formations circulaires appelées circinées. Ils se divisent en exsudats mous (bords flous) et durs (bords nets et brillants) et sont localisés au pôle postérieur du fond d'œil. Les exsudats durs peuvent provoquer un épaississement rétinien, entraînant un œdème maculaire diabétique ou une cécité. Ils constituent un signe clé de la progression de la RD après les microanévrismes [171].

Les taches cotonneuses (CWSs) sont les structures les plus grandes et irrégulières, avec

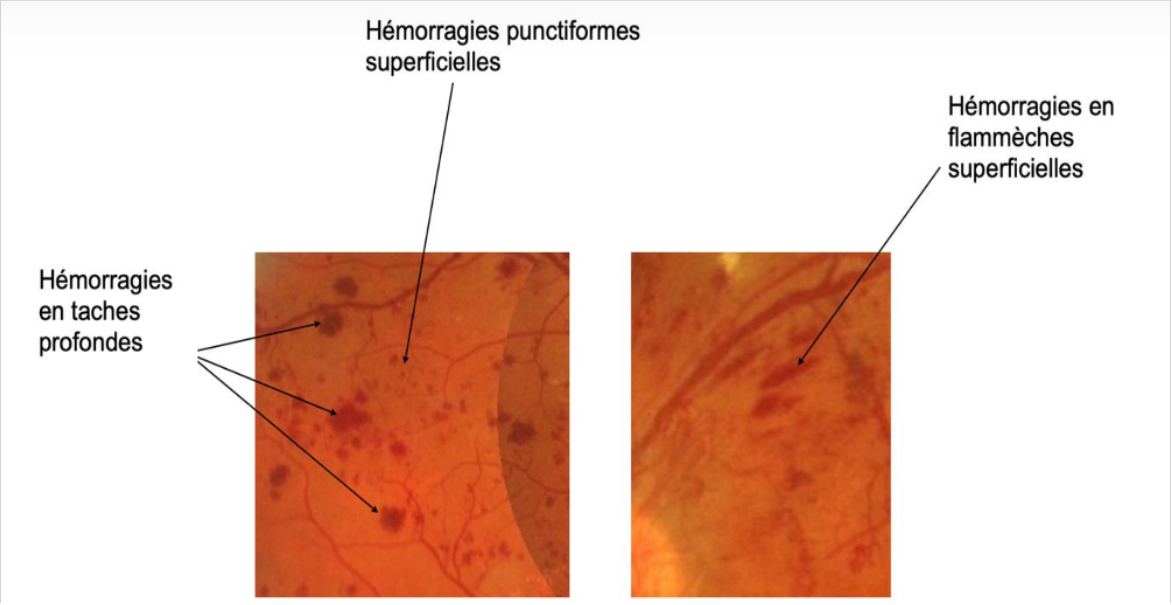


FIGURE 1.5 – Signe d'Hémorragies [70]

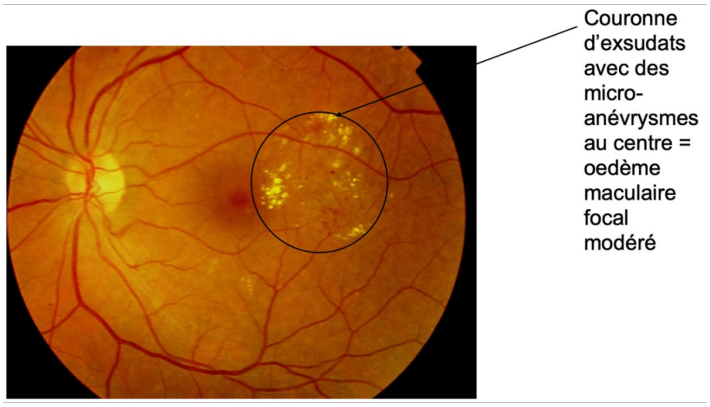


FIGURE 1.6 – Signe : Exsudats [70]

des contours flous, en comparaison aux MA et aux EXs. Elles sont le résultat d'un infarctus rétinien causé par une thrombose et une obstruction des vaisseaux sanguins. Ce sont des plaques grisâtres-blanchâtres de décoloration localisées dans la couche des fibres nerveuses. Elles résultent d'une ischémie locale qui perturbe le flux axoplasmique. La présence de plusieurs taches cotonneuses, comme environ six ou plus dans un seul œil, peut indiquer une ischémie rétinienne généralisée, conduisant au stade de rétinopathie diabétique pré-proliférative. La Figure 1.6 illustre des taches cotonneuses dans une image du fond d'œil atteinte de rétinopathie diabétique.

1.4 Différents Stades de la Rétinopathie Diabétique

Comme nous pouvons le voir sur la Figure 1.3, la RD progresse selon plusieurs stades en raison des altérations vasculaires induites par l'hyperglycémie chronique au niveau de la rétine.

La classification de la RD élaborée sous l'égide de la Société américaine d'ophtalmologie [175] se presente selon 4 stades. La RD est d'abord classée en deux stades : la Rétinopathie Diabétique non Proliférative (NPDR) et la Rétinopathie Diabétique Proliférative (PDR). La NPDR est ensuite subdivisée en trois niveaux de gravité : légère, modérée et sévère [67]. Les caractéristiques standardisées de tous les types de RD sont résumées dans le Tableau 1.1. La sévérité de l'ischémie rétinienne est évaluée par le nombre et l'étendue des signes cliniques visibles sur les photographies du fond d'œil : hémorragies rétiniennes, anomalies veineuses moniliformes, anomalies micro- vasculaires intrarétiniennes.

— Stade 1 : Rétinopathie Diabétique non Proliférente Légère (RDNP-L) Ce stade initial se caractérise par l'apparition de micro-anévrismes, correspondant à des dilatations focales des capillaires rétiniens. Ces lésions vasculaires peuvent entraîner des extravasations minimes de plasma ou d'hématies dans le tissu rétinien. À ce stade, la fonction visuelle demeure généralement préservée, bien que la présence de ces premières anomalies témoigne d'une susceptibilité accrue à l'évolution vers des formes plus avancées de la pathologie.

— Stade 2 : Rétinopathie Diabétique non Proliférente Modérée (RDNP-M) À ce stade, les altérations vasculaires deviennent plus marquées et plus étendues, se traduisant par une augmentation des œdèmes et des hémorragies rétiniennes. L'irrigation sanguine de la rétine est compromise, entraînant une insuffisance trophique du tissu rétinien. Si l'accumulation de sang et de fluides atteint la macula, une baisse d'acuité visuelle sous forme de flou visuel peut survenir. Une surveillance ophtalmologique plus fréquente est généralement préconisée afin de suivre l'évolution de la maladie. Ce stade est parfois qualifié de rétinopathie pré-proliférante.

— Stade 3 : Rétinopathie Diabétique NON Proliférente Sévère (RDNP-S) Les lésions microvasculaires s'aggravent, avec une augmentation significative du nombre de vaisseaux sanguins endommagés et obstrués, exacerbant ainsi l'ischémie rétinienne. Face à ce déficit en oxygène, la rétine émet des signaux pro-angiogéniques en réponse à l'hypoxie, favorisant le développement de nouveaux vaisseaux. À ce stade, l'atteinte visuelle devient plus significative, annonçant la transition vers une phase proliférative.

— Stade 4 : Rétinopathie Diabétique Proliférente Il s'agit du stade avancé de la maladie, caractérisé par l'apparition de néovaisseaux pathologiques à la surface de la rétine et du disque optique. Ces vaisseaux, particulièrement fragiles, présentent un risque élevé d'hémorragie intra-vitréenne et s'accompagnent souvent d'une prolifération de tissu fibreux. La contraction de ces membranes fibrovasculaires peut induire un décollement tractionnel de la rétine, menant à une perte visuelle sévère, voire irréversible.

En complément de ces stades, une maculopathie diabétique peut survenir lorsque les vaisseaux sanguins de la macula deviennent perméables ou obstrués. Cette atteinte peut se manifester à n'importe quel stade de la rétinopathie diabétique et constitue une cause majeure de déficience visuelle centrale.

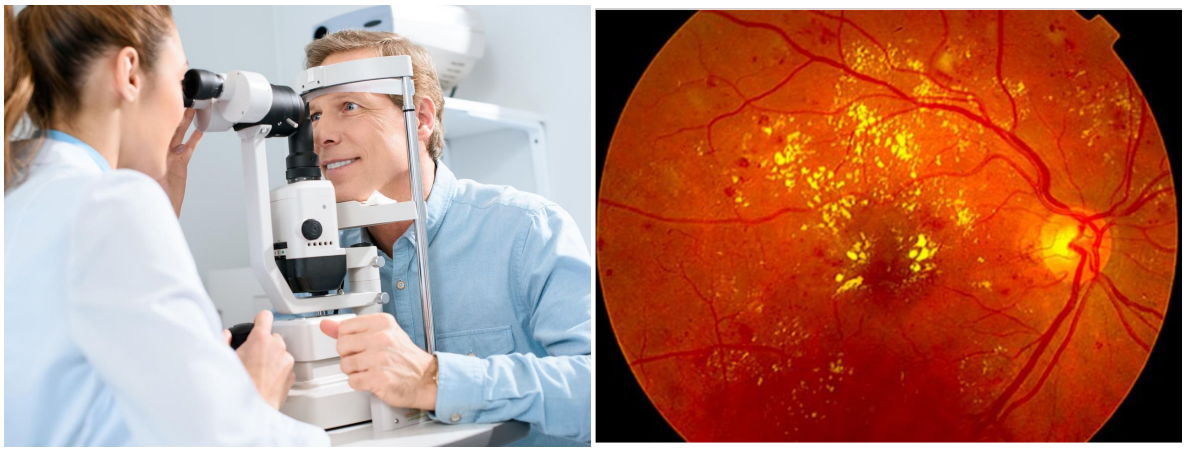
TABLE 1.1 – Classification des différents stade de la RD [175]

Stade	Signes	Niveau de sévérité
0	Pas d'anomalies	Pas de DR
1	Microanévrismes seuls	RDNP minime
2	Nombre de microanévrismes plus élevé	RDNP Modéré
3	Un ou plus des items suivants : - plus de 20 hémorragies intrarétiniennes - AMIRs nombreuses et pas de signe de RD proliférante	RDNP Sévère
4	Neovascularization, vitreous/preretinal HM	PDR

1.5 Examens Nécessaires au Diagnostic

La rétinopathie diabétique (RD) est principalement diagnostiquée au moyen d'un examen ophtalmologique complet avec dilatation pupillaire. L'instillation de collyres mydriatiques permet d'élargir la pupille, offrant ainsi au clinicien une meilleure visualisation des structures intraoculaires, notamment la rétine. La surveillance régulière par des examens spécialisés est essentielle, car la RD demeure asymptomatique à ses premiers stades et peut évoluer vers une perte visuelle irréversible avant même l'apparition de signes cliniques perceptibles [10, 11]. Parmi les méthodes couramment utilisées pour le diagnostic et le dépistage de la RD, on retrouve [64] :

— **L'examen du fond d'œil** (FO) : Cet examen constitue une étape fondamentale permettant une observation directe des structures rétiniennes afin d'identifier d'éventuelles anomalies vasculaires ou tissulaires. L'ophtalmoscope, un dispositif optique spécialisé, est employé pour examiner en détail la rétine et détecter d'éventuelles lésions caractéristiques de la RD (voir Figure 1.7).

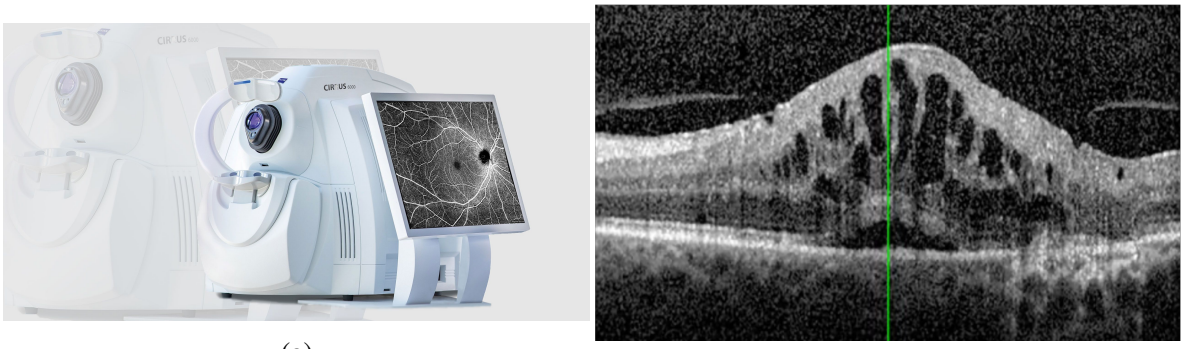


(a) (b)

FIGURE 1.7 – Examen du Fond d'oeil :(a) Ophtalmoscope et (b)Image du Fond d'oeil obtenue

— **La tomographie à cohérence optique** (OCT) : L'OCT est une technique d'imagerie haute résolution permettant une visualisation en coupe des différentes couches réti-

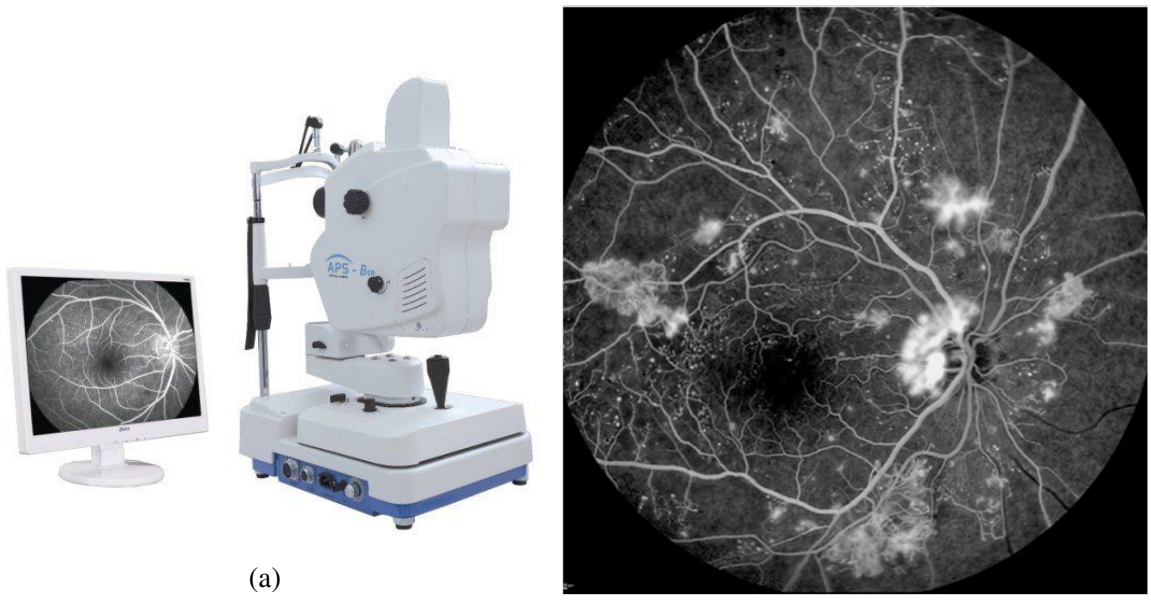
niennes. Elle mesure l'épaisseur des structures rétiniennes, facilitant ainsi la détection précoce d'anomalies telles que l'œdème maculaire diabétique. En outre, l'OCT constitue un outil précieux pour le suivi thérapeutique, en évaluant la réponse aux traitements administrés. L'OCT joue un rôle essentiel dans le dépistage, le diagnostic et le suivi de la rétinopathie diabétique et de ses complications (voir Figure 1.8).



(a) (b)

FIGURE 1.8 – La tomographie à cohérence optique (OCT : (a) Appareil pour l'OCT et Image obtenue par OCT

— **L'angiographie rétinienne** : Cet examen consiste en l'injection intraveineuse d'un colorant fluorescent dans une veine du bras. La circulation du colorant à travers les vaisseaux rétiniens est ensuite capturée par une série d'images, permettant d'identifier avec précision les zones d'ischémie, les microanévrismes, ainsi que les fuites vasculaires caractéristiques de la RD (voir Figure 1.9).



(a) (b)

FIGURE 1.9 – L'angiographie rétinienne : (a) Angiographe et (b) Image obtenue par Angiographie

La rétinopathie diabétique est indolore et n'induit pas toujours de baisse visuelle, même dans certains stades avancés de la maladie. C'est pourquoi il est impératif de réaliser un dépistage de la maladie en effectuant régulièrement des photographies de la rétine.

1.6 Conclusion

Les méthodes traditionnelles de dépistage de la rétinopathie diabétique (RD), fondées sur l'analyse manuelle des images rétiniennes par des ophtalmologistes, se révèlent non seulement très chronophages et intensives en main-d'œuvre, mais elles dépendent également de la disponibilité de professionnels hautement qualifiés [99]. De surcroît, l'interprétation des images du fond d'œil est fortement tributaire de l'expertise, de l'expérience et du jugement subjectif de chaque spécialiste, ce qui peut conduire à des évaluations divergentes, en particulier concernant la classification de la sévérité de la maladie ou la détection de certaines lésions. Cette subjectivité induit une variabilité dans les diagnostics, susceptible d'affecter négativement la prise en charge thérapeutique.

Dans ce contexte, détecter la maladie à un stade précoce et poser un diagnostic précis sont des enjeux majeurs pour prévenir une perte de vision irréversible. Pourtant, les approches traditionnelles, bien qu'efficaces, restent complexes et sujettes aux erreurs humaines [150].

Face à ces défis, l'intelligence artificielle (IA) ouvre de nouvelles perspectives en automatisant le diagnostic avec une rapidité et une fiabilité accrues. Le chapitre suivant explorera les différentes architecture basées sur l'apprentissage automatique qui permettront le dépistage de la rétinopathie diabétique.

CHAPITRE 2

Apprentissage Profond, Transfert de Connaissances et Vision Transformers

2.1 Introduction

Ces dernières années, le domaine de l'Intelligence Artificielle (IA) a connu une progression fulgurante. En effet, l'IA s'est imposée comme une technologie incontournable non seulement pour les entreprises, mais également pour les particuliers. De l'agriculture à la santé en passant par les industries de production, l'intelligence artificielle a un impact sur quasiment tous les domaines de la vie.

2.2 Definition de l'Intelligence Artificielle

Il n'y a pas de définition unique de l'IA. Dès 1950, Alan Turing posait les bases théoriques de l'IA avec son célèbre article « Computing Machinery and Intelligence », où il proposait ***le test de Turing comme critère d'intelligence machine*** [162]. Cependant, le terme « intelligence artificielle » a été officiellement introduit en 1956 par John McCarthy lors d'un workshop organisé à l'Université de Dartmouth (États-Unis) [108] en le définissant comme ***« la science et l'ingénierie de la création de machines intelligentes »***. Les différentes définitions s'accordent sur le fait que l'objectif de l'IA est de créer des systèmes intelligents, mais elles diffèrent significativement dans leur façon de définir l'intelligence. Certaines se foca- lisent sur le comportement du système, tandis que d'autres considèrent que c'est le fonctionnement interne (le raisonnement) du système qui importe. C'est pourquoi la plus part des documents [134, 65] ainsi que Wikipedia définissent l'IA comme suit :

L'IA est la capacité des machines à effectuer des tâches typiquement associées à l'intelligence humaine, comme l'apprentissage, le raisonnement, la résolution de problème, la perception ou la prise de décision.

L'IA fait appel à différentes disciplines scientifiques comme les mathématiques, l'informatique, la logique, la robotique, les neurosciences, la linguistique ou encore la psychologie cognitive. Pour se rapprocher le plus possible du fonctionnement du cerveau humain, l'IA fait appel à un certain nombre d'éléments dont les plus importants sont :

— Des algorithmes informatiques ;

— De grandes bases de données ;

— Des systèmes et matériels informatiques performants.

L'utilisation de ces trois éléments permet aux systèmes d'IA d'apprendre et de s'améliorer de manière itérative en analysant et en intégrant les informations qui leur sont fournies. L'IA peut être retrouvée dans plusieurs domaines dont : L'automobile (voiture autonome), la médecine, l'industrie, le marketing, l'éducation et bien plus encore. Dans notre cas nous nous intéressons au domaine de la santé ou autrement dit au domaine médical. Les principaux champs d'application de l'IA en matière de santé sont : la médecine prédictive notamment la prédiction d'une maladie mais aussi l'aide à la décision pour établir un diagnostic médical [73].

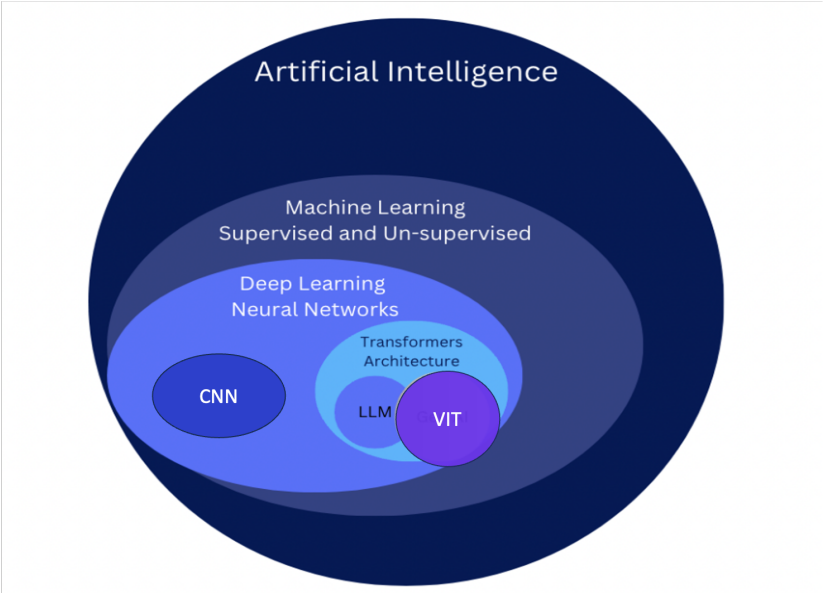


FIGURE 2.1 – Domaines de l'IA

2.3 Apprentissage Automatique

La figure 2.1 présente une vue d'ensemble hiérarchisée des principales techniques de l'IA. Dans cette section, nous passons en revue les approches traditionnelles utilisées en IA

pour l'analyse d'images médicales.

Cette exploration pose les bases de l'étude, en montrant les niveaux d'abstraction croissants et en justifiant la pertinence de l'apprentissage par transfert et des transformers pour la classification des images.

2.4 Définition et Fondements

L'apprentissage automatique (*machine learning*, ML) est un sous-domaine central de l'intelligence artificielle (IA), dont l'objectif est de concevoir des systèmes capables d'améliorer leur performance sur une tâche donnée à partir de données, sans que les règles de décision soient explicitement programmées [136]. Contrairement aux approches logicielles classiques, où le comportement est entièrement codé manuellement, le ML permet à un algorithme de s'ajuster dynamiquement à son environnement via des processus d'optimisation.

Le ML s'appuie sur une approche interdisciplinaire, à l'interface des mathématiques, de l'informatique et des statistiques. Son enjeu principal est le développement de modèles généralisables capables d'inférer des connaissances à partir de données observées, puis de les réutiliser dans des contextes nouveaux ou non observés [188].

Selon les références fondatrices, leML est défini :

Définition de Mitchell (1997)

Un programme apprend à partir d'une expérience $\mathcal{E}$ relativement à une classe de tâches $\mathcal{T}$ et une mesure de performance $\mathcal{P}$, si sa performance s'améliore avec l'expérience :

$$\forall \epsilon > 0, \exists n \in \mathbb{N} \text{ tel que } \mathcal{P}_{\mathcal{T}}(\mathcal{E}_n) > \mathcal{P}_{\mathcal{T}}(\mathcal{E}_{n-1}) + \epsilon$$

[109].

Définition de Goodfellow et al (2016)

L'apprentissage est modélisé comme un processus d'optimisation différentiable des paramètres θ minimisant une fonction de perte $\mathcal{L}$:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}(\theta_t, \mathcal{D})$$

où η est le taux d'apprentissage et $\mathcal{D}$ le jeu de données. [65]

Définition Synthétique

Définition : Apprentissage automatique

L'apprentissage automatique désigne un processus par lequel un système algorithmique $\mathcal{A}$:

1. améliore itérativement sa performance $\mathcal{P}$ sur une tâche $\mathcal{T}$ à partir d'une expérience $\mathcal{E}$;
2. extrait des structures latentes $\mathcal{S}$ à partir de $\mathcal{E}$ via une fonction d'abstraction Φ : $\mathcal{S} = \Phi(\mathcal{E})$;
3. apprend sans programmation explicite : $\nexists\, \Psi_{\text{man}}$ telle que $\mathcal{P}_{\mathcal{T}} = \Psi_{\text{man}}(\mathcal{E})$.

Formulation Mathématique de l'Apprentissage

Un système d'apprentissage est défini par le quintuplet

$(\mathcal{T}, \mathcal{P}, \mathcal{E}, f_\theta, \mathcal{L})$, où :

$$f_\theta : \mathcal{X} \to \mathcal{Y} \quad ; \quad \mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$$

L'objectif est de déterminer :

$$\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\mathcal{L}\big(f_\theta(\mathbf{x}), y\big) \right],$$

avec amélioration de la performance si :

$$\mathcal{P}_{\mathcal{T}}(f_{\theta^*}) > \mathcal{P}_{\mathcal{T}}(f_{\theta_0}) \quad \forall \theta_0 \notin \mathcal{B}(\theta^*)$$

où $\mathcal{B}(\theta^*)$ désigne un voisinage sous-optimal de θ^*.

2.4.1 Algorithmes de Machine Learning

Les algorithmes de Machine Learning peuvent être divisés en 3 types (voir Figure 2.2) :

2.4.2 Apprentissage Supervisé

L'apprentissage supervisé est une approche de l'apprentissage automatique dans laquelle un modèle est entraîné sur un ensemble de données étiquetées, chaque exemple d'entraînement étant associé à une sortie cible [26] (voir Figure 2.3). L'objectif principal est d'apprendre

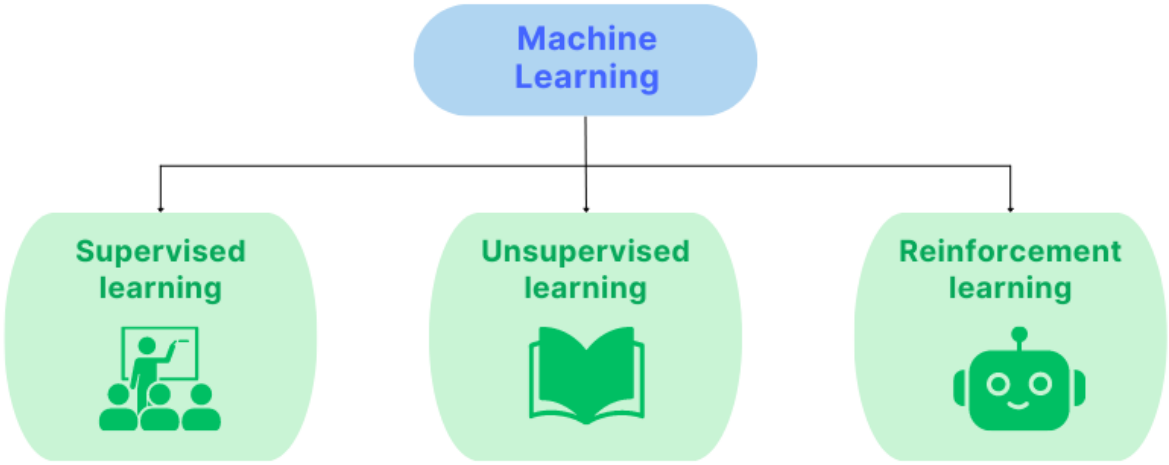


FIGURE 2.2 – Différents Algorithmes de ML

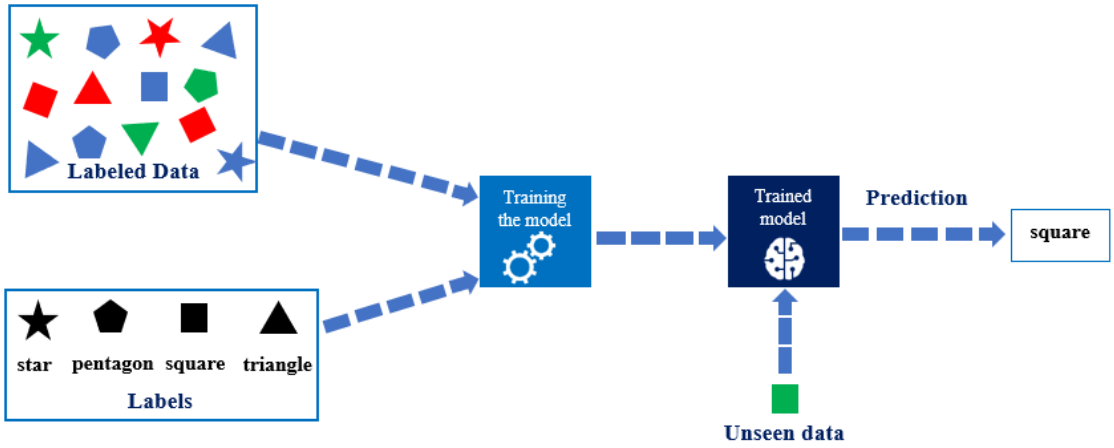


FIGURE 2.3 – Apprentissage Supervisé

une fonction de correspondance (mapping) qui permettra de prédire avec précision les étiquettes des nouvelles données, jamais vues auparavant [100]. Divers algorithmes, tels que la régression linéaire, les machines à vecteurs de support (SVM) [38], les arbres de décision [95] et les forêts aléatoires [27], sont utilisés en fonction des caractéristiques du problème à résoudre. Ce processus repose sur l'optimisation des paramètres du modèle à l'aide de fonctions de coût et de techniques d'optimisation, comme la descente de gradient, afin de minimiser l'écart entre les prédictions et les valeurs réelles. L'apprentissage supervisé trouve ainsi de larges applications dans divers domaines, notamment la classification d'images, la détection de fraudes et le diagnostic médical, où il contribue à améliorer la précision et l'efficacité des prises de décision [100].

2.4.3 Apprentissage Non Supervisé

L'apprentissage non supervisé se focalise sur l'extraction d'informations pertinentes à partir de données non annotées [26]. Contrairement à l'apprentissage supervisé, aucune étiquette de sortie n'est fournie a priori. L'objectif principal est alors de découvrir des motifs

cachés, des structures ou des relations intrinsèques aux données (voir Figure 2.4). Parmi les techniques courantes, le clustering permet de regrouper des points de données similaires en fonction de leurs propriétés intrinsèques. D'autres méthodes incluent la réduction de dimensionnalité et la modélisation générative [66]. Dans ce cadre, les modèles apprennent directement à partir des données d'entrée, sans supervision explicite, afin de capturer la structure sous-jacente et identifier des schémas significatifs. Ces approches offrent des perspectives précieuses pour comprendre les caractéristiques inhérentes des jeux de données et sont particulièrement utiles dans des domaines tels que les systèmes de recommandation, la détection d'anomalies et le prétraitement des données, surtout lorsque l'obtention de données étiquetées est limitée ou coûteuse [26].

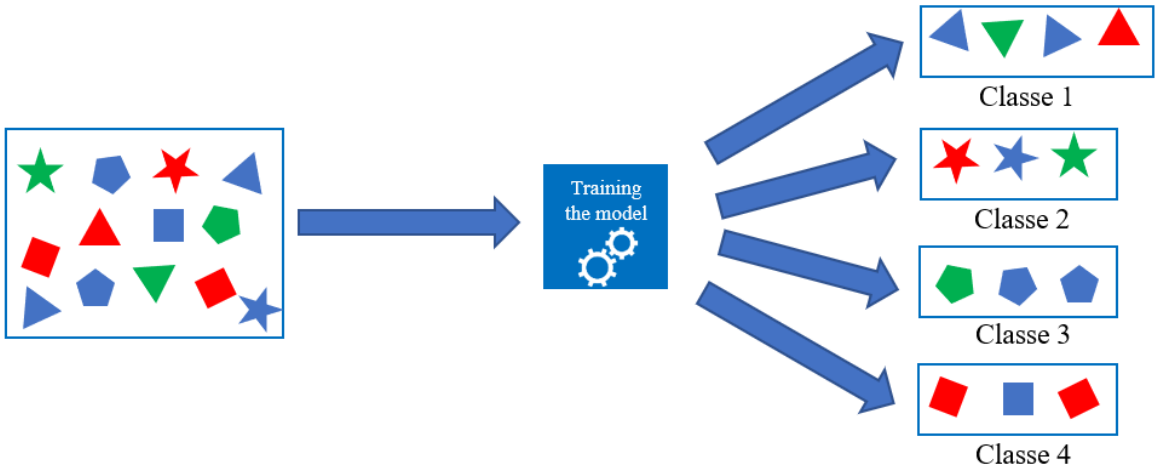


FIGURE 2.4 – Apprentissage non supervisé

2.4.4 Apprentissage par Renforcement

L'apprentissage par renforcement permet à un agent d'apprendre à prendre des décisions dans un environnement en maximisant une récompense cumulée, en s'inspirant du processus d'essais et erreurs observé chez les êtres vivants [155, 186]. Comme l'illustre la Figure 2.5, l'agent interagit avec son environnement en observant son état, en réalisant des actions et en recevant en retour des récompenses ou pénalités, ce qui le guide vers des comportements optimaux. En équilibrant exploration et exploitation, il développe une politique (mapping états-actions) visant à maximiser la récompense à long terme. [155].

Cette approche a démontré son efficacité dans des domaines variés comme la robotique, les jeux, les systèmes de recommandation et les véhicules autonomes.

2.4.5 Le Deep Learning

L'apprentissage profond constitue une sous-discipline de l'apprentissage automatique basée sur l'utilisation de réseaux de neurones artificiels (RNA) multicouches. Ces architectures (Figure 2.6), comportant au minimum deux couches cachées, visent à approximer une

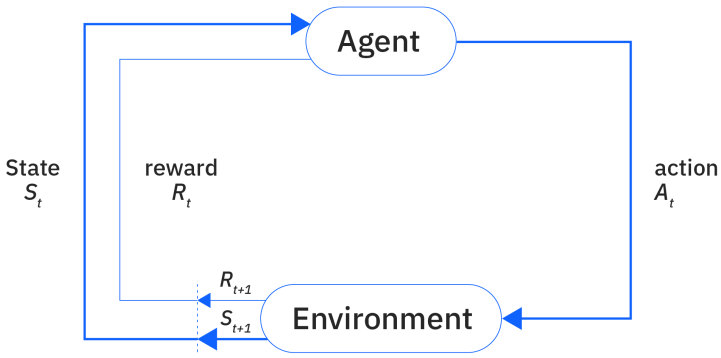


FIGURE 2.5 – Apprentissage par Renforcement

fonction f transformant les données d'entrée en nouvelles représentations ou générant des prédictions [85]. La capacité d'approximation de ces réseaux est formellement garantie par le théorème d'approximation universelle de Cybenko [45].

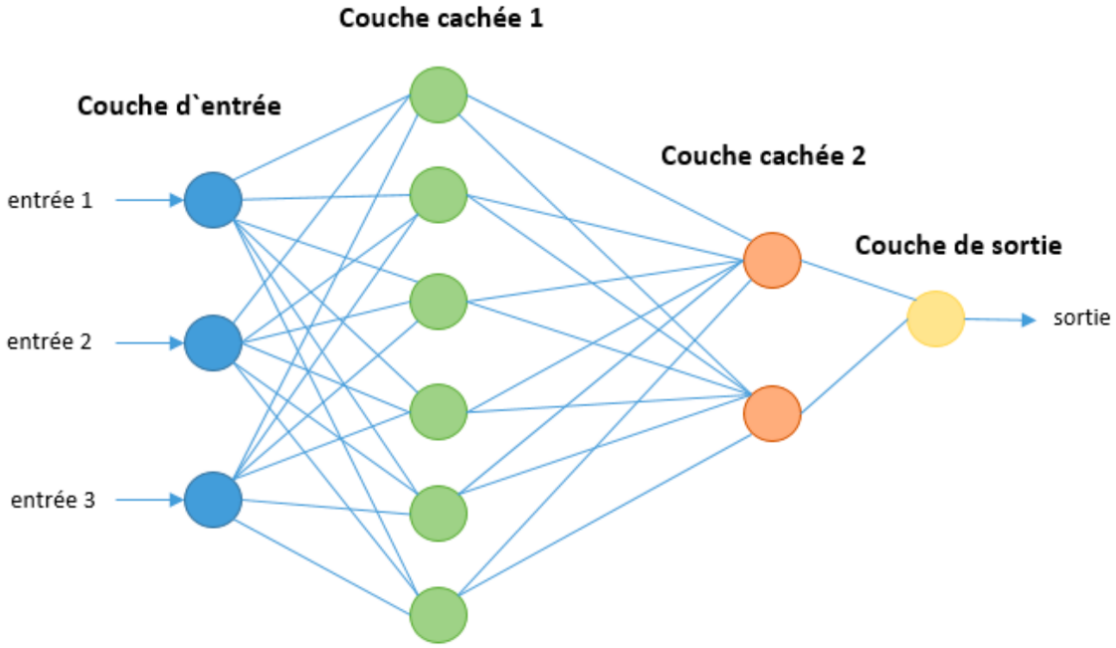


FIGURE 2.6 – Réseau de neurones profond

Théorème 2.4.1 – (Théorème d'Approximation Universelle (Cybenko, 1989)])

Soit $\sigma : \mathbb{R} \to \mathbb{R}$ une fonction d'activation **continue**, **bornée** et **non constante**. Alors, pour tout entier $n \geq 1$, l'ensemble des réseaux de neurones à une couche cachée de la forme :

$$f(\mathbf{x}) = \sum_{j=1}^{N} \alpha_j \sigma\left(\mathbf{w}_j^\top \mathbf{x} + b_j\right)$$

est **dense** dans l'espace des fonctions continues $C(I_n)$ muni de la norme uniforme $\|\cdot\|_\infty$, où $I_n = [0,1]^n$ désigne l'hypercube unité de $\mathbb{R}^n$.
Plus formellement :

$$\forall g \in C(I_n),\ \forall \epsilon > 0,\ \exists f \text{ de la forme ci-dessus} : \sup_{\mathbf{x} \in I_n} |f(\mathbf{x}) - g(\mathbf{x})| < \epsilon$$

Inspirés de la structure neuronale biologique, les RNA sont composés de neurones interconnectés par des connexions pondérées. Chaque neurone transforme ses entrées en une sortie unique via l'application d'une fonction d'activation non linéaire à la somme pondérée de ses entrées (Figure 2.7).

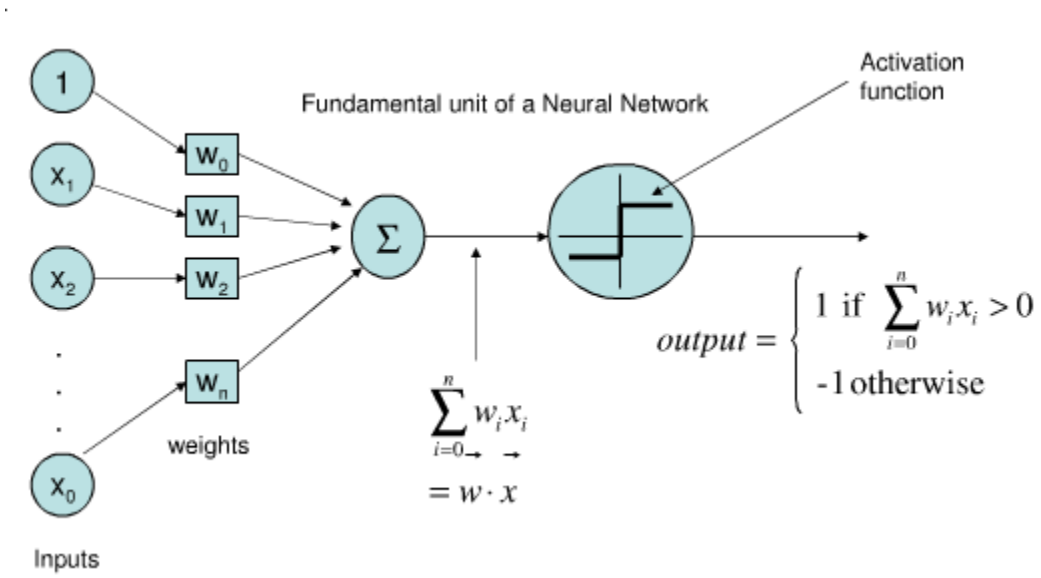


Figure 2.7 – Architecture du Perceptron [166].

Mathématiquement, la sortie d'un neurone j est donnée par :

$$y_j = \sigma\left(\sum_{i=1}^{n} w_{ij}x_i + b_j\right) \tag{2.1}$$

où x_i représente l'entrée i, w_{ij} le poids de la connexion entre l'entrée i et le neurone j, b_j le biais du neurone j, et σ la fonction d'activation.

Dans les réseaux à propagation avant, les données se propagent séquentiellement à travers les couches successives, chaque couche cachée recevant la sortie de la couche précédente. La sortie finale dépend des données d'entrée, de la fonction d'activation et des paramètres de pondération. L'optimisation de ces paramètres s'effectue par descente de gradient [101], méthode itérative minimisant la fonction de perte qui quantifie l'écart entre les sorties attendues et obtenues [79]. L'algorithme de descente de gradient met à jour les poids selon :

$$w_{ij}^{(t+1)} = w_{ij}^{(t)} - \eta\frac{\partial L}{\partial w_{ij}} \tag{2.2}$$

où η est le taux d'apprentissage, L la fonction de perte, et t l'itération courante.

La fonction de perte couramment utilisée pour les problèmes de régression est l'erreur quadratique moyenne :

$$L = \frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{y}_k)^2 \tag{2.3}$$

où N est le nombre d'échantillons, y_k la sortie attendue et $\hat{y}_k$ la sortie prédite pour l'échantillon k.

2.4.6 Le Réseau de Neurones Convolutionnel

Le réseau de neurones convolutionnel (CNN) constitue une architecture spécialisée pour le traitement de données multidimensionnelles, particulièrement adaptée à l'analyse d'images [101]. Inspiré par le cortex visuel, le CNN trouve ses origines dans le Neocognitron de Fukushima (1980) [62].

Comme, l'illustre la figure 2.8, une architecture CNN comprend trois types de couches principales :

— **Couches convolutives** : Ces couches extraient les caractéristiques locales en appliquant des filtres à l'image d'entrée. L'opération de convolution mathématique s'effectue selon :

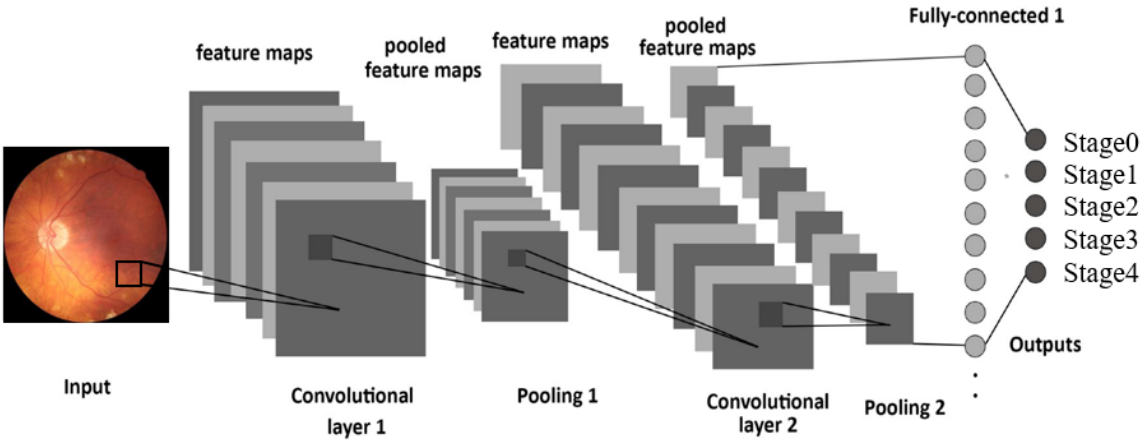


FIGURE 2.8 – Architecture CNN

$$(I * K)i, j = \sum m \sum_{n} I_{i+m,j+n} \cdot K_{m,n} \tag{2.4}$$

où I représente l'image d'entrée, K le noyau de convolution, et $*$ l'opération de convolution.

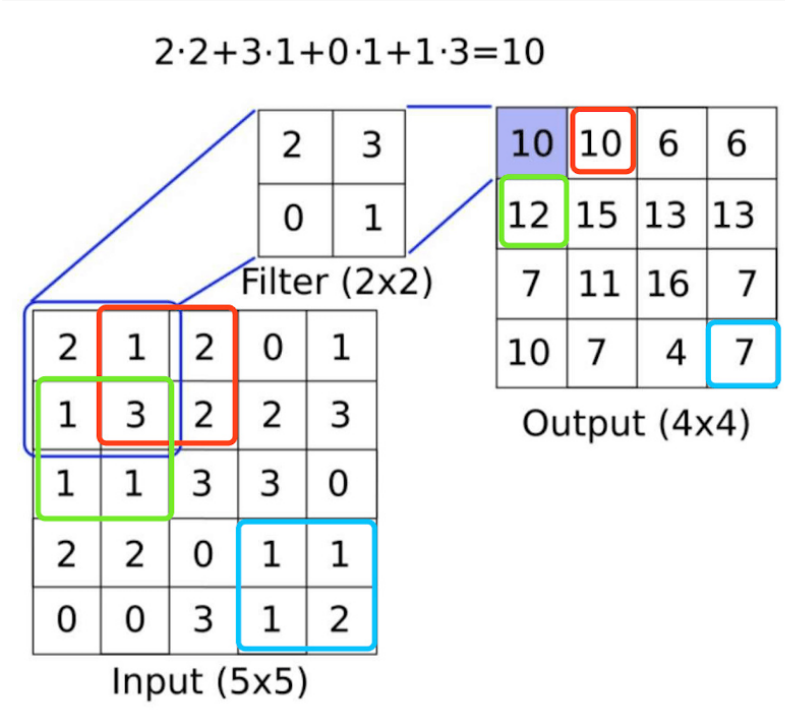


FIGURE 2.9 – Principe de convolution

— **Couches de pooling** (sous-échantillonnage) : Elles réduisent la dimensionnalité spatiale des cartes de caractéristiques, diminuant la complexité computationnelle tout en préservant l'information essentielle. L'opération de max-pooling s'exprime comme :

$$P_{i,j} = \max_{(m,n)\in R_{i,j}} F_{m,n} \tag{2.5}$$

où $P_{i,j}$ est la valeur de sortie et $R_{i,j}$ la région de pooling considérée.

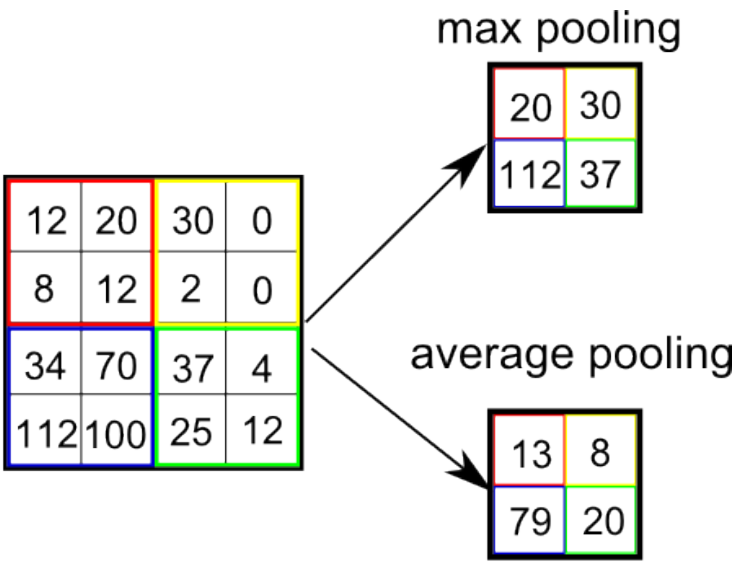


FIGURE 2.10 – Le Principe du Pooling

Comme l'illustre la Figure 2.10, les opérations de pooling les plus courantes sont le max-pooling et l'average-pooling.

— **Couches entièrement connectées** : Similaires aux couches d'un perceptron multicouche, elles effectuent la classification finale en combinant les caractéristiques extraites par les couches précédentes.

Grâce à leur capacité à apprendre automatiquement des représentations hiérarchiques, les CNN ont démontré des performances remarquables dans la classification d'images, la détection d'objets et d'autres tâches de vision par ordinateur [68]

Cette architecture hiérarchique permet l'apprentissage automatique de représentations à plusieurs niveaux, rendant les CNNs particulièrement efficaces pour la classification d'images médicales [68]. Comme l'illustre la Figure 2.11, les couches convolutionnelles extraient automatiquement différentes hiérarchies de caractéristiques, depuis des éléments génériques (bords, coins, textures) jusqu'aux patterns spécifiques des lésions rétiniennes.

Les cartes de caractéristiques sont ensuite aplaties pour obtenir un vecteur de caractéristiques unidimensionnel, puis traitées par des couches entièrement connectées avant la classification finale. La couche de sortie utilise par exemple la fonction d'activation softmax qui convertit les scores en probabilités de classe :

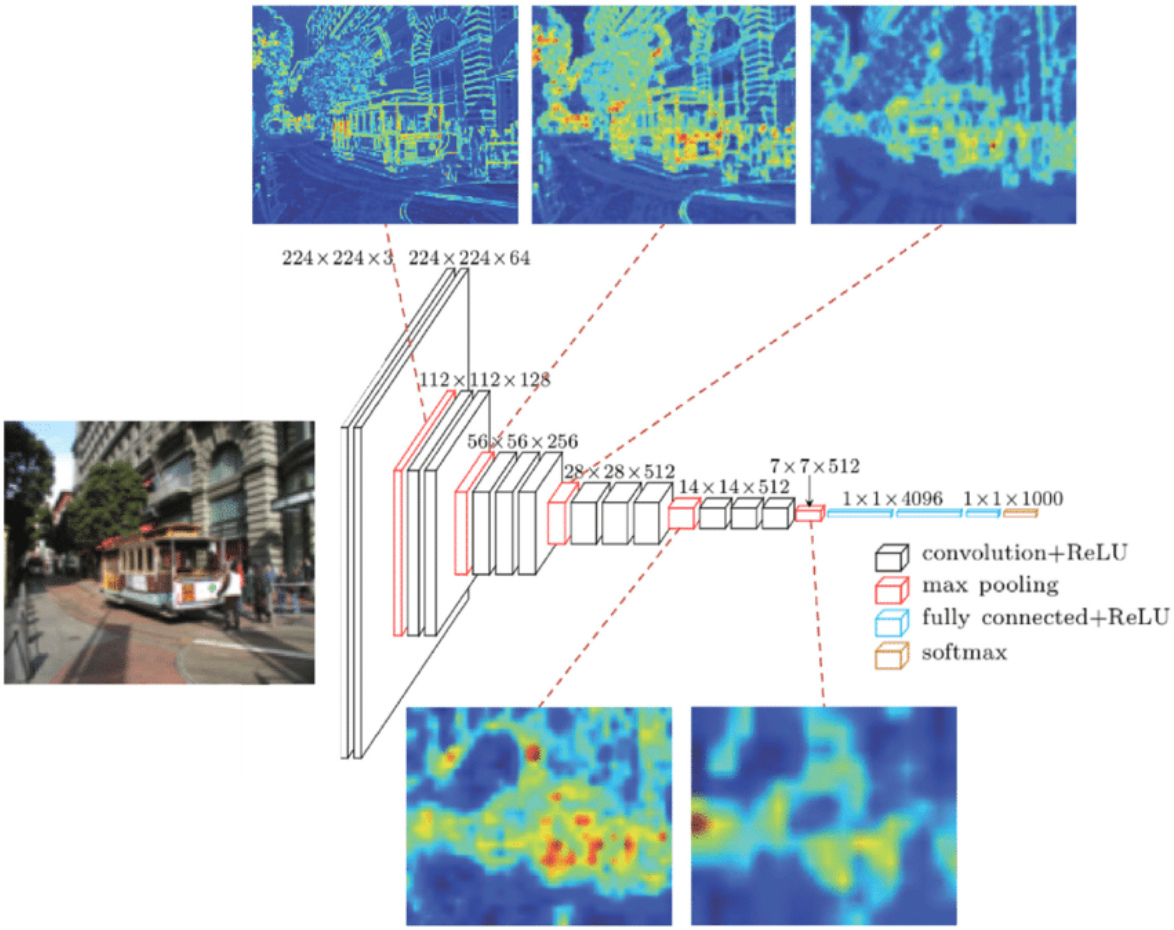


FIGURE 2.11 – Visualisation des caracteristiques des différentes couches [137]

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{2.6}$$

où z_i représente le score de la classe i et K le nombre total de classes.

2.4.7 Conclusion

Bien que plusieurs approches d'apprentissage automatique aient été étudiées pour la détection et la classification de la RD [28, 127, 171], les modèles d'apprentissage profond (DL) se sont distingués par leurs performances remarquables dans l'analyse et le traitement des images médicales [48, 12].

Toutefois, ces modèles présentent des limitations importantes : ils requièrent de vastes ensembles de données annotées, une puissance de calcul élevée et une mémoire conséquente, constituant un obstacle majeur à leur déploiement clinique [41]. L'un des défis fondamentaux réside dans la nécessité de disposer d'un volume important d'images rétiniennes annotées, alors que leur acquisition et leur annotation représentent des processus longs et coûteux. Les bases de données accessibles souffrent fréquemment de déséquilibres de classes, avec un déficit d'échantillons représentant certains stades de la maladie.

Face à ces limitations, l'apprentissage par transfert émerge comme une solution prometteuse, permettant de tirer parti de modèles pré-entraînés pour améliorer les performances tout en réduisant les exigences en données d'entraînement.

2.5 Transfer Learning (TL)

Face à la rareté, au coût prohibitif et à la complexité de l'acquisition et de l'annotation des données d'entraînement, le financement de leur collecte est devenu difficilement réalisable. Cette contrainte a favorisé le développement de techniques alternatives, notamment le transfert de connaissances entre tâches [161], qui constitue le fondement de l'apprentissage par transfert (TL).

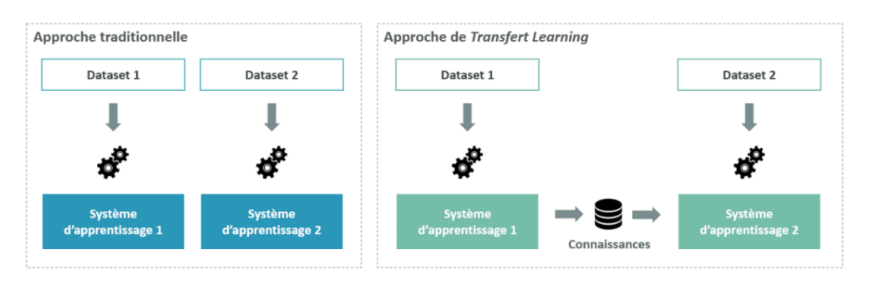


FIGURE 2.12 – Approche traditionnelle vs. Approche de Transfert Learning

2.5.1 Principe

Principe

Le transfer learning repose sur l'hypothèse que les représentations apprises par un réseau de neurones profond sur un domaine source peuvent être bénéfiques pour un domaine cible. Cette approche tire parti du fait que les couches inférieures des réseaux convolutionnels apprennent généralement des caractéristiques de bas niveau (contours, textures, formes géométriques) qui sont transférables entre différents domaines visuels (Voir Figure 2.12) [65, 174].

2.5.2 Formalisation Mathématique

Soit $\mathcal{D}_S$ le domaine source et $\mathcal{D}_T$ le domaine cible, avec leurs tâches respectives $\mathcal{T}_S$ et $\mathcal{T}_T$. L'objectif du transfer learning est d'améliorer l'apprentissage de la fonction de prédiction cible $f_T(\cdot)$ dans $\mathcal{D}_T$ en utilisant les connaissances acquises de $\mathcal{D}_S$ et $\mathcal{T}_S$, où $\mathcal{D}_S \neq \mathcal{D}_T$ ou $\mathcal{T}_S \neq \mathcal{T}_T$.

Le concept de transfert d'apprentissage repose sur plusieurs notions clés :

— **Modèle pré-entraîné** : Un modèle qui a été entraîné sur un vaste ensemble de données, généralement pour une tâche générale. Par exemple, un modèle pré-entraîné sur ImageNet pour la classification d'images naturelles.

— **Domaine source et tâche source** : Le domaine et la tâche sur lesquels le modèle pré-entraîné a été initialement entraîné (ex : classification d'objets sur ImageNet).

— **Domaine cible et tâche cible** : Le domaine et la tâche spécifiques pour lesquels on souhaite adapter le modèle pré-entraîné (ex : classification de la rétinopathie diabétique).

— **Transfert de connaissances** : Le processus d'adaptation du modèle pré-entraîné à la nouvelle tâche, impliquant généralement un ajustement des paramètres (fine-tuning) ou l'utilisation des caractéristiques extraites comme point de départ.

2.5.3 Types de Transfer Learning

Selon la relation entre les domaines et tâches source et cible, on distingue plusieurs types de transfer learning :

— Transfer Learning Inductif : Dans ce cas, les domaines source et cible sont identiques ($\mathcal{D}_S = \mathcal{D}_T$), mais les tâches diffèrent ($\mathcal{T}_S \neq \mathcal{T}_T$). Les données étiquetées sont disponibles dans le domaine cible.

— Transfer Learning Transductif : Les domaines source et cible sont différents ($\mathcal{D}_S \neq \mathcal{D}_T$) mais les tâches sont identiques ($\mathcal{T}_S = \mathcal{T}_T$). Également appelé adaptation de domaine.

— Transfer Learning Non-supervisé : Similaire au cas inductif, mais les tâches cibles ne disposent pas de données étiquetées.

2.5.4 Mécanismes du Transfer Learning

Le transfert d'apprentissage repose sur l'idée fondamentale de réutiliser des connaissances acquises lors de l'entraînement d'un modèle sur une tâche source pour améliorer les performances sur une tâche cible, souvent caractérisée par une quantité limitée de données annotées. Ce processus s'appuie sur plusieurs mécanismes principaux [161] :

Réutilisation de Poids (Weight Transfer)

Cette méthode fondamentale consiste à utiliser les poids d'un modèle pré-entraîné pour initialiser un nouveau modèle. Formellement, si θ_S représente les paramètres optimaux du modèle source, alors :

$$\theta_T^{(0)} = \theta_S$$

où $\theta_T^{(0)}$ constitue l'initialisation des paramètres pour la tâche cible.

Cette approche permet de conserver les représentations hiérarchiques acquises lors de l'entraînement sur le domaine source, particulièrement les filtres de bas niveau qui capturent des motifs universels (contours, textures, formes géométriques simples).

Adaptation Fine (Fine-tuning)

Après avoir réutilisé les poids, le modèle est affiné sur le domaine cible en continuant l'entraînement avec un ensemble de données spécifique. Cette étape implique généralement l'utilisation d'un taux d'apprentissage réduit pour préserver les connaissances acquises :

$$\theta_T^{(k+1)} = \theta_T^{(k)} - \alpha_{fine} \nabla_\theta \mathcal{L}_T(\theta_T^{(k)})$$

où $\alpha_{fine} \ll \alpha_{scratch}$ (avec $\alpha_{scratch}$ le taux d'apprentissage pour un entraînement from scratch), et $\mathcal{L}_T$ la fonction de perte sur le domaine cible.

Les différentes Stratégies de fine-tuning :

— **Fine-tuning global** : Ajustement de l'ensemble des paramètres du réseau

— **Fine-tuning partiel** : Gel des couches initiales et ajustement des couches finales uniquement

— **Fine-tuning progressif** : Dégelage graduel des couches du réseau pendant l'entraînement

Extraction de Caractéristiques (Feature Extraction)

Le modèle pré-entraîné sert d'extracteur fixe, seul le classificateur final est entraîné : :

$$f_T(x) = g_T(\phi_S(x))$$

où $\phi_S : \mathcal{X}_S \to \mathcal{H}$ représente l'extracteur de caractéristiques pré-entraîné (gelé), et $g_T : \mathcal{H} \to \mathcal{Y}_T$ est un nouveau classificateur entraîné spécifiquement pour la tâche cible.

Adaptation de Domaine (Domain Adaptation)

Cette technique vise à réduire l'écart de distribution entre les domaines source et cible. Elle peut être formalisée comme la minimisation de :

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda \mathcal{L}_{domain}$$

où $\mathcal{L}_{task}$ est la perte de classification et $\mathcal{L}_{domain}$ mesure la discordance entre les distributions des domaines source et cible.

Ces mécanismes exploitent la structure hiérarchique des réseaux profonds, où les couches initiales apprennent des représentations génériques et universelles de bas niveau, transférables à diverses tâches, tandis que les couches finales capturent des caractéristiques spécifiques [180, 113]. Cette approche optimise l'efficacité de l'apprentissage, particulièrement dans des contextes à données limitées [123]. Les principales approches de Transfer Learning utilisées communément en Deep Learning sont récapitulées dans la Figure 2.13.

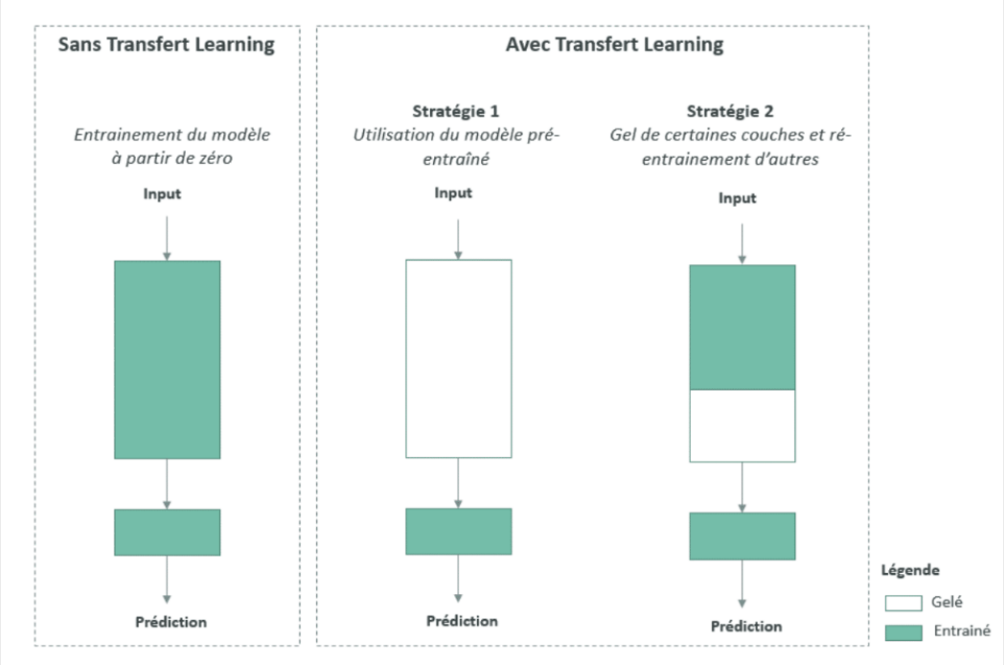


FIGURE 2.13 – Approche de Transfer Learning en Deep Learning

2.5.5 Historique et Contexte

L'évolution des modèles préentraînés en vision par ordinateur a connu des progrès architecturaux significatifs, dont les contributions importantes sont décrites chronologiquement ci-dessous.

1. **Fondements du transfert de connaissances (2000–2012)** : Les premières applications reposaient sur des descripteurs de caractéristiques manuels tels que SIFT (Scale-Invariant Feature Transform) [104] et HOG (Histogram of Oriented Gradients)[46]. Ces approches extrayaient des représentations locales robustes aux variations d'échelle et d'illumination, qui étaient ensuite utilisées dans des classificateurs traditionnels (SVM, Random Forest, etc.). Durant cette période, des recherches théoriques sur le *transfer*

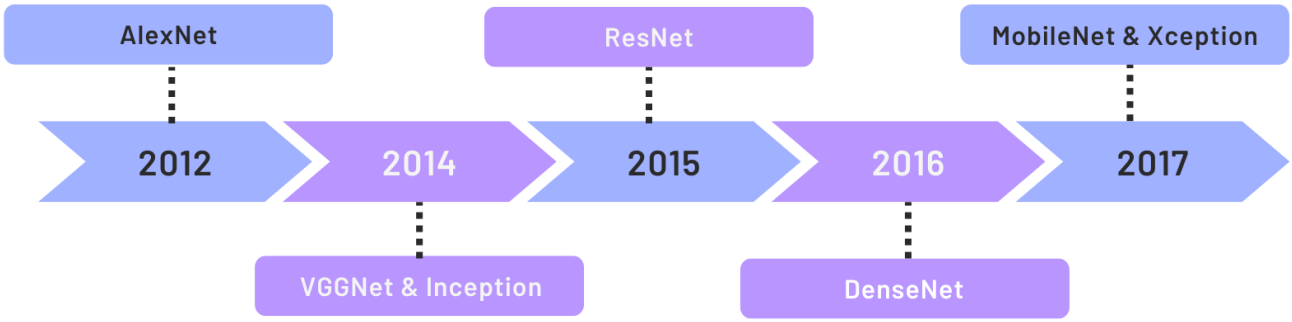


FIGURE 2.14 – Chronologie des principales étapes du développement des modèles de TL en imagerie médicale

learning se développaient [124, 161], posant les bases conceptuelles du transfert de connaissances entre domaines et et espaces de représentation distincts.

2. **Avènement d'AlexNet et des CNN profonds (2012-2014)** : L'introduction d'AlexNet [96] constitue une étape déterminante dans l'histoire de l'apprentissage profond. L'introduction des CNNs entraîné sur des grandes base de données d'images (ImageNet), a ouvert la voie à l'utilisation des CNN pré-entraînés comme extracteurs de caractéristiques génériques, facilitant ainsi l'émergence du transfert d'apprentissage à grande échelle, notamment vers des domaines spécialisés tels que l'imagerie médicale.
3. **Premières applications des CNN pré-entraînés en imagerie médicale (2014-2016)** : Les travaux de Tajbakhsh et al. [157] ont systématiquement évalué l'efficacité du transfer learning pour diverses tâches d'imagerie médicale. Cette période a vu l'émergence des premières applications spécifiques à la rétinopathie diabétique avec les travaux de Gulshan et al. [71].
4. **Architectures spécialisées (2016-2018)** : Le développement de CNN plus profonds (ResNet, DenseNet, Inception) a permis une meilleure capture des caractéristiques complexes des images rétiniennes. Les travaux de Abràmoff et al. [2] ont démontré l'efficacité clinique de ces approches pour le dépistage automatisé.
5. **Optimisation pour l'imagerie rétinienne (2018-2020)** : Cette période a vu l'émergence de techniques spécialisées : augmentation de données adaptée aux particularités rétiniennes, architectures multi-échelles pour capturer les lésions de différentes tailles, et approches ensemblistes. Les travaux de Ting et al. [160] ont établi de nouveaux standards de performance.
6. **L'arrivée des Transformers et approches hybrides (2020-présent)** : L'introduction des Vision Transformers (ViT) [52] a ouvert de nouvelles perspectives. Les approches hybrides CNN-Transformer montrent des performances prometteuses pour la classification fine des stades de rétinopathie diabétique, avec une meilleure interprétabilité des décisions diagnostiques.

Cette évolution reflète une maturation progressive du domaine, passant d'adaptations directes

de modèles génériques vers des approches spécialement conçues pour les défis de l'imagerie rétinienne : variabilité des conditions d'acquisition, subtilité des lésions précoces, et nécessité d'une classification précise des stades de sévérité.

2.5.6 Différentes Architectures de CNN pour le Transfer Learning

L'évolution des architectures de réseaux de neurones convolutionnels a considérablement influencé l'efficacité du transfer learning. Chaque architecture apporte des innovations spécifiques qui impactent directement les performances de transfert vers des domaines spécialisés comme l'imagerie médicale. Cette section présente les architectures les plus influentes utilisées comme modèles pré-entraînés.

1. **AlexNet** : AlexNet a été introduit en 2012 par Alex Krizhevsky et al. [96]. Comme illustré à la Figure 2.15.a, AlexNet se compose de huit couches, dont cinq couches convolutionnelles et trois couches entièrement connectées. L'importance d'AlexNet réside dans l'utilisation de l'unité linéaire rectifiée (ReLU) comme fonction d'activation, au lieu de la fonction sigmoïde ou tangente hyperbolique et du dropout pour la régularisation. AlexNet possède 60 millions de paramètres, c'est une architecture relativement simple mais efficace pour l'époque. Son impact sur le transfer learning réside dans sa capacité à apprendre des représentations génériques robustes dans les premières couches.
2. **GoogLeNet ou Inception v1** : Introduit en 2014 par Szegedy et al. [156], GoogLeNet a révolutionné l'architecture des CNNs grâce à l'intégration du module Inception. Ce module permet de traiter l'information à différentes échelles simultanément, en utilisant des convolutions parallèles (1x1, 3x3, 5x5) et du max-pooling (comme illustré à la Figure 2.15.c). Une autre innovation majeure réside dans le remplacement des couches entièrement connectées traditionnelles par un Global Average Pooling, ce qui a entraîné une réduction drastique du nombre de paramètres, passant de 60 millions pour AlexNet à seulement 7 millions. Cependant, un inconvénient de cette architecture est qu'elle peut saturer la précision du réseau à mesure que sa profondeur augmente.
3. **VGG16 :Visual Geometry Group networks** : Développé en 2015 par l'équipe du Visual Geometry Group de l'Université d'Oxford [145], VGG16 a démontré l'importance de la profondeur dans les CNNs. Cette architecture se caractérise par sa simplicité architecturale avec des filtres de convolution uniformément petits (3×3). Cette architecture se caractérise par sa profondeur, composée de 16 couches, dont 13 couches convolutionnelles et 3 couches entièrement connectées. VGG-16 est réputé pour sa simplicité et architecturale avec des filtres de convolution uniformément petits (3×3). VGG16 est lent pendant l'entraînement (138 millions de paramètres) et nécessite un espace de stockage plus important, ce qui rend son déploiement fastidieux (voir Figure 2.15.b).

4. **Xception**, proposé par Chollet en 2017 [42], Xception est une evolution de l'architecture Inception, basée sur l'innovation des convolutions séparables en profondeur. Sa principale innovation réside dans le remplacement des modules Inception classiques par des convolutions séparables en profondeur (convolution depthwise), où le filtrage spatial est effectué indépendamment pour chaque canal (convolution depthwise), suivie d'une convolution séparable ponctuelle (convolution pointwise). Cette architecture, composée de 71 couches et comptant 22,9 millions de paramètres, offre une efficacité computationnelle supérieure à celle d'Inception v3 tout en maintenant des performances comparables. L'inconvénient majeur est que ce réseau nécessite de très grands ensembles de données pour pouvoir entraîner tous ses paramètres.
5. **ResNet : Residual Neural Network** : ResNet a été mis au point en 2015 par Kaiming He et al.[75]. L'apport majeur de ResNet réside dans l'utilisation de connexions résiduelles comme l'illustre la Figure 2.15.d, également appelées « skip connections », qui permettent de contourner certaines couches du réseau (Voir Figure 2.15.f). Cette stratégie facilite l'entraînement de réseaux extrêmement profonds. Plutôt que d'apprendre directement une fonction de transformation, chaque bloc résiduel apprend le résidu, c'est-à-dire la différence entre l'entrée du bloc et la sortie désirée. Grâce à cette approche, ResNet a considérablement amélioré la performance, établissant de nouvelles références en termes de précision pour diverses tâches de vision par ordinateur.

   Principe des connexions résiduelles :

   $$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x}$$

   où $\mathcal{F}(\mathbf{x}, \{W_i\})$ représente la fonction résiduelle à apprendre.

6. **DenseNet**, DenseNet, ou réseaux de neurones densément connectés, a été présenté en 2017 par Huang et al. [78]. DenseNet a introduit dans sa structure des blocs densément connectés, où chaque couche est directement connectée à toutes les couches précédentes, favorisant ainsi la réutilisation des caractéristiques et réduisant le nombre de paramètres à apprendre. Grâce à ces connexions denses, chaque couche reçoit non seulement les signaux de la couche précédente, mais l'ensemble des cartes de caractéristiques extraites antérieurement, ce qui facilite la propagation des gradients et atténue le problème de la disparition des gradients dans les réseaux très profonds. DenseNet a démontré des performances remarquables dans des tâches de classification d'images et s'est imposé comme une alternative efficace aux architectures traditionnelles, offrant un excellent compromis entre complexité et précision.

Principe des connexions denses :

$$\mathbf{x}_l = H_l([\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_{l-1}])$$

où $[\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_{l-1}]$ représente la concaténation des cartes de caractéristiques.

7. **MobileNet : Mobile Networks** , proposé par Howard et al. en 2017 [76], est une architecture de réseau de neurones convolutionnel légère, spécialement conçue pour les appareils mobiles et embarqués. Il introduit les convolutions séparables en profondeur, qui factorisent l'opération de convolution standard en une convolution "depthwise" suivie d'une convolution "pointwise". Cette approche réduit la complexité computationnelle ainsi que la taille du modèle tout en maintenant une bonne précision. MobileNet parvient à un compromis efficace entre taille du modèle et précision, ce qui le rend particulièrement adapté aux environnements disposant de ressources limitées. Il est largement adopté dans diverses applications nécessitant une inférence en temps réel ou directement sur l'appareil.

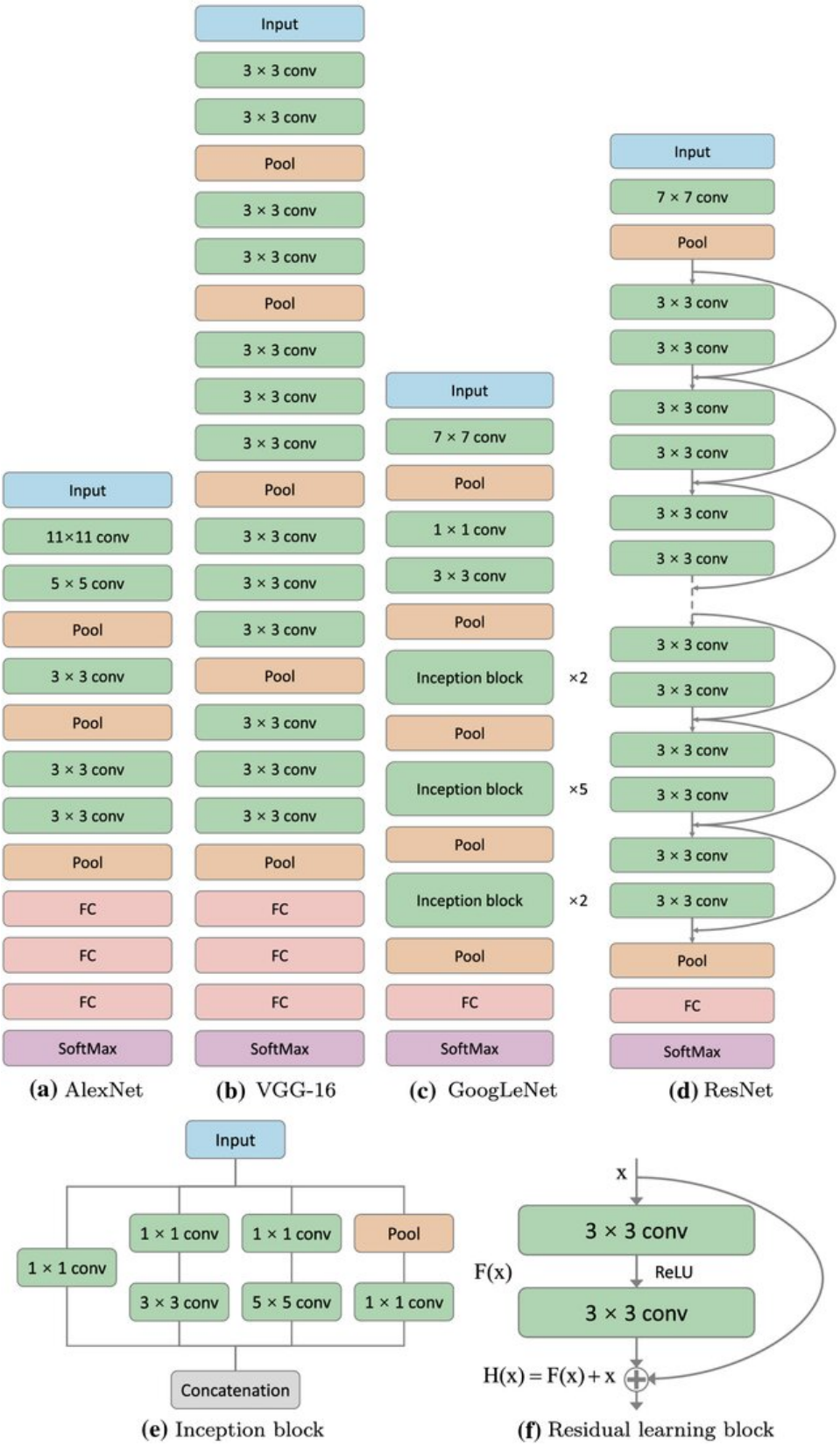


FIGURE 2.15 – Exemples d'architectures basées sur le transfer learning [184]

2.5.7 Comparaison des Modèles Pré-entrainés

Le tableau 2.1 présente une synthèse comparative des principales architectures de réseaux de neurones convolutifs utilisées pour le transfer learning. Cette comparaison porte sur les aspects architecturaux, computationnels et de performance, mesurés sur le jeu de données ImageNet-1K [133]. Les métriques de performance incluent la précision top-1 (classification exacte en une seule classe) et top-5 (étiquette correcte parmi les cinq premières prédictions). Les temps d'inférence sont mesurés sur GPU Tesla V100 avec une résolution d'entrée de 224×224 pixels.

> La précision est calculée en divisant le nombre d'observations correctement classées par le nombre total d'observations. La mesure « top-k » correspond à la précision obtenue en considérant les k premières étiquettes prédites par le modèle.

la précision top-5 indique la proportion d'observations pour lesquelles l'étiquette correcte figure parmi les cinq premières prédictions, tandis que la précision top-1 représente celle d'une classification en une seule classe. Ainsi, pour k = 5, on vérifie si l'étiquette réelle est présente parmi les cinq étiquettes prédites, alors que pour k = 1, la précision top-k correspond à la précision de facto[87].

TABLE 2.1 – Comparaison des architectures de réseaux de neurones convolutifs pour le transfer learning.

Architecture	Paramètres	FLOPs	Top-1	Top-5	Inférence
AlexNet$^{(2012)}$	60M	714M	57.1%	84.7%	127ms
VGG-16$^{(2014)}$	138M	15.5G	71.6%	90.6%	198ms
Inception-v3$^{(2015)}$	24M	5.7G	77.5%	93.3%	89ms
ResNet-50$^{(2015)}$	26M	4.1G	76.1%	92.9%	76ms
DenseNet-121$^{(2016)}$	8M	2.9G	74.4%	92.2%	67ms
Xception$^{(2016)}$	23M	8.4G	79.0%	94.5%	91ms
MobileNet-v2$^{(2018)}$	3.4M	300M	72.0%	91.0%	31ms
EfficientNet-B0$^{(2019)}$	5.3M	390M	77.3%	93.5%	43ms
ConvNeXt-T$^{(2022)}$	29M	4.5G	82.1%	96.0%	64ms

L'analyse révèle plusieurs tendances importantes dans l'évolution des architectures.

Les modèles (AlexNet, VGG) présentent un coût computationnel élevé avec des performances modestes, en raison de leur architecture dense et de leur grand nombre de paramètres [35, 159].

Les architectures intermédiaires (Inception, ResNet, DenseNet) optimisent l'équilibre

précision-efficacité grâce à des innovations structurelles : modules parallèles pour Inception, connexions résiduelles pour ResNet, et réutilisation de features pour DenseNet. Xception améliore encore cet équilibre via les convolutions depthwise séparables. Les architectures modernes se distinguent par leur spécialisation : MobileNet privilégie l'efficacité pour les environnements contraints (IoT, mobile), EfficientNet optimise le scaling composé, tandis que Vision Transformer et ConvNeXt représentent les approches les plus récentes avec des performances de pointe.

Le choix du modèle dépendra donc des exigences spécifiques de la tâche, de l'ensemble de données et des ressources disponibles.

2.6 Vision Transformer (Vit)

L'avènement des Vision Transformers (ViT) a marqué une rupture significative dans le domaine de la vision par ordinateur, remettant en question la domination des réseaux de neurones convolutionnels (CNN) qui prévalait depuis plusieurs années.

> Le ViT est une architecture d'apprentissage profond conçue pour traiter les données visuelles en utilisant la même architecture de transformateur qui a révolutionné le traitement du langage naturel (NLP), introduite en 2017 par Vaswani et al. [165]. Cette architecture, initialement proposée par Dosovitskiy et al. en 2020 [52], transpose avec succès les mécanismes d'attention du domaine textuel vers l'analyse d'images.

Contrairement aux CNN, qui s'appuient sur des convolutions pour capturer les caractéristiques spatiales locales, les ViT ont introduit une approche novatrice, traitant les images comme des séquences de patchs et exploitant le mécanisme d'auto-attention pour capturer les dépendances globales [165]. Cette transition a ouvert de nouvelles perspectives, offrant des avantages considérables en termes de modélisation du contexte global, d'évolutivité et de flexibilité.

Le modèle ViT s'est imposé dans de nombreuses tâches de vision par ordinateur : classification [52], détection d'objets [37], segmentation [187], et analyse d'images médicales [40].

Dans ce qui suit, nous allons explorer en détail la structure et les composants de l'architecture de Vision Transformer.

2.6.1 Architecture des Transformers

Les Transformers constituent une révolution dans le traitement des données séquentielles, introduits par [165] avec l'article «Attention is All You Need». Comme nous pouvons le voir sur la Figure 2.16, cette architecture repose sur le mécanisme d'auto-attention, qui permet de modéliser de manière efficace les relations à longue distance entre les éléments

d'une séquence, sans recourir aux mécanismes de récursivité ou de convolution présents dans les approches antérieures. Cette capacité a conduit à l'adaptation de ces modèles au-delà du traitement du langage naturel, notamment dans le domaine de la vision par ordinateur .

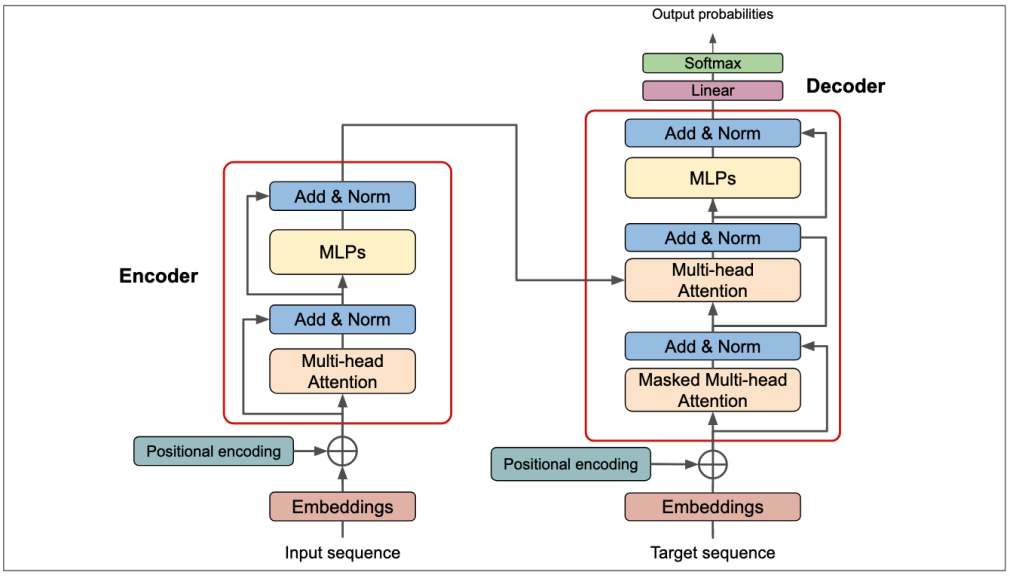


FIGURE 2.16 – Architecture d'un transformer[165]

Le mécanisme d'auto-attention constitue la clé du succès des Transformers. En permettant le calcul simultané des interactions entre tous les éléments d'une séquence, il offre une modélisation parallèle et efficace des dépendances à longue distance. Cette caractéristique est particulièrement avantageuse dans des tâches telles que la traduction automatique ou la modélisation du langage [165, 49], et s'avère également déterminante pour la classification d'images dans le cadre des ViT en minimisant les biais inductifs des CNN, l'auto-attention globale des ViT permet ainsi d'exploiter pleinement les données, notamment lorsque celles-ci sont volumineuses et hétérogènes [52].

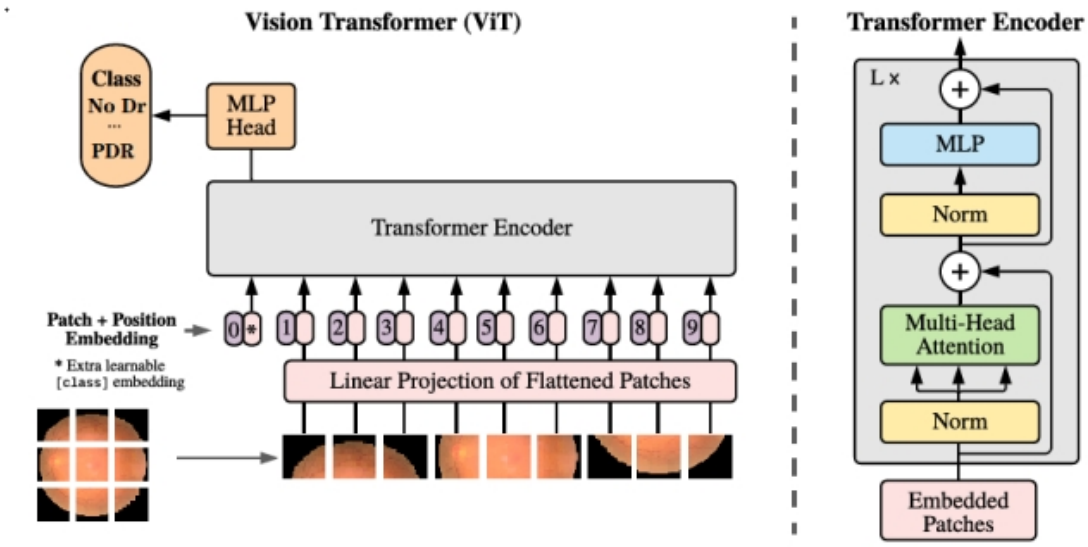


FIGURE 2.17 – Architecture Vision Transformer [52]

2.6.2 Architecture des Vision Transformers

Les Vision Transformers (ViTs) adaptent l'architecture des Transformers, initialement conçue pour le traitement du langage naturel, à l'analyse d'images. Leur force réside dans leur capacité à modéliser des interactions globales entre toutes les régions d'une image, sans se limiter aux voisinages locaux comme le font les réseaux de neurones convolutionnels (CNN). La Figure 2.17 illustre l'architecture de base des ViTs, qui se décompose en quatre parties principales :

— Découpage en patches : Division de l'image en fragments non-chevauchants

— Projection linéaire : Transformation des patches en embeddings de dimension fixe

— Ajout d'informations positionnelles : Injection de l'information spatiale

— Traitement par l'encodeur Transformer : Application des couches d'attention et de feed-forward

Découpage de l'image en patches

La première étape consiste à diviser l'image en une séquence de patchs d'image de manière analogue aux tokens dans un modèle NLP (voir figure 2.18).

Pour une image de dimensions H×W×C (hauteur, largeur, canaux), on partitionne l'image en N patches carrés de taille P×P. Formellement, le nombre de patches est calculé par :

$$N = HW/P^2 \tag{2.7}$$

Par example, une image de taille 224x224 peut être divisée en 16x16 patches, ce qui donne 224/16 * 224/16 =14×14=196 patches.

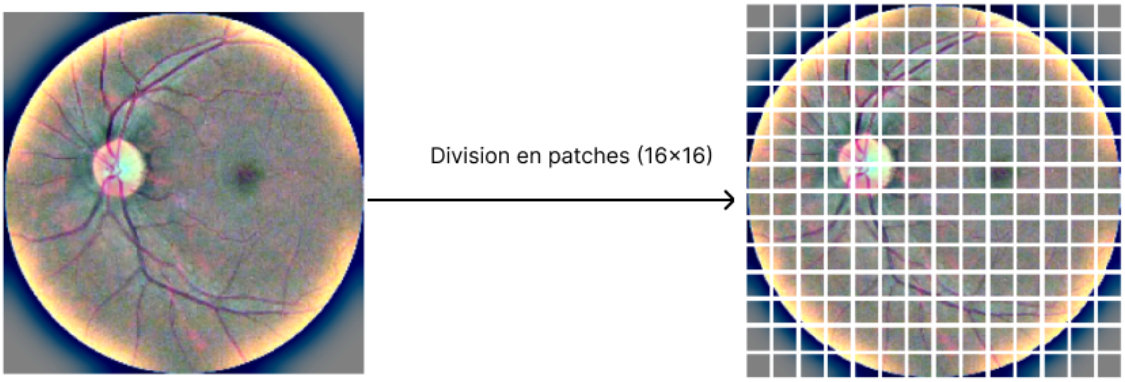


FIGURE 2.18 – Division en patches de taille fixe

Projection linéaire des patches (Embedding)

Chaque patch x_i de taille (P×P×C) est aplati en vecteur de dimension $P^2.C$. Ces vecteurs sont ensuite projetés dans un espace latent de dimension D via une transformation linéaire, définie par :

$$z_i = W.x_i + b \tag{2.8}$$

— $x_i \in \mathbb{R}^{P^2 \cdot C}$ représente le vecteur aplati du $i^{\text{ème}}$ patch ;

— $W \in \mathbb{R}^{D \times (P^2 \cdot C)}$ est la matrice de projection ;

— $b \in \mathbb{R}^D$ est un biais optionnel.

Ainsi, chaque patch est encodé sous forme d'un vecteur de dimension D, compatible avec le traitement par le Transformer [52].

Token de classe [CLS] et embeddings de position

Afin d'agréger les informations de l'ensemble des patches, un token de classe [CLS] de dimension D est concaténé à la séquence d'embeddings. Parallèlement, des embeddings positionnels $E_{pos} \in \mathbb{R}^{n+1 \times D}$ sont ajoutés pour injecter l'information spatiale :

$$Z = \begin{bmatrix} [\text{CLS}] \\ z_1 \\ \vdots \\ z_N \end{bmatrix} + E_{\text{pos}}$$

Ces embeddings permettent au modèle d'apprendre la localisation relative de chaque patch dans l'image.

Encodeur Transformer et Tête MLP

Comme l'illustre la figure 2.17, la séquence ainsi obtenue est ensuite traitée par une pile de L couches d'encodeur Transformer. Chaque couche comprend deux modules principaux :

1. **Multi-Head Self-Attention (MSA)**

   Le mécanisme d'auto-attention multi-têtes permet à chaque élément de la séquence d'établir des connexions pondérées avec l'ensemble des autres éléments, modélisant ainsi les dépendances globales au sein de la représentation. Pour une entrée $\mathbf{z}_{l-1}$, les matrices de requêtes **Q**, clés **K** et valeurs **V** sont calculées :

$$\mathbf{Q} = \mathbf{z}_{l-1}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{z}_{l-1}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{z}_{l-1}\mathbf{W}_V$$

où $\mathbf{W}_Q$, $\mathbf{W}_K$, et $\mathbf{W}_V \in \mathbb{R}^{D\times D}$ sont des matrices de projection apprises.

L'attention pour chaque tête h (avec $h = 1, \ldots, H$) est définie par :

$$\text{Attention}_h(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$$

où $d_k = D/H$ est la dimension des vecteurs de requêtes et clés pour chaque tête, et H est le nombre total de têtes.

Les sorties de toutes les têtes sont ensuite concaténées puis projetées à nouveau :

$$\text{MSA}(\mathbf{z}_{l-1}) = \text{Concat}\left(\text{Attention}_1, \ldots, \text{Attention}_H\right)\mathbf{W}_O$$

où $\mathbf{W}_O \in \mathbb{R}^{D\times D}$ est une matrice de projection linéaire finale.

2. **Réseau feed-forward (FFN)** : Appliqué de manière indépendante à chaque vecteur, il réalise une transformation non linéaire définie par :

$$\text{FFN}(x) = W_2 \cdot \text{GELU}(W_1 \cdot x + b_1) + b_2.$$

où W1,W2 sont des matrices de poids et GELU une fonction d'activation.

Chaque sous-couche est encadrée par des connexions résiduelles et une normalisation (LayerNorm LN), garantissant ainsi la stabilité de l'apprentissage.

3. **Tête de Classification** Finalement, après le passage à travers les L couches de l'encodeur, seule la représentation du token CLS $\mathbf{z}_L^{cls}$ est utilisée pour la prédiction finale :

$$y = \ln(z_{L_0})\mathbf{W}_{\text{head}} \tag{2.9}$$

où $\mathbf{W}_{\text{head}} \in \mathbb{R}^{D\times K}$ est la matrice de projection vers les K classes de sortie.

2.6.3 Comparaison entre CNN et ViT

Bien que les ViT et les CNN soient tous deux utilisés pour le traitement d'images, ils diffèrent fondamentalement dans leur approche de l'analyse visuelle. Cette distinction archi-

tecturale entraîne des implications importantes pour leurs performances et leur applicabilité dans différents contextes, notamment en imagerie médicale.

Réseaux de Neurones Convolutionnels (CNN) Les CNN exploitent la structure spatiale locale des images en appliquant des opérations de convolution sur des fenêtres glissantes de taille réduite. Cette approche hiérarchique permet une extraction progressive des caractéristiques, des détails locaux (contours, textures) vers des concepts de plus haut niveau. Les CNN possèdent un biais inductif spatial fort, incorporant naturellement des propriétés telles que l'invariance translationnelle et la localité spatiale.

Vision Transformers (ViT) À l'inverse, les ViT adoptent une approche globale en traitant l'image comme une séquence de patches et en utilisant le mécanisme d'auto-attention pour capturer les relations entre toutes les régions simultanément [52]. Cette architecture permet de modéliser directement les dépendances à long terme, mais nécessite un apprentissage explicite des relations spatiales locales, contrairement aux CNN qui les intègrent structurellement [94].

Le tableau 2.2 synthétise les principales différences entre ces deux approches :

TABLE 2.2 – Comparaison entre les architectures CNN et Vision Transformers (ViT)

Aspect	**Vision Transformers (ViT)**	**Réseaux de Neurones Convolutionnels (CNN)**
Mécanisme principal	Auto-attention multi-têtes permettant la modélisation des dépendances globales	Convolutions locales avec fenêtres glissantes pour l'extraction hiérarchique
Structure des données	Traitement séquentiel de patches d'image comme tokens	Préservation de la structure spatiale 2D native
Capacité de modélisation	Relations à long terme dès les premières couches	Construction progressive de représentations hiérarchiques
Robustesse	Sensible aux perturbations adverses mais robuste aux occlusions	Robustesse éprouvée aux variations locales
Transfer Learning	Excellentes performances après pré-entraînement sur de larges corpus	Efficace même avec des modèles pré-entraînés sur des jeux de données modestes
Complexité mémoire	$O(n^2)$ avec le nombre de patches	$O(n)$ avec la taille de l'image

2.6.4 Conclusion

Dans ce chapitre nous avons établi les fondements architecturaux des approches modernes de vision artificielle, analysant la transition des réseaux de neurones convolutifs (CNN)

aux Vision Transformers (ViTs). L'apprentissage profond y a radicalement transformé l'extraction de caractéristiques, supplantant les méthodes manuelles traditionnelles. Les CNN démontrent leur efficacité par l'exploitation de sous-réseaux locaux (contours, textures), atteignant des performances optimales pour la classification et la détection en imagerie médicale. À l'inverse, les ViTs instaurent un paradigme attentionnel global modélisant les relations inter-patchs, offrant ainsi une modélisation holistique adaptée aux structures anatomiques complexes. Le chapitre suivant appliquera ces architectures à la classification de la rétinopathie diabétique (RD), évaluant leur pertinence clinique pour la détection précoce des lésions rétiniennes.

CHAPITRE 3

Revue de Littérature

3.1 Introduction

La rétinopathie diabétique (RD) constitue l'une des principales causes de cécité évitable dans la population active mondiale. L'hyperglycémie chronique induit des altérations microvasculaires rétiniennes, conduisant à l'apparition progressive de lésions telles que microanévrismes, hémorragies et néovaisseaux pathologiques. Un dépistage précoce et un suivi régulier sont dès lors cruciaux pour prévenir la perte de vision irréversible et améliorer le pronostic fonctionnel des patients diabetiques.

La détection précoce de ces anomalies, repose sur l'expertise médicale mobilisée lors de l'examen du fond d'œil. Néanmoins, ce mode de dépistage, coûteux, chronophage et dépendant de l'expérience clinique, limite considérablement l'accès rapide aux soins pour de nombreux patients. Face à ces contraintes, les méthodes d'apprentissage automatique, et plus récemment d'apprentissage profond (Deep Learning), offrent des perspectives nouvelles pour l'automatisation du diagnostic. Dans ce contexte, dans le présent chapitre, nous proposons, dans un premier temps, une présentation des principales bases de données publiques mises à disposition pour l'entraînement et la validation des modèles de classification de la RD, ainsi qu'une revue des métriques d'évaluation standards permettant de mesurer leur performance. Nous analysons ensuite les contributions majeures du transfert learning appliqué pour la détection de la RD, avant d'étudier l'émergence des Vits weight decaycomme alternative prometteuse.

3.2 Bases de Données et Métriques

Dans cette section nous allons présenté les datasets ainsi que les métriques d'évaluation les plus utilisés

3.2.1 Bases de Données

Afin d'évaluer les performances des modèles de classification de la RD, plusieurs datasets publics ont été largement utilisés :

— **EyePACS (Kaggle Diabetic Retinopathy Detection)** : constitue l'une des bases les plus utilisées pour la classification de la RD. Il comprend environ 88 700 images couleur du fond d'œil, capturées à l'aide de divers dispositifs d'acquisition [44]. Les images sont annotées selon cinq stades de sévérité de la rétinopathie (niveaux 0 à 4). Ce jeu présente une forte variabilité en termes de résolution (allant de 1024×1024 à 4288×2848 pixels) et de luminosité, ce qui en fait un ensemble de données particulièrement exigeant pour les modèles d'apprentissage automatique.

— **Messidor 1 et 2** :

  — Messidor-1 : Messidor-1 comporte 1200 images (1440×960 pixels), annotées à la fois de manière binaire (présence ou absence de RD) et selon quatre niveaux de gravité [47].

  — Messidor-2 : contient 1744 images de même résolution, enrichies par un grading expert et accompagnées de métadonnées cliniques, fournissant ainsi un contexte diagnostique plus riche.

— **APTOS 2019 Blindness Detection** : élaboré pour une compétition Kaggle entre octobre et décembre 2019, propose 3662 images annotées selon cinq niveaux de sévérité de la RD (0 à 4) [90]. Provenant de cliniques indiennes, ces images présentent une forte hétérogénéité en termes de qualité, d'exposition et de conditions d'acquisition, ce qui en fait un ensemble particulièrement pertinent pour évaluer la robustesse des modèles de classification.

— **DDR (Diabetic Retinopathy Debrecen)** : constitue l'un des ensembles les plus volumineux disponibles, avec 13 673 images annotées selon cinq niveaux de sévérité de la RD [185]. Elle se distingue par la variété des appareils d'acquisition utilisés et par l'inclusion de données multimodales (formats JPEG et TIFF), permettant d'évaluer la capacité des modèles à généraliser sur des modalités d'imagerie différentes.

— **FGADR (Fine-Grained Annotated Diabetic Retinopathy)** : est plus restreinte, comprenant 1842 images [72]. Elle se caractérise par des annotations fines au niveau des lésions élémentaires telles que les microanévrismes, les hémorragies et les exsudats. Conçue pour l'étude de la capacité des modèles à localiser précisément les anomalies, FGADR constitue une ressource précieuse pour les approches orientées segmentation et intention.

— **OCTA (Optical Coherence Tomography Angiography)** : plusieurs petits dataset (300–1000 images) d'angiographies OCT, fournissant une vue microvasculaire [103]. Ces ensembles exploitent l'angiographie par cohérence optique pour fournir une cartographie microvasculaire détaillée de la rétine.

L'analyse des bases de données en imagerie rétinienne révèle plusieurs problématiques majeures affectant la qualité des images et la robustesse des modèles.

— **Variabilité inter-appareils** : différences de résolution, de champs de vue et de calibrage des capteurs impactent la distribution des pixels et la qualité des textures.

— **Artefacts et bruit** : presence de reflets, flous et artefacts de compression JPEG exigent des étapes de *pre-processing* (CLAHE, filtres médian, débruitage).

— **Déséquilibre de classes** : stades sévères (3–4) sous-représentés, pénalisant l'entraînement des modèles.

— **Annotations approximatives** : certaines images peuvent souffrir de labels erronés ou imprécis qui complique la convergence.

TABLE 3.1 – Résumé des principales bases de données utilisées pour la classification de la rétinopathie diabétique et leurs limitations.

Base de données	**Nombre d'images**	**Annotations**	**Spécificités**	**Limitations**
EyePACS (Kaggle)	88 700	5 niveaux	Variabilité élevée de résolution et de qualité ; multiples appareils d'acquisition [44]	Annotations parfois imprécises ; hétérogénéité forte des images ; artefacts et bruit
Messidor-1	1 200	Présence/absen RD et 4 niveaux	Haute qualité d'images ; cohérence clinique [47]	Taille limitée pour l'apprentissage profond ; distribution déséquilibrée
Messidor-2	1 744	4 niveaux (grading expert)	Inclut métadonnées cliniques ; annotations expertes [47]	Diversité démographique limitée ; faible variabilité d'acquisition
APTOS 2019	3 662	5 niveaux	Variabilité de qualité et d'exposition ; données issues de cliniques indiennes [90]	Flous fréquents ; bruit numérique ; annotations parfois subjectives
DDR	13 673	5 niveaux	Données multimodales (JPEG/TIFF) ; diversité d'appareils	Variabilité de qualité importante ; déséquilibre inter-classes (moins de cas sévères)
FGADR	1 842	Annotations fines de lésions	Localisation précise des anomalies ; segmentation des lésions [72]	Petite taille du corpus ; centré sur la localisation plutôt que sur la classification globale
OCTA	300–1000 par corpus	Images microvasculaires (angiographie OCT)	Analyse fine de la perfusion rétinienne ; détection précoce des signes vasculaires [103]	Bases réduites ; interprétation complexe ; non comparables aux fonds d'œil standards

Nous avons constaté que le jeu de données APTOS 2019 a connu un usage croissant dans la littérature récente, en raison de son accessibilité publique via la plateforme Kaggle, de sa taille intermédiaire compatible avec les architectures profondes, de ses annotations cliniques standardisées en cinq classes, ainsi que de la variabilité des images collectées en contexte réel. Ces caractéristiques en font un support privilégié pour le développement, l'entraînement et la validation de modèles robustes.

3.2.2 Prétraitement des Données

La phase de prétraitement est essentielle pour la détection et la classification des images DR. Elle consiste en un large éventail de techniques et de procédures qui contribuent à améliorer la qualité globale et l'interprétabilité des images. En réduisant le bruit, en éliminant les artefacts et en traitant d'autres éléments indésirables, le prétraitement vise à améliorer la précision du système de classification. En général, le redimensionnement de l'image, l'amélioration du contraste et la réduction du bruit des images rétiniennes sont effectués sur les images du fond d'œil des différents ensembles de données. L'augmentation des données est couramment appliquée en tant qu'étape de prétraitement dans de nombreux algorithmes basés sur l'apprentissage profond pour résoudre le problème du déséquilibre des classes dans l'ensemble de données lors de la détection et de la classification.

Redimensionnement des images

Le redimensionnement des images à une taille fixe constitue une étape préalable indispensable, permettant de standardiser les données issues de différentes sources ou dispositifs d'acquisition. L'imposition d'une dimension uniforme facilite l'intégration des images dans les modèles de classification sans nécessiter d'opérations de redimensionnement supplémentaires lors des phases d'entraînement et d'évaluation. Cette uniformisation permet également de s'assurer que toutes les images sont traitées selon les mêmes contraintes spatiales, éliminant ainsi les biais liés aux variations de résolution ou de cadrage et permettant aux algorithmes de se concentrer sur les caractéristiques intrinsèques des structures rétiniennes et les motifs associés à la RD.

Recadrage des images (Cropping)

Le recadrage [130] des images est une opération de prétraitement visant à éliminer les zones non informatives ou non pertinentes de l'image, afin de recentrer l'analyse sur la région d'intérêt (voir Figure 4.8c). Dans le cadre de la classification de la RD, cette étape permet de focaliser l'attention du modèle sur les structures anatomiques essentielles de la rétine. Le recadrage permet d'exclure les bordures sombres ou les artefacts liés au champ visuel de l'appareil de capture, tout en conservant la région centrale de la rétine, notamment la macula et la papille optique où les manifestations cliniques de la rétinopathie sont fréquemment observées.

Filtrage des images

Le filtrage des images constitue une étape fondamentale dans le traitement de l'imagerie médicale, visant à améliorer la qualité visuelle des images, à réduire le bruit. Cette technique consisyte à onvoluer l'image avec différents filtres spécifiques qui modifient les valeurs des pixels pour mettre en évidence les structures d'intérêt tout en atténuant les éléments

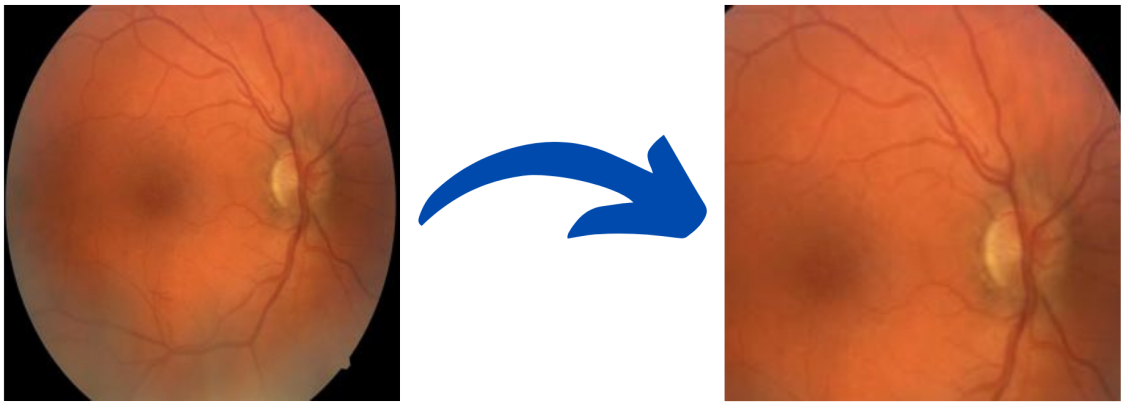


FIGURE 3.1 – Exemple d'image de fond d'oeil recadrer

parasites [34, 60]. Dans le contexte de la RD, un filtrage approprié peut significativement renforcer la précision de la classification automatique, en facilitant la détection des lésions caractéristiques telles que les microanévrismes, exsudats ou hémorragies. Parmi les filtres les plus couramment utilisés sont :

— Le filtre gaussien, qui permet un lissage efficace de l'image tout en conservant les structures globales,

— Le filtre médian, particulièrement efficace pour la suppression du bruit impulsionnel tout en préservant les contours,

— , est un filtre utilisé pour améliorer le contraste local de l'image tout en limitant l'amplification du bruit dans les zones homogènes.

La Figure 3.2 visualise les résultats obtenus suite à l'application de ces filtres.

Augmentation de l'ensemble de données

L'augmentation des données est une technique essentielle en apprentissage profond, consistant à appliquer un ensemble de transformations spécifiques aux images d'entraînement, dans le but d'enrichir artificiellement la diversité de l'ensemble d'apprentissage [50]. Dans le contexte de la classification de la RD, cette méthode s'avère particulièrement cruciale pour pallier les limitations liées à la faible quantité de données annotées et au déséquilibre fréquent entre les classes. L'augmentation des données permet de réduire le risque de sur-apprentissage et à équilibrer la représentation des différentes classes (notamment les stades rares de la RD par exemple). L'augmentation consiste concrètement à générer des variantes artificielles des images originales à l'aide de transformations géométriques aléatoires telles

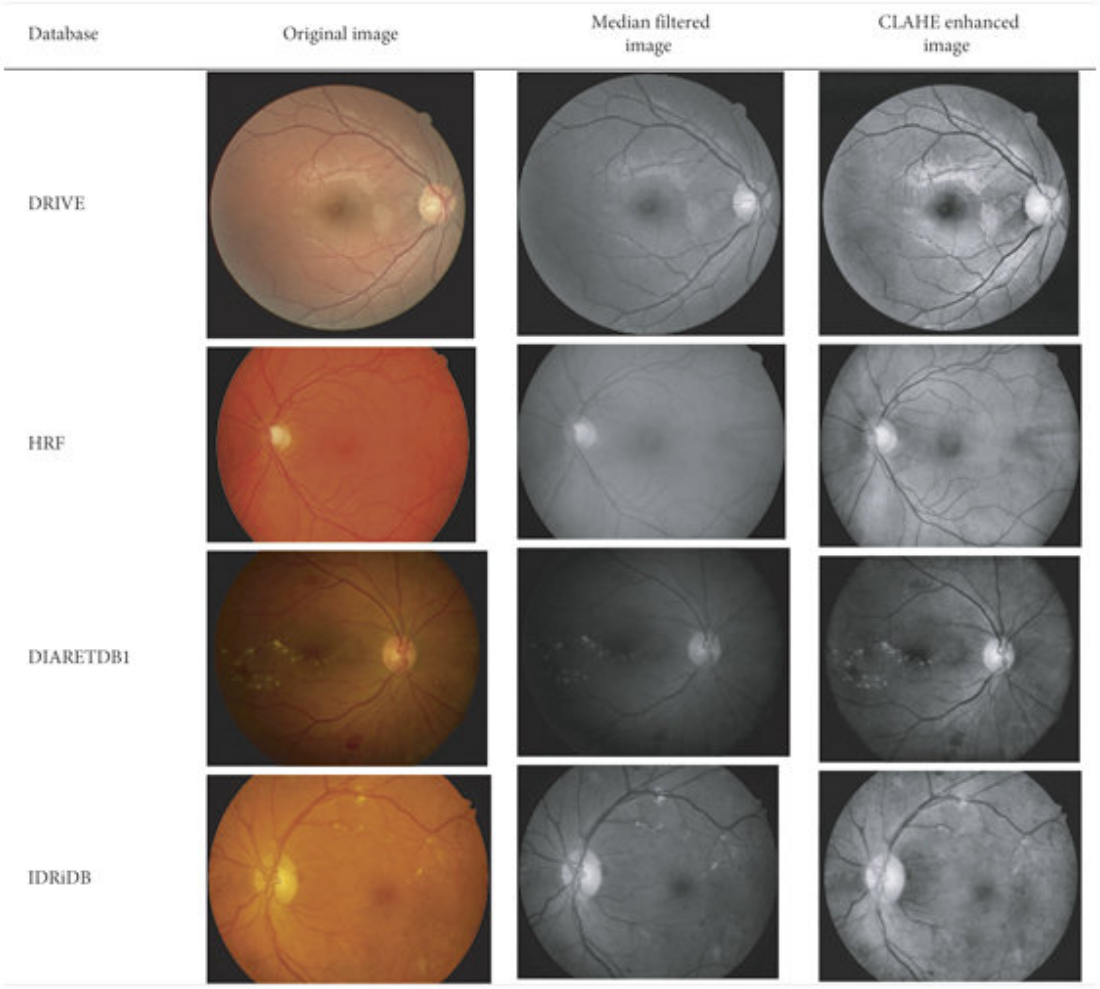


FIGURE 3.2 – Exemple d'images prétraitées

que le retournement horizontal ou vertical, la rotation, le zoom ou encore le recadrage (voir Figure 3.3). Ces opérations conservent l'intégrité des informations cliniques pertinentes tout en augmentant la diversité apparente du jeu de données.

Normalisation des Images

La normalisation des images comme l'illustre la Figure 3.4, vise à corriger les variations d'intensité lumineuse, de contraste et de dynamique entre les images issues de différentes sources. Une méthode couramment utilisée consiste à mettre à l'échelle les valeurs des pixels dans une plage normalisée, typiquement entre 0 et 1. Cette opération permet d'homogénéiser les niveaux d'intensité tout en conservant les relations relatives entre les structures anatomiques et pathologiques présentes dans les images. En assurant une distribution cohérente des valeurs d'entrée, la normalisation favorise la stabilité de l'optimisation durant l'entraînement des réseaux de neurones [80].

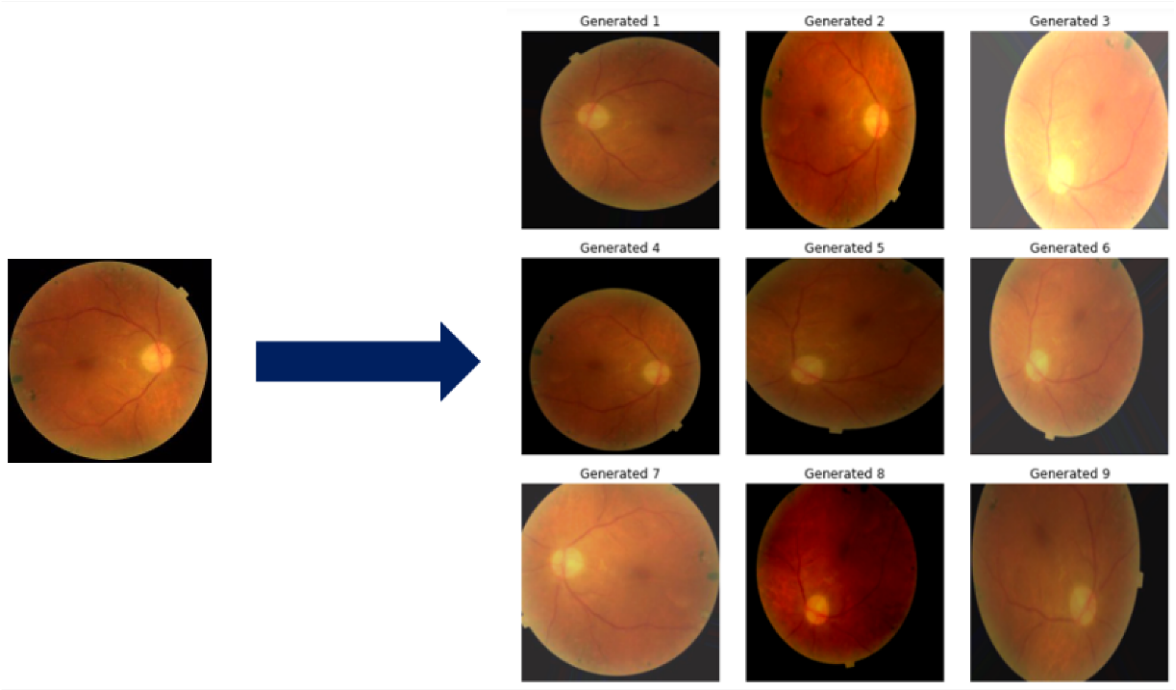


FIGURE 3.3 – Exemple d'augmentation d'images

3.2.3 Métriques d'Evaluation

L'évaluation des performances des modèles de deep learning est une étape fondamentale visant à mesurer leur capacité à produire des prédictions fiables sur des données non vues. Dans les tâches de classification, cette évaluation repose principalement sur l'analyse de la matrice de confusion, qui compare les prédictions du modèle aux étiquettes réelles. Soient :

— TP (True Positives) : vrais positifs,

— TN (True Negatives) : vrais négatifs,

— FP (False Positives) : faux positifs,

— FN (False Negatives) : faux négatifs.

Les métriques les plus couramment employées sont [149, 126] :

— **Accuracy ou exactitude (Acc)** :Mesure la proportion de prédictions correctes, toutes classes confondues. Elle Fournit une évaluation globale de la performance du modèle de classification.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

L'exactitude est une bonne mesure lorsque les classes sont équilibrées.

— **Précision** :La précision mesure la proportion de prédictions positives correctes. Elle représente la fiabilité du modèle lorsqu'il prédit une classe positive.

$$\text{Precision} = \frac{TP}{TP + FP}$$

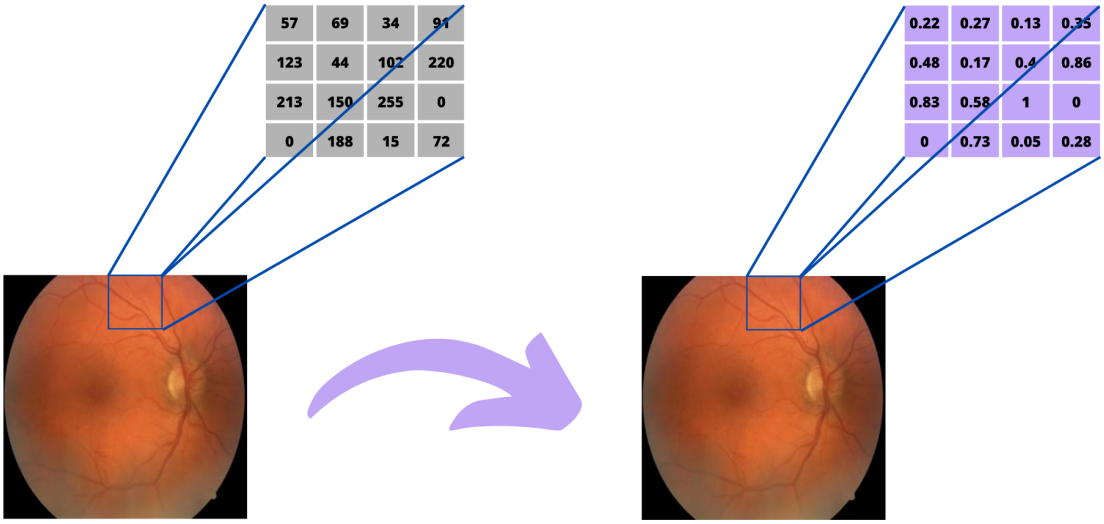


FIGURE 3.4 – Exemple de normalisation [80]

— **AUC–ROC** : aire sous la courbe ROC, mesure la capacité du modèle à séparer classes positives et négatives.

— **Cohen's Kappa** : indique le degré d'accord corrigé du hasard, utile pour évaluer la classification multi-classes ordinales.

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

où p_0 est la probabilité d'accord observé, c'est-à-dire la proportion d'instances pour lesquelles les prédictions du modèle coïncident avec les classes réelles (exactitude globale). p_e est la probabilité d'accord attendu par hasard, calculé à partir des distributions marginales des classes prédites et réelles.

— **Sensibilité (Recall)** : Mesure la proportion d'échantillons positifs correctement classés. Elle Évalue la capacité du modèle à détecter toutes les instances positives du jeu de données.

$$\text{Recall} = \frac{TP}{TP + FN}$$

— **Spécificité** : Mesure la proportion d'échantillons négatifs correctement identifiés. Évalue la capacité du modèle à éviter les faux positifs.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

— **F1-score** : moyenne harmonique de la Précision et le Recall. Il Offre un compromis

équilibré entre fiabilité des prédictions positives et couverture des instances positives.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

L'ensemble de ces métriques résumées dans le table 3.2, vise à quantifier la performance globale des modèles et leur sensibilité aux stades précoces de la RD.

TABLE 3.2 – Résumé des métriques d'évaluation pour les modèles de classification

Métrique	**Formule**	**Description**
Accuracy (Exactitude)	$\frac{TP + TN}{TP + TN + FP + FN}$	Mesure globale de la performance. Pertinente pour des classes équilibrées.
Précision (Precision)	$\frac{TP}{TP + FP}$	Proportion de prédictions positives correctes. Évalue la fiabilité des détections positives.
Rappel (Recall ou Sensibilité)	$\frac{TP}{TP + FN}$	Proportion de classes positives correctement détectées. Mesure la couverture des vrais positifs.
Spécificité	$\frac{TN}{TN + FP}$	Proportion de classes négatives correctement identifiées. Évalue la capacité à éviter les faux positifs.
F1-score	$2 \times \frac{\text{Prcision} \times \text{Rappel}}{\text{Prcision} + \text{Rappel}}$	Moyenne harmonique entre précision et rappel. Utile en cas de déséquilibre entre les classes.
AUC–ROC	—	Aire sous la courbe ROC. Mesure la capacité du modèle à distinguer les classes positives et négatives.
Cohen's Kappa	$\kappa = \frac{p_o - p_e}{1 - p_e}$	Mesure l'accord entre les prédictions et les classes réelles, corrigé du hasard. Pertinent pour la classification multi-classes.

3.3 Techniques d'Ajustement des Hyperparamètres en Deep Learning

Les hyperparamètres, sont des paramètres de configuration qui contrôlent le processus d'apprentissage. Ils peuvent être impliqués dans la construction de la structure du modèle, comme le nombre de couches cachées et la fonction d'activation, ou dans la détermination de l'efficacité et de la précision de l'apprentissage du modèle, comme le taux d'apprentissage ou la taille du lot (Batch size). Leur optimisation est donc cruciale pour obtenir des modèles performants et généralisables. Compte tenu de leur influence sur la précision et la vitesse d'entraînement, ils doivent être soigneusement configurés avant que le processus d'entraînement ne commence [131].

L'ajustement automatique des hyperparamètres consiste à déterminer la combinaison de valeurs qui permet d'optimiser les performances du modèle pour une tâche spécifique. Afin d'atteindre cet objectif, de nombreuses stratégies d'optimisation des hyperparamètres commencent à être proposé dans la littérature [58, 183, 61].

La technique la plus utilisée est l'ajustement manuel qui permet un réglage précis des hyperparamètres en fonction des observations issues de l'expérimentation. Cette approche devient trés couteuse en temps car les réseaux de neurones deviennent de plus en plus profonds et complexes d'ou la nécessité d'automatiser ce processus. Plusieurs techniques ont été proposées pour l'ajustement automatique des hyperparamètres, parmi lesquelles :

3.3.1 Recherche par Grille et Recherche Aléatoire

— **Méthode de la Recherche par Grille (Grid Search)** : La recherche par grille est une méthode de base pour l'optimisation des hyperparamètres. Elle effectue une recherche exhaustive sur l'ensemble d'hyperparamètres spécifié par les utilisateurs [86]. Bien que simple à mettre en œuvre, cette méthode souffre d'une explosion combinatoire lorsque le nombre de paramètres croît. Elle est lente et gourmande en ressources de calcul. Ce qui la rend difficilement applicable aux grands ensembles de données ou lorsque le nombre de paramètres à explorer est élevé (voir Figure 3.5). La recherche par grille reste la méthode la plus utilisée en raison de sa simplicité mathématique.

— **La Recherche Aléatoire (Random Search)** :La recherche aléatoire est une amélioration fondamentale de la recherche par grille [25]. Elle procède par un tirage aléatoire des combinaisons dans l'espace des d'hyperparamètres. Cette méthode est plus efficace que la recherche par grille, car elle permet de découvrir des configurations de performance élevée sans nécessiter une exploration complète. Cependant, elle présente la limite de rester "aveugle", car elle n'utilise pas les informations des performances déjà observées pour guider les sélections futures (voir Figure 3.5).

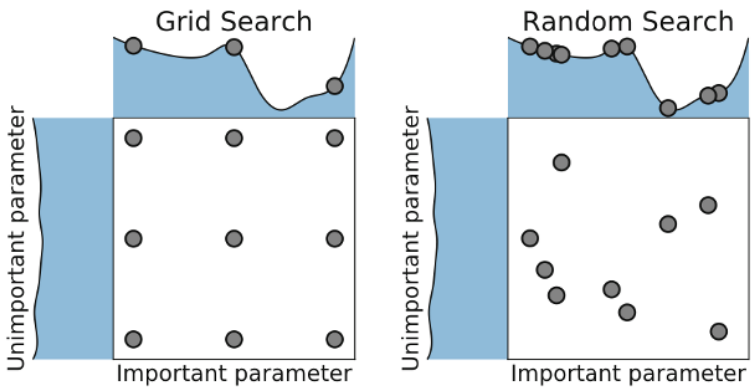


FIGURE 3.5 – Comparaison de la disposition entre la recherche par grille et la recherche aléatoire [25]

.

3.3.2 L'optimisation Bayésienne

L'optimisation bayésienne (OB) est un algorithme développé par Mockus[110], c'est une technique d'optimisation globale séquentielle utilisée pour trouver le maximum ou le minimum d'une fonction complexe, coûteuse à évaluer et dont la structure interne est inconnue. Son principe repose sur deux composantes : (i) un modèle de substitution probabiliste destiné à approximer la fonction objectif, et (ii) une fonction d'acquisition qui guide la sélection des points d'échantillonnage en équilibrant exploration et exploitation [58, 183]. L'exploitation est le processus permettant de prendre la meilleure décision en fonction des informations actuelles, tandis qu'avec l'exploration, le modèle collectera plus d'informations (voir Figure 3.6).

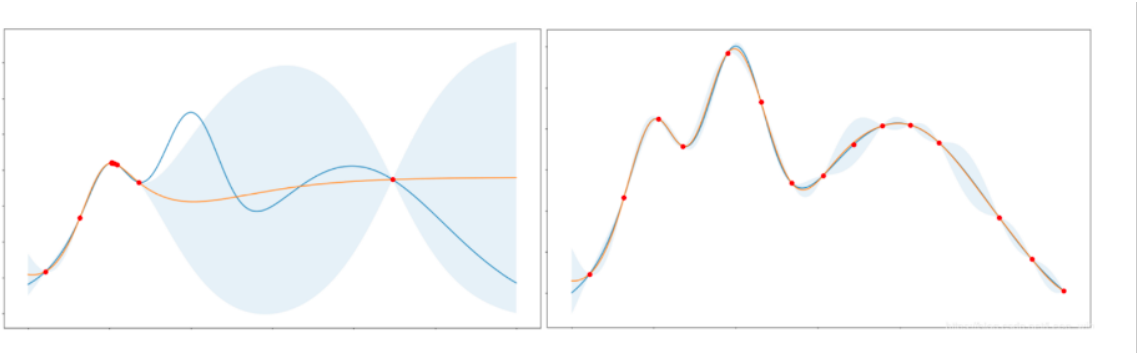


FIGURE 3.6 – Optimisation bayésienne : à gauche l'exploration, et à droite l'exploitation ; l'ombre indique une incertitude [58]

Le processus de l'optimisation bayésienne est itératif, un nouveau point est choisi, à chaque étape en maximisant une fonction d'acquisition, puis en utilisant les observations obtenues, le modèle probabiliste est mis à jour, jusqu'à convergence vers une configuration optimale.

Plusieurs modèles de probabilité peuvent être utilisés pour approximer la fonction ob-

jectif, mais le plus utilisé est le processus gaussien (GP)[183, 61]. Le Gaussian Process (processus gaussien) (GP) par défaut comme modèle de substitution pour les fonctions objectif pour deux raisons principales : la première, pour un modèle non paramétrique, le nombre de paramètres est déterminé par la taille de l'ensemble de données, sans nécessité de le fixer préalablement il s'adapte automatiquement à la quantité et à la nature des observations disponibles, ce qui lui confère une grande flexibilité pour modéliser des fonctions complexes et la deuxième, le GP utilise une distribution gaussienne multivariée comme distribution a priori pour un nombre infini de variables aléatoires à valeurs réelles. Ce cadre bayésien permet non seulement de prédire une valeur pour la fonction objectif en chaque point, mais aussi d'associer à cette prédiction une incertitude bien calibrée [58, 183]. Comparée à la recherche par grille ou à la recherche aléatoire, l'Optimisation Bayesienne (OB) se distingue par son efficacité computationnelle, nécessitant moins d'essais pour identifier des solutions optimales. Elle est particulièrement pertinente pour l'ajustement des hyperparamètres en apprentissage automatique, y compris dans des contextes où la fonction objectif est stochastique, discrète ou non convexe.

3.3.3 Méthodes à Base de Gradients

L'optimisation par gradient [105, 58] est une. méthode d'optimisation basée sur la différentiabilité de la fonction perte par rapport aux hyperparamètres. Cela permet de mettre à jour ces derniers de manière analogue à l'entraînement des poids du réseau. (voir Figure 3.7). Ces methodes permettent d'ajuster dynamiquement certains hyperparamètres pendant l'entraînement. Elles supposent donc que les hyperparamètres sont continus et et différentiables (par ex. le taux d'apprentissage, les coefficients de régularisation), ce qui n'est pas le cas pour tous les hyperparamètres. Calculer les hypergradients nécessite souvent de différencier à travers plusieurs étapes d'optimisation (parfois même sur tout l'entraînement). Cela implique une consommation mémoire importante et des algorithmes sophistiqués pour l'approximation.

3.3.4 Méthodes évolutionnaires

Les méthodes évolutionnaires (telles que les algorithmes génétiques, les stratégies d'évolution et l'optimisation par essaim de particules) reposent sur le principe de la sélection naturelle et de l'évolution biologique, où une population de solutions évolue par sélection, croisement et mutation comme illustré dans la Figure 3.8. De nature stochastique et populationnelle, elles ne nécessitent pas de modèle explicite de la fonction objectif et exploitent la diversité de la population pour favoriser l'exploration, ce qui leur permet d'éviter efficacement les minima locaux [144, 74, 129] Toutefois, leur convergence peut être lente en raison de l'absence de mécanismes statistiques guidant directement l'échantillonnage, et elles requièrent souvent un grand nombre d'évaluations, ce qui augmente leur complexité computationnelle. Elles deviennent néanmoins particulièrement adaptées dans des environnements

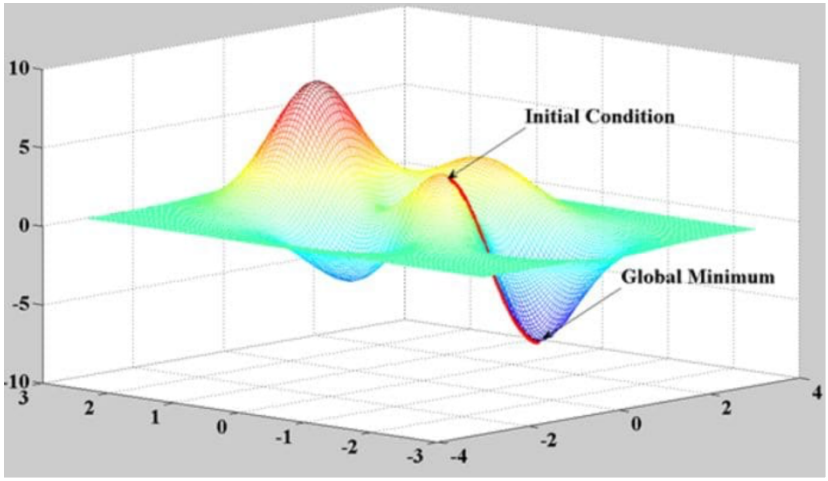


FIGURE 3.7 – Méthodes à base de gradient : chemin suivi par l'optimiseur pour atteindre le minimum global [39]

.

massivement parallélisés (clusters, TPUs) et se montrent robustes face à des espaces de recherche discrets, bruités ou non différentiables [179, 61]. Ces approches incluent notamment la neuroévolution [152], qui fait évoluer conjointement architectures et poids des réseaux , ainsi que la Neural Architecture Search (NAS) [129], qui vise à découvrir automatiquement des architectures performantes, en particulier dans des contextes nécessitant une exploration radicale de l'espace de recherche, comme la génération de topologies inédites.

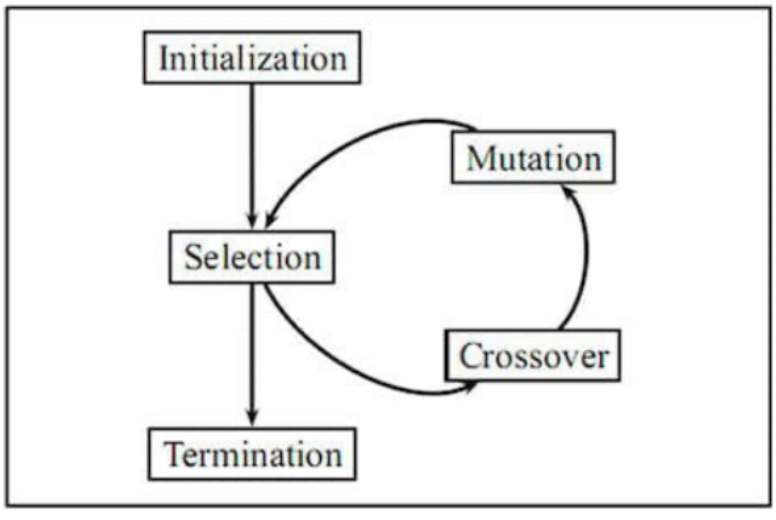


FIGURE 3.8 – Processus de l'optimisation evolutionaire [39]

.

3.3.5 Conclusion

Comme l'illustre le tableau 3.3, l'optimisation bayésienne se distingue des autres approches par sa capacité à modéliser probabilistiquement l'espace des hyperparamètres et à guider la recherche au moyen d'une fonction d'acquisition équilibrant exploration et exploi-

tation. Contrairement aux méthodes locales basées sur le gradient ou aux recherches par grill et aléatoires, elle permet d'éviter les minima locaux tout en réduisant le coût computationnel. Dans le contexte du deep learning appliqué à la classification médicale, cette approche offre ainsi un compromis optimal entre rigueur, efficacité et performance, en la plaçant parmi les méthodes les plus pertinentes pour l'optimisation.

TABLE 3.3 – Comparaison des principales méthodes de recherche d'hyperparamètres

Méthode	Principe	Avantages	Limites	Cas d'usage typiques
Grid Search	Exploration exhaustive d'une grille discrète prédéfinie d'hyperparamètres.	Simple, systématique, facilement parallélisable.	Explosion combinatoire, peu efficace dans les grands espaces.	Espaces de recherche petits et bien bornés.
Random Search	Tirage aléatoire de combinaisons d'hyperparamètres dans un espace donné.	Plus efficace que Grid Search (explore davantage de configurations).	Ne profite pas d'un retour statistique, inefficace si le nombre d'évaluations est limité.	Espaces larges, recherche exploratoire initiale.
Optimisation bayésienne	Modélisation probabiliste de la fonction objectif (souvent via GP) + fonction d'acquisition.	Échantillonnage intelligent, équilibre exploration/exploitation, nécessite moins d'évaluations.	Moins adaptée aux espaces très hauts dimensionnels ou discrets ; coût de mise à jour du modèle.	Hyperparamètres coûteux à évaluer (ex. deep learning, simulations lourdes).
Optimisation par gradient	Ajustement continu des hyperparamètres différentiables en suivant le gradient (hypergradient descent).	Rapide, efficace pour paramètres différentiables, bonne convergence locale.	Nécessite dérivabilité, sensible aux minima locaux, non applicable aux espaces discrets.	Réglage fin de taux d'apprentissage, régularisation continue.
Méthodes évolutionnaires	Algorithmes inspirés de l'évolution (sélection, mutation, croisement) appliqués aux hyperparamètres ou architectures.	Très exploratoires, robustes aux espaces non convexes, bruités, discrets ou non différentiables.	Convergence lente, coûteux en calcul (beaucoup d'évaluations).	NAS (Neural Architecture Search), problèmes combinatoires complexes.

3.4 Travaux Connexes

L'évolution des méthodes de détection automatique de la RD suit une progression naturelle : des approches classiques fondées sur l'apprentissage automatique, aux réseaux de neurones convolutifs issus de l'apprentissage profond, jusqu'aux architectures récentes basées sur les Vision Transformers. La présente revue de littérature examine successivement l'ensemble de ces contributions.

3.4.1 Algorithmes Classiques d'Apprentissage Automatique

Avant l'avènement des approches profondes, la classification de la RD s'appuyait principalement sur des techniques d'apprentissage automatique dites classiques dont le principe est représenté dans la Figure 3.9.

Ces méthodes sont bien répertoriées et bien détaillées par Ishtiaq et al. [82], incluent notamment le Support Vector Machine (SVM) [116, 147, 151, 153, 36, 3], Random Forest

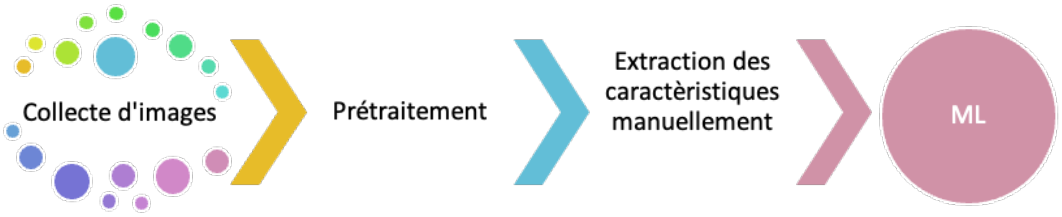


FIGURE 3.9 – Processus ML pour la classification de RD

(RF) [16, 177, 169, 43], k-plus proches voisins (kNN [170, 169, 118, 127], Naïve Bayes (NB)[171, 127], les arbres de décision (DT) [127] ; ainsi que certaines approches non supervisées [30, 185]. Elles ont été utilisées pour la détection et la classification de lésions spécifiques (microanévrismes, hémorragies, exsudats), à partir de caractéristiques extraites manuellement ou semi-automatiquement.

Ces approches présentent toutefois un certain nombre de limitations structurelles :

— Elles nécessitent une étape préalable d'extraction de caractéristiques, souvent fondée sur l'expertise humaine, ce qui limite leur capacité de généralisation.

— Les performances des modèles dépendent fortement de la qualité et de la pertinence des caractéristiques extraites ; or, dans le cas de la RD, les signes pathologiques (microanévrismes, exsudats) sont souvent subtils et de faible contraste, rendant difficile une extraction fiable.

— Ces méthodes sont extrêmement sensibles aux variations inter-patients, aux conditions d'acquisition (luminosité, bruit, artefacts) et aux déséquilibres de classes, fréquemment rencontrés dans les bases de données ophtalmologiques.

Au-delà de ces limitations spécifiques, ces techniques se révèlent également insuffisantes pour modéliser la complexité visuelle des images médicales. Leur incapacité à extraire automatiquement des représentations discrimantes à partir d'images, rend leur usage peu adapté aux contextes réels de dépistage massif ou de diagnostic assisté par ordinateur.

Ces constats ont conduit à l'émergence d'une nouvelle génération d'approches basées sur l'apprentissage profond, et en particulier sur les CNN, adaptés à la classification d'images médicales, en exploitant leur capacité à apprendre des représentations hiérarchiques complexes à partir des données visuelles.

3.4.2 Deep learning pour la Classification de la RD

Avec l'émergence du deep learning, les architectures de CNN ont progressivement supplanté les méthodes classiques fondées sur la segmentation manuelle. Ces architectures sont capables d'apprendre automatiquement des représentations discriminantes à partir d'images brutes, supprimant ainsi la nécessité d'une extraction manuelle explicite de caractéristiques (Figure 3.10).

Face aux limitations des méthodes traditionnelles, l'apprentissage profond (Deep Learning, DL) s'impose comme une alternative puissante, permettant de substituer aux étapes

FIGURE 3.10 – Processus d'apprentissage profond pour la classification de la RD

manuelles d'extraction des caractéristiques, des processus d'apprentissage completement automatique. Les architectures CNN traitent directement des images prétraitées (amélioration du contraste, correction de l'illumination, redimensionnement) afin d'atténuer le bruit et les artefacts, puis extraient, de manière hiérarchique, des descripteurs optimaux tout en apprenant simultanément les règles de classification.

L'efficacité de ces approches a été démontrée dans plusieurs études cliniques, notamment dans le cadre de programmes de dépistage à grande échelle, soulignant leur potentiel dans un contexte médical réel. De nombreuses revues de la littérature [114, 143, 148] ont synthétisé les avancées majeures de l'application du deep learning à la détection automatisée de la rétinopathie diabétique.

Les réseaux de neurones artificiels ont d'abord été utilisés pour établir un pronostic de la rétinopathie diabétique (RD) à partir de l'analyse des lésions caractéristiques de la pathologie. L'avènement des réseaux de neurones convolutifs (CNN) a ensuite permis d'importants progrès, en particulier dans la segmentation des images du fond d'œil, en ciblant des anomalies spécifiques telles que les microanévrismes [53, 138], les hémorragies [164, 146], ou encore les exsudats cotonneux [107, 182]. Parallèlement, le deep learning a été largement exploité pour la classification de la RD [20, 81]. Les méthodes proposées diffèrent selon plusieurs critères : le jeu de données utilisé,le type de classification abordé (binaie ou multi-classes), ainsi que la complexité et la profondeur des architectures mises en œuvre [115, 143, 29].

Cependant, malgré leurs performances remarquables, les CNN présentent plusieurs limitations.

— D'une part, leur capacité à généraliser dépend fortement de la disponibilité de bases de données de grande taille, rigoureusement annotées, Or, en ophtalmologie, l'annotation experte des images demeure une tâche complexe, coûteuse et chronophage.

— D'autre part, les modèles CNN sont souvent sensibles aux variations de distribution des données (biais de domaine, hétérogénéité des conditions d'acquisition, diversité inter-patients), ce qui peut compromettre leur robustesse en contexte réel.

— Enfin, l'entraînement d'un modèle CNN à partir de zéro nécessite des ressources computationnelles importantes et un temps d'apprentissage considérable, ce qui limite leur adoption dans les environnements cliniques à ressources limitées.

Dans ce contexte, l'apprentissage par transfert (TL) apparaît comme une alternative particulièrement pertinente. Cette approche consiste à exploiter des modèles pré-entraînés sur

de vastes dataset d'images génériques (comme ImageNet), puis à les adapter à des tâches spécifiques comme la classification de la RD. En tirant parti des représentations visuelles déjà apprises (textures, formes, contrastes), le TL permet non seulement de réduire considérablement le besoin en données annotées, mais aussi d'accélérer la convergence des modèles, et de limiter le surapprentissage (overfitting).

Le recours à l'apprentissage par TL constitue ainsi une réponse directe aux limites des CNN entraînés à partir de zéro, en assurant une meilleure robustesse, une efficacité computationnelle accrue, et une adaptabilité renforcée aux spécificités des images médicales. Ces avantages justifient pleinement son adoption dans les systèmes d'aide au diagnostic pour la détection automatisée de la rétinopathie diabétique. Conformément aux standards cliniques, l'analyse de la rétinopathie diabétique repose typiquement sur cinq étapes clés : (i) le dépistage des patients à risque, (ii) la segmentation des vaisseaux rétiniens, (iii) la détection des lésions caractéristiques, (iv) la classification du stade de la maladie, et (v) la validation du diagnostic assisté. Le développement de systèmes d'aide au diagnostic fondés sur le transfert learning, capables d'adresser efficacement chacune de ces étapes, apparaît dès lors comme une priorité pour améliorer la prise en charge précoce de la RD.

3.4.3 Apprentissage par Transfert pour la classification de la RD

L'apprentissage par TL s'est imposé comme une approche particulièrement efficace pour la classification des images de RD, en permettant l'adaptation de modèles pré-entraînés à des jeux de données de petite taille, tout en maintenant des performances élevées [167, 29, 23, 91]. Les bases de données publiques d'imagerie rétinienne, telles qu'IDRiD [125], EyePACS [44], Messidor [47] et APTOS [90], ont joué un rôle essentiel dans l'entraînement et l'évaluation de ces modèles, en fournissant des ensembles annotés indispensables à la détection et à la classification de la RD [114, 143, 148]. Parmi ces jeux de données, APTOS 2019 présente des avantages notables qui justifient son utilisation dans cette étude :

— Le dataset comprend un grand nombre d'images (plus de 3 600) issues de conditions cliniques réelles, capturées dans des centres ophtalmologiques indiens, ce qui en fait un jeu de données représentatif des situations de dépistage à large échelle.

— Les images sont annotées en cinq niveaux de gravité, respectant ainsi les standards cliniques internationaux.

— Le dataset présente une hétérogénéité importante en termes de qualité d'image (présence d'artefacts, flou, sous/surexposition), ce qui constitue un défi réaliste pour les modèles de deep learning.

— Enfin, APTOS est librement accessible, bien documenté, et largement utilisé dans les travaux récents, ce qui en facilite la reproductibilité et la comparaison avec l'état de l'art.

C'est pourquoi, dans cette section, nous nous concentrons spécifiquement sur les avancées récentes du TL appliquées à la classification de la RD, en mettant l'accent sur les études ayant utilisé le jeu de données APTOS. Cette mise en perspective vise à positionner notre propre méthode dans le contexte des approches actuelles de l'apprentissage profond appliqué à la RD.

Kassani et al. [92] ont proposé un modèle basé sur l'architecture Xception, entraîné sur les images du jeu de données APTOS, pour une classification en cinq niveaux de gravité de la rétinopathie diabétique. Afin de remédier au déséquilibre des classes, les auteurs ont appliqué des techniques de régression Lasso et Ridge. Cette approche a permis d'atteindre une précision de 83%. Bien que le modè améliore la précision de classification, le problème réside dans la complexité accrue du modèle liée à l'agrégation des couches profondes

D'autre part, Dekhil et al. [48] ont proposé un modèle personnalisé de CNN fondé sur l'apprentissage par transfert pour une classification en cinq classes. Le modèle repose sur un prétraitement, suivi de l'utilisation de VGG16 et de couches entièrement connectées. En réentraînant l'ensemble des couches du modèle pré-entraîné, les auteurs ont obtenu une accuracy de validation de 77%.

Par ailleurs, Rao et al. [128] ont proposé une étude expérimentale comparativecomparative de cinq classificateurs CNN, à savoir Inception-V3, VGG19, VGG16, ResNet50 et InceptionResNetV2. Le modèle ResNet50 a obtenu la meilleure précision (95,59%) pour une classification binaire. En revanche, InceptionResNetV2 a donné les meilleurs résultats pour la classification multi-classes, atteignant une accuracy de 88,14% pour une classification en trois classes, et 85% pour une classification en cinq classes.

De même, Bodapati et al. [32] ont présenté un travail basée sur la fusion de caractéristiques multi-modales (Blended multi-modal deep convnet features). Leur méthode combine des caractéristiques extraites de plusieurs CNN préentraînés (VGG16, ResNet, et Inception-v3). Ces caractéristiques alimentent deux réseaux profonds. Le premier pour la classification binaire ; ce modèle a atteint une accuracy de 97,41% et un score de kappa de 94,82%. Le second pour la classification en 5 classes a permis d'atteindre une accuracy de 81,7% et un score de kappa de 71,1%.

De leur côté, Karki et Kulkarni [89] ont présenté une approche par apprentissage profond pour classifier la sévérité de la RD en cinq niveaux. Ils ont mené des expérimentations à l'aide de différentes variantes du modèle EfficientNet. Le modèle le plus performant, entraîné sur APTOS, a atteint un score kappa quadratique de 92,43%.

De plus, Sugeno et al.[154] ont développé un système pour la classification de la sévérité de la RD. Le système a été développé sur la base de l'apprentissage par transfert avec l'architecture EfficientNet-B3. Ils ont entraîné et testé leur modèle sur le jeu de données APTOS, où ils ont obtenu une précision de 84,2% pour la tâche de classification de la sévérité de la RD.

Une hybridation entre VGG16 et un réseau de capsules pour la classification de la RD,

dans un modèle appelé DRISTI (Diabetic Retinopathy classification by analySing reTinal pictures) a été proposé par Kumar et al. [98]. Pour une classification binaire, le modèle a atteint une accuracy de 96,24% ; alors que pour une classification en cinq classes, l'accuracy est de 82,06%. Le modèle a été entraîné sur APTOS et validé sur plusieurs bases de données publiques. Toutefois, la performance du modèle est affectée par une distribution déséquilibrée des classes dans le jeu de données utilisé, ce qui constitue une limitation importante.

Gangwar et Rav [63] quant eux, ont proposé un modèle hybride combinant un bloc CNN personnalisé au modèle Inception-ResNet-v2 pré-entraîné. Les modèles ont été entraînés sur les jeux de données Messidor-1 et APTOS 2019, atteignant respectivement des précisions de 72,33% et 82,18%.

Une hybridation entre un CNN et un réseau pré-entrainé a été développé par Wejdan et al. [13]. Le modèle utilise le Le CNN pour classifier RD selon les cinq stades de sévérité, et YOLOv3 pour détecter et localiser les lésions associées à la RD. Leur système a atteint une précision de 89,6%, en s'appuyant sur les bases de données DDR et APTOS.

Par ailleurs, Hu et al. [77] ont proposé une approche innovante combinant apprentissage profond et graphes pour la détection de la RD. Leur modèle repose sur un entraînement auto-supervisé, exploite un réseau de neurones à graphes (Graph Neural Network, GNN) pour extraire des caractéristiques latentes pertinentes, et intègre un mécanisme d'entraînement adversaire afin d'améliorer la robustesse aux variations inter-individuelles. Testée sur le jeu de données APTOS 2019, la méthode a atteint une accuracy de 83,5% pour la classification multi-classes et de 94,3% pour la classification binaire. Toutefois, cette approche présente certaines limitations, notamment en raison de la petite taille des lésions et de leur forte similarité visuelle, ce qui entrave la classification fine des stades de la RD. En plus de la complexité du mod, la transformation de l'image en graphe peut mener à une perte d'information.

Islam et al. [83] ont proposé une approche d'apprentissage contrastif supervisé (Supervised Contrastive Learning, SCL) pour la classification de la RD. Le modèle Xception pré-entraîné a été utilisé comme encodeur, dans le cadre d'un apprentissage par transfert. En amont, une étape de prétraitement a été appliquée à l'aide de la méthode CLAHE (Contrast Limited Adaptive Histogram Equalization), afin d'améliorer la qualité des images du fond d'œil. L'approche a été validée sur les jeux de données APTOS et Messidor-2. Les résultats obtenus sont prometteurs, avec une précision de 98,36% pour la classification binaire et de 84,36% pour la classification multi-classe.

Oulhadj et al.[120] ont proposé une méthode automatique de classification de la rétinopathie diabétique en cinq stades, combinant un recalage déformable pour supprimer le fond à une fusion de caractéristiques extraites de cinq modèles préentraînés (Xception, InceptionV3, VGG16, DenseNet121, ResNet50). Leur approche a atteint une précision de 85,28% et un score kappa de 77,78%. Dans une étude complémentaire[121], ils ont comparé cinq approches de transfert learning appliquées à la même tâche, en introduisant un vote majoritaire entre

les prédictions des modèles. Cette stratégie d'ensemble a permis d'améliorer la précision par rapport à chaque modèle pris individuellement.

Enfin, Shakibania et al. [142] ont utilisé l'apprentissage par transfert en combinant deux modèles préentraînés (ResNet50 et EfficientNetB0) en tant qu'extracteurs de caractéristiques. Le modèle a été entraîné sur plusieurs jeux de données, y compris APTOS 2019, et a atteint une précision de 98,50% pour la classification binaire, et 89,60% pour la classification par stade. L'optimisation bayésienne a été utilisée pour ajuster uniquement le taux d'apprentissage et le momentum.

Dans la continuité des avancées en apprentissage profond, plusieurs travaux récents ont introduit des modèles hybrides qui combinent les bénéfices du transfert learning (TL) avec ceux des mécanismes d'attention, afin d'améliorer la classification des stades de la rétinopathie diabétique (RD). Par exemple, Fan et al.[56] ont classifié la RD en utilisant la fusion de caractéristiques multi-échelles avec une opération de pondération adaptative, où les caractéristiques extraites avec différentes couches de convolution étaient ensuite fusionnées par une opération de pondération adaptative. Ils ont obtenu une précision de 85,32% sur le jeu de données APTOS. La méthode proposée présente un inconvénient majeur lié à des résultats inexplicables obtenus à partir de caractéristiques non pertinentes fournies par le modèle d'attention.

Bodapati et al. ont exploré des modèles hybrides combinant réseaux de neurones profonds (DNN) et mécanismes d'attention dirigée (gated attention) pour la classification de la rétinopathie diabétique. Dans une première étude [33], ils ont proposé une architecture composite enrichie par une couche d'attention, et intégrant le transfert learning. Ce modèle, évalué sur le jeu de données APTOS-2019, a atteint une accuracy de 82,54%, confirmant la pertinence de l'intégration de l'intention aux architectures basées sur le TL . Dans une étude ultérieure [31], ils ont étendu leur approche en introduisant une architecture empilée (stacked) de DNNs attentifs, permettant une meilleure modélisation des relations complexes entre les caractéristiques extraites. Cette version améliorée, également testée sur le même jeu de données, a conservé des performances compétitives, à savoir, une accuracy de 86,06% .

Par ailleurs, Shaik et Cherukuri [141] ont présenté une méthode de classification des niveaux de sévérité de la rétinopathie diabétique reposant sur une architecture attentionnelle en plusieurs étapes, nommée Hinge Attention Networks (HA-Net). Leur modèle a démontré de bonnes performances, atteignant une précision de 85,54% sur le jeu de données APTOS et de 66,41% sur le jeu de données IDRiD.

Enfin, Al-Antary et Arafa[9] ont proposé un réseau à attention multi-échelle s'appuyant sur un encodeur préentraîné, chargé de projeter les images rétiniennes dans un espace de représentation de haut niveau. Cette représentation est enrichie via une pyramide de caractéristiques multi-échelles, suivie d'un mécanisme d'attention permettant de renforcer le pouvoir discriminant du modèle. En complément, une tâche auxiliaire de classification binaire (sain vs non-sain) est intégrée à l'entraînement pour affiner la capacité du modèle à détecter les

images pathologiques. Les performances obtenues sont prometteuses, avec une précision de 84,6% sur le jeu APTOS et de 79,9% sur EyePACS.

Alyoubi et al. [14] ont rapporté que près de 73% des études se sont concentrées sur la classification binaire de la RD, tandis que seulement 27% ont tenté une classification plus fine en plusieurs stades. En effet, la classification en cinq niveaux reste difficile, notamment en raison des caractéristiques subtiles des formes légères à modérées, comme la présence de microanévrysmes, difficiles à détecter de manière fiable.

La classification automatique de la rétinopathie diabétique repose majoritairement sur le transfert de modèles préentraînés, ajustés par fine-tuning pour s'adapter aux spécificités des bases médicales. Toutefois, la performance de ces modèles dépend fortement du choix des hyperparamètres (taux d'apprentissage, taille du batch, nombre d'epochs, etc.), dont la sélection est souvent empirique. Or, peu d'études ont abordé l'optimisation conjointe de ces paramètres en complément du fine-tuning (voir Tableau 3.4).

Les architectures fondées sur l'apprentissage par transfert ont démontré une efficacité remarquable pour l'extraction de caractéristiques locales discriminantes, surpassant ainsi les méthodes traditionnelles fondées sur des descripteurs manuels. Toutefois, en raison de leur champ réceptif intrinsèquement restreint, ces modèles convolutionnels peinent à capturer les dépendances globales à l'échelle de l'image, limitant ainsi leur capacité à appréhender pleinement la complexité structurelle des lésions rétiniennes caractéristiques de la rétinopathie diabétique.

Pour surmonter cette limitation, les Vision Transformers (ViT) ont récemment émergé comme une alternative prometteuse. En tirant parti de mécanismes d'attention globale, ces architectures sont capables de modéliser explicitement les relations à longue portée entre les différentes régions d'une image, ce qui les rend particulièrement adaptées à l'analyse fine et contextualisée des images de fond d'œil.

3.4.4 Vision Transformer pour la classification de la RD

Inspirés par le succès des transformeurs dans le traitement du langage naturel [165], Dosovitskiy et al. [52] ont introduit le Vision Transformer (ViT) pour la classification d'images, marquant une rupture majeure avec les architectures convolutionnelles traditionnelles. En représentant les images sous forme de séquences de patches et en exploitant le mécanisme d'auto-attention, les ViTs ont montré des performances remarquables dans la modélisation des dépendances globales, surpassant les CNN traditionnels dans plusieurs tâches de vision par ordinateur.

Récemment, les capacités remarquables de représentation des transformers ont suscité un intérêt croissant dans le domaine de l'analyse d'images médicales [40, 172, 54]. Bien que leur potentiel soit prometteur, l'application des ViTs à la classification de la rétinopathie diabétique reste relativement récente, et les les études spécifiques à cette pathologie demeurent limitées. Pour la classification de la RD, Wu et al. [176] ont utilisé des ViTs pour démontrer

leur performance supérieure par rapport aux réseaux convolutifs.

Par ailleurs, Mohan et al. [111] ont montré que la division des images du fond d'œil en patch non chevauchantes préserve les informations relatives à la position spatiale de chaque patch.

Concernant spécifiquement la RD, Wu et al.[176] ont montré que les Vision Transformers surpassent les CNN en termes de performance de classification, tandis que Mohan et al.[111] ont exploré leur usage dans la segmentation des lésions rétiniennes, en les combinant à des réseaux neuronaux convolutionnels.

Différents jeux de données ont été utilisés pour évaluer l'efficacité des ViTs dans la classification de la RD. Citons, Nazih et al. [117] qui proposent une architecture basée sur les ViTs pour identifier les stades de sévérité de la RD. Les ViTs nécessitant de grands volumes de données pour un apprentissage efficace, ils ont utilisé le jeu de données FGADR (Fine-Grained Annotated Diabetic Retinopathy), comprenant 1 842 images du fond d'œil, pour entraîner leur modèle. atteignant des résultats encourageants avec une accuracy, un rappel et un F1-score de 82,5% chacun.

De leur côté, Gu et al. [69], ont classifié la RD à l'aide d'un ViT sur le jeu de données DDR, obtenant 82,45% de spécificité, 81,40% de sensibilité et 82,35% de précision.

Khan et al. [93] ont proposé un modèle basé sur le Compact Convolutional Transformer (CCT), combinant des couches convolutives avec des mécanismes de transformer. Entraîné sur un vaste ensemble de données fusionnant cinq bases (APTOS, IDRiD, Messidor-2, DDR, Kaggle DR), leur modèle a atteint une accuracy de 84,5%, surpassant à la fois le ViT standard (81,56%) et le Swin Transformer (82,23%).

Dans une perspective comparative, Karkera et al. [88] ont comparé différentes architectures de ViTs, notamment ViT, DeiT, CaiT et BEiT, entraînées sur la base DBtr, ont montré que la combinaison de ces modèles permet d'atteindre une précision de 94,63%, surpassant les performances individuelles de chaque modèle.

Plus récemment, Oulhadj et al. [122] ont proposé une architecture hybride combinant un Vision Transformer ajusté par fine-tuning à un réseau de capsules pour la prédiction automatique de la sévérité de la RD. Le modèle utilise les architecture en parallèle. Évaluée sur quatre bases (APTOS, Messidor-2, DDR et EyePACS), cette méthode a obtenu sa meilleure performance sur le jeu de données APTOS avec une précision de 88,18%.

Alors que, Lian et Liu [102] ont combiné un réseau de neurones convolutif (Inception-ResNet-v2) avec un vision transformer. Leur modèle hybride a atteint une précision de 93,2% en classification binaire sur Messidor-1 et de 89,1% en classification à cinq stades sur le jeu de données APTOS.

Par contre, Yang et al. [178] ont développé un modèle de Transformer basé sur l'apprentissage multi-instances (Multiple Instance Learning, MIL) pour classer la rétinopathie diabétique (RD). Leur approche découpe les images rétiniennes haute résolution en patches de 224 × 224 pixels, traités ensuite par un Vision Transformer (ViT) pour extraire des carac-

téristiques locales. Un module de calcul global (Global Instance Computing Block, GICB) agrège les informations de ces patches, améliorant la capacité du modèle à comprendre les relations contextuelles au sein de l'image. Le modèle a atteint une accuracy de 93,2% en classification binaire sur le jeu de données Messidor-1 et de 85,65% en classification à cinq stades sur APTOS, surpassant le modèle MIL-ViT proposé par Yu et al. [181]. Dihin et al.[51] use a combination of Wavelet and multi-Wavelet transforms with the Swin Transformer model. The study highlights the innovative use of the multi-Wavelet transform for feature extraction, integrated into the Swin Transformer. The model obtained 96% accuracy for binary classification on the Kaggle APTOS 2019 dataset. The Swin-T model with multi-Wavelet transformation achieved a 98% recall and 96% F1 score for binary classification. However, the model's accuracy decreased in multiclass classification (82%). Le tableau 3.5 résume l'ensemble de ces travaux.

TABLE 3.4 – Synthèse des travaux de classification de la RD basés sur le TL

Authors	Year	Classes	Method	Performance	Remarks
Kassani et al.[92]	2019	5	Xception Modifié avec régularisation Lasso et Ridge	Accuracy :83.06% Sensitivity :88.24% Specificity :87%	Déséquilibre des classes, Hyperparamètres initialisés
Dekhil et al.[48]	2019	5	Transfert learning avec VGG16 et CNN personnalisé	Accuracy : 77% Sensitivity :% Specificity :%	Pas d'augmentation des données
Rao et al.[128]	2020	2,3 et 5	Comparaison entre plusieurs CNN (VGG, ResNet, etc.)	accurracy 2 : ResNet50 : 96,59%, 3 : InceptionResNetV2 : 88.14% 5 : InceptionResNetV2 : 85,02%	Hyperparamètres initialisés
Bodapati et al. [32]	2021	2 et 5	Fusion multi-CNN + DNN pour la classification	accurracy 2 : 97,41%, 5 : 81,7%	Complexité et surapprentissage, Pas d'augmentation des données, Hyperparamètres initialisés

suite du tableau					
Authors	**Year**	**Classes**	**Method**	**Performance**	**Remarks**
Kumar et al.[98]	2021	2 et 5	VGG16 combiné avec réseaux capsule (DRISTI)	accurracy 2 : 96,24% (binaire), 5 : 82,06% (5 classes)	Complexité de l'architecture, Leur méthode était centrée sur la distribution de classe déséquilibrée du jeu de données utilisé, Hyperparamètres initialisés
Karki et Kulkarni [89]	2021	5	Transfert learning avec EfficientNet	Kappa : 92,43%	pas de valeurs des autres métriques, Pas de matrice de confusion
Gangwar et Rav [63]	2021	5	CNN personnalisé + InceptionResNet-v2 sur deux jeux de données	Accuracy : 72,33% (Messidor), Accuracy : 82,18% (APTOS)	Hyperparamètres initialisés
Al AYoubi et al.[13]	2021	5	Classification + détection avec YOLOv3	Accuracy : 89,6%	modèle complexe, Hyperparamètres initialisés
Fan et al. [56]	2021	5	Fusion adaptative multi-échelle pondérée	Accuracy : 85,32%	Interprétabilité faible Learning rate initialisé
Sugeno et al. [154]	2021	5	EfficientNet-B3	Accuracy : 84,2%	Une limitation majeure du modèle proposé réside dans ses performances insuffisantes pour les niveaux de sévérité de la rétinopathie diabétique

suite du tableau

Authors	Year	Classes	Method	Performance	Remarks
Bodapati et al. [33]	2021		DNN avec attention spatiale dirigée (gated attention)	Accuracy : 82,54% , Kappa : 97%	Une limitation majeure du modèle proposé réside dans ses performances insuffisantes pour les niveaux de sévérité de la rétinopathie diabétique. Pas d'augmentation des données
Bodapati et al.[31]	2022	5	Autoencodeurs empilés + attention spatiale	APTOS : 84,17%, IDRiD : 63,24%	Abscence de prétraitement, L'utilisation d'un autoencodeur incomplet n'est pas idéale pour des tâches nécessitant une localisation précise des lésions, Hyperparametres initialisés.
Hu et al.[77]	2022	2 et 5	GNN + apprentissage adversarial auto-supervisé	Binaire : 94,3%, 5 classes : 83,5%	Faible détection des petites lésions, L'inconvénient de leur méthode se réfère au problème de la petite taille et de la similarité des lésions DR, ce qui rend leur méthode incapable de prendre en compte la classification fine des lésions.
Shaik et Cherukuri[141]	2022	5	HA-Net avec attention hiérarchique (Hinge Attention)	APTOS : 85,54%, IDRiD : 66,41%	Complexité architecturale élevée, Hyperparametres initialisés

suite du tableau

Authors	Year	Classes	Method	Performance	Remarks
Islam et al. [83]	2022	2 et 5	Xception + CLAHE + supervised contrastive learning	Accuracy : 98,36% (binaire), 84,36% (multi)	Seul l'hyperparamètre marge a été ajusté avec des valeurs comprises entre 0.1 et 0.9, Modèle complexe necessitant des batch de grandes tailles, Difficultés à reconnaître les classes rares.
Oulhadj et al. [120]	2022	5	Ensemble de 5 modèles + recalage déformable	Accuracy : 85,28%, Kappa : 77,78%	Une déformation excessive ou mal contrôlée peut altérer la forme ou la taille réelle des lésions, ce qui peut induire en erreur le modèle de classification ou rendre l'image biologiquement incohérente.
Athira T R et al.[22]	2023	5	Auto-ajustement RMSprop + exploration de 5 modèles	Accuracy : 99,8% (binaire), 94,7% (3 classes)	Pas de test en 5 classes, Performances non spécifiés sur la detection des différents statdes de la RD

suite du tableau					
Authors	**Year**	**Classes**	**Method**	**Performance**	**Remarks**
Oulhadj et al. [121]	2023	5	Ensemble de 5 modèles + vote	Accuracy : 85,28%, Kappa : 77,78%	difficulté de l'enregistrement des images déformées dans le cas d'images bruitées, Un grand nombre de paramètres d'apprentissage pour le vote d'ensemble
Shakibania et al.[142]	2024	2 et 5	Fusion ResNet50 + EfficientNetB0 avec BO	98,50% (binaire), 89,60% (multi)	Optimisation partielle des hyperparamètres, ajustement du learning rate et du momentum

TABLE 3.5 – Résumé des travaux de classification de la RD utilisant les Vision Transformers (ViTs))

Authors	Year	Classes	Method	Performance	Remarks
Wu et al. [176]	2021	5	DUtilisation directe de ViT pour démontrer sa supériorité par rapport aux CNN classiques	Non specifiées	L'architecture hiérarchique nécessite deux phases d'entraînement distinctes. Cela complique la mise en œuvre et augmente le temps de développement et de validation. Hyperparametres initialisés
Nazih et al. [117]	2023	5	ViT entrainé sur le dataset FGADR	F1, Precision, Recall : 82.5%	Plusieurs experimentations ont été effectuées pour choisir les meilleurs hyperparametres et le meilleur Vit (base ou Large)
Gu et al. [69]	2023	5	ViT entrainé sur le dataset DDR	Specificity : 82.45%, Sensitivity : 81.40%, Precision : 82.35%	Performances faible dans le cas de détection des classes précoces de la RD Hyperparametres initialisés

Continued from previous page

Authors	Year	Classes	Method	Performance	Remarks
Khan et al. [93]	2023	5	Compact Convolutional Transformer CCT (CNN + ViT)	Accuracy : 84.5%, ViT : 81.56%, Swin : 82.23%	Le modèle est entraîné sur un jeu de données non équilibré, cela peut biaiser l'apprentissage et entraîner une dégradation des performances sur les classes rares. Plusieurs experimentations ont été effectuées pour choisir les meilleurs hyperparametres
Karkera et al. [88]	2024	5	Fusion de multiple pré-entrainé ViTs (ViT, DeiT, CaiT, BEiT) sur le dataset DBtr	Accuracy : 94.63% (fusion)	Pas d'augmentation des données Hyperparametres initialisés
Oulhadj et al. [122]	2024	5	Hybrid ViT + Capsule Network sur APTOS dataset	APTOS : 88.18%	Sensibilité au bruit et la complexité computationnelle liée aux capsules Hyperparametres initialisés
Lian & Liu [102]	2024	2 and 5	Inception-ResNet-v2 + ViT	Messidor-1 : 93.2% (2 classes), APTOS : 89.1%	Inception-ResNet-v2 cette architecture hybride est très gourmande en mémoire GPU et en temps de calcul Hyperparametres initialisés

3.5 Conclusion

Les avancées récentes en apprentissage profond, notamment à travers l'utilisation de l'apprentissage par transfert et des Vision Transformers (ViTs), ont considérablement amélioré les performances des modèles de classification de la rétinopathie diabétique.

Les modèles pré-entraînés, en particulier les CNN profonds et les ViTs, ont montré leur capacité à extraire automatiquement des caractéristiques discriminantes à partir d'images du fond d'œil, même à partir de bases de données limitées. En particulier, les ViTs ont offert une capacité accrue à capturer les dépendances globales au sein des images, souvent négligées par les réseaux convolutionnels classiques, renforçant ainsi l'identification des motifs pathologiques diffus.

Cependant, malgré ces avancées notables, la détection précoce de la RD, cruciale pour prévenir la progression vers des formes sévères, demeure un défi. Les stades précoces présentent des signes subtils, parfois difficilement discernables, même pour les modèles les plus performants. Cette difficulté souligne la nécessité de concevoir des approches plus adaptées, intégrant à la fois des architectures puissantes et des stratégies d'ajustement fin. Cette difficulté souligne la nécessité de concevoir des approches plus adaptées, capables de combiner la puissance de modèles complexes avec des stratégies d'optimisation d'optimisation automatique des hyperparamètres.

Dans cette perspective, les chapitres suivants présentent des modèles de classification fondés sur le transfert learning et les Vision Transformers (ViTs), intégrant une stratégie d'autotuning des hyperparamètres. L'objectif est de concevoir un système adaptatif, capable d'ajuster automatiquement sa configuration aux spécificités des données, afin d'améliorer la précision de la classification, notamment pour la détection précoce des cas de rétinopathie diabétique.

CHAPITRE 4

Classification de la RD par Modèles Profonds avec Optimisation Automatique des Hyperparamètres

4.1 Introduction

L'apprentissage par transfert exploite les représentations apprises sur de vastes ensembles de données d'images afin de pallier à l'insuffisance des données médicales annotées [87, 139]. Cependant, les approches existantes souffrent de trois inconvénients majeurs aaaaà savoir :

1. La majorité des études se concentrent sur une seule architecture CNN, sans comparaison systématique de leurs capacités d'extraction de caractéristiques ;
2. L'optimisation des hyperparamètres repose encore sur un ajustement manuel, à la fois coûteuse en ressources et sous-optimale ;
3. Les systèmes de classification sont généralement conçus pour un niveau de granularité fixe, ce qui limite leur adaptabilité aux divers besoins cliniques en matière de dépistage.

Afin de surmonter ces limitations, le présent travail propose un cadre automatisé et complet d'apprentissage par transfert pour la classification de la RD, s'appuyant sur quatre apports majeures [17, 8]"

1. Optimisation automatisée des hyperparamètres :nous éliminons le processus de réglage fin manuel chronophage en intégrant l'optimisation bayésienne qui découvre automatiquement les hyperparamètres optimaux sur différentes architectures, réduisant ainsi la surcharge informatique tout en maximisant les performances.
2. Évaluation de 4 architectures distinctes : nous comparons quatre modèles CNN standard sous optimisation hyperparamétrique uniforme pour identifier l'architecture la plus efficace pour la classification de la RD.

3. Différentes Classification : nous développons une approche permettant de traiter la classification binaire (présence/absence de RD), ternaire et à cinq classes, avec une attention particulière portée à la détection précoce. Notamment, **la configuration à trois classes, encore peu explorée, constitue un compromis pertinent entre la détection globale et la gradation fine de la maladie.** Cette classification intermédiaire améliore la sensibilité aux signes pathologiques précoces tout en réduisant les confusions inter-classes, ce qui s'avère précieux dans un contexte clinique où le diagnostic précoce est essentiel à la préservation de la vision.

Ce chapitre présente le pipeline méthodologique de notre approche, validé sur les jeux de données APTOS et EyePACS, permettant ainsi de choisir la meilleure architecture et le meilleur dataset pour la classification de la RD.

4.2 Méthodologie

Cette section présente un modèle automatisé d'apprentissage profond pour la classification de la RD. L'approche combine l'apprentissage par transfert et l'optimisation bayésienne des hyperparamètres pour améliorer les performances sur les tâches de classification binaire, à 3 classes et à 5 classes.

4.2.1 Processus du Modèle Proposé

Notre pipeline de bout en bout répond aux défis de la classification de la RD grâce à une évaluation des architectures et à l'optimisation bayésienne des hyperparamètres (Voir Figure 4.1). La méthodologie élimine l'ajustement manuel des paramètres tout en garantissant une représentation optimale des caractéristiques à travers diverses architectures CNN par le biais de quatre composants intégrés :

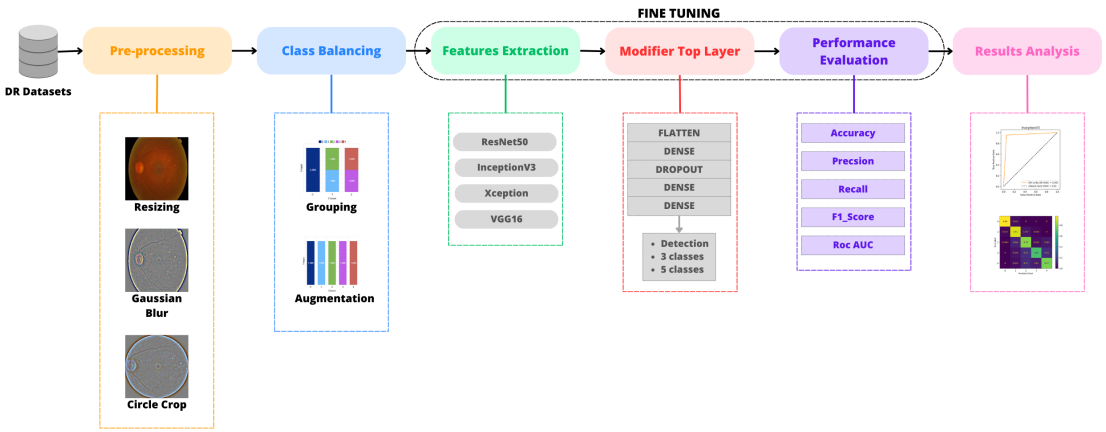


FIGURE 4.1 – Pipeline Complet de l'approche proposée : du prétraitement à la classification de la RD

— Préparation des données : Prétraitement des images avec stratégies d'augmentation équilibrées par classe pour les jeu de données APTOS et EyePACS, incluant le débruitage d'images et le redimensionnement à une résolution RGB de 224×224.

— Optimisation bayésienne des hyperparamètres : Exploration autonome de l'espace des hyperparamètres utilisant l'optimisation bayésienne, permettant un ajustement efficace des paramètres clés tels que le taux d'apprentissage, la taille de lot et le taux de dropout. Ce processus améliore la stabilité d'entraînement et la généralisation du modèle tout en réduisant significativement l'intervention manuelle.

— Extraction de caractéristiques : Évaluation comparative de quatre architectures CNN de base (ResNet50, InceptionV3, VGG16, Xception) sous des conditions d'optimisation identiques. Les têtes de classification sont supprimées et les caractéristiques extraites des blocs convolutifs finaux pour générer des vecteurs de caractéristiques de longueur fixe.

— Classification adaptative aux tâches : Perceptron multicouche personnalisé avec couches de sortie configurables supportant de multiples granularités diagnostiques, utilisant l'activation softmax pour l'estimation de probabilité.

4.2.2 Modèles de Classification

Sur la base du pipeline, nous développons trois modèles de classification de la RD (voir Figure 4.2) :

1. AtRD (Auto-tuned Retino Detection) : Modèle de classification binaire pour le dépistage de la RD (classes : 0, 1) , distinguant les rétines saines de tout stade d'atteinte de la RD.
2. AtR3C (Auto-tuned Retino 3-Class) : Modèle de détection précoce spécialement conçu pour identifier les manifestations subtiles de la RD avant le développement de complications menaçant la vision, catégorisant les cas en absence de RD, RD précoce (légère/modérée), et RD avancée (sévère/proliférante) (classes : 0, 1, 2). Cette configuration maximise la sensibilité pour détecter la progression asymptomatique de la RD.
3. AtR5C (Auto-tuned Retino 5-Class) : Modèle complet de stadification de sévérité correspondant à la classification clinique standard (classes : 0-4), permettant un suivi détaillé de la progression de la RD.

Pour chaque modèle, nous évaluons systématiquement les quatre architectures CNN sous des configurations optimisées bayésiennes identiques en utilisant les deux datasets APTOS et EyePACS. Les combinaisons optimales architecture-hyperparamètres sont sélectionnées sur la base des performances de validation utilisant les métriques de Précision, Rappel, F1-score et exactitude (accuracy). Ce processus de sélection rigoureux assure la validation

empirique tant du choix architectural que de l'optimisation des hyperparamètres, démontrant la supériorité de notre approche automatisée par rapport aux stratégies d'ajustement manuel.

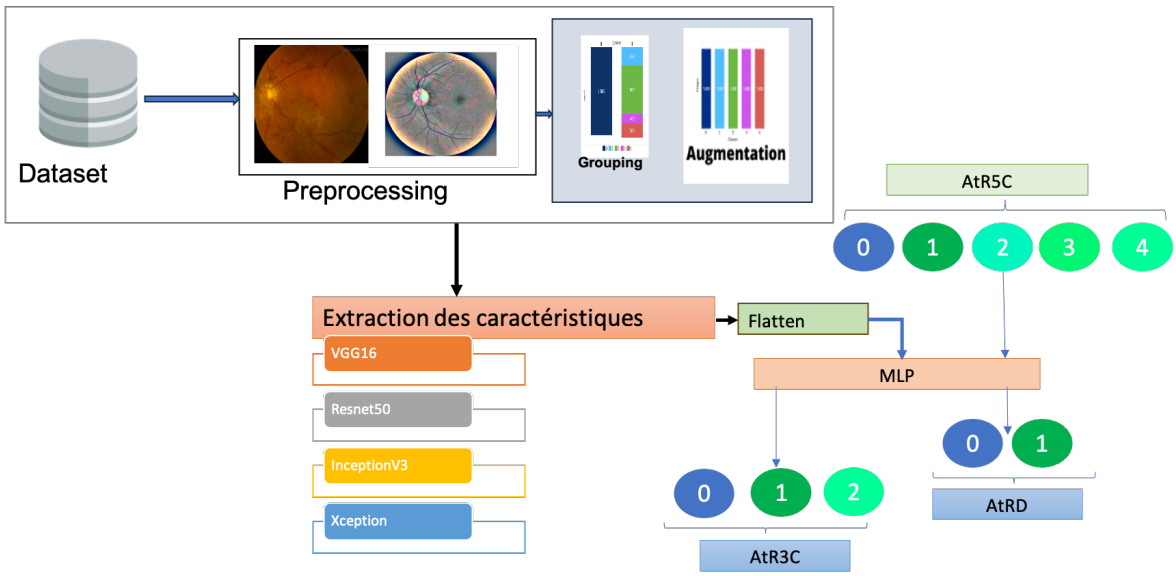
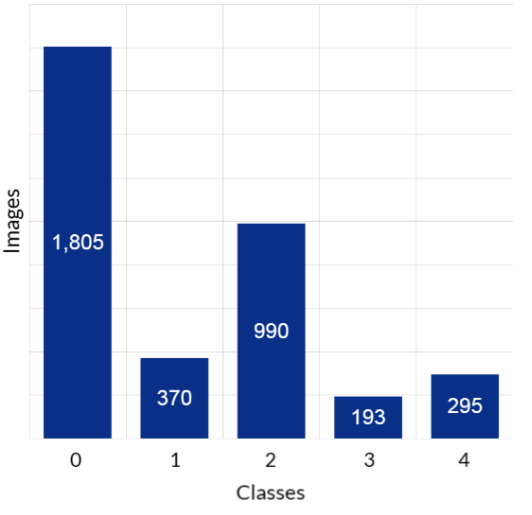


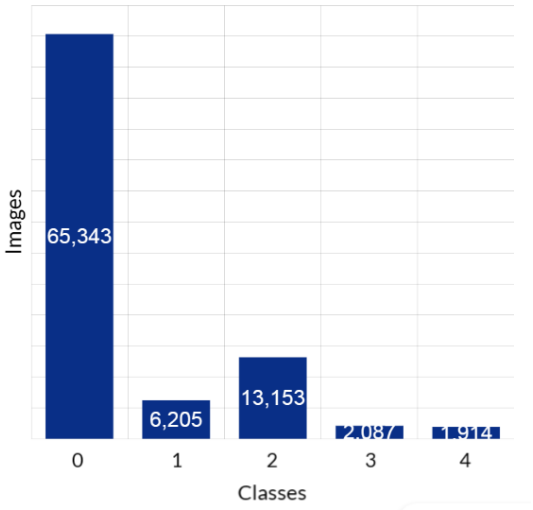
FIGURE 4.2 – Architectures des trois modèles proposés : AtRD, AtR3C et AtR5C

4.2.3 Préparation des Jeux de Données

Tous les modèles que nous avons développés ont été entrainés à l'aide des bases de données publiques APTOS[90] et EyePACS [44], qui contiennent respectivement 3662 et 88507 images. Les deux ensembles de données sont divisés en cinq classes : No RD (0), légère (1), modérée (2), sévère (3) et proliférative (4). La figure 4.3a et 4.3b montrent les distributions par classe des bases de données.



(a) APTOS



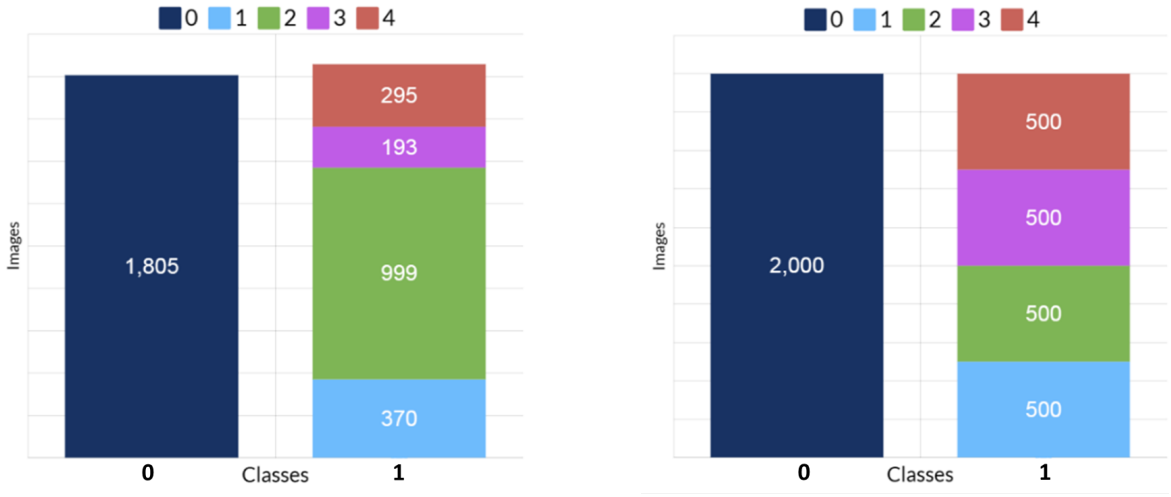
(b) EyePACS

FIGURE 4.3 – Représentation des Classes des Datasets

Les jeu de données présentent un défi majeur lié au déséquilibre prononcé des classes. La majorité des échantillons appartiennent aux classes 0 et classe 2, tandis que les autres classes 1, 3 et 4 ne contiennent qu'un nombre relativement faible d'images (voir Figure 4.3). De plus, les images présentent des variations significatives en termes de résolution.

Afin de permettre une évaluation dans divers contextes de classification, les jeux de données ont été réorganisés selon trois configurations distinctes :

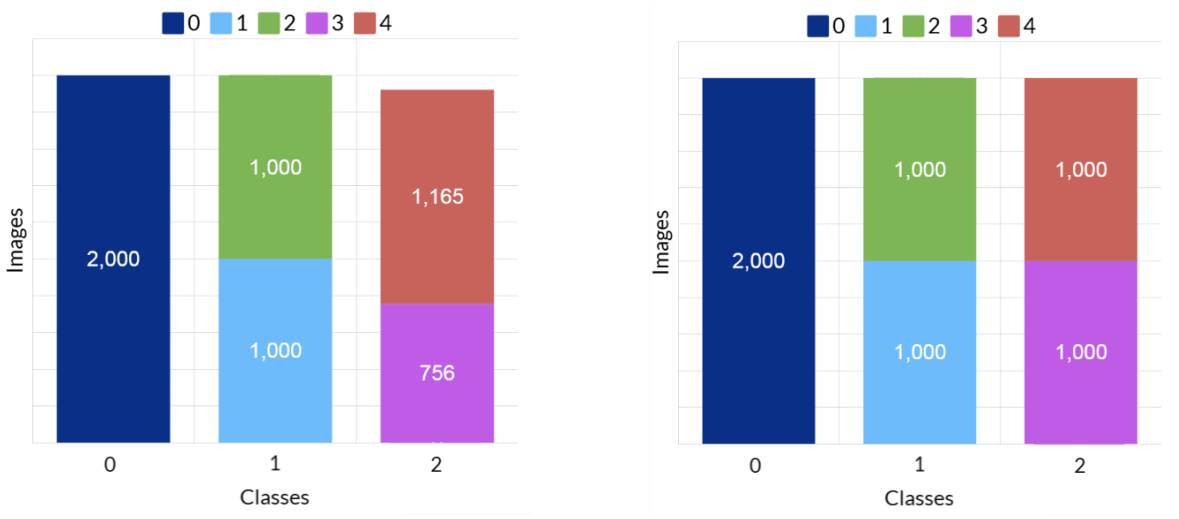
— **Classification binaire** : Les échantillons de la classe 0 (« No DR ») ont été conservés tels quels. Les échantillons des classes 1 à 4 ont été fusionnés en une seule classe, représentant la présence de rétinopathie diabétique (classe 1). Ce réétiquetage a permis d'obtenir un jeu de données binaire équilibré, rendant l'augmentation de données non nécessaire dans cette configuration (Figure 4.4).



(a) APTOS (b) EyePACS

FIGURE 4.4 – Distribution en deux classes après regroupement

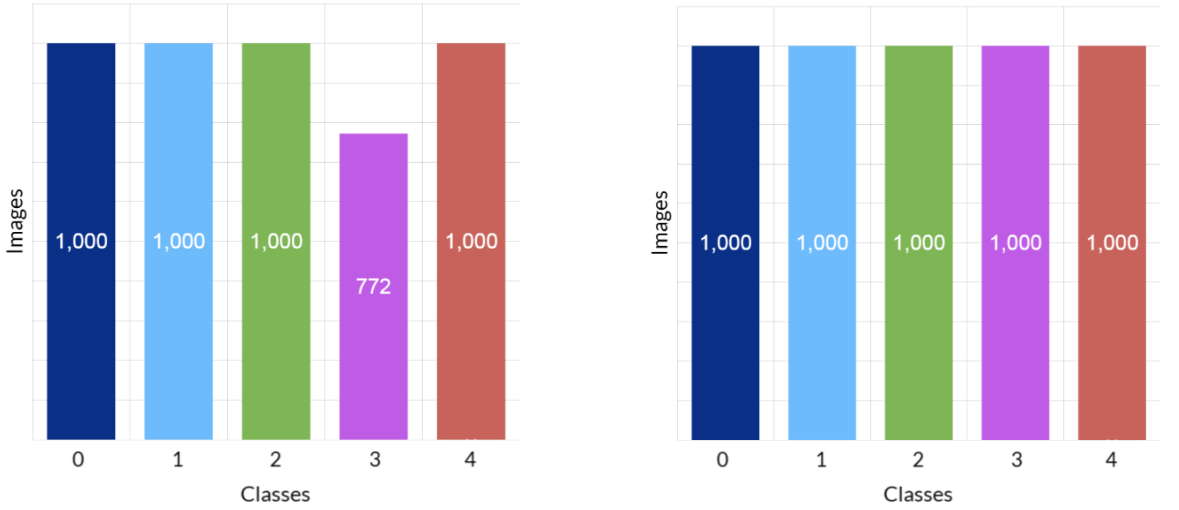
— **Classification en 3 classes** : Pour faciliter la détection précoce de la RD, les jeux de données ont été réétiqueté comme suit : « Healthy » (classe 0 : No DR), « Early DR » (classe 1 : Mild et Moderate), et « Advanced DR » (classe 2 : Sévère et Proliférative). Cette catégorisation reflète des distinctions cliniquement pertinentes dans la progression de la maladie (Figure 4.5).



(a) APTOS (b) EyePACS

FIGURE 4.5 – Distribution en trois classes après regroupement

— Classification en 5 classes : La structure originale des classes a été maintenue. Cependant, en raison du déséquilibre des classes, une augmentation de données s'est avérée nécessaire (Figure 4.6).



(a) APTOS (b) EyePACS

FIGURE 4.6 – Distribution en cinq classes après augmentation

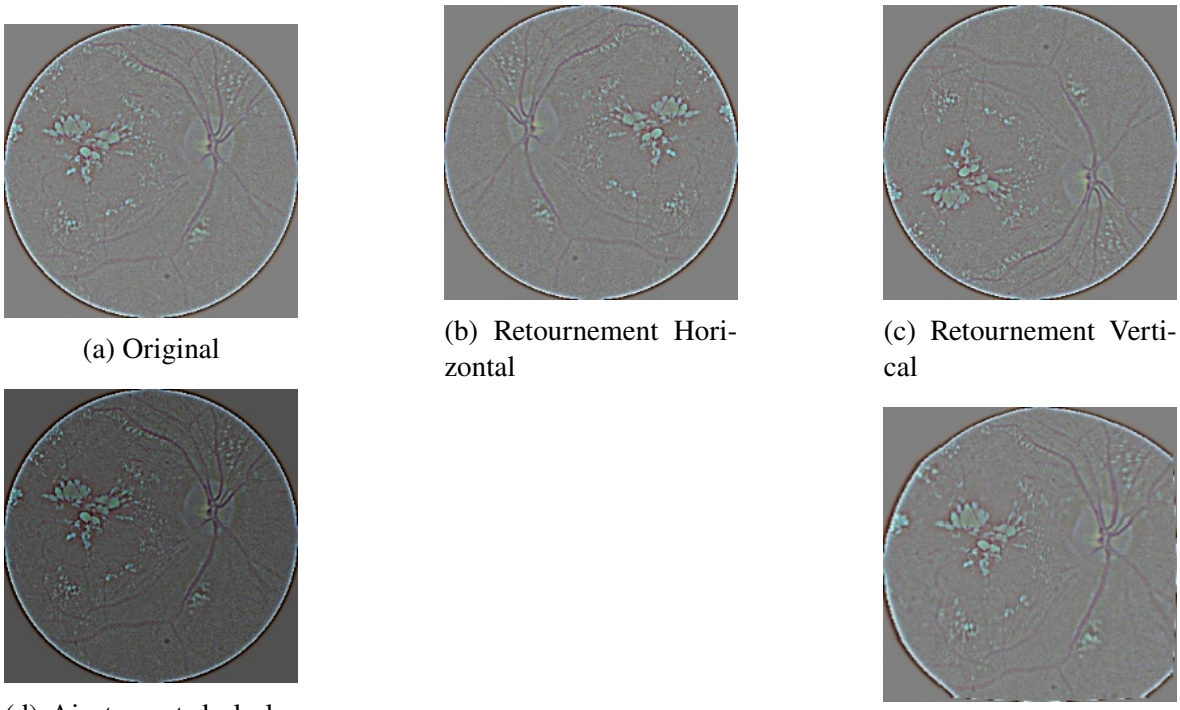
4.2.4 Augmentation des Données

Pour atténuer les effets du déséquilibre des classes, particulièrement dans les configurations à trois et cinq classes, nous avons mis en œuvre une stratégie d'augmentation des données. Ce processus vise à améliorer la généralisabilité du modèle et à garantir sa robustesse face à la variabilité des images due aux conditions d'acquisition. Pour les classes sous-représentées, chaque image originale a été augmentée cinq fois à l'aide de diverses techniques de transformation :

— **Transformations géométriques :** Retournement horizontal et vertical, rotation et distorsion en grille ont été appliqués afin de simuler des variations d'orientation du globe oculaire et des angles d'acquisition de l'image.

— **Ajustements photométriques :** Des modifications aléatoires de la luminosité ont été introduites pour tenir compte des variations d'éclairage.

Ce processus d'augmentation préserve le contenu sémantique des lésions rétiniennes tout en augmentant la diversité du jeu de données, ce qui conduit à une amélioration des performances pendant l'entraînement. Des exemples d'images augmentées sont présentés dans la Figure 4.7.



(a) Original

(b) Retournement Horizontal

(c) Retournement Vertical

(d) Ajustement de la luminosité

(e) Distorsion de la grille

FIGURE 4.7 – Examples d'augmentation d'images

4.2.5 Prétraitement des Images

Le prétraitement développé comprend quatre étapes séquentielles visant à standardiser et à améliorer la qualité visuelle des images rétiniennes.

1. **Standardisation dimensionnelle** : Toutes les images sont redimensionnées à une résolution uniforme de 224×224 pixels par interpolation bilinéaire. Cette standardisation assure la compatibilité avec les architectures de réseaux de neurones convolutifs tout en préservant les caractéristiques morphologiques essentielles.
2. **Conversion en niveaux de gris** : Les images RGB sont converties en niveaux de gris.

La luminance d'un pixel en niveau de gris $I_{gray}(x, y)$ est calculée en utilisant les composantes de couleur rouge, verte et bleue du pixel. L'équation de conversion est :

$$I_{gray}(x, y) = 0.299 \cdot R(x, y) + 0.587 \cdot G(x, y) + 0.114 \cdot B(x, y) \tag{4.1}$$

où :

- $I_{gray}(x, y)$ est l'intensité du pixel en niveau de gris aux coordonnées (x, y).
- $R(x, y)$, $G(x, y)$ et $B(x, y)$ sont les intensités des composantes rouge, verte et bleue du pixel aux coordonnées (x, y).
- Les coefficients 0.299, 0.587, et 0.114 sont des valeurs standard utilisées pour convertir l'espace de couleur RGB en niveau de gris. Ces valeurs sont basées sur la perception de la luminosité par l'œil humain, qui est plus sensible au vert et moins au bleu.

Cette conversion réduit la dimensionnalité tout en préservant l'information structurelle pertinente pour le diagnostic.

3. **Filtrage gaussien et amélioration du contraste** : Un filtre gaussien [135] avec écart-type σ=10 est appliqué pour réduire le bruit haute fréquence :

Le flou d'une image en niveau de gris est obtenu par convolution avec un noyau gaussien. Cette opération est définie par l'équation suivante :

$$I_{blurred}(x, y) = I_{gray}(x, y) * G_{\sigma}(x, y) \tag{4.2}$$

où :

- $I_{blurred}(x, y)$ est l'intensité du pixel flouté aux coordonnées (x, y).
- $I_{gray}(x, y)$ est l'intensité du pixel d'origine en niveau de gris aux coordonnées (x, y).
- $*$ désigne l'opération de convolution.
- $G_{\sigma}(x, y)$ est le noyau de Gauss, calculé avec l'équation suivante :

$$G_{\sigma}(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \tag{4.3}$$

où :

- σ est l'écart-type de la distribution gaussienne, qui contrôle l'intensité du flou. Une valeur de σ plus grande entraîne un flou plus important.

L'amélioration du contraste est réalisée par combinaison linéaire pondérée :

L'amélioration du contraste d'une image peut être réalisée en combinant l'image originale avec une version floutée. Le résultat est une image améliorée par une accentuation des bords. L'équation utilisée pour cette opération est :

$$I_{enhanced}(x, y) = \alpha \cdot I_{gray}(x, y) + \beta \cdot I_{blurred}(x, y) + \gamma \tag{4.4}$$

où :

- $I_{enhanced}(x, y)$ est l'intensité du pixel de l'image améliorée aux coordonnées (x, y).
- $I_{gray}(x, y)$ est l'intensité du pixel de l'image originale en niveau de gris aux coordonnées (x, y).
- $I_{blurred}(x, y)$ est l'intensité du pixel de l'image floutée aux coordonnées (x, y).
- α, β et γ sont des coefficients ajustables pour contrôler le niveau d'amélioration.

Dans cet exemple, les coefficients sont fixés aux valeurs suivantes :

- $\alpha = 5$
- $\beta = -5$
- $\gamma = 128$

Cette technique, basée sur la soustraction du masque flou (unsharp masking), amplifie les détails pathologiques tels que les exsudats et microanévrismes (voir Figure 4.8(b)).

4. **Recadrage circulaire (Circle Crop)** : Pour concentrer l'analyse sur les régions pertinentes pour le diagnostic, une méthode de recadrage circulaire a été appliquée [1]. Cette technique élimine les bords noirs périphériques et le bruit de fond, tout en conservant la zone rétinienne centrale 4.8(c). Un masque circulaire centré est appliqué avec rayon $R = \min(H/2, W/2)$, où H et W sont les dimensions de l'image. Cette approche géométrique est particulièrement adaptée à l'anatomie rétinienne, préservant la papille optique et la région maculaire tout en éliminant les zones périphériques.

Le masque binaire circulaire, utilisé pour isoler une région d'intérêt dans une image, est défini par l'équation suivante :

$$M(x,y) = \begin{cases} 1 & \text{si } \sqrt{(x-x_c)^2 + (y-y_c)^2} \leq R \\ 0 & \text{sinon} \end{cases} \tag{4.5}$$

où :

- $M(x,y)$ est la valeur du pixel du masque binaire aux coordonnées (x,y).
- (x_c, y_c) sont les coordonnées du centre du cercle.
- R est le rayon du cercle.
- La condition $\sqrt{(x-x_c)^2 + (y-y_c)^2} \leq R$ représente la distance euclidienne de chaque point (x,y) par rapport au centre du cercle (x_c, y_c). Si cette distance est inférieure ou égale au rayon R, le pixel appartient au cercle.

Cette équation attribue une valeur de 1 (souvent blanc) à tous les pixels à l'intérieur du cercle et 0 (noir) à tous les pixels à l'extérieur.

5. **Normalisation photométrique** : La normalisation consiste à appliquer une transformation linéaire, pour ajuster les valeurs de pixel pour qu'elles se situent dans l'intervalle [0,1].

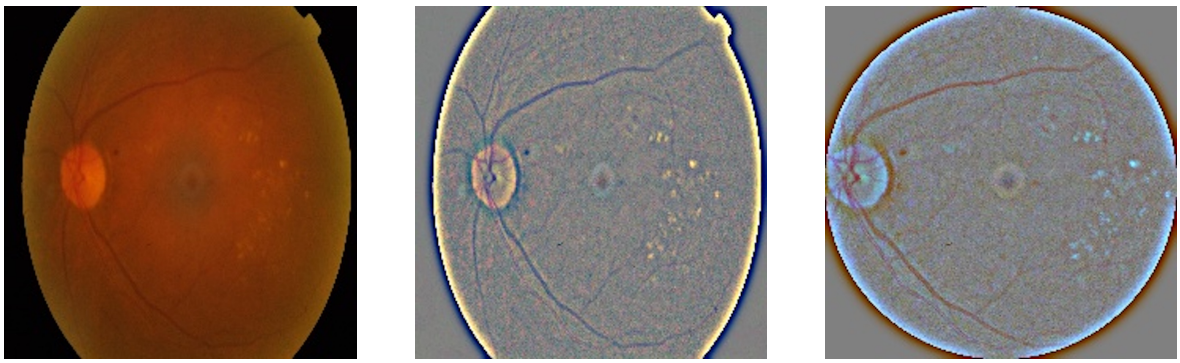
L'équation de normalisation est la suivante :

$$I_{normalized}(x,y) = \frac{I_{circular}(x,y) - I_{min}}{I_{max} - I_{min}} \tag{4.6}$$

où :

- $I_{normalized}(x,y)$ est l'intensité du pixel normalisé aux coordonnées (x,y).
- $I_{circular}(x,y)$ est l'intensité du pixel de l'image circulaire aux coordonnées (x,y).
- I_{min} est la plus petite valeur de pixel dans l'image.
- I_{max} est la plus grande valeur de pixel dans l'image.

Cette standardisation homogénéise la dynamique des images, réduisant l'impact des variations d'éclairage et optimisant la convergence lors de l'entraînement.



(a) Original image (b) Gaussian blur (c) Circular Cropping

FIGURE 4.8 – Illustration des étapes de pré-traitement des images

4.2.6 Processus d'Ajustement Automatique des Hyperparamètres (Fine Tuning)

Les modèles pré-entraînés doivent être adaptés aux exigences spécifiques de la tâche de détection et de classification de la rétinopathie diabétique (RD). À cette fin, nous avons eu recours au transfert d'apprentissage en utilisant des modèles pré-entraînés, à savoir : ResNet50, VGG16, Xception et InceptionV3, initialement entraînés sur ImageNet [97]. Les procédures traditionnelles d'ajustement fin (fine-tuning) reposent souvent sur un réglage manuel des hyperparamètres, ce qui peut entraîner une convergence sous-optimale et une généralisation limitée. Pour surmonter cette limitation, nous avons mis en œuvre un cadre d'optimisation automatique en deux étapes (voir Figure 4.9).

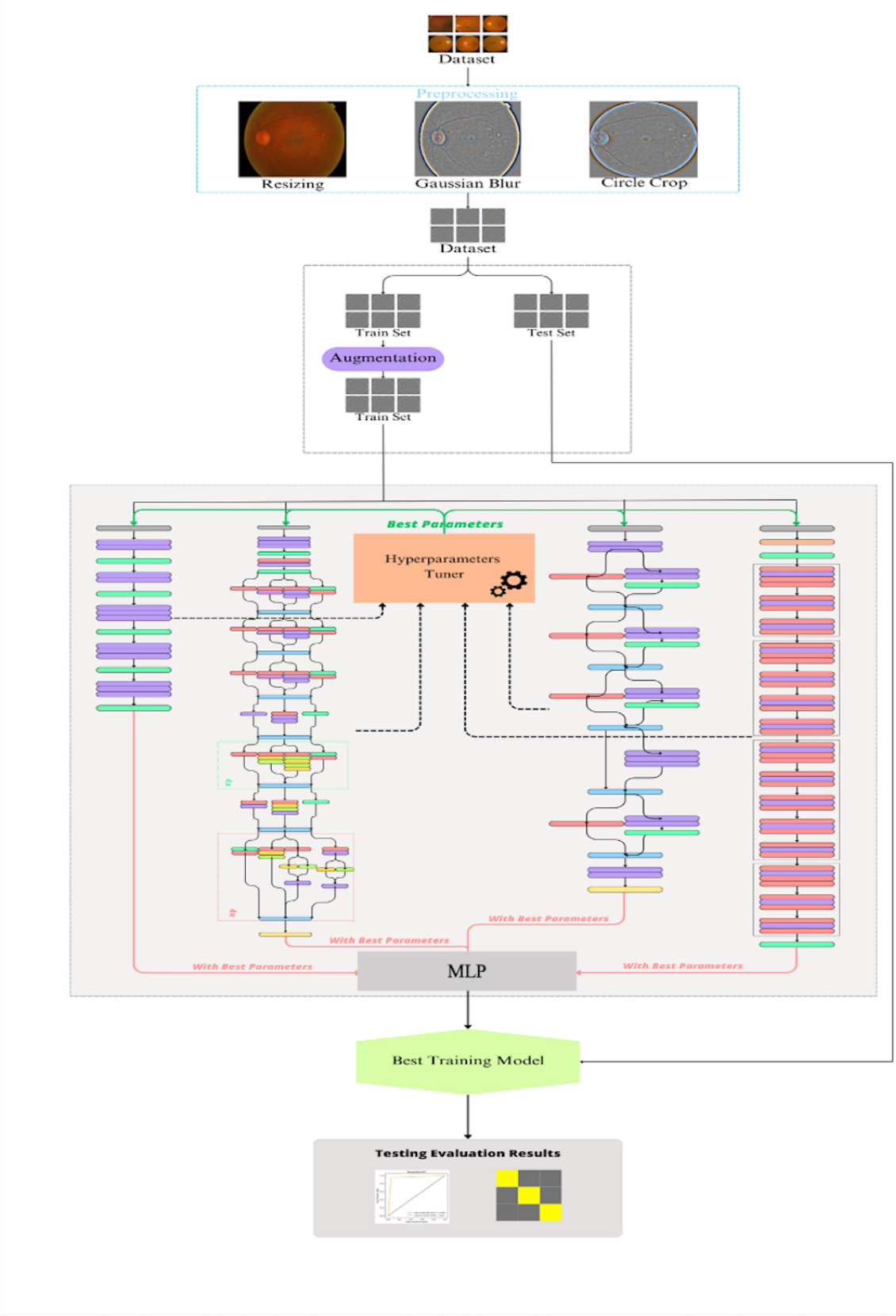


FIGURE 4.9 – Architecture avec le processus Auto-Tuning en évidence

Dans la première étape, l'optimisation bayésienne a été employée pour explorer l'espace des hyperparamètres et identifier les configurations optimales, incluant le taux d'apprentissage, le taux de décroissance (dropout rate) et la taille du lot (batch size), qui maximisent

la précision de validation. Dans la deuxième étape, la configuration sélectionnée a été utilisée pour ajuster finement les réseaux pré-entraînés sur les jeux de données, leur permettant d'adapter leurs paramètres aux caractéristiques visuelles spécifiques des images rétiniennes.

Cette stratégie d'ajustement fin est illustrée dans la Figure 4.10.

Pour renforcer davantage la précision et la robustesse, nous avons appliqué cette stratégie d'optimisation à nos architectures personnalisées pré-entraînées (AtRD, AtR3C et AtR5C). La recherche d'hyperparamètres a été menée en utilisant un algorithme d'optimisation bayésienne basé sur le processus gaussiens, implémenté via la fonction `gp_minimize` de la bibliothèque scikit-optimize.

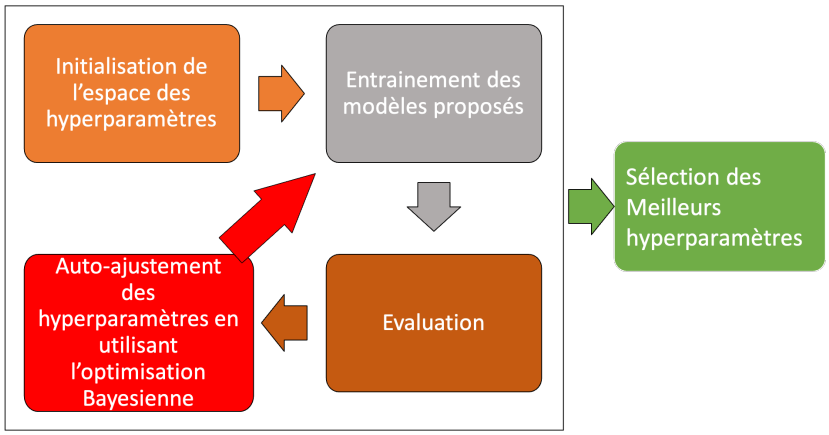


FIGURE 4.10 – Processus d'ajustement automatique des hyperparamètres

Processus d'ajustement des hyperparamètrese

L'ajustement des hyperparamètres s'effectue selon les étapes suivantes :

1. **Espace de recherche** : l'optimisation des hyperparamètres a été réalisée sur un espace de recherche multidimensionnel couvrant l'ensemble des paramètres critiques pour l'entraînement du modèle. L'espace de recherche $\mathcal{H}$ couvre un ensemble exhaustif d'hyperparamètres. Le Tableau 4.1 fournit une description détaillée de l'espace de recherche utilisé.
2. **Stratégie d'optimisation des hyperparamètres** : l'optimisation a été conduite sur 40 itérations selon une approche d'optimisation bayésienne. Pour chaque essai $t \in \{1, 2, \ldots, 40\}$, la configuration d'hyperparamètres $\mathbf{h}_t \in \mathcal{H}$ a été évaluée selon le protocole suivant :

$$\mathbf{h}_t = \arg\max_{\mathbf{h}\in\mathcal{H}} \alpha(\mathbf{h} \mid \mathcal{D}_{1:t-1}) \tag{4.7}$$

où $\alpha(\mathbf{h} \mid \mathcal{D}_{1:t-1})$ représente la fonction d'acquisition basée sur les observations précédentes $\mathcal{D}_{1:t-1} = \{(\mathbf{h}_1, y_1), \ldots, (\mathbf{h}_{t-1}, y_{t-1})\}$.

TABLE 4.1 – Espace des Hyperparamètres à ajuster

Hyperparamètre	**Plage**	**Description**
Batch Size	16 to 64	Contrôle le nombre d'échantillons traités par étape de mise à jour.
Learning Rate	5×10^{-5} to 1.5×10^{-4}	Influe sur la vitesse de convergence pendant l'entraînement.
Dropout Rate	0.2 to 0.7	Régularise le modèle en abandonnant aléatoirement des connexions entre les couches.
Hidden Units	64, 128, 256	Nombre d'unités dans les couches denses, contrôlant la capacité du modèle.
Number of Layers	1 to 4	Définit la profondeur du réseau neuronal, influençant sa capacité à apprendre des caractéristiques complexes.
Learning Rate Schedule	`constant`, `step`	Détermine comment le taux d'apprentissage est ajusté pendant l'entraînement.
Warmup Epochs	1, 2, 5	Nombre d'époques avant que le taux d'apprentissage ne commence à diminuer.
Decay Drop	0.1 to 0.5	Contrôle le taux auquel le taux d'apprentissage décroît pendant l'entraînement.

3. **Protocole d'évaluation** : chaque configuration a été évaluée selon un entraînement limité à 7 époques, permettant un compromis optimal entre précision d'évaluation et efficacité computationnelle. La fonction objective $f(\mathbf{h})$ correspond à la précision de validation :

$$f(\mathbf{h}) = \frac{1}{|\mathcal{V}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{V}} \mathbb{1}[\hat{y}_i(\mathbf{h}) = y_i] \tag{4.8}$$

où $\mathcal{V}$ représente l'ensemble de validation, $\hat{y}_i(\mathbf{h})$ la prédiction du modèle paramétré par $\mathbf{h}$ pour l'échantillon $\mathbf{x}_i$, et $\mathbb{1}[\cdot]$ la fonction indicatrice.

Cette méthodologie d'optimisation permet une identification efficace des configurations performantes, équilibrant la vitesse de convergence et les performances de généralisation. L'approche bayésienne assure une exploration intelligente de l'espace des hyperparamètres, concentrant les évaluations sur les régions les plus prometteuses tout en maintenant une capacité d'exploration suffisante pour éviter les optima locaux.

Algorithme d'Ajustement des Hyperparamètres

L'algorithme d'optimisation bayésienne des hyperparamètres est formalisé dans l'Algorithme 1. Le processus itératif débute avec une configuration initiale $\mathbf{H}_0$ et procède à l'amélioration progressive selon les étapes suivantes :

Algorithm 1: Optimisation Bayésienne des Hyperparamètres

Input: H_0 : hyperparamètres initiaux
$n_{\max}$: nombre d'itérations
Output: H_{best} : configuration optimale
$Score_{\text{best}}$: performance maximale obtenue

$Score_0 \leftarrow$ `ÉvaluerModèle(`H_0`)`
$H_{\text{best}} \leftarrow H_0$
$Score_{\text{best}} \leftarrow Score_0$
for $n \leftarrow 1$ **to** n_{max} **do**
    $H \leftarrow$ `OptimisationBayésienne(`H_{best}`)`
    $Score \leftarrow$ `ÉvaluerModèle(`H`)`
    if $Score > Score_{best}$ **then**
        $Score_{\text{best}} \leftarrow Score$
        $H_{\text{best}} \leftarrow H$
    end
end
return $H_{best}, Score_{best}$

4.2.7 Processus d'Extraction des Caractéristiques

Nous avons appliqué le transfert d'apprentissage pour extraire des caractéristiques pertinentes à partir d'images rétiniennes. Cette étape représente une phase de grande importance dans la chaîne de traitements, car la qualité des caractéristiques extraites aura une incidence directe sur la capacité du modèle à bien distinguer les niveaux de sévérité de la RD. Ainsi nous avons évalué quatre des CNN préentraînés les plus utilisés comme extracteurs de caractéristiques : VGG16, ResNet50, InceptionV3 et Xception. Chaque modèle préentraîné a été alimenté avec des images redimensionnées et prétraitées. Par la suite, des caractéristiques ont été extraites des couches convolutionnelles finales de chaque réseau préentrainé pour obtenir une représentation globale des images rétiniennes, comme illustré à la Figure 4.11. Les caractéristiques extraites sont ensuite utilisées pour l'étape ultérieure de classification DR. Chaque modèle a été ajusté avec précision sur le jeu de données APTOS et EyePACS, avec des hyperparamètres optimisés à l'aide de la modélisation bayésienne, pour déterminer l'architecture qui produit les représentations de caractéristiques les plus robustes pour la classification de la RD.

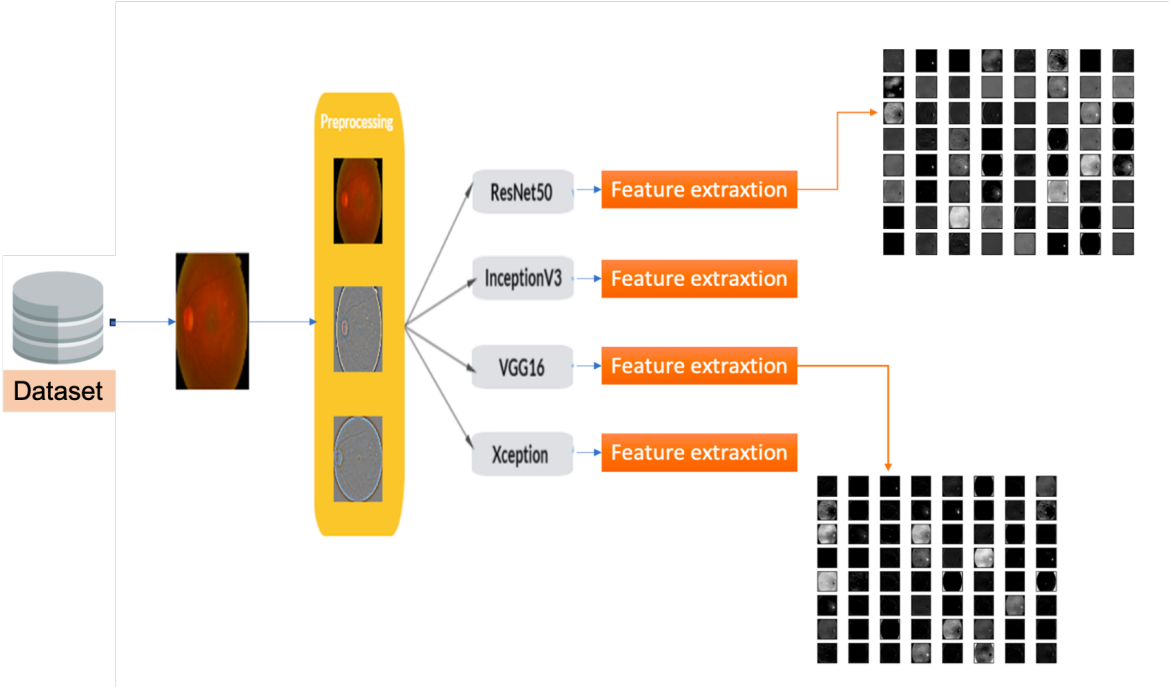


FIGURE 4.11 – Visualisation des caractéristiques.

4.2.8 Processus de Classification

Le processus de classification utilise les caractéristiques extraites des modèles préentraînés. Ces caractéristiques sont ensuite aplaties en un vecteur unidimensionnel qui est ensuite introduit dans une série de couches denses entièrement connectées appelées MLP. Plus précisément, il y a trois couches denses : la première est composée de 1024 neurones et de la fonction d'activation ReLU, suivie d'un dropout de 0.3 pour prévenir le surapprentissage. La seconde couche dense comprend 512 neurones et repose également sur la fonction d'activation ReLU. La couche de sortie varie en fonction de la tâche de classification considérée. Pour la classification binaire (absence ou présence de rétinopathie), la couche de sortie comprend deux neurones activés par une fonction Softmax, produisant une distribution de probabilité sur les deux classes. Dans le cas des classifications multiclasse, la sortie est adaptée en conséquence : elle comprend trois neurones pour la tâche à 3 classes, et cinq neurones pour la tâche à 5 classes (Voir Figure 4.2).

Cette tête de classification unifiée est entraînée conjointement avec l'extracteur de caractéristiques affiné, en utilisant la fonction de perte d'entropie croisée. L'optimisation repose sur les hyperparamètres optimaux identifiés par l'algorithme d'optimisation bayésienne. Le processus d'apprentissage est répliqué indépendamment pour chaque configuration de classification (binaire, 3 classes et 5 classes), en conservant la même architecture générale mais en adaptant la configuration de sortie.

4.3 Résultats et Analyse

Cette section présente l'évaluation expérimentale des différentes architectures que nous proposons pour la classification de la rétinopathie diabétique (RD). Nous évaluons de manière systématique les performances des modèles de détection et de classification en nous appuyant sur plusieurs architectures CNN (VGG16, InceptionV3, Xception et ResNet50) et deux jeux de données (APTOS et EyePACS), dans le but d'identifier la meilleure architecture pour la classification de la RD en 2, 3 et 5 stades de sévérité ainsi que le dataset qui a mené au modèle le plus performant.

Les expérimentations ont été réalisées sur la plateforme Kaggle, à l'aide d'un processeur Intel Xeon @ 2,20 GHz et de 13 Go de mémoire RAM. Les datasets APTOS 2019 et EyePACS ont été partitionnés selon un ratio de 90 :10 pour l'entraînement et le test, avec une sélection aléatoire de 220 images issues de l'ensemble de test, utilisées pour la validation et la sélection des modèles.

Tous les modèles ont été entraînés en utilisant les hyperparamètres optimaux obtenus par optimisation bayésienne, et qui sont illustrés dans le Tableau 4.2.

TABLE 4.2 – Les Meilleurs Hyperparamètres

Parameters	Values
Image size	224×224
Batch size	32
Warmup epochs	5
Warmup learning rate	1×10^{-5}
Epochs	50
Learning rate	1×10^{-4}
Weight decay	2×10^{-2}
Early stopping patience	15
Patience for learning rate reduction	5
Regularizer	2×10^{-2}

Dans le but d'identifier l'architecture CNN la plus adaptée à la classification de la RD, nous avons utilisées 4 extracteurs de caractéristiques préentraînés (VGG16, Xception, InceptionV3 et ResNet50). Afin d'analyser l'impact du jeu de données sur les performances des modèles, deux bases de données distinctes ont été utilisées pour l'évaluation : APTOS et EyePACS. Pour chaque cas de classification nous allons présentées les performances fournies par le meilleur extracteur pour chaque dataset.

4.3.1 Performance de Détection de la RD : Classification Binaire

Les Tableaux 4.3 et 4.4 présentent les performances comparées des quatre architectures CNN préentraînées (ResNet50, InceptionV3, VGG16 et Xception), évaluées sur les deux jeux

de données publics APTOS et EyePACS pour la classification binaire (détection de la RD). La Figure 4.12 montre la meilleure architecture pour chaque jeux de données

TABLE 4.3 – Performances des architectures CNN sur le jeu de données APTOS

Métrique	ResNet50	InceptionV3	VGG16	Xception
Accuracy	**0.9922**	0.9805	0.9903	0.9883
Precision	0.9961	0.9807	0.9885	**0.9984**
Recall	**0.9923**	0.9807	**0.9923**	0.9884
F1 Score	**0.9941**	0.9807	0.9804	0.9884

TABLE 4.4 – Performances des architectures CNN sur le jeu de données EyePACS

Métrique	ResNet50	InceptionV3	VGG16	Xception
Accuracy	0.7075	0.6693	0.6929	**0.7179**
Precision	0.7026	0.6302	**0.7429**	0.7117
Recall	0.6377	**0.7125**	0.5693	0.6496
F1 Score	0.6686	0.6655	0.6446	**0.6792**

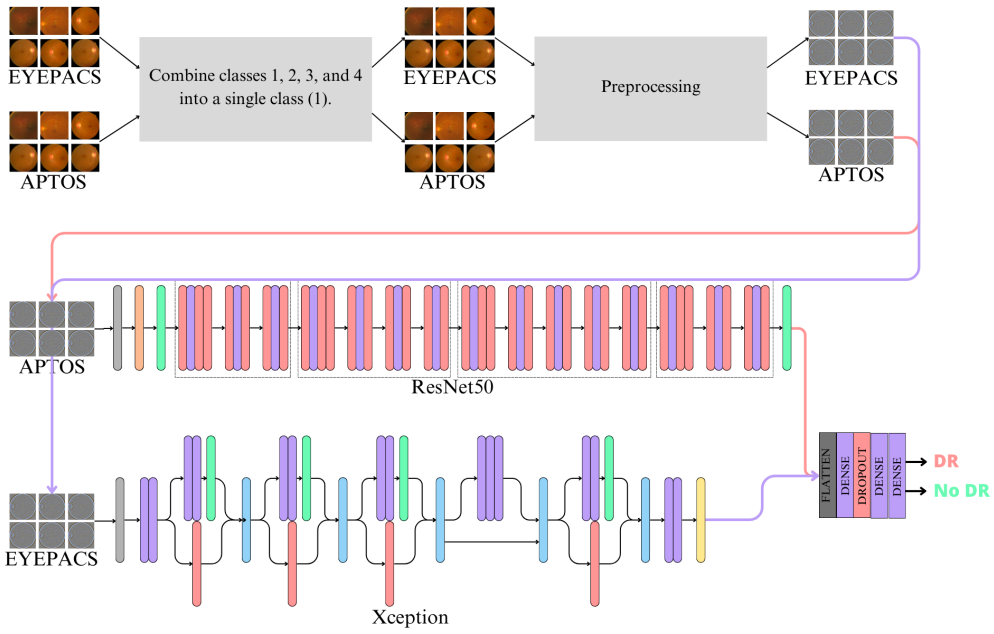


FIGURE 4.12 – AtRD Architecture

- **Performances sur le jeu de Données APTOS**

Les résultats obtenus sur APTOS indiquent que l'architecture ResNet50 surpasse nettement les autres modèles sur l'ensemble des métriques considérées. Elle atteint une accuracy de 99,22%, une précision de 99,61%, un Recall de 99,23% et un F1-score de 99,41%. Ces scores extrêmement élevés témoignent d'une excellente capacité de généralisation du modèle

dans ce contexte. Bien que l'architecture VGG16 produise de très bons résultats, notamment un Rappel (Recall) comparable à celui de ResNet50, ce qui souligne sa capacité à bien identifier les cas positifs, sa précision et son F1-score s'avèrent légèrement inférieurs. Le modèle Xception, malgré une très haute précision de 99,84%, affiche un Rappel (Recall) légèrement inférieur (98,84%). Cela pourrait indiquer une plus grande tendance à générer des faux négatifs comparé à ResNet50.

- **Performances sur le jeu de données EyePACS**

Sur EyePACS, les performances sont nettement plus faibles que sur APTOS, ce qui révèle une sensibilité importante au changement de données (distribution des données, qualité des annotations, diversité des patients, conditions d'acquisition, etc.).

Le modèle Xception s'impose comme le plus performant sur ce jeu de données, avec une accuracy de 71,79% et un F1-score de 67,92%. Il surpasse les autres architectures sur toutes les métriques, à l'exception de la précision, où VGG16 atteint la meilleure valeur (74,29%).

- **Analyse discriminative des modèles par matrice de confusion et courbe ROC**

Dans ce qui suit, nous présentons l'évaluation des performances du modèle à travers la matrice de confusion et la courbe ROC. La Figure 4.13 et la Figure 4.14 illustrent les résultats obtenus sur le jeu de données APTOS, tandis que la Figure 4.15 et la Figure 4.16 présentent les performances sur le jeu de données EyePACS.

La matrice de confusion permet d'analyser la capacité du modèle à distinguer correctement chaque classe, en identifiant les erreurs de classification spécifiques. La courbe ROC, quant à elle, met en évidence le compromis entre le taux de vrais positifs (sensibilité) et le taux de faux positifs, en fonction des différents seuils de décision.

Ces visualisations offrent une évaluation complémentaire et interprétable de la performance des modèles, en particulier pour la détection automatique de la rétinopathie diabétique.

Les résultats indiquent clairement une supériorité des performances sur le jeu de données APTOS, notamment avec l'architecture ResNet50, qui affiche une meilleure capacité de généralisation et de detection de la RD.

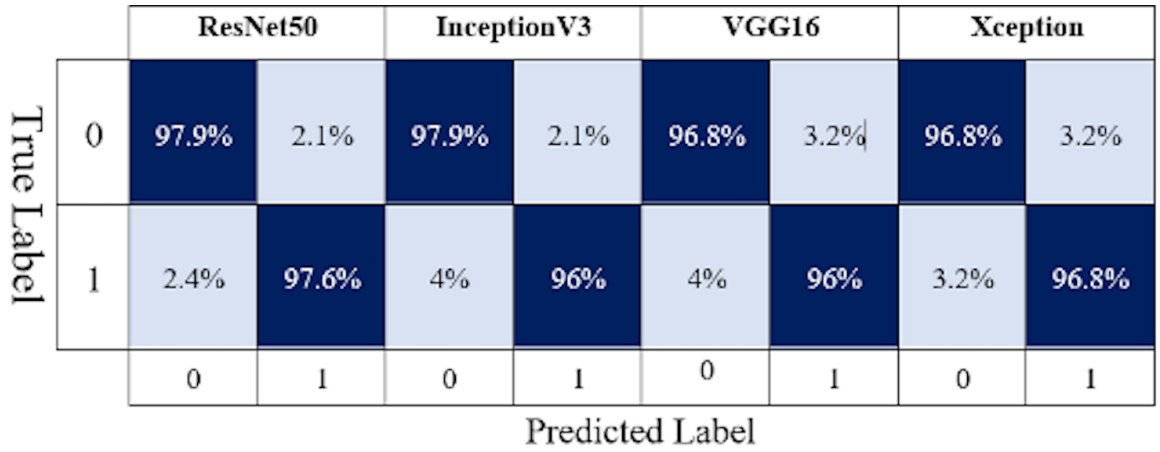


FIGURE 4.13 – Matrices de confusion de chaque modèle CNN sur le jeux de données APTOS

La matrice de confusion illustrée à la Figure4.13 met en évidence une très bonne capacité discriminante du modèle ResNet50 pour distinguer les images normales (classe 0) de celles présentant des signes de rétinopathie diabétique (classe 1).

— Classe 0 (absence de RD) : Le modèle a correctement identifié 97,9% des cas négatifs, avec seulement 2,1% de faux positifs, ce qui reflète une spécificité élevée. Cela signifie que très peu d'images saines ont été faussement classées comme pathologiques.

— Classe 1 (présence de RD) : La sensibilité atteint 97,6%, avec seulement 2,4% de faux négatifs. Autrement dit, la grande majorité des cas atteints de RD ont été correctement détectés. Cela est crucial dans un contexte de dépistage, car les faux négatifs représentent un risque médical important (patients non diagnostiqués à temps).

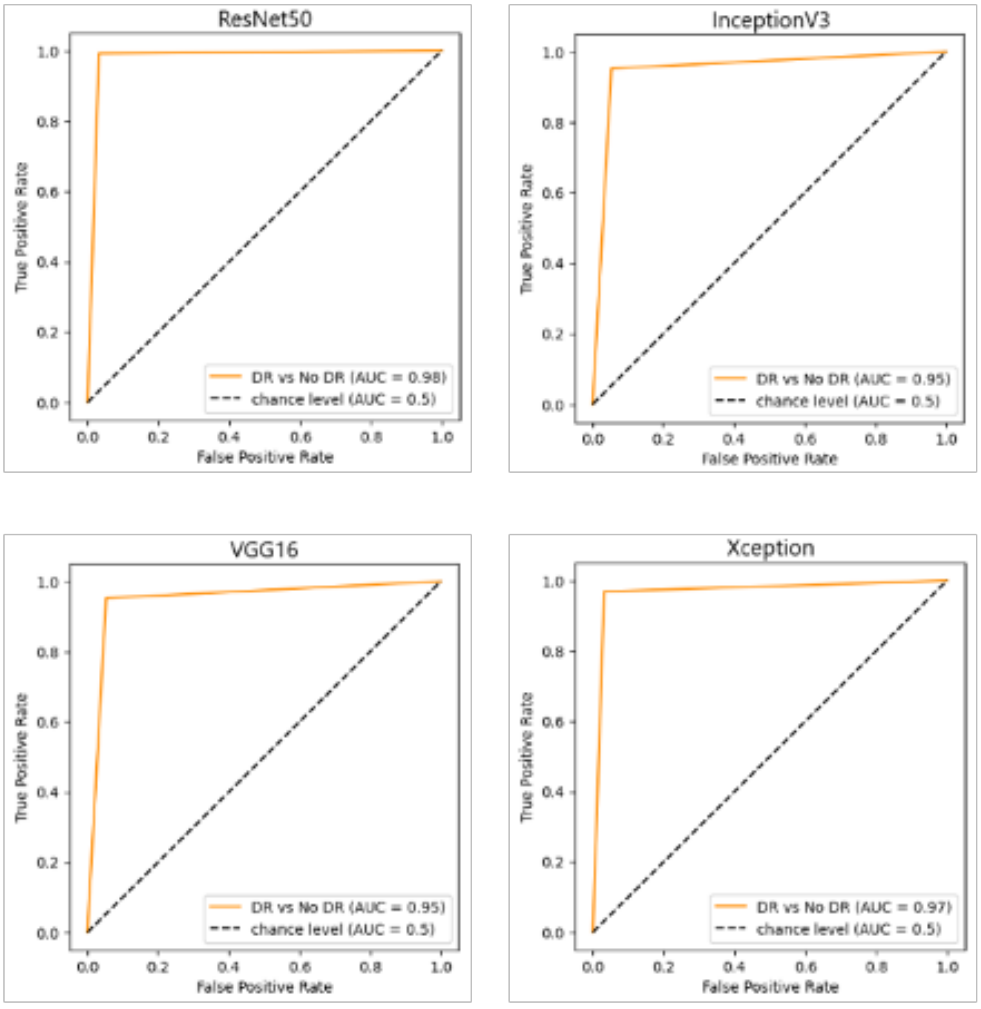


FIGURE 4.14 – Courbes AUC-ROC des modèles CNN : Classification binaire (APTOS)

L'analyse de la courbe ROC (Receiver Operating Characteristic) (voir Figure 4.14) confirme la robustesse du modèle avec une AUC de 0,98, indiquant une excellente séparabilité des classes pour les applications de dépistage clinique.

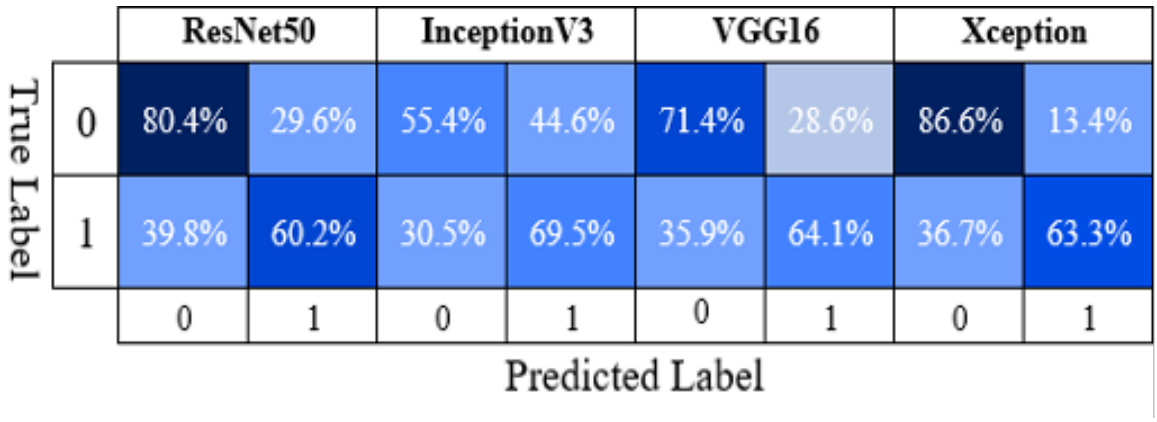


FIGURE 4.15 – Matrices de confusion de chaque modèle CNN sur le jeux de données EyePACS

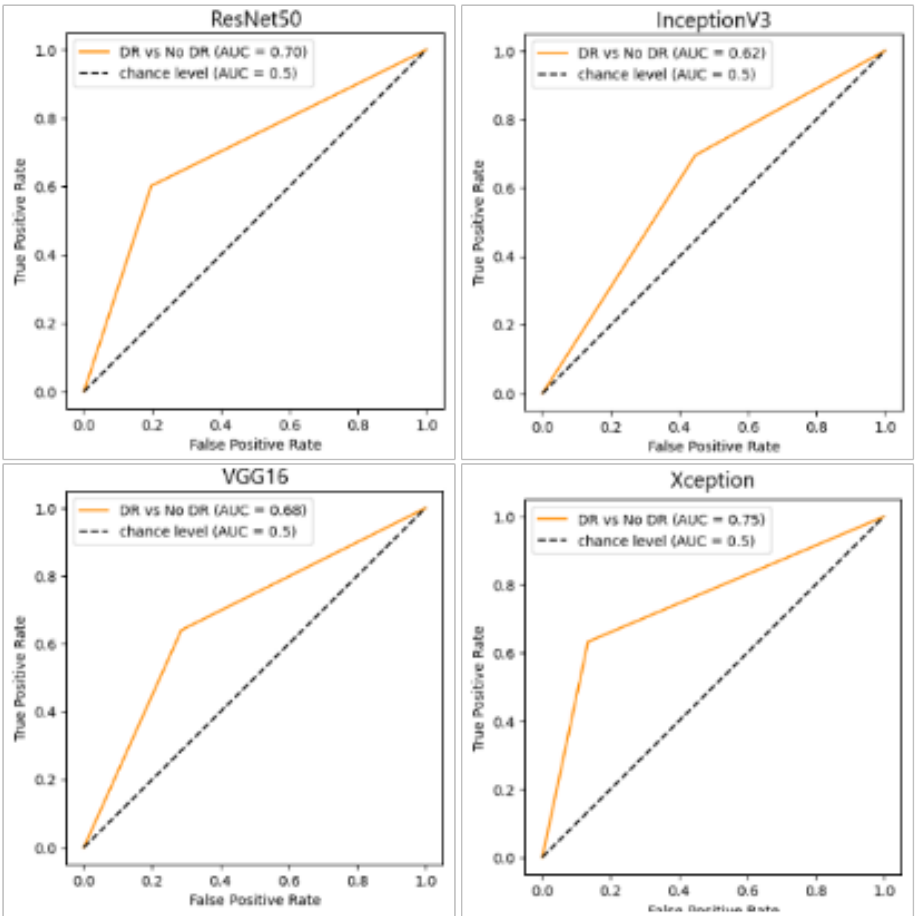


FIGURE 4.16 – Courbes AUC-ROC des modèles CNN : Classification binaire (EyePACS)

La matrice de confusion obtenu par l'architecture Xception sur le dataset Eypacs indique des performances contrastées du modèle Xception sur ce jeu de données, avec une asymétrie importante entre la classification des images saines (classe 0) et pathologiques (classe 1).

— Classe 0 (absence de RD) : Le modèle identifie correctement 86,6% des cas sains. Cependant, 13,4% des images sans signes de RD sont faussement classées comme positives (faux positifs). Cela peut entraîner un taux de surdiagnostic non négligeable, potentiellement source de stress ou de procédures médicales inutiles.

— Classe 1 (présence de RD) : La sensibilité est relativement basse, avec seulement 63,3% de détection correcte, contre 36,7% de faux négatifs. Ce taux d'erreur est préoccupant en contexte clinique, car il signifie qu'un tiers des patients atteints ne sont pas détectés par le modèle, ce qui compromet sérieusement le rôle du système comme outil de dépistage précoce.

- **Conclusion**

— La comparaison des performances entre les jeux de données APTOS et EyePACS révèle un écart significatif de 20 à 30 points de pourcentage, toutes architectures confondues. Cette divergence souligne l'importance cruciale de la variabilité des jeux de données et son impact direct sur les performances des modèles de détection.

— Les résultats expérimentaux démontrent que l'architecture ResNet50 surpasse nettement les autres modèles en termes de performance sur le jeu de données APTOS, tant en précision qu'en équilibre entre sensibilité et spécificité. Ainsi, l'architecture que nous retenons comme modèle final, désignée AtRD, correspond à une configuration comme extracteur de caractéristiques ResNet50 entraînée spécifiquement sur le jeu de données APTOS. Ce choix s'appuie sur sa capacité à détecter efficacement la rétinopathie diabétique.

TABLE 4.5 – Comparaison des modèles proposés selon différentes métriques d'évaluation (%)

Metrics	ResNet50	Inception V3	VGG16	Xception
Accuracy	99.22	98.05	99.03	98.83
Precision	99.61	98.05	98.85	99.84
Recall	99.23	98.05	98.04	98.84
F1_Score	99.41	98.07	98.04	98.84

Ces résultats démontrent l'efficacité de l'optimisation automatisée des hyperparamètres pour atteindre des niveaux de performance cliniquement pertinents. Les capacités supérieures d'extraction de caractéristiques du modèle AtRD basé sur ResNet50, combinées à des hyperparamètres optimisés, établissent une base solide pour le dépistage de la RD au niveau de la population.

4.3.2 Performance de Classification Multi-Classes de la Rétinopathie Diabétique

Cette section présente l'évaluation des modèles AtR3C et AtR5C pour la classification des stades de gravité de la RD. Les deux modèles s'appuient sur ResNet50 comme architecture de base, entraînés sous des hyperparamètres optimisés bayésiens, et démontrent une

performance supérieure par rapport aux autres architectures CNN testées sur le dataset Aptos
.

Cette expérience a été menée en utilisant deux bases de données différentes pour évaluer les deux modèles pour leurs capacités de classification de la RD en 3 ou 5 classes. Un ensemble de test composé de 1184 images du jeu de données APTOS et 1200 images du jeu de données EyePACS ont été choisis pour tester les modèles.

a) Classification en 3 classes

Comme indiqué dans les tableaux 4.6, les performances ont atteint une précision impressionnante de 94,26% sur l'ensemble de données APTOS en utilisant ResNet50, tandis que sur l'ensemble de données EyePACS, la précision était inférieure à 67% en utilisant Xception (voir Figure 4.17).

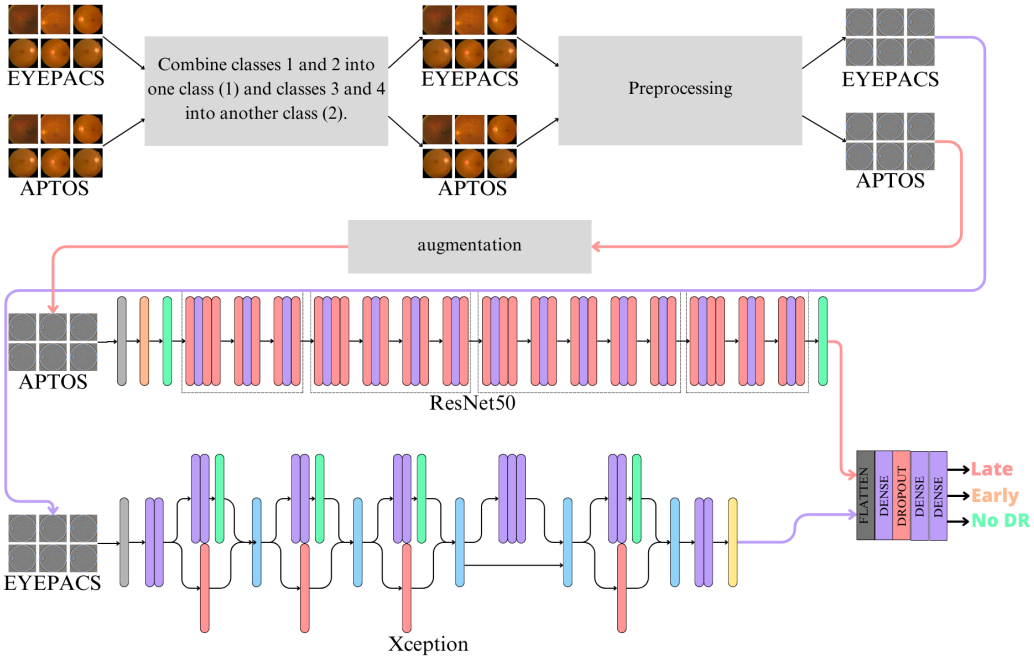


FIGURE 4.17 – AtR3C Architecture

TABLE 4.6 – Résultats obtenus par les meilleurs modèles pour la classification en 3 classes

	ResNet50	
APTOS	Accuracy	**0.9426**
	Precision	**0.9441**
	Recall	**0.9409**
	F1_score	**0.9424**

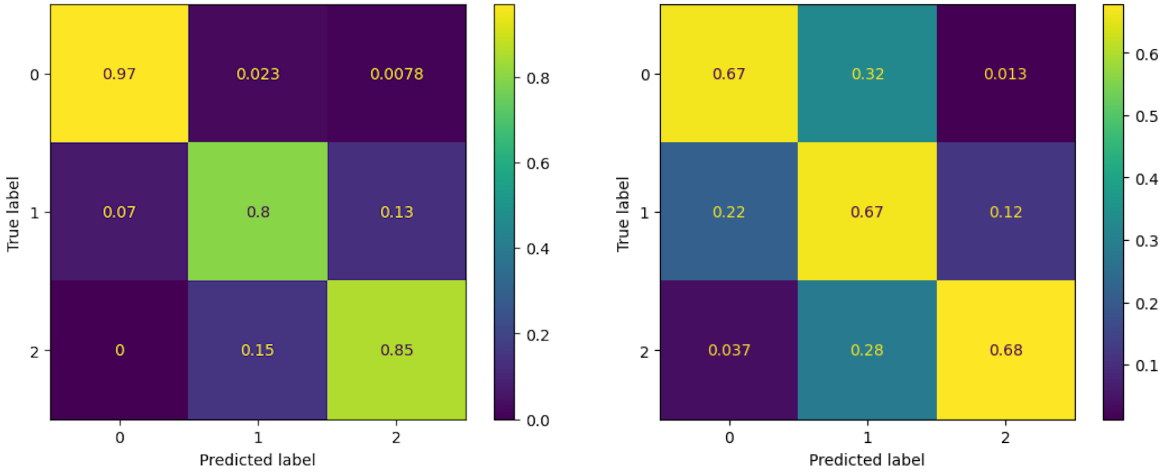
	Xception	
EyePACS	Accuracy	**0.6700**
	Precision	**0.6713**
	Recall	**0.6858**
	F1_score	**0.6735**

Pour fournir un aperçu de la performance du modèle en termes de prédiction, une matrice de confusion et un tableau de métriques ont été générés pour chaque modèle, comme indiqué dans la table 4.6 et figures 4.18a et 4.18b. Nous pouvons constater que l'architecture basée

ResNet50 entrainé sur le dataset APTOS fournit de meilleurs résultats que ce soit en terme de métriques génerales ou par classes, donc le modèle Atr3C est le modèle basé su Resnet50 et entrainé sur le datset APTOS.

TABLE 4.7 – Mesures de performance obtenues avec le dataset APTOS pour la classification en 3 classes

Classe	Précision (%)	Rappel (%)	F1-score (%)
0	93	97	95
1	82	80	81
2	86	85	85



(a) APTOS (b) EyePACS

FIGURE 4.18 – Matrices de confusion pour le modèle 3 classes

TABLE 4.8 – Mesures de performance obtenues avec le dataset EyePACS pour la classification en 3 classes

Classe	Précision (%)	Rappel (%)	F1-score (%)
0	72	67	69
1	52	67	59
2	84	68	75

L'analyse de la matrice de confusion normalisée révèle une performance globalement satisfaisante du modèle de classification sur les trois classes considérées. La classe 0 (No RD) est prédite avec une précision remarquable de 97%, avec un taux de confusion négligeable vis-à-vis des autres classes (moins de 3%). La classe 2, correspondant aux cas sévères, est également bien identifiée avec un taux de classification correcte de 85%, bien que 15% des cas soient incorrectement attribués à la classe 1. En revanche, la classe 1, qui correspond à un stade intermédiaire, présente une performance moins élevée avec un taux de bonne

classification de 80 % et des confusions notables à la fois vers la classe 0 (7%) et vers la classe 2 (13 %). Cette asymétrie dans les erreurs, particulièrement marquée entre les classes 1 et 2, suggère une difficulté du modèle à distinguer des frontières diffuses entre les stades modérés et sévères de la pathologie. Plusieurs facteurs peuvent expliquer cette limitation, tels qu'un déséquilibre des classes dans les données d'apprentissage, une similarité visuelle accrue entre les échantillons des classes voisines, ou encore une incertitude liée à la qualité des annotations.

b) Classification en 5 Classes

Comme illustré à la Figure 4.19, pour la classification en cinq stades, un bloc de couches entièrement connectées est ajouté à l'extrémité de chaque réseau préentraîné. Après le prétraitement des images et l'optimisation des hyperparamètres, les modèles ResNet50 et VGG16 atteignent les meilleures performances sur les jeux de données APTOS et EyePACS.

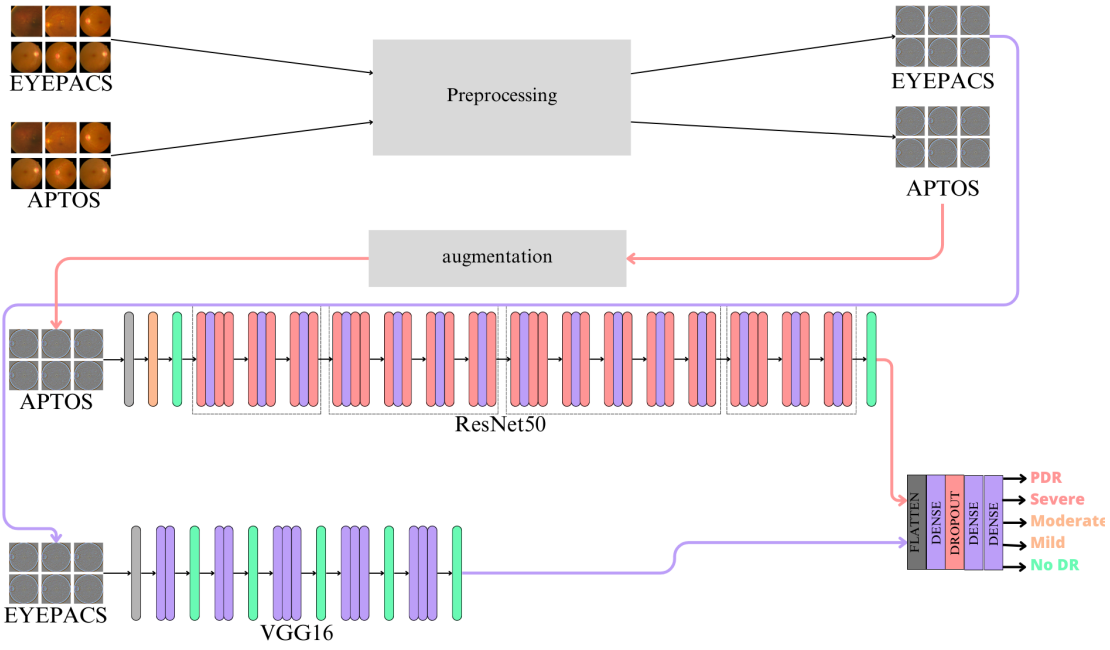


FIGURE 4.19 – AtR5C architecture

Comme indiqué dans les tableaux 4.9, les résultats obtenus ont atteint une précision de 85,42% sur l'ensemble de données APTOS en utilisant ResNet50, tandis que sur l'ensemble de données EyePACS, la précision était inférieure à 47,4% en utilisant VGG16.

TABLE 4.9 – Résultats obtenus par les meilleurs modèles pour la classification en 5 classes

	ResNet50	
APTOS	Accuracy	**0.8542**
	Precision	**0.8567**
	Recall	**0.8508**
	F1_score	**0.8537**

	VGG16	
EyePACS	Accuracy	**0.4740**
	Precision	**0.5136**
	Recall	**0.3780**
	F1_score	**0.4354**

Un aperçu de la performance prédictive de chaque modèle est fourni dans le tableau de métriques de performance par classe 4.10.

TABLE 4.10 – Mesures de performance obtenues avec le dataset APTOS pour la classification de la RD en 5 classes

Classe	Précision (%)	Rappel (%)	F1-score (%)
0	98	96	97
1	77	91	83
2	81	79	80
3	75	69	72
4	82	77	79

TABLE 4.11 – Mesures de performance obtenues avec le dataset EyePACS pour la classification de la rétinopathie diabétique en 5 classes

Classe	Précision (%)	Rappel (%)	F1-score (%)
0	45	30	36
1	37	58	45
2	33	37	35
3	57	45	50
4	69	61	65

c) Conclusion

Nous pouvons très bien constater la supériorité des résultats obtenus sur le dataset APTOS par rapport au dataset EyePACS. D'où le choix du dataset APTOS pour la suite des modèles proposés pour la classification de la RD. Les modèle AtR3C et AtR5C sont les modèles basés sur Resnet50 entrainés sur le dataset APTOS et en utilisant un ajustement automatique des hyper-paramètres.

4.3.3 Etude Comparative dans le Cas du Multi-classes

Cette section présente l'évaluation des modèles AtR3C et AtR5C pour la classification multiclasses de la sévérité de la rétinopathie diabétique (RD). Ces deux modèles s'appuient sur l'architecture de base ResNet50, entraînée avec des hyperparamètres optimisés par approche bayésienne, et démontrent des performances supérieures à celles obtenues avec les autres configurations de réseaux de neurones convolutifs testées. Le tableau 4.12 présente les métriques d'évaluation globales pour les deux tâches de classification. Le modèle AtR3C, dédié à la détection précoce de la DR via un schéma à 3 classes (Non DR, Early DR, Advanced DR), atteint une Accuracy de 94,26%, avec des valeurs équilibrées et élevées à travers la précision (94,41%), le rappel (94,09%) et le F1-score (94,24%). En parallèle, Le modèle AtR5C, conçu pour la classification complète de la RD, atteint une Accuracy de 85,42%,

avec des performances constantes sur la précision (85,67%), le rappel (85,08%) et le F1-score (85,37%).

L'écart de performance entre les modèles AtR3C et AtR5C (d'environ neuf points de pourcentage sur l'ensemble des métriques) reflète la complexité accrue de la classification en cinq classes, laquelle exige une discrimination plus fine entre les différents niveaux de sévérité de la rétinopathie diabétique. Néanmoins, la capacité du modèle AtR3C à distinguer avec une grande précision les cas sains, les stades précoces et les formes avancées de la maladie confirme sa pertinence pour une possible intégration dans des systèmes automatisés de détection précoce de la RD.

À l'inverse, le modèle AtR5C, bien que légèrement moins performant, permet une Classification fine et cliniquement pertinente. Sa capacité à différencier les stades de la RD Légère, Modérée, Sévère et Prolifération offre des informations cruciales pour la planification de traitements personnalisés dans des contextes cliniques spécialisés.

La robustesse des deux modèles se manifeste par un équilibre entre précision et rappel. Cette performance, qui limite à la fois les fausses alertes et les omissions, constitue un atout essentiel dans le domaine médical.

TABLE 4.12 – Valeurs de Performance pour les meilleures architectures AtR3C et AtR5C basées ResNet50 (unité %)

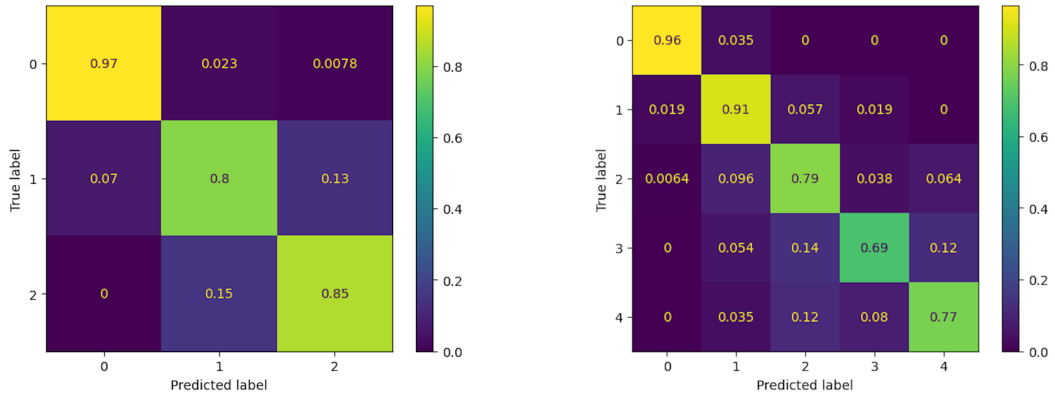
Metrics	AtR3C	AtR5C
Accuracy	94.26	85.42
Precision	94.41	85.67
Recall	94.09	85.08
F1_score	94.24	85.37

Une analyse détaillée des performances par classe est présentée dans le Tableau 4.13. Pour le modèle AtR3C, les taux de précision atteignent respectivement 93%, 82% et 86% pour les classes 0 (absence de RD), 1 (RD précoce) et 2 (RD avancée). Les valeurs de rappel correspondantes, 97%, 80% et 85%, mettent en évidence la capacité du modèle à réduire les faux négatifs, un aspect particulièrement crucial dans les contextes de dépistage.

Pour le modèle AtR5C plus granulaire, la précision (98%) et le rappel (96%) les plus élevés sont obtenus pour la classe 0 (No DR), ce qui souligne la précision du modèle dans l'identification des cas sains. Une performance modérée est observée pour les classes 1 (RD légère) et 2 (RD modérée), avec des scores F1 de 83% et 80%, respectivement. Les classes 3 (RD sévère) et 4 (RD proliférative) sont plus difficiles, comme prévu, avec des scores F1 de 72% et 79%, respectivement. Ces résultats confirment l'utilité clinique du modèle pour détecter les stades précoces et intermédiaires de la maladie, qui sont essentiels pour une intervention en temps opportun.

TABLE 4.13 – Évaluation des performances par classe des modèles AtR3C et AtR5C basés sur ResNet50 (en%)

Classes	Precision	Recall	F1_Score
3-class classification			
0	93.00	97.00	95.00
1	82.00	80.00	81.00
2	86.00	85.00	85.00
5-class classification			
0	98.00%	96.00	97.00
1	77.00%	91.00	83.00
2	81.00%	79.00	80.00
3	75.00%	69.00	72.00
4	82.00%	77.00	79.00



(a) (b)

FIGURE 4.20 – Matrices de Confusion de (a) AtR3C (b) AtR5C

La Figure 4.20 présente une vue détaillée des performances de classification par classe, illustrée par les matrices de confusion des modèles AtR3C et AtR5C. Pour le modèle AtR3C, les erreurs de classification sont minimales : seuls 2% des cas sains sont à tort étiquetés comme RD précoce et, point crucial, aucun cas sain n'est confondu avec une RD avancée. Ce résultat met en évidence la forte capacité discriminante du modèle ainsi qu'un taux particulièrement faible de faux positifs pour les classes critiques, essentiel dans les applications de dépistage.

Dans le cas de l'AtR5C, les erreurs de classification sont principalement confinées aux stades adjacents, ce qui est cliniquement justifiable compte tenu de la continuité morphologique entre les grades RD successifs dans les images rétiniennes. Plus précisément, seul 3,5% des cas de RD légère (classe 1) sont mal classés comme sains, et aucune erreur de classification n'est observée pour les stades les plus sévères (classes 3 et 4). Ce schéma traduit à la fois la sensibilité du modèle à des indices cliniques subtils et sa robustesse dans la détection

précoce des altérations pathologiques.

En résumé, les modèles AtR3C et AtR5C affichent tous deux des performances de classification élevées. Le modèle AtR3C se prête particulièrement bien au dépistage global de la RD grâce à sa sensibilité et sa spécificité élevées dans la distinction des stades précoces de la maladie. Inversement, le modèle AtR5C offre une discrimination plus nuancée à travers tous les grades de RD, ce qui en fait un outil précieux pour un diagnostic plus précis et un meilleur suivi des patients sur le long terme.

4.3.4 Comparaison avec les Méthodes de l'Etat de l'Art

Afin d'évaluer les performances de nos modèles proposés (AtRD, AtR3C, AtR5C), nous avons réalisé une comparaison avec les méthodes de l'état de l'art appliquées au jeu de données APTOS 2019. Les méthodes de référence retenues couvrent à la fois les réseaux de neurones convolutifs (CNN), les architectures hybrides et les modèles récents basés sur des transformeurs. Les Tableaux 5.14 et 5.15 présentent en détail les résultats obtenus pour les tâches de classification binaire, en 3 classes et en cinq classes. L'analyse ci-après se concentre sur quatre métriques : la précision globale (accuracy), la précision (precision), le rappel (recall) et le F1-score, qui reflètent conjointement les performances de classification.

a) Classification Binaire

Comme le montre le Tableau 5.14, plusieurs méthodes récentes ont atteint des performances élevées pour la classification binaire de la rétinopathie diabétique. L'apprentissage contrastif supervisé [84] a atteint une précision globale (accuracy) de 98,36%, tandis que l'architecture à double branche ResNet50–EfficientNetB0 [142] a rapporté une précision de 98,50% et un rappel de 99,46%. D'autres approches, incluant les ensembles de CNN [32], les réseaux à capsules [98] et les Swin Transformers [51], ont produit des résultats compétitifs mais légèrement inférieurs.

Notre modèle proposé AtRD, basé sur ResNet50 combiné à un ajustement automatisé des hyperparamètres par optimisation bayésienne, a surpassé toutes les méthodes précédentes avec une précision globale de 99,22%, une précision (precision) de 99,60%, un rappel (recall) de 99,23% et un F1-score de 99,41%. Ces gains illustrent le double effet bénéfique de l'apprentissage par transfert et de l'ajustement des hyperparamètres. En outre, l'équilibre observé entre les métriques renforce la pertinence clinique du modèle en réduisant simultanément le nombre de faux positifs et de faux négatifs.

b) Classification en Trois Classes

Peu d'études se sont intéressées à la classification de la RD en trois niveaux de sévérité à partir du jeu de données APTOS. Parmi elles, l'étude de Rao et al. [128] a atteint une précision globale (accuracy) de 88,00%, une précision de 88,00% et un rappel (recall) de

TABLE 4.14 – Comparaison de l'approche proposée avec des travaux antérieurs pertinents (classification en deux stades)

Architecture	**Année**	**Approche**	**Précision globale**	**Précision**	**Rappel**	**Score F1**
Rao et al. [128]	2020	Comparaison entre différents CNN	96.56	97.00	97.00	96.56
Bodapati et al. [32]	2020	Fusion de plusieurs CNN avec DNN	96.10	–	–	–
Kumar et al. [98]	2021	VGG16 combiné à des réseaux à capsules (DRISTI)	96.24	–	–	–
Islam et al. [84]	2022	Apprentissage contrastif supervisé	98.36	98.37	98.36	98.37
Dihin et al. [51]	2023	Approche basée sur Swin Transformer	96.00	–	98.00	96.00
Shakibania et al. [142]	2024	Fusion de ResNet50 et EfficientNetB0	98.50	97.61	99.46	–
Karthika et al. [91]	2024	Généralisation empilée de modèles profonds	97.77	100.00	96.00	98.00
Notre AtRD	2024	ResNet50 avec réglage automatique des hyperparamètres	**99.22**	**99.60**	**99.23**	**99.41**

88,02%. Toutefois, leur méthode reposait sur une approche CNN classique, sans intégrer les avancées récentes en matière de stratégies d'affinement (fine-tuning) ou d'optimisation des modèles.

En comparaison, notre modèle AtR3C a atteint une précision globale (accuracy) de 94,26%, avec des valeurs de précision, de rappel et de F1-score toutes supérieures à 94%. Cette amélioration significative des performances traduit une meilleure capacité à reconnaître les cas de RD qui, un défi bien connu dans le cadre de la classification en trois classes et valide la robustesse apportée par l'optimisation bayésienne des hyperparamètres.

c) Classification en Cinq Classes

La tâche de classification en cinq classes est particulièrement difficile en raison d'une variabilité intra-classe accrue et d'un déséquilibre des données. Comme l'illustre le Tableau 5.15, les premières méthodes basées sur des CNN, telles que ResNet50, InceptionV3 et Xception modifié [92], ont atteint des niveaux de précision globale (accuracy) modérés (variant de 74,64% à 79,59%), mais ont fourni des métriques incomplètes, omettant souvent le rappel ou le F1-score, ce qui limite leur interprétabilité clinique.

Les travaux ultérieurs, tels que ceux de Rao et al. [128] et de Bodapati et al. [33], ont proposé des architectures plus complexes présentant de meilleures performances (avec une précision globale (accuracy) atteignant 84,36%), mais sans fournir pour autant une

évaluation complète sur l'ensemble des métriques liées à la sensibilité. En revanche, les méthodes récentes reposant sur l'apprentissage par ensembles [121], l'apprentissage contrastif supervisé [84] et les architectures hybrides de type ViT [178, 122] rapportent des valeurs de précision globale plus élevées (≥ 85%), bien que certaines présentent un déséquilibre entre la précision et le rappel. Les méthodes récentes fondées sur les transformeurs ont permis d'améliorer les performances globales : par exemple, Oulhadj et al. [122] ont obtenu la précision globale (accuracy) la plus élevée (88,18%), mais avec un rappel relativement faible de 76%, traduisant une sensibilité réduite aux cas pathologiques.

En comparaison, notre modèle AtR5C a atteint une précision globale (accuracy) de 85,42%, une précision de 85,67%, un rappel (recall) de 85,08% et un F1-score de 85,37%, affichant ainsi un profil plus équilibré et cliniquement pertinent. Bien que légèrement inférieur en précision globale au meilleur résultat rapporté, notre modèle préserve un équilibre métrique essentiel, évitant ainsi les déséquilibres susceptibles de compromettre la fiabilité du diagnostique et du dépistage automatisé.

TABLE 4.15 – Comparaison de l'approche proposée avec des travaux antérieurs pertinents : classification en cinq classes de la RD (unité : %)

Architecture	**Année**	**Approche**	**Précision globale**	**Précision**	**Rappel**	**Score F1**
ResNet50 [92]	2019	ResNet50 de base	74.64	–	56.52	–
InceptionV3 [92]	2019	InceptionV3 de base	78.72	–	63.64	–
Kassani et al. [92]	2019	Xception modifié	79.59	–	82.35	–
Rao et al. [128]	2020	DL sur APTOS	84.36	70.51	73.84	–
Bodapati et al. [33]	2021	DNN + attention avec portes (gated-attention)	82.54	–	–	–
Kumar et al. [98]	2021	VGG16 + CapsuleNet (DRISTI)	75.50	–	–	–
Gangwar et al. [63]	2021	CNN + InceptionResNet-v2	75.50	–	–	–
Oulhadj et al. [121]	2023	Vote par ensemble (ensemble voting)	85.00	86.00	85.00	84.00
Islam et al. [84]	2022	Apprentissage contrastif supervisé	85.00	86.00	85.00	84.00
Notre AtR5C	2024	ResNet50 avec réglage automatique des hyperparamètres	**85.42**	**85.67**	**85.08**	**85.37**

4.4 Discussion

Les résultats expérimentaux démontrent clairement l'efficacité des modèles proposés pour la détection et la stadification de la rétinopathie diabétique (RD) à travers les différents types de classification. Le modèle AtRD, dédié à la classification binaire, a atteint des performances exceptionnelles avec une précision globale (accuracy) de 99,22% et un profil précision–rappel équilibré, le rendant particulièrement adapté à un dépistage précoce fiable. Ses faibles taux de faux positifs et de faux négatifs, confirmés par l'analyse des courbes ROC et des matrices de confusion, constituent un atout majeur pour une utilisation clinique, où la minimisation des erreurs diagnostiques est cruciale.

Pour la classification multiclasse, le modèle AtR3C, structuré autour de trois stades cliniquement pertinents (absence de RD, RD précoce, RD avancée), a présenté des performances élevées et équilibrées pour l'ensemble des métriques, avec une précision globale (accuracy) de 94,26% et une confusion minimale entre les cas sains et pathologiques. Ce modèle offre ainsi un outil efficace et interprétable pour le dépistage précoce et la priorisation des patients nécessitant un examen ophtalmologique approfondi.

Le modèle AtR5C, bien que présentant une précision globale (accuracy) légèrement inférieure (85,42%), propose une classification plus fine, conforme aux standards diagnostiques cliniques. L'analyse de sa matrice de confusion révèle que la plupart des erreurs de classification se situent entre des stades adjacents, ce qui est cohérent avec la similitude visuelle des images du fond d'œil dans les cas frontières. Il est important de noter que le modèle maintient sa robustesse même lors de la détection des stades sévères de RD, ce qui renforce son potentiel pour un suivi de la maladie.

Dans l'ensemble, les trois modèles proposés partagent des points forts clés :

(i) l'intégration d'architectures pré-entraînées robustes telles que **ResNet50**,

(ii) l'optimisation systématique des hyperparamètres d'apprentissage via une recherche bayésienne.

Cette combinaison aboutit à des modèles stables, dotés d'une forte capacité discriminante dans divers contextes de classification. Les résultats mettent en évidence un équilibre entre sensibilité et spécificité, ce qui est crucial en imagerie médicale pour éviter à la fois les faux positifs et les diagnostics manqués, qui peuvent avoir de graves conséquences pour les patients.

Ces résultats confirment que notre approche se généralise efficacement à des niveaux de complexité de classification variés et peut s'adapter à la fois aux applications de dépistage et de stadification de la RD dans des contextes diagnostiques réels.

4.5 Conclusion

Dans ce chapitre, nous avons proposé un ensemble de trois architectures d'apprentissage profond pour la détection et la classification de la rétinopathie diabétique (RD) en trois et cinq classes de sévérité, en tirant parti de l'apprentissage par transfert avec ResNet50 comme extracteur de caractéristiques robuste. ResNet50 s'est imposé comme le meilleur modèle sur le jeu de données APTOS, confirmant sa robustesse dans un contexte homogène.

L'une des contributions majeures de cetravail réside dans l'intégration d'une optimisation automatique des hyperparamètres via une recherche bayésienne, permettant à chaque modèle d'atteindre ses performances optimales sans ajustement manuel.

Nous avons mené une évaluation approfondie de chaque architecture, mettant en évidence leur capacité respective à identifier avec précision la présence et le degré de sévérité de la RD. Nos résultats indiquent que l'ajustement automatique des paramètres d'entraînement améliore significativement les performances des modèles, notamment en termes de précision globale (accuracy), de précision et de rappel (recall). Parmi les différentes architectures testées, les modèles basés sur ResNet50 se sont systématiquement distingués par leur précision de classification la plus élevée et par des profils métriques parmi les plus équilibrés.

La comparaison empirique avec les méthodes récentes de référence confirme l'efficacité et la compétitivité de notre approche. Nos modèles surpassent ou égalent les techniques de pointe appliquées au jeu de données APTOS, non seulement en termes de précision globale (accuracy), mais également en préservant un compromis optimal entre sensibilité et spécificité qui est un critère essentiel en diagnostic médical.

Dans le prochain chapitre, nous explorons l'intégration des Vision Transformers (ViT) pour proposer un modèle hybride d'amélioration de la classification de la rétinopathie diabétique. L'objectif est de tirer parti des caractéristiques locales et globales de l'image, en exploitant leurs relations mutuelles, pour obtenir une performance de classification supérieure.

CHAPITRE 5

Classification de la Rétinopathie Diabétique à l'Aide des Vision Transformers et d'une Approche Hybride CNN–ViT

5.1 Introduction

Les réseaux de neurones convolutifs (CNN), qui constituent l'ossature de nombreux modèles existants, se focalisent principalement sur l'extraction de caractéristiques locales à partir des images d'entrée. Cette approche, bien que performante pour la détection de motifs spatiaux restreints, limite leur capacité à modéliser efficacement les dépendances à longue portée et les relations contextuelles globales. Les Vision Transformers (ViTs) se sont imposés comme une alternative révolutionnaire en surmontant ces limitations grâce à l'utilisation de mécanismes d'auto-attention (self-attention), capables de capturer simultanément les dépendances à longue distance et les associations contextuelles globales à l'échelle complète de l'image.

Bien que les approches reposant sur l'apprentissage par transfert [115, 82] soient largement adoptées pour la classification de la RD, les méthodes actuelles peinent encore à atteindre une précision diagnostique optimale aux stades précoces de la maladie. En effet, la détection des lésions subtils (tels que les micro-anévrismes ou les hémorragies légères) exige à la fois une extraction fine et détaillée des caractéristiques locales et une compréhension contextuelle globale de l'image rétinienne.

Pour relever ces défis et évaluer l'efficacité des ViTs pour la classification de la RD, nous proposons dans ce chapitre :

1. Un modèle de classification de la RD basé sur les ViTs
2. Un nouveau modèle hybride qui est née de la fusion du du CNN (auto-tuned Resnet50)

et ViT, combinant leurs forces complémentaires.

3. Le modèle basé sur les ViTs et le modèle hybride ont été développés et évalués pour les trois types de classification de la RD : binaire, 3 classes et 5 classes. Il est à noter que, dans la littérature existante, l'approche de classification à 3 classes n'a pas encore été explorée, ce qui représente une contribution supplémentaire de notre travail [4, 6, 5].

5.2 Méthodologie

Cette section présente deux architectures d'apprentissage profond pour la classification de la RD. Chaque modèle a été entraîné pour la détection et la classification en trois et cinq stades de gravité. La première architecture proposée utilise les ViTS pour l'extraction de caractéristiques. ViRD, ViR3C et ViR5C traitent respectivement de la classification binaire, 3 et 5 classes. Puis, nous proposons des architectures Hybrides, ReVi-RD, ReVi-3C et ReVi-5C, pour la détection et la classification en 3 et 5 classes respectivement, combinant les forces des modèles auto-tuned basés Resnet50 (AtRd, AtR3C et AtR5C) développés au chapitre 5 et des Vits. Comme illustré à la figure 5.1, chaque modèle suit un pipeline similaire composé de plusieurs processus :

— Processus de prétraitement qui équilibre le jeu de données et améliore la qualité des images d'entrée.

— L'extraction de caractéristiques est effectuée en utilisant l'architecture choisie (Rsnet50 et Vit).

— Un réseau de neurones multicouches classe l'image en deux, trois ou cinq classes selon l'architecture .

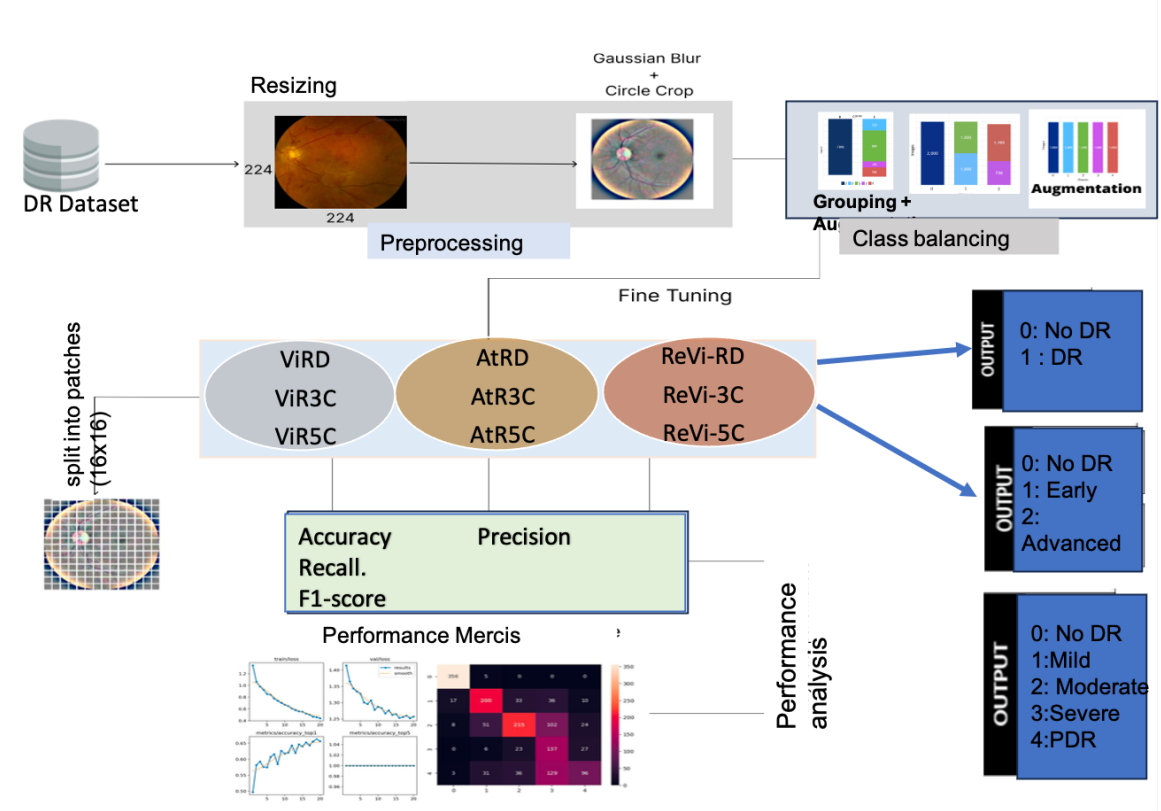


FIGURE 5.1 – Pipeline de l'approche proposée, du prétraitement des données à la prédiction de classe

Le processus de prétraitement a été discuté au chapitre 5 dans les sections Preparation des données, Augmentation des données et Prétraitement des images.

Dans ce chapitre nous allons uniquement nous focaliser sur l'extraction des caracteristiques et la classification pour les deux modèles Basés Vits et l'approche Hybride que nous avons développés.

5.2.1 Classification de la RD en utilisant ViRD, ViR3C et ViR5C : Approche Basée sur les Vits

Après avoir exploité les caractéristiques local des CNNs préentrainés, nous nous sommes intéressés à la capacité des ViTs à modéliser les dépendances à longue portée afin d'améliorer la détection et la classification de la RD. Nous avons proposé ViRD, ViR3C et ViR5C, trois architectures basées sur ViT pour la détection et la classification des RD. La figures 5.2 illustrent le modèle proposée pour les 3 types de classification.

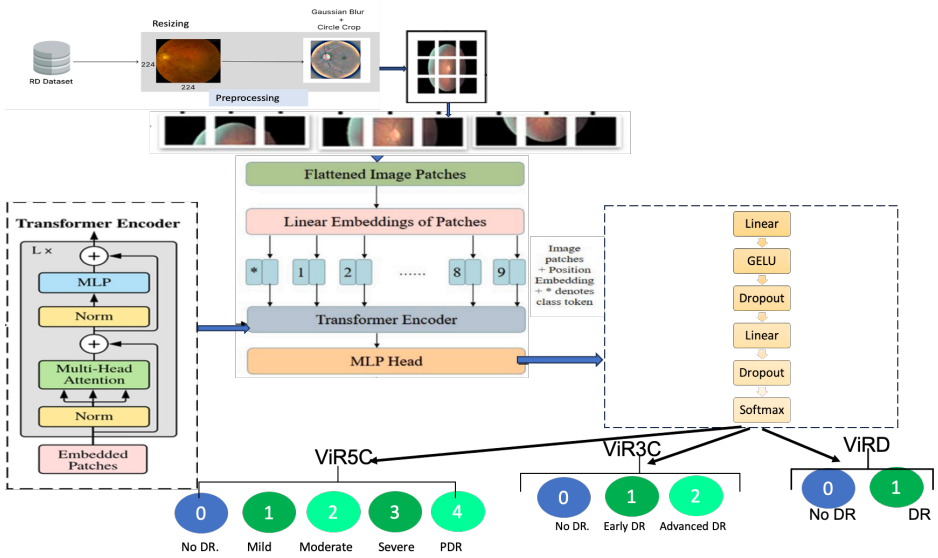


FIGURE 5.2 – Architecture proposée basée ViT : ViRD, ViR3C et ViR5C

Les composantes essentielles d'un Transformer sont le mécanisme d'attention multi-têtes (Multi-Head Self-Attention, MSA) et le perceptron multicouche (Multi-Layer Perceptron, MLP) (Voir chapitre 3). L'attention multi-têtes constitue le cœur opérationnel de l'architecture Transformer.

Le modèle *Vision Transformer* (ViT) traite une image en la décomposant en une série de *patchs* (sous-images) avant de les soumettre au pipeline d'apprentissage.

Les principales étapes de fonctionnement sont les suivantes :

1. **Découpage de l'image en patchs** : Après prétraitement et redimensionnement en 224×224, l'image d'entrée I est divisée en une série de patchs aplatis X_{ip} (pour $i = 1, 2, \ldots, n_p$), chacun de taille $p \times p \times C$, avec $C = 3$ correspondant aux trois canaux RGB de l'image I et $p = 16$, ce qui donne $n_p = \frac{224 \times 224}{16 \times 16} = 196$ patchs. Chaque patch X_{ip} est aplati et transformé en un vecteur 1D X_0 de dimension $p \times p \times 3 = 16 \times 16 \times 3 = 768$ via un *linear embedding*.

$$X_0 = \left[x_1, x_2, \ldots, x_{n_p}\right] \in \mathbb{R}^{196 \times 768} \tag{5.1}$$

2. **Projection linéaire des patchs (Patch Embedding)** : Chaque patch aplati est projeté dans un espace de dimension D au moyen d'une matrice d'apprentissage $\mathbf{E} \in \mathbb{R}^{768 \times D}$. Pour le i-ème patch $\mathbf{x}_i$, l'encodage est donné par $\mathbf{z}_i = \mathbf{x}_i \cdot \mathbf{E}$. La matrice $\mathbf{E}$ représente les poids de projection, où 768 est la dimension du patch aplati et D est la dimension de l'espace de projection. Cette dimension D correspond à celle des jetons d'entrée du Transformer, base du mécanisme d'auto-attention. Dans les ViT de base, D est couramment fixé à 768.

$$Z_0 = \left[z_1, z_2, \ldots, z_{n_p}\right] \in \mathbb{R}^{196 \times D} \tag{5.2}$$

3. **Initialisation du jeton de classification et des embeddings positionnels** : Comme illustré dans la Figure 5.2, les informations positionnelles $\mathbf{Pos} \in \mathbb{R}^{197\times D}$ sont ajoutées à chaque patch encodé, permettant au ViT de mieux comprendre les relations spatiales au sein des données d'entrée. Un jeton de classification ($z[\text{cls}]$) est inséré parmi les patchs encodés. Ce jeton est un paramètre aléatoirement initialisé et apprend. Il est utilisé pour agréger l'information globale en vue de la classification, agissant comme un décodeur.

   L'entrée de l'encodeur Transformer est construite comme suit :

$$Z = [z[\text{cls}], Z_0] \in \mathbb{R}^{(196+1)\times D} \tag{5.3}$$

   Après ajout de l'encodage positionnel, l'entrée finale devient :

$$Z_f = Z + \mathbf{Pos} \in \mathbb{R}^{197\times D} \tag{5.4}$$

   La matrice Z_f, enrichie des informations visuelles et positionnelles, est ensuite transmise à une pile d'encodeurs Transformer.

4. **Encodeurs Transformer** : L'encodeur Transformer est composé de deux couches principales : *Multi-Head Self-Attention* (MSA) et *Multi-Layer Perceptron* (MLP). La matrice Z_f est injectée dans une pile de six blocs encodeurs, chacun comprenant un module MSA à huit têtes d'attention, suivi d'un MLP. Une normalisation de couche et des connexions résiduelles sont appliquées avant et après chaque sous-couche.

Le mécanisme MSA est une forme d'auto-attention permettant au modèle de se concentrer simultanément sur différentes sous-espaces de représentation à diverses positions. Pour calculer les scores d'attention, le MSA emploie plusieurs mécanismes d'attention à produit scalaire normalisé.

$$MSA(Q, K, V) = \text{Concat}(h_1, h_2, \ldots, h_n) \cdot W_0 \tag{5.5}$$

où *Concat* désigne la concaténation des sorties des têtes d'attention ; n est le nombre de têtes, h_i étant la sortie de la i-ème tête. La sortie concaténée est ensuite projetée dans l'espace d'encodage original par la matrice W_0.

La sortie de chaque tête h_i est calculée comme suit :

$$h_i = \text{Attention}(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \tag{5.6}$$

où d_k est la dimension du vecteur clé K. La fonction softmax normalise les poids d'attention. Ces poids sont ensuite multipliés par la matrice de valeurs V_i pour obtenir la sortie d'auto-attention h_i.

Les vecteurs Q_i, K_i et V_i de chaque tête ($i \in \{1, \ldots, n\}$) sont obtenus en multipliant la

matrice d'entrée Z_f par trois matrices de poids distinctes :

$$Q_i = Z_f W_i^Q, \qquad K_i = Z_f W_i^K, \qquad V_i = Z_f W_i^V$$

Les sorties de toutes les têtes sont fusionnées et transmises à un MLP pour un traitement complémentaire. Chaque bloc MSA et MLP est précédé et suivi par des couches de normalisation et des connexions résiduelles afin de garantir stabilité et performance. Le MLP comprend deux couches linéaires entièrement connectées, séparées par une fonction d'activation non linéaire, souvent la fonction *Gaussian Error Linear Unit* (GELU), connue pour sa continuité et sa capacité à modéliser des motifs complexes [165] :

$$\text{GeLU}(x) = 0.5\,x \left[1 + \tanh\left(\sqrt{\frac{2}{\pi}}\left(x + 0.0447x^3\right)\right)\right] \tag{5.7}$$

Nous avons inséré deux couches de *dropout* pour régulariser le modèle et éviter le surapprentissage.

Enfin, le jeton $[Cls]$ est extrait de la sortie de l'encodeur Transformer et envoyé à une tête de classification pour obtenir les prédictions y. Pour la classification de la RD en 2, 3 ou 5 niveaux de sévérité, nous utilisons une couche de sortie comportant respectivement 2, 3 ou 5 neurones pour ViRD, ViR3C et ViR5C.

Une fonction softmax est appliquée afin de produire une distribution de probabilité pour la classification des images du fond d'œil :

$$y = \text{softmax}(z[\text{Cls}]) \tag{5.8}$$

5.2.2 Classification de la RD à l'aide de ReVi-RD, ReVi-3C et Revi-5C : une Approche Hybride Novatrice

Afin d'améliorer la précision de la classification de la DR, nous proposons une nouvelle architecture hybride qui combine les avantages des Vision Transformers (ViT) et de l'auto-tuned Resnet50. Les modèles ReVi-RD, ReVi-3C et ReVi-5C permettent de capturer simultanément les caractéristiques locales et globales des images rétiniennes en combinant les modèles préentraînés ViRD,ViR3C et ViR5C avec les modèles préentraînés AtRD, AtR3C et AtR5C.

Pour construire ce modèle hybride, nous utilisons les poids des modèles AtRD, AtR3C et AtR5C préentraînés afin d'extraire des caractéristiques locales. Nous avons supprimé le MLP (couche finale) de ces modèles et nous l'avons remplaçé par les modèles préentraînés ViRD, ViR3C ou ViR5C, comme décrit dans la Figure 5.3. Le composant ViT fonctionne donc comme un mécanisme de raisonnement global, opérant sur les représentations spatialement localisées fournies par le modèle basé sur ResNet50.

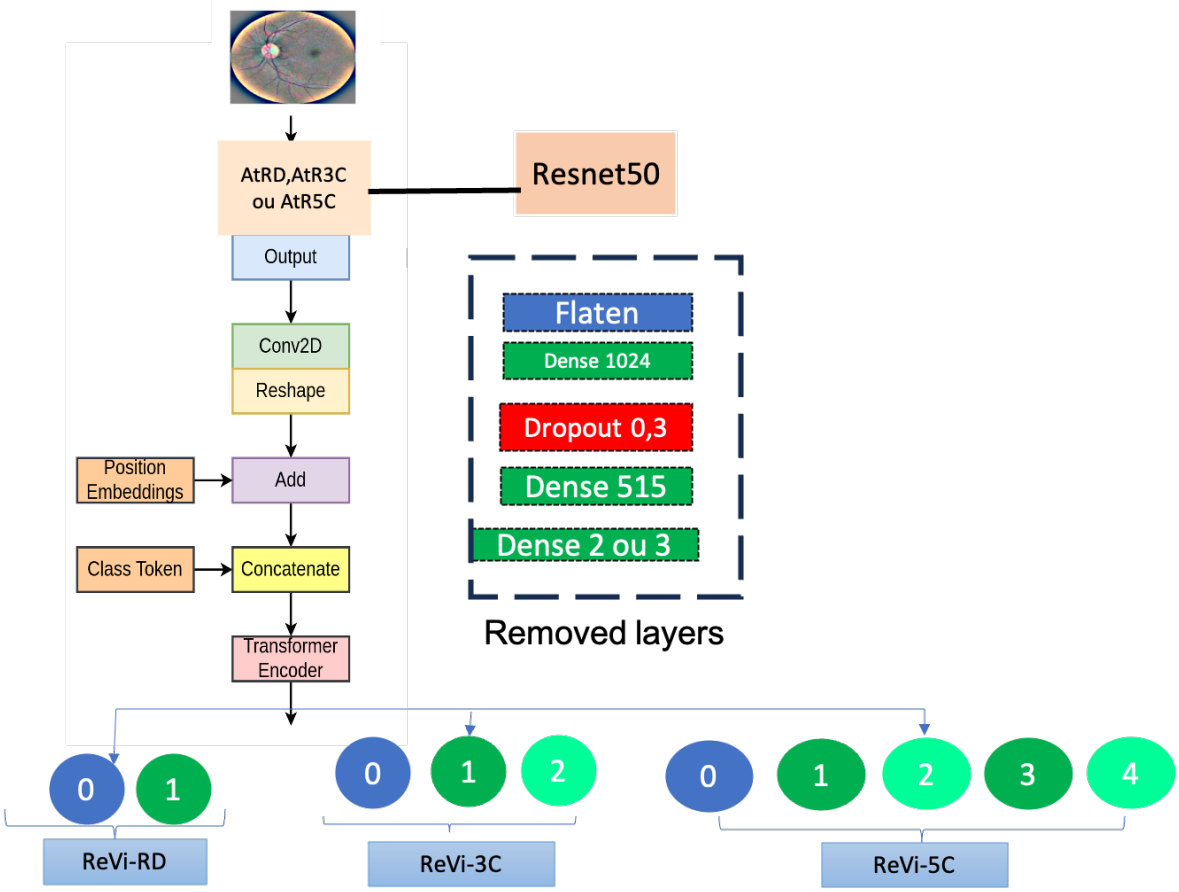


FIGURE 5.3 – Architectures hybrides : ReVi-RD,ReVi-3C et ReVi-5C

Pour aborder la complémentarité entre CNN et transformeur, notre architecture hybride est organisée en deux étapes principales. Dans ce qui suit, nous fournissons une description détaillée de cette approche hybride, illustrée à la Figure 5.4, couvrant l'ensemble du pipeline depuis l'image d'entrée jusqu'à la sortie de classification finale.

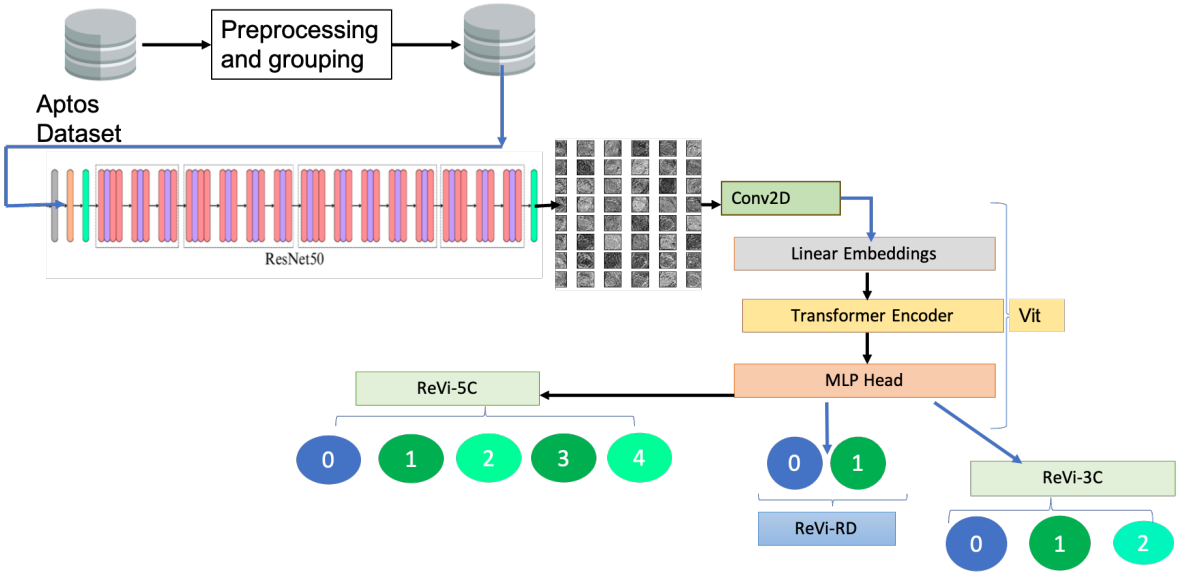


FIGURE 5.4 – Architecture détaillée de ReVi-RD, ReVi-3C et Revi-5C

1. **Étape 1 : Extraction locale de caractéristiques avec les modèles basés ResNet50 (AtRD, AtR3C ou AtR5C)**
   - **Entrée :** Une image RVB de 224×224 pixels, représentée par un vecteur de dimensions $[224, 224, 3]$ (taille d'entrée standard pour le modèle ResNet50), est introduite dans le modèle pré-entraîné après une phase de prétraitement.
   - **AtRD, AtR3C ou AtR5C :** Ces modèles sont utilisés pour extraire des caractéristiques spatiales locales à partir des images d'entrée. Les couches finales de classification d'AtRD, AtR3C ou AtR5C sont supprimées et remplacées par une tête basée sur un transformeur.
   - **Caractéristiques intermédiaires :** La sortie provenant d'une couche intermédiaire (plus précisément, la 7^{e} couche à partir de la fin) du modèle ResNet50 modifié est extraite. La carte de caractéristiques résultante présente des dimensions $7 \times 7 \times 768$. Elle conserve des représentations de haute dimension des motifs localisés tout en comprimant la résolution spatiale en une grille de 7×7, chaque cellule étant associée à 768 canaux.
2. **Étape 2 : Raisonnement global avec ViRD, ViR3C ou ViR5C**
   - **Restructuration pour le Vision Transformer (ViT) :** La carte de caractéristiques de dimensions $7 \times 7 \times 768$ est remodelée en une séquence de 49 patchs aplatis, produisant un tenseur de dimensions $[49, 768]$. La grille spatiale 7×7 est ainsi réinterprétée comme 49 patchs non chevauchants, chacun représenté par un vecteur de 768 dimensions. Cette étape prépare la sortie convolutionnelle pour un traitement par transformeur.
   - **Position embeddings et jeton de classification [CLS] :** Afin d'intégrer l'information spatiale dans le transformeur, des embeddings de position apprenables sont ajoutés aux 49 patchs. Un jeton de classification [CLS] apprenable est ensuite concaténé à la séquence, portant sa longueur à 50 ($[50, 768]$). La séquence résultante est ensuite traitée par un encodeur de type Transformer.
   - **Encodeur Transformer :** La séquence de longueur 50 est traitée par une série de 6 blocs encodeurs Transformer. Chaque bloc est constitué d'un mécanisme d'auto-attention multi-têtes (8 têtes), suivi d'un perceptron multicouche (MLP).
   - **Tête de classification :** Après passage dans l'encodeur Transformer et application d'une normalisation finale, la sortie correspondant au jeton [CLS] est extraite. Cette représentation est projetée dans l'espace des classes (2 classes pour ReVi-RD, 3 classes pour ReVi-3C et 5 classes pour ReVi-5C) via une couche dense. Les scores obtenus (logits) sont convertis en probabilités de classe à l'aide d'une fonction softmax.

Cette hybridation en deux étapes permet au modèle d'extraire dans un premier temps des descripteurs locaux robustes par convolution, puis d'effectuer une intégration globale des

caractéristiques et la classification via des mécanismes d'attention. Ce pipeline séquentiel assure une abstraction progressive des caractéristiques, allant du local au global, et produit des représentations à la fois sémantiquement enrichies et contextualisées spatialement.

5.3 Résultats d'Expérimentation

Cette section présente les résultats expérimentaux obtenus, en mettant en évidence le modèle le plus performant pour la classification de la rétinopathie diabétique (RD). Les expériences ont été réalisées dans un environnement Python sur la plateforme Kaggle, sur une machine équipée d'un processeur Intel Xeon @ 2,20 GHz, de 29 Go de RAM et d'un GPU NVIDIA Tesla P100 de 16 Go.

Le modèle a été entraîné sur la base de données APTOS, segmentée selon un ratio de 80% pour l'entraînement et 20% pour le test. Pour résoudre le problème de déséquilibre des classes, l'augmentation des données a été appliquée uniquement à l'ensemble d'entraînement, en veillant à ce que les échantillons générés artificiellement ne soient pas infiltrés dans les ensembles de validation ou de test.

Pour le modèle basé sur ResNet50, nous avons utilisé l'optimiseur Adam, tandis que le modèle basé sur ViT a fait appel à l'optimiseur AdamW. Nous avons employé l'entropie croisée catégorielle comme fonction de perte, ce qui est approprié pour notre tâche de classification multi-classe avec une activation softmax. Le taux d'apprentissage a été sélectionné automatiquement via une optimisation des hyperparamètres, et la valeur optimale obtenue a été de 0,0001 pour le modèle basé sur ResNet50 et de 0,00002 pour le modèle basé sur ViT. Cette valeur a été fixée durant l'entraînement afin d'assurer une convergence stable.

5.3.1 Métriques d'Evaluation

Pour évaluer la performance de détection des modèles proposés, nous utilisons les métriques les plus couramment utilisées : accuracy, précision, sensibilité, spécificité ou rappel(Recall) et le F1-score. Leurs expressions mathématiques sont données dans la Table 5.1 du chapitre "Revue de Littérature".

5.3.2 Performances des Modèles Basés Vit : ViRD, ViR3C et ViR5C

Après le prétraitement de l'image, nous avons affiné les modèles pour obtenir les meilleurs hyperparamètres, qui sont présentés dans la table 5.2.

Toutes les architectures proposées sont entraînées en utilisant leurs hyerparameters obtenus.

Table 5.1 – Performance Metrics

Métriques	**Formules**
Accuracy (Acc)	$Acc = \frac{TP+TN}{TP+TN+FP+FN}$
Precision (Positive Predictive Value)	$Precision = \frac{TP}{TP+FP}$
Recall (Sensitivity)	$Recall = \frac{TP}{TP+FN}$
F1 Score	$F1\_Score = \frac{2 \times Precision \times Recall}{Precision+Recall}$
Specificity (True Negative Rate)	$Specificity = \frac{TN}{TN+FP}$

Table 5.2 – Résultats de l'ajustement des Hyperparamètres pour les modèles basés Vit

Paramètres ter	Valeur
Taille de l'image	224x224
Taille des patches	16x16
Train Batch Size	32
Test Batch Size	64
warmup steps	500
Warmup learning rate	0.00001
Epochs	20
Learning rate	0.00002
weight decay	0.01

Performances des Modèles Basés Vits

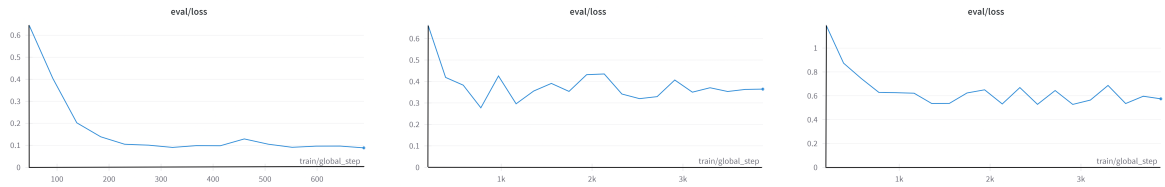
D'après les résultats présentés dans le Tableau 5.3, pour la détection et la classification de la RD, les modèles ViRD, ViR3C et ViR5C ont atteint une Accuracy remarquable de 97,73%, 92.97 et 87,33%, respectivement.

Table 5.3 – Performances de ViRD, ViR3C et ViR5C

Metrics	ViRD (%)	ViR3C (%)	ViR5C (%)
Accuracy	97.73	92.97	87.33
Précision	97.72	93.77	87.17
Recall	97.73	93.22	85.66
F1_score	97.73	93.46	86.26
Specificity	98.00	96.60	92.29

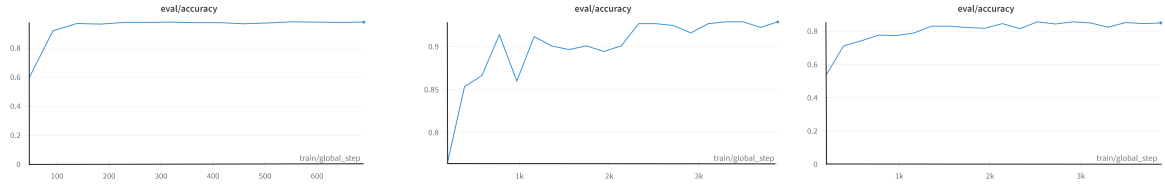
Les figure 5.5 et 5.6 indiquent que les modèles s'adaptent bien aux données et ne sont pas surajustés.

Le tableau 5.4 présente les métriques clés : précision, rappel, F1-score et spécificité, par classe, pour chaque architecture. Nous allons analyser et examiner successivement chaque modèle pour souligner ses points forts, identifier ses limites et comparer leurs performances.



(a) (b) (c)

FIGURE 5.5 – Courbes de perte des modèles (a) ViRD. (b) ViR3C. (c) ViR5C



(a) (b) (c)

FIGURE 5.6 – Les courbes de precision globale. (a) ViRD. (b) ViR3C. (c) ViR5C

ViRD (2 classes : No DR/DR) : l'architecture ViRD obtient d'excellentes performances pour les deux classes, dépassant les 97,7 % en termes de précision, de rappel et de F1-score par classe. La spécificité est également très élevée, confirmant que le modèle est tout aussi excellent pour identifier correctement les négatifs que les positifs. En terme clinique, cela signifie que les faux négatifs (patients malades identifiés comme sains) sont extrêmement rares, un aspect crucial pour éviter des diagnostics manqués et préserver la prise en charge en temps utile. Ce modèle est très fiable pour une tâche binaire ce qui lui permet d'être une base robuste pour un système de dépistage automatisé de la RD.

VIR3C (Classification en 3 stades de gravité : ViR3C affichent une performance bien équilibrée en termes de précision, de rappel et de F1-score, avec des scores avoisinant les 94% et 93% respectivement sur l'ensemble des métriques et pour toutes les catégories de sévérité. Pour la classe « No DR », les métriques sont quasi parfaites, précision, rappel et F1-score s'établissent à 0,98, avec une spécificité de 0,99, offrant une assurance solide sur la fiabilité diagnostique pour les cas sains. Pour la classe « Early DR », la précision atteint 0,91, le rappel 0,94, et le F1-score 0,92, traduisant une capacité robuste à détecter les formes précoces, souvent les plus difficiles à reconnaître . Quant à la classe « Advanced DR », elle se distingue également avec précision de 0,93, rappel de 0,88 et F1-score de 0,90, mettant en évidence la capacité du modèle à reconnaître les formes les plus graves malgré leur complexité visuelle. Cette performance équilibrée reflète une architecture bien calibrée capable de capter à la fois les détails fins et les signes plus marqués de la RD.

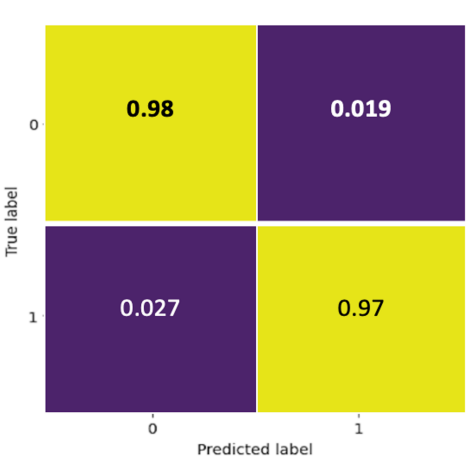
TABLE 5.4 – Performance par classe des modèles ViRD, ViR3C et ViR5C

	ViRD		ViR3C			ViR5C				
Metrics	0	1	0	1	2	0	1	2	3	4
Precision	0.97	0.98	0.98	0.91	0.93	0.97	0.87	0.82	0.85	0.85
Recall	0.98	0.97	0.98	0.94	0.88	0.98	0.89	0.88	0.71	0.82
F1-score	0.98	0.98	0.98	0.92	0.90	0.98	0.88	0.85	0.77	0.84
Specificity	0.98	0.98	0.99	0.92	0.97	0.98	0.95	0.87	0.98	0.98

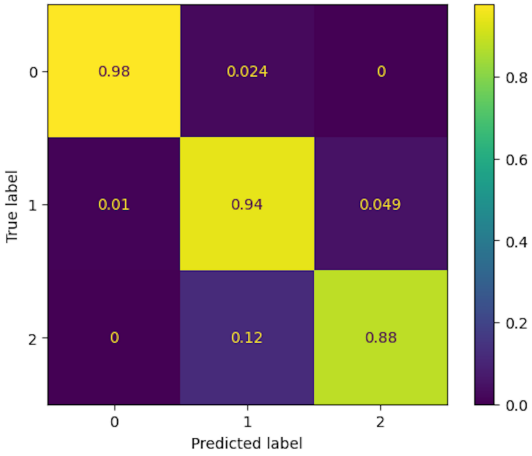
VIR5C (Classification en 5 stades de gravité) : Avec un niveau de détail accru, ViR5C excelle dans la détection des patients sains (classe 0), avec une précision, un rappel, un F1-score et une spécificité tous supérieurs à 0,97(classe 0). Ces résultats garantissent une confiance dans l'identification des cas sans pathologie, minimisant les risques de surtraitement clinique. Pour les stades précoces (stade 1 et 2), les résultats sont performants. Le stade 1 obtient une précision de 0,87, un rappel de 0,89 et un F1-score de 0,88, tandis que le stade 2 atteint une précision de 0,82, un rappel de 0,88 et un F1-score de 0,85. Ces valeurs indiquent que le modèle est meilleur pour détecter les cas précoces (rappel élevé), avec une légère sur-alarme (précision plus modeste), ce qui est acceptable en contexte clinique puisqu'un suivi vigilant peut rapidement corriger les faux positifs. Les stades sévères (3 et 4) présentent des performances un peu plus modérées. Le stade 3 affiche une précision de 0,85, un rappel de 0,71 et un F1-score de 0,77, tandis que le stade 4 atteint un rappel plus élevé de 0,82 (précision de 0,85, F1-score de 0,84). Compte tenu du fait que ces stades sont à la fois rares et souvent confondus en raison de signes visuels similaires (prolifération néovasculaire). Le modèle reste pertinent, notamment dans le dépistage des cas les plus urgents, tout en identifiant correctement les formes intermédiaires dans la majorité des cas.

Afin d'évaluer l'efficacité des modèles (ViRD, ViR3C et ViR5C), nous avons examiné les matrices de confusion (voir Figure 5.7), pour fournir des détails sur la distribution des erreurs et la précision de classification à travers les différentes classes de gravité.

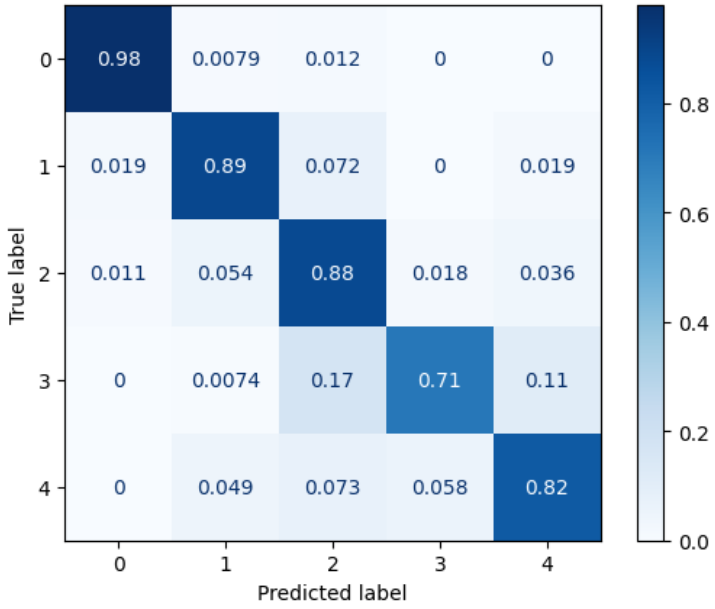
La matrice de confusion, qui illustre les prédictions de notre modèle ViRD (voir Figure 5.7a, montre des résultats très encourageants pour la détection de la rétinopathie diabétique. La faible proportion (2%) de patients malades classés à tort comme sains est particulièrement remarquable. Cela souligne l'efficacité du modèle à minimiser les erreurs critiques et à garantir un diagnostic fiable pour la majorité des patients.



(a) Matrice de confusion de ViRD



(b) Matrice de confusion de ViR3C



(c) Matrice de confusion de ViR5C

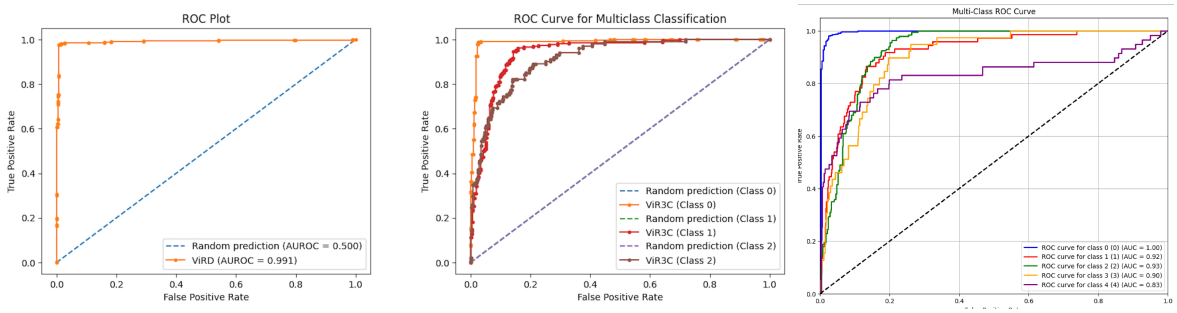
FIGURE 5.7 – Matrices de confusion des modèles (a) ViRD, (b) ViR3C, et (c) ViR5C.

Par contre pour le modèle ViR3C, nous pouvons constater que la matrice de confusion (voir Figure 5.7b révèle une grande fiabilité, en particulier pour l'identification des patients en bonne santé ou à un stade précoce de la maladie. Le modèle excelle à détecter les patients sains, en classant correctement 98% des images saines. Cette performance est cruciale, car elle limite considérablement les traitements inutiles. Pour les cas de RD à un stade précoce (classe 1), le modèle identifie correctement 94% des images. Un faible pourcentage de 1% est cependant mal classé comme sain, ce qui constitue un risque de diagnostic manqué. En ce qui concerne les stades avancés (classe 2), le modèle a une précision de 88%. Bien que 12% des cas soient confondus avec le stade précoce, aucun n'est classé à tort comme sain. Ces résultats confirment que le modèle est très prometteur pour le diagnostic précoce de la RD. Les erreurs de classification sont concentrées sur les stades où les symptômes sont les

plus subtils.

La matrice de confusion pour le modèle ViR5C (voir Figure 5.7c révèle une grande précision dans la détection des patients ne souffrant pas de rétinopathie diabétique, classant correctement 98% des images de la classe 0. Pour la détection précoce, le modèle a correctement identifié 89% des patients avec une RD de classe 1 et 88% de ceux de la classe 2. Le modèle a mal classé 1,9% des patients de classe 1 comme étant sains et 1,1% des patients de classe 2 comme étant sains, sans aucune erreur pour les autres stades.

Ces résultats mettent en évidence le potentiel du modèle à diagnostiquer avec précision la RD à un stade précoce et démontrent que la faible erreur de classification concerne principalement les stades 1 et 2, où les symptômes sont difficiles à identifier.



(a) (b) (c)

FIGURE 5.8 – Les courbes Auc-Roc. (a) ViRD. (b) ViR3C. (c) ViR5C

Les courbes ROC fournissent une confirmation visuelle de la performance des modèles. Les courbes Auc-Roc dans la Figure 5.8, montre la performance des modèles ViRD, ViR3C et ViR5C est clairement validée. Le modèle ViRD excelle dans la détection binaire de la maladie avec une AUC de 0,991, ce qui confirme sa fiabilité pour un dépistage initial. Les modèles ViR3C et ViR5C se distinguent par leur capacité à offrir une stadification plus fine : le ViR3C, avec des AUC supérieures à 0,90 pour chaque classe, démontre sa robustesse dans la classification en trois stades, tandis que le ViR5C, malgré une légère baisse de performance pour les cas les plus sévères, reste très efficace pour les stades précoces. Dans l'ensemble, ces résultats démontrent la pertinence clinique des trois architectures, chacune répondant à des besoins de diagnostic distincts, du dépistage à la surveillance détaillée de la progression de la maladie.

Comparaison de ViRD, ViR3C et ViR5C avec d'autres travaux

La performance des modèles ViRD et ViR5C, proposés pour la détection et la classification de la rétinopathie diabétique (RD) à l'aide de ViT, a été évaluée et comparée aux modèles récents de la littérature. Par contre le cas de classification en 3 classes n'a pas été abordé dans la littérature.

Nous avons comparé notre méthode à celle de Dihin et al. [51], qui ont utilisé le transformeur Swin-T pour la classification binaire de la RD. Dans la phase d'amélioration, ils

ont appliqué les filtres CLAHE et de lissage gaussien sur des images de 224x224, atteignant une précision, un rappel et un F1-score de 96%. Pour la classification de la RD, Yang et al. [178] ont proposé un modèle de transformeur avec apprentissage multi-instances (TMIL), qui a obtenu une précision de 85,6% et un rappel de 73,7%. Ils ont démontré que leur modèle améliore les résultats obtenus en utilisant ViT avec une taille d'image de 224, atteignant une précision de 83,5% et un rappel de 76,7%. Pour la classification binaire, ils ont utilisé le jeu de données Messidor-1 et ont atteint une précision de 93,1%.

Les (ViT) nécessite généralement de grands ensembles de données pour apprendre efficacement, ce qui a limité son application dans la RD. Pour résoudre ce problème, Nazih et al. [117] ont pré-entraîné leur modèle ViT en utilisant le jeu de données ImageNet et un très grand jeu de données de RD annoté (FGADR). Ils ont atteint une performance de 82,5% sur l'ensemble des métriques d'évaluation.

Les résultats affichés dans le Tableau 5.5 indiquent clairement que nos approches proposées (ViRD, ViR3C et ViR5C) obtiennent de meilleures performances sur toutes les métriques par rapport aux méthodes existantes.

TABLE 5.5 – Comparaison des performances des modèles ViRD, ViR3C et ViR5C avec des travaux récents pour la classification binaire et multi-classes de la RD (unités en %).

Modèle	**Classes**	**Précision**	**Rappel**	**F1-score**	**Exactitude**
Classification binaire					
Dihin et al. [51]	2	96.00	–	96.00	96.00
Yang et al. [178]	2	93.20	–	86.90	–
Lian et al. [102]	2	95.30	–	94.20	–
Notre ViRD	2	**97.72**	**97.73**	**97.73**	**97.73**
Classification multi-classes					
Classification à 3 classes					
Notre ViR3C	3	textbf92.97	**93.77**	**93.22**	**93.46**
Classification à 5 classes					
ViT [178]	5	83.50	–	67.30	–
Nazih et al. [117]	5	82.50	82.25	82.25	82.50
Yang et al. [178]	5	85.60	–	73.70	–
Notre ViR5C	5	**87.33**	**87.17**	**85.66**	**86.26**

5.3.3 Performances de nos Mdèles Hybrides : ReVi-RD, ReVi-3C et ReVi-5C

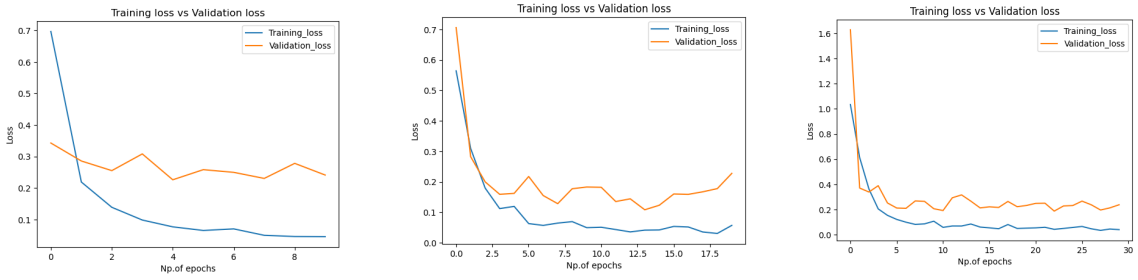
Pour évaluer la capacité des modèles ReVi-RD, ReVi-3C et ReVi-5C à détecter et à classer la RD, nous avons mené une expérimentation en utilisant un ensemble de test de 1545 images du jeu de données APTOS. Comme le présente le Tableau 5.6, notre approche a atteint une exactitude (accuracy) impressionnante de 99,55% pour la détection de la RD, de 98.26%

pour la classification en 3 stades de gravité et de 96,21% pour la classification en 5 stades de gravité. Les scores élevés et équilibrés de précision, de rappel et F1-score confirment la fiabilité globale des modèles.

TABLE 5.6 – ReVi-RD,ReVi-3C et ReVi-5C Performance

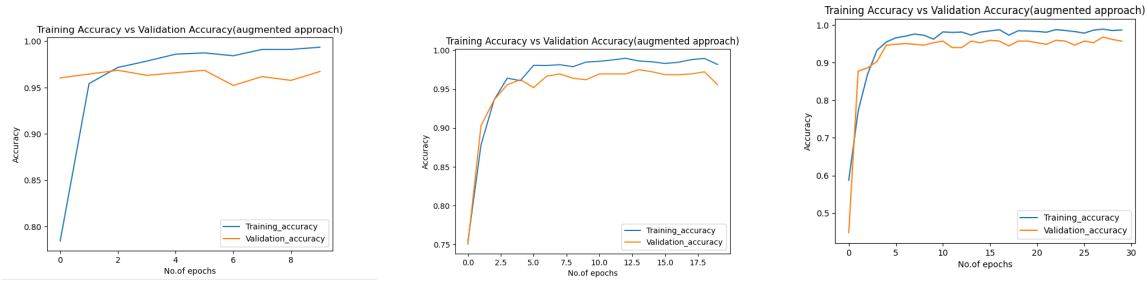
Metrics	ReVi-RD (%)	ReVi-3C (%)	ReVi-5C (%)
Accuracy	99.55	98.26	96.21
Precision	99.51	98.43	96.08
Recall	99.58	98.21	95.59
F1_score	99.54	98.32	95.80
Specificity(Average)	99.00	99.00	98.97

Les figure 5.9 et 5.10 indiquent que les modèles s'adaptent bien aux données et ne sont pas surajustés.



(a) (b) (c)

FIGURE 5.9 – Courbes de perte des modèles (a) ReVi-RD. (b) ReVi-3C. (c) ReVi-5C



(a) (b) (c)

FIGURE 5.10 – Courbes precision globale (Accuracy) des modèles : (a) ReVi-RD. (b) ReVi-3C et (c) ReVi-5C

Le tableau 5.7 présente les métriques obtenues par classe, pour chaque architecture hybride développée. Nous allons analyser et examiner successivement chaque modèle pour souligner ses points forts, identifier ses limites et comparer les performances.

Les trois modèles démontrent des performances exceptionnelles sur la tâche de classification de la RD, avec légère diminution des performances malgré l'augmentation de la

complexité du problème (5 classes). Cette robustesse illustre l'efficacité de la conception hybride.

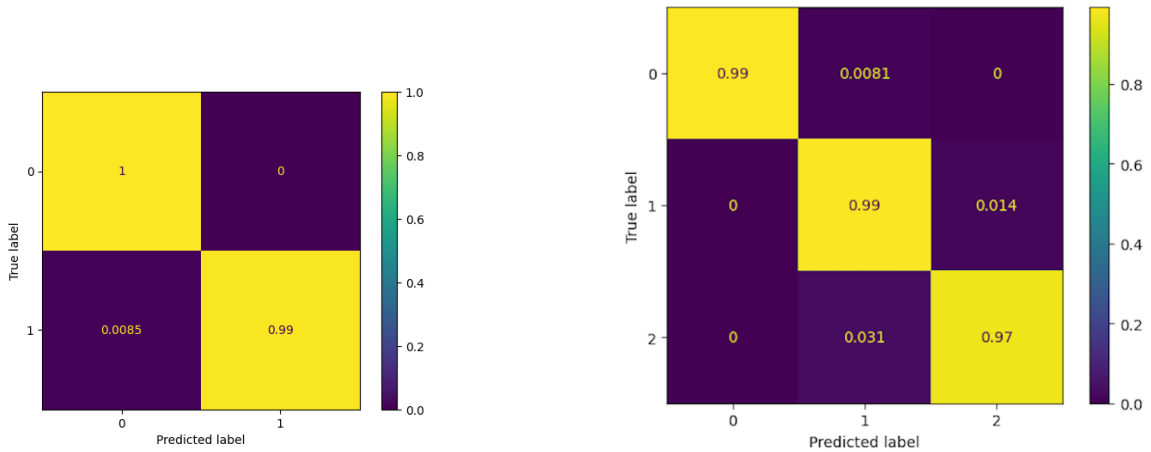
Le modèle ReViRD démontre une performance exceptionnelle pour la classification binaire de la RD. Nous observons que les performances obtenues sur les deux classes sont quasiment optimales, sans aucune erreur de classification. Concernant la classe 0, le modèle affiche une précision de 0,99 et un Rappel de 1,00, aboutissant à un F1-score parfait de 1,00. Cela signifie que le modèle ne classe pratiquement aucun patient sain comme malade. Pour la classe 1, les performances sont tout aussi impressionnantes, avec une symétrie parfaite qui confirme la robustesse de l'approche.

Le modèle ReVi-3C démontre une performance exceptionnelle pour la classification de la rétinopathie diabétique en trois classes. Il s'impose dans la détection des cas sains, atteignant un taux de précision exceptionnel de 100%. Il conserve également une précision élevée (0,98 pour les deux stades) et un taux de rappel notable (0,99 et 0,97 respectivement) pour les classe 1 et 2, ce qui se manifeste par d'excellents résultats du F1-score (0,98 et 0,97). Cette robustesse s'étend même à sa capacité de distinguer ces classes avec des spécificités de 0,97 et 0,99. Ainsi, le modèle est non seulement fiable pour le dépistage, mais il est aussi capable de fournir une stadification précise, même pour des cas cliniquement complexes.

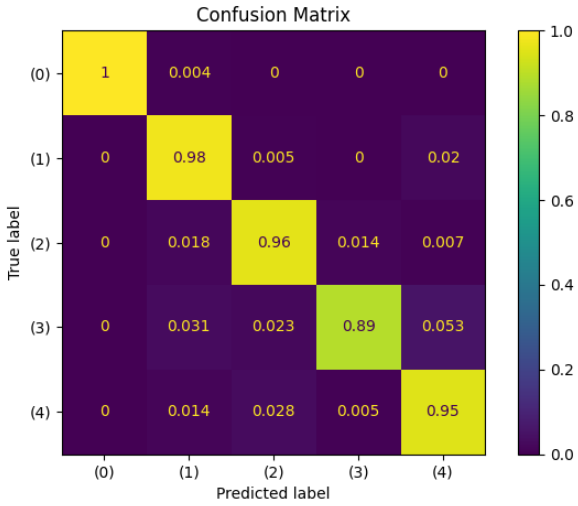
TABLE 5.7 – Performance par classe des modèles en (%) : ReviViRD, ReVi-3C et ReVi-5C

	ReViRD		ReVi-3C			ReVi-5C				
Metrics	0	1	0	1	2	0	1	2	3	4
Précision	0.99	1.00	1.00	0.98	0.98	1.00	0.94	0.96	0.96	0.94
Rappel	1.00	0.99	0.99	0.99	0.97	1.00	0.98	0.96	0.89	0.95
F1-score	1.00	1.00	1.00	0.98	0.97	1.00	0.96	0.96	0.92	0.95
Spécificité	1.00	0.99	1.00	0.97	0.99	1.00	0.99	0.97	1.00	0.99

Le modèle ReVi-5C démontre une performance exceptionnelle pour la classification de la RD en cinq stades distincts. Pour la classe 0, il atteint des scores parfaits sur toutes les métriques, ce qui garantit une exactitude totale dans le diagnostic des cas sains. Les stades précoces (1 et 2), affichent une performance robuste avec un F1-score de 0,96 pour les deux classes. Bien que le rappel pour le stade 1 (0,98) soit légèrement supérieur à celui du stade 2 (0,96), la spécificité reste très élevée (0,99 et 0,97 respectivement), ce qui témoigne d'une excellente capacité à différencier ces stades des autres. Les stades avancés (3 et 4) attestent de sa fiabilité avec des scores F1 de 0,92 et 0,95, présentant une spécificité impeccable pour le stade 3. Ces résultats soulignent l'aptitude du modèle à fournir une classification précise, indispensable pour le suivi et la prise en charge individualisée des patients.



(a) (b)



(c)

FIGURE 5.11 – Matrices de confusion des modèles hybrides (a) ReVi-RD, (b) ReVi-3C, et (c) ReVi-5C.

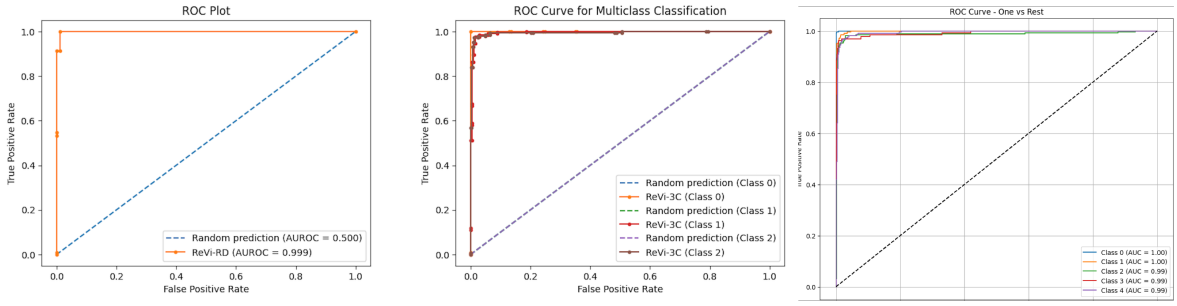
Afin d'évaluer l'efficacité des modèles développés (ReVi-RD, ReVi-3C et ReVi-5C), nous avons examiné les matrices de confusion (voir Figure 5.11), pour fournir des détails sur la distribution des erreurs et la précision de classification.

La matrice de confusion du modèle ReVi-RD (voir Figure 5.11a, montre une efficacité quasi-parfaite pour le dépistage binaire, avec un taux de faux négatifs quasi nul, ce qui est crucial pour ne pas omettre un diagnostic. Comme l'illustre la matrice de confusion de la Figure 5.11b, le modèle hybride ReVi-3C parvient à une classification quasi sans faille pour tous les niveaux de gravité : il atteint une précision de 99% pour les classes 0 et 1, et de 97% pour la classe 2, marquant ainsi une augmentation substantiel en exactitude. Les erreurs de classification sont réduites à des niveaux négligeables, avec seulement 3% des cas de classe 2 incorrectement étiquetés comme classe 1, tandis que la confusion entre les classes 0 et 1 est pratiquement éliminée. Ces résultats soulignent le rôle crucial des architectures hybrides dans la résolution des défis de classification multi-classe, où des différences subtiles entre les

classes exigent une discrimination précise.

De la même manière, la matrice de confusion du modèle ReVi-5C (voir Figure 5.11c démontre l'exactitude exceptionnelle du modèle dans l'identification des patients sans rétinopathie diabétique (RD). Seulement 0,4% des patients sains ont été mal classés, ce qui souligne le risque minimal de traitement inutile pour les patients non atteints de RD. En d'autres termes, nous pouvons voir qu'aucun patient malade n'a été classé comme sain. Les résultats montrent un taux d'erreur de classification de seulement 0,5 % entre les classes 1 et 2, ce qui suggère que le modèle est très précis pour distinguer ces deux stades. Cependant, nous observons un taux d'erreur de 2% pour les patients de classe 2 qui sont incorrectement classés comme classe 1 ou 3. Cette petite divergence pourrait être attribuée à l'annotation des images au sein du jeu de données. Ces résultats illustrent la capacité du modèle hybride à détecter la rétinopathie diabétique à un stade précoce, ce qui est crucial pour un traitement précoce et efficace de la maladie.

Les courbes ROC fournissent une confirmation visuelle de la performance des modèles. la Figure 5.12, montre la performance des modèles ReVi-RD, ReVi-3C et ReVi-5C est clairement validée. Le modèle ReVi-RD excelle dans la détection binaire de la maladie avec une AUC de 99.9%, ce qui confirme sa fiabilité pour un dépistage initial. Quant aux modèles plus fins, ReVi-3C atteint une AUC supérieure à 0,99 pour chacune des trois classes, preuve de sa robustesse dans la classification en trois stades de gravité. Enfin, ReVi-5C confirme son efficacité sur les cinq stades de gravité avec des AUC supérieures à 0,97, attestant de sa capacité à différencier avec succès même les stades avancés



(a) (b) (c)

FIGURE 5.12 – Les courbes Auc-Roc. (a) ReVi-RD. (b) ReVi-3C. (c) ReVi-5C

5.3.4 Comparaison entre Nos Différents Modèles Développés

Dans cette section nous allons faire une étude Comparative entre les différents modèles que nous avons proposés.

a) Cas de detection de la RD

Dans une première expérience, nous avons comparé les performances des modèles AtRD, ViRD et ReVi-RD afin d'évaluer leur efficacité dans la détection de la rétinopathie diabétique (RD) et d'analyser l'impact des caractéristiques extraites par chacun d'eux. Les

résultats présentés dans le Tableau 5.8 résument les métriques d'évaluation obtenues pour cette tâche.

TABLE 5.8 – Evaluation des performances des modèles proposés pour la classification de la RD en 2 classes

Metrics	**AtRD**	**ViRD**	**ReVi-RD**
Accuracy	99.22%	97.73%	99.55%
Precision	99.66 %	97.72%	99.51%
Recall	99.23%	97.73%	99.58%
F1_Score	99.41%	97.73%	99.54%

La performance de détection des modèles AtRD, ViRD et ReVi-RD est comparée à l'aide de leurs matrices de confusion et des métriques d'évaluation résumées dans le Tableau 5.8. Le modèle AtRD atteint une sensibilité élevée pour la détection de la rétinopathie (un taux de vrais positifs de 99,2%), mais présente une spécificité de 96,81%, correspondant à un taux de faux positifs de 3,2% pour les patients sains. Bien que cela souligne son efficacité à identifier les cas pathologiques, le taux élevé de diagnostics erronés pour les patients sains met en évidence ses limitations à distinguer des variations non pathologiques subtiles. En revanche, le modèle ViRD démontre une spécificité plus équilibrée (98%), ainsi qu'un taux de faux négatifs légèrement réduit de 2,7% pour les cas de rétinopathie, le modèle ViT excelle à capturer le contexte global grâce à l'auto-attention, il peut parfois manquer des caractéristiques locales subtiles qui sont pourtant essentielles à l'identification de la RD. Cette dépendance au contexte global implique que, dans les cas où les signes pathologiques sont très localisés ou subtils, le modèle pourrait ne pas les distinguer suffisamment des variations normales. L'architecture hybride ReVi-RD résout les limitations inhérentes aux modèles autonomes en combinant de manière synergique l'extraction de caractéristiques locales basée sur les CNN (AtRD) et la modélisation des dépendances globales des ViT (ViRD).

Les métriques spécifiques par classe (Tableau 5.9) confirment ces distinctions : l'AtRD présente une harmonisation modérée de la précision et du rappel (scores F1 de 97,7% pour les deux classes) en raison de sa focalisation sur les textures localisées, tandis que le ViRD améliore l'équilibre avec des scores F1 de 98,00% via l'attention globale, tout en restant vulnérable aux oublis de détails localisés.

Le modèle hybride ReVi-RD résout ces compromis, atteignant des métriques quasi parfaites, avec un F1-score de 100% pour les deux classes, une précision et un rappel de 99 à 100%. Sa supériorité provient de la synergie entre l'extraction de caractéristiques localisées d'AtRD et la modélisation du contexte global de ViRD. Cette intégration aboutit à une classification quasi-parfaite : un taux de faux négatifs de 1,0% et un taux de faux positifs de 0,0% pour les cas sains. Avec une spécificité de 99,50%, ReVi-RD minimise les diagnostics inutiles, surpassant la robustesse de l'AtRD et du ViRD.

b) Classification en 3 stades de gravité

TABLE 5.9 – Performance par classe des modèles proposés pour la détection DR (%)

	AtRD		ViRD		ReVi-RD	
Metrics	Class 0	Class 1	Class 0	Class 1	Class 0	Class 1
Precision	97.60	97.90	97.00	98.00	99.00	100.00
Recall	97.90	97.60	98.00	97.00	100.00	99.00
F1_score	97.70	97.70	98.00	98.00	100.00	100.00
Specificity	99.21	96.81	98.00	98.00	100.00	99.00

D'après le Tableau5.10, les modèles AtR3C et ViR3C présentent une performance bien équilibrée pour la stadification de la rétinopathie diabétique en trois classes. Les deux modèles affichent des résultats performants et constants, avec environ 94% et 93% sur l'ensemble des métriques (précision, rappel et F1-score). De plus, le modèle ReVi-3C obtient des résultats encore plus remarquables, avec une moyenne de près de 98% sur l'ensemble des métriques et des classes. Ces métriques équilibrées indiquent que les prédictions des modèles sont à la fois robustes et fiables, ce qui est essentiel pour les tâches de classification médicale.

TABLE 5.10 – Evaluation des performances des modèles proposés pour la classification de la RD en 3 classes (%)

Metrics	**AtR3C**	**ViR3C**	**ReVi-3C**
Accuracy	94.26	92.97	98.26
Precision	94.41	93.77	98.43
Recall	94.09	93.22	98.21
F1_Score	94.24	93.46	98.32
Specificity(Average)	93.70	96.60	98.67

Comme l'illustre le Tableau5.11, le modèle AtR3C excelle à identifier les cas de classe 0, atteignant une précision de 97%, ce qui signifie que la quasi-totalité des prédictions pour cette catégorie sont exactes. Cependant, une spécificité de 91,40% indique que le modèle rencontre des difficultés avec la classe 1. Plus précisément, 13% de ces cas sont mal étiquetés comme classe 2 et 7% sont incorrectement classés comme classe 0. De même, 15% des cas de classe 2 sont attribués par erreur à la classe 1. Ces schémas révèlent une limitation critique : le modèle a du mal à différencier les niveaux de gravité adjacents, en particulier à distinguer la classe 1 de la classe 2. Cette confusion suggère que l'AtR3C manque de la nuance nécessaire pour séparer des catégories étroitement liées, une lacune qui pourrait affecter sa fiabilité dans des cas nécessitant une stadification précise de la gravité. En revanche, le modèle ViR3C est très précis dans la détection des patients sans rétinopathie diabétique, classant correctement 98% des cas de la classe 0. Seuls 2,4% sont mal classés comme étant au stade précoce (classe 1) et aucun n'est classé comme étant au stade avancé (classe 2), ce qui souligne sa grande spécificité pour les patients sains. Pour les cas de patients malades, la RD précoce (classe 1) est correctement identifiée dans 94% des cas, bien qu'une erreur de classification de 1%

comme étant sain pose un risque de mauvais diagnostic. La RD avancée (classe 2), quant à elle, montre une exactitude de 88%, avec 12% des cas confondus avec le stade précoce, mais aucun n'est mal classé comme sain. Cela met en évidence la robustesse du modèle pour les cas sévères, malgré un certain chevauchement dans la gravité de la stadification. Ces résultats soulignent le potentiel du modèle à diagnostiquer avec précision la RD à un stade précoce et montrent que les erreurs de classification concernent principalement les stades 1 et 2.

TABLE 5.11 – Performance par classe des modèles proposés pour la Classification en 3 classes(%)

	AtR3C			ViR3C			REVi-3C		
Metrics	Class 0	Class 1	Class 2	Class 0	Class 1	Class 2	Class 0	Class 1	Class 2
Precision	98.00	82.00	86.00	98.00	91.00	93.00	100.00	98.00	98.00
Recall	97.00	80.00	85.00	98.00	94.00	88.00	99.00	99.00	97.00
F1_score	95.00	81.00	85.00	98.00	92.00	90.00	100.00	98.00	97.00
Specificity	96.50	91.40	93.20	99.5	92.80	97.50	100.00	97.00	99.00

Comparé à AtR3C, le modèle ViR3C améliore la détection des cas sains en réduisant leur taux d'erreur de classification en tant que cas non sains de 3% à 2%. Cette amélioration souligne l'efficacité des ViT à mieux analyser les caractéristiques à l'échelle de l'image entière.

Le modèle hybride ReVi-3C surpasse considérablement ses prédécesseurs, AtR3C et ViR3C, en atteignant une classification quasi parfaite pour tous les niveaux de gravité. Il obtient une précision de 99% pour les classes 0 et 1 et de 97% pour la classe 2, marquant un bond substantiel en exactitude. Les erreurs de classification sont réduites à des niveaux négligeables, avec seulement 3% des cas de classe 2 étant mal étiquetés comme classe 1, tandis que la confusion entre la classe 0 et la classe 1 est pratiquement éliminée. Ces résultats soulignent le rôle crucial des architectures hybrides pour répondre aux défis de la classification multi-classe, où des différences inter-classes subtiles exigent une discrimination précise.

c) Classification en 5 stades de gravité

Pour la classification en 5 classes de la RD, le tableau 5.12 compare la performance de nos trois modèles : AtR5C, ViR5C et ReVi-5C en utilisant les métriques d'Accuracy, de précision, de rappel et de F1-score. Le modèle ReVi-5C se distingue des autres modèles avec une exactitude de 96,21%, une précision de 96,08%, un rappel de 95,59% et un F1-score 95,80%. Ces résultats montrent que ReVi-5C est le modèle le plus efficace, surpassant AtR5C et ViR5C dans toutes les catégories évaluées. Le modèle ReVi-5C fait preuve d'une exactitude élevée, d'une meilleure capacité à identifier les instances positives et d'un excellent équilibre global entre exactitude et rappel.

D'après le tableau 5.13, le modèle AtR5C présente des performances élevées pour la classe 0 (patients sains), avec une précision de 98%, un rappel de 96% et une spécificité de 99%, traduisant une excellente aptitude à identifier correctement les images sans signe pathologique. Toutefois, ses résultats se dégradent sur les stades plus avancés, en particulier

TABLE 5.12 – Evaluation des performances des modèles proposés pour la classification de la RD en 5 classes

Metrics	**AtR5C**	**ViR5C**	**ReVi-5C**
Accuracy	85.42%	87.33%	96.21%
Precision	85.67%	87.17%	96.08%
Recall	85.08%	85.66}	95.59%
F1_Score	85.37%	86.26%	95.80%

pour la classe 3 (stade sévère), où le rappel chute à 69% et le F1-score à 72%. Dans le cas d'AtR5C, les erreurs de classification concernent principalement les stades adjacents, ce qui est cliniquement justifiable étant donné la continuité morphologique entre les stades successifs de la RD. Plus précisément, seulement 3,5% des cas de RD légère (classe 1) sont mal classés comme sains, et aucune erreur de classification n'est observée parmi les stades les plus sévères (classes 3 et 4). Ceci montre que AtR5C est sensibile aux indices cliniques à grain fin et sa robustesse à détecter les changements pathologiques précoces.

Le modèle ViR5C est très précis pour détecter les patients sans RD, classant correctement 98% des images de classe 0. Ce faible taux d'erreur minimise le traitement inutile pour les patients sans RD. Le modèle ViR5C améliore sensiblement l'équilibre entre les classes. Les résultats obtenus pour les stades précoces (classes 1 et 2) sont supérieurs à ceux de AtR5C, avec des précisions respectives de 87% et 82%, et des rappels avoisinant 89% et 88%. Cette amélioration traduit une meilleure capacité de généralisation pour la détection des formes initiales de la pathologie.

Le modèle a mal classé 1,9% des patients de classe 1 atteints de RD comme étant en bonne santé et 1,1 % des patients de classe 2 atteints de RD comme étant en bonne santé, sans aucune erreur pour les autres stades. Ces résultats mettent en évidence la pertinence du modèle pour diagnostiquer avec précision la RD à un stade précoce et montrent que la faible classification erronée concerne principalement les stades 1 et 2 où les symptômes sont difficiles à identifier. Ce faible taux d'erreur minimise le traitement inutile pour les patients sans RD. Néanmoins, la classe 3 demeure problématique, le rappel ne dépassant pas 71%, ceci dénote un handicape dans la reconnaissance des stades sévères.

Le modèle REVi-5C surpasse nettement les deux approches précédentes. Il obtient des résultats remarquables et stables pour l'ensemble des classes, avec des précisions et rappels supérieurs à 94% dès la classe 1, et atteignant des scores parfaits (100%) pour la classe 0, éliminant totalement les faux positifs. pour identifier les patients sans RD. Seulement 0,4% des patients sains ont été mal classés, soulignant le risque minimal de traitement inutile pour les patients sans RD. Dans un autre sens, nous pouvons voir **qu'aucun patient malade n'a été classé comme sain.** Les résultats montrent un taux d'erreur de classification de seulement 0,5% entre les classes 1 et 2, ce qui suggère que le modèle est très précis pour distinguer ces deux classe. Cependant, nous observons un taux d'erreur de 2% pour les patients de classe 2 qui sont à tort classés dans la classe 1 ou 3. Cette petite différence pourrait être attribuée à

l'annotation des images dans le jeu de données. Par contre, les performances sur les stades critiques (classe 3 et classe 4) sont significativement meilleures : la précision atteint 96% et le rappel 89% pour la classe 3, tandis que la classe 4 présente un équilibre optimal avec 94% de précision et de rappel. Ce comportement indique que REVi-5C est capable de distinguer de manière fiable non seulement les cas sains et précoces, mais également les stades les plus avancés, minimisant ainsi le risque de faux négatifs dans les situations cliniques les plus critiques.

TABLE 5.13 – Performance par classe des modèles proposés pour la classification en 5 classes (%)

Classe	AtR5C				ViR5C				REVi-5C			
	Prec.	Rec.	F1	Spec.	Prec.	Rec.	F1	Spec.	Prec.	Rec.	F1	Spec.
0	98.00	96.00	97.00	99.00	97.00	98.00	98.00	97.50	100.00	100.00	100.00	100.00
1	77.00	91.00	83.00	99.00	87.00	89.00	88.00	100.00	94.00	98.00	96.00	99.00
2	81.00	79.00	80.00	90.00	82.00	88.00	87.00	97.00	96.00	96.00	96.00	97.00
3	75.00	69.00	72.00	99.00	85.00	71.00	77.00	99.00	96.00	89.00	92.00	100.00
4	82.00	77.00	79.00	95.00	85.00	82.00	84.00	99.00	94.00	94.00	95.00	99.00

Abréviations : Prec. = Précision, Rec. = Rappel, F1 = F-mesure, Spec. = Spécificité.

De manière globale, ces résultats mettent en lumière une progression graduelle des performances entre les trois modèles. AtR5C, bien qu'efficace sur la détection des cas sains, montre une sensibilité insuffisante pour les stades sévères. ViR5C améliore cet équilibre, mais demeure limité sur certaines classes avancées. Enfin, REVi-5C constitue le modèle le plus robuste et le plus fiable, offrant une couverture homogène des cinq classes et une meilleure adéquation avec les besoins du dépistage clinique de la rétinopathie diabétique.

Discussion

En s'appuyant sur les comparaisons quantitatives de nos modèles, l'analyse qui suit explorera les raisons fondamentales derrière la performance supérieure de l'approche hybride proposée par rapport aux architectures individuelles basées sur CNN ou Transformeur. Nous discuterons également de la manière dont ces améliorations se traduisent par des bénéfices cliniques concrets, en particulier pour la détection précoce de la rétinopathie diabétique (RD), et examinons les implications plus larges des gains de performance obtenus par nos modèles hybrides, ReVi-RD, ReVi-3C et ReVi-5C. Les résultats obtenus, et leur interprétation ultérieure, démontrent que les architectures hybrides proposées (Revi-RD, Revi-3C et Revi-5C) atteignent des performances remarquablement élevées dans les tâches de classification de la RD. Ces Résultats remarquables sont dûs à l'exploitation efficace des forces complémentaires de l'extraction de caractéristiques locales (par Resnet50) et de la modélisation globale des dépendances spatiales (par ViTs). Cette double capacité permet à nos modèles hybrides de détecter des signes pathologiques subtils et spatialement diffus caractéristiques de la RD précoce, qui sont fréquemment ignorés par les modèles conventionnels à base de CNN ou de transformateurs.

En particulier, le modèle ReVi-3C et ReVi-5C montre une amélioration du Rappel (Recall) dans la détection des signes pathologiques à un stade précoce. Cela facilite une intervention clinique rapide, susceptible d'arrêter la progression de la maladie et de préserver la fonction visuelle à long terme.

Quant au modèle ReVi-RD, il atteint une performance de classification binaire exceptionnelle (99,5% sur l'ensemble des métriques), ce qui peut contribuer à réduire les consultations inutiles et à alléger la charge sur les systèmes de santé débordés, en particulier dans les contextes aux ressources limitées.

5.3.5 Comparaison avec d'Autres Travaux de la Littérature

Afin d'évaluer notre approche, nous avons comparé nos résultats avec ceux d'autres méthodes de pointe ayant exploité le transfert d'apprentissage sur le jeu de données APTOS pour la classification des niveaux de sévérité de la rétinopathie diabétique. Nos modèles ont été mis en concurrence avec des réseaux de neurones convolutionnels (CNN) [55, 128], des approches basées sur l'apprentissage par transfert ensembliste [22], l'apprentissage contrastif supervisé [83], un modèle à double branche profonde (Deep Dual Branch) [142], un Swin Transformer [51], ainsi que des modèles hybrides combinant un Multiple Instance Vision Transformer (Milv4) [181] et un Vision Transformer couplé à Inception [102]. La comparaison a été effectuée en utilisant des paramètres de performance standards, incluant l'accuracy, la précision, le rappel (ou sensibilité) et le F1-score, aussi bien pour les tâches de classification binaire que pour celles à trois ou cinq classes. L'ensemble de ces méthodes, présentées dans le Tableau 5.14, est décrit en détail dans le chapitre 3. Les résultats du benchmarking montrent clairement que nos modèles surpassent ces approches de l'état de l'art, offrant de meilleures performances et une robustesse accrue sur l'ensemble des métriques d'évaluation.

— **Classification en deux stades de gravité**

Le modèle AtRD présente une performance équilibrée avec une exactitude de 99,22%, une précision de 99,60% et un F1-score de 99,41%, dépassant ainsi les travaux récents de Shakibania et al. [142] (98,50% d'exactitude) et d'Islam et al. [83] (98,36%). En revanche, les modèles AtRD (99,22%) et ReVi-RD (99,55%) se distinguent nettement en surpassant la quasi-totalité des travaux antérieurs. Plus particulièrement, le modèle hybride ReVi-RD, avec une exactitude de 99,55%, une précision de 99,51% et un F1-score de 99,54%, établit un nouveau seuil de référence, surclassant toutes les approches existantes.

— **Classification en 3 stades de gravité**

Peu d'études ont abordé la classification de la RD en trois classes à l'aide du jeu de données APTOS. Parmi elles, l'étude de Rao et al. a atteint une exactitude de 88,00%, une précision de 88,00% et un rappel de 88,02%. Cependant, leur méthode reposait sur une approche par CNN conventionnels, sans intégrer les avancées récentes en matière de stratégies de fine-tuning ou d'optimisation des hyperparamètres.

TABLE 5.14 – Comparaison des approches proposées avec les travaux antérieurs pertinents : classification binaire et en 3 classes (unité%)

Architecture	Year	Approach	Accuracy	Precision	Recall	F1-Score
Rao et al. [128]	2020	Comparison between different CNNs	96.56	97.00	97.00	96.56
Bodapati et al. [32]	2020	Fusion of multiple CNNs with DNN	96.10	–	–	–
Kumar et al. [98]	2021	VGG16 combined with capsule networks (DRISTI)	96.24	–	–	–
Islam et al. [84]	2022	Supervised contrastive learning	98.36	98.37	98.36	98.37
Shakibania et al. [142]	2024	Fusion of ResNet50 and EfficientNetB0	98.50	97.61	99.46	–
Karthika et al. [91]	2024	Stacked generalization of deep models	97.77	100.00	96.00	98.00
Our AtRD	2024	ResNet50 avec automatic hyperparameter tuning	**99.22**	**99.60**	**99.23**	**99.41**
Dihin et al. [51]	2023	Swin Transformer-based approach	96.00	–	98.00	96.00
Yang et all[178]	2024	Vit	93.2	/	86.9	/
Lian and Liul[102]	2024	Vit+Inception	95.3	/	94.2	/
Our ViRD	2024	Vit+Hyperparamètres tuning	97.73	97.72	97.73	97.73
Our ReVi-RD	Hybride :AtRD+ViRD	2024	**99.55**	99.51	**99.58**	**99.54**
3-class classification						
Rao et al [128]	2020	Comparison between different CNNs	/	88.00	88.00	88.02
Our AtR3C	2024	ResNet50 avec automatic hyperparameter tuning	94.26	94.41	94.09	94.24
Our ViR3C	2024	Vit avec Hyperparamètres tuning	92.97	93.77	93.22	93.46
Our ReVi-3C	2024	Hybride :Atr3C+ViR3C	98.26	98.431	98.21	98.32

En revanche, notre modèle AtR3C a montré de bonnes performances dans la tâche de classification à trois classes, atteignant une exactitude de 94,41%, un rappel de 94,09% et un F1-score de 94,24%. Cette amélioration significative de la performance reflète une meilleure capacité à séparer les cas limites de DR, un défi connu dans le cadre des 3 classes, et valide la robustesse introduite par l'optimisation hyperparamétrique bayésienne.

Par ailleurs, le modèle ViR3C atteint un F1-score de 93,46%, confirmant le potentiel des Vision Transformers (ViTs) dans la classification de la rétinopathie diabétique (RD), bien que ces modèles nécessitent davantage de données que les CNNs basés sur l'apprentissage

par transfert. En revanche, l'architecture hybride ReVi-3C obtient un F1-score remarquable de 98,32%, soit une amélioration absolue de 10,3% par rapport à Rao et al. Cette progression significative valide l'efficacité des modèles hybrides, dans lesquels les CNNs excellent dans l'extraction de caractéristiques locales, tandis que les ViTs permettent de capturer des motifs contextuels globaux. **L'importance de notre contribution est renforcée par le fait que peu de travaux se sont intéressés à la classification en trois classes de la RD**. Les performances prometteuses de ReVi-3C mettent en évidence son potentiel pour la détection de la RD, en particulier à ses premiers stades, ce qui conduità de meilleurs résultats diagnostiques.

— **Classification en cinq stades de gravité**

Nous avons également comparé nos modèles AtR5C, ViR5C et ReVi-5C avec des travaux antérieurs basés sur l'apprentissage par transfert, tels que ceux de [128], [121], [84], [112], ainsi qu'avec des études récentes exploitant les Vision Transformers (ViTs) pour la classification de la rétinopathie diabétique, notamment [117], [178], [51], [122], [102] et [88]. Notre modèle hybride ReVi-5C surpasse de manière significative les approches existantes, comme l'illustre le Tableau 5.15. Les performances obtenues par nos approches de classification, en particulier les modèles AtR5C et ReVi-5C, dépassent celles des méthodes de pointe, tout comme c'est le cas pour la détection. Notre approche fondée sur ResNet50 (AtRD et AtR5C) a démontré une supériorité notable par rapport aux méthodes de l'état de l'art appliquées sur le jeu de données APTOS. De plus, notre approche hybride (ReVi-RD et ReVi-5C) a produit des résultats remarquables en tirant parti à la fois du transfert d'apprentissage et des mécanismes d'attention des Vision Transformers (ViTs), atteignant un score de précision de 96,21%.

TABLE 5.15 – Comparaison des approches proposées avec les travaux antérieurs pertinents : classification en 5 classes (unit %)

Architecture	Year	Approach	Dataset	Accuracy	Precision	Recall	F1-Score
Rao et al [128]	2020	DL	Aptos	84.36	70.51	73.84	/
Oulhadj et al [120]	2022	Ensemble voting	Aptos	85	86	85	84
Islam et al [84]	2022	Supervised contrastive learning	Aptos	85	86	85	84
Mondal et al [112]	2022	Ensemble deep-learning technique	Aptos and DIARETDB1s	86	76	82	/
Our AtR5C	2024	ResNet50 avec automatique hyperparamétres tuning	Aptos	85.42	85.67	85.08	85.37
Nazih et al [117]	2023	ViT	fine-grained annotated DR (FGADR)	82.50	82.25	82.25	82.50
Yang et al [178]	2024	TMILv4 :Multiple instance ViT	Aptos	85.6	/	73.7	/
Oulhadj et al [122]	2024	ViT+ CapsNet+ PLT+CLAHE	Aptos	88.18	80	76	78
Lian and Liul [102]	2024	ViT+Inception	Aptos+RFID	89.1	/	/	87.8
Karkera et al [88]	2024	ViT+Deit+Beit+Cait	Aptos	94.63	90.55	92.88	91.67
Our ViR5C	2024	ViT avec Hyperparamètres tuning	Aptos	87.33	87.17	85.66	86.26
Our ReVi-5C	2024	AtR5C+ViR5C	Aptos	**96.21**	**96.08**	**95.80**	**95.80**

5.3.6 Discussion

Au-delà des performances quantitatives, nous allons mettre en évidence les limites des approches antérieures que notre méthode cherche à surmonter.

Plus précisément : Les modèles fondés sur les CNN (Rao et al. [128], Islam et al.[84], Mondal et al. [112], Karthika et al.[91]) présentent de grandes capacités d'extraction de caractéristiques locales, mais demeurent limités par l'absence de mécanismes permettant de modéliser les relations spatiales globales. Cette faiblesse restreint leur aptitude à détecter précocement des lésions diffuses et de petite taille, telles que les microanévrismes.

Les modèles basés sur les Transformers (Dihin et al.[51], Yang et al.[178], Oulhadj et al.[122]) excellent dans la capture des dépendances à longue portée, mais manquent les indices locaux, ce qui réduit leur sensibilité aux caractéristiques discriminantes fines, comme les microanévrismes. De plus, les ViTs exigent généralement de vastes jeux de données annotés et des ressources computationnelles considérables, limitant ainsi leur applicabilité clinique.

Les modèles hybrides CNN–ViT (Lian et Liu [102], Karkera et al. [88]) combinent le raisonnement local et global, mais reposent souvent sur des architectures CNN figées et n'intègrent pas de mécanismes d'optimisation automatique des hyperparamètres, ce qui restreint leurs capacités de généralisation.

Pour pallier ces limites, notre modèle hybride intègre trois composantes essentielles :

1. les modèles basés sur ResNet50 permettant de capturer des caractéristiques locales fines, telles que les hémorragies et les exsudats ;
2. les modèles basés sur les ViTs aptes à modéliser l'information contextuelle globale, y compris la distribution spatiale des lésions ;
3. une optimisation bayésienne pour l'ajustement automatique des hyperparamètres, renforçant ainsi la robustesse et la capacité de généralisation du modèle.

Notre architecture combinée démontre une performance stable aussi bien pour les tâches de classification binaire que multi-classes (voir. Tables 5.14 et 5.15), et offre des possibilités d'applicabilité en conditions réelles.

5.4 Conclusion

Dans ce chapitre, nous avons abordé le problème de la détection et la classification de la RD à partir d'images du fond d'œil. Nous avons proposés dans ce chapitre deux nouvelles architectures pour les trois types de classification de la RD : une architecture basée sur les visions transformers (ViRD,ViR3C et ViR5C) et une architecture hybride (ReVi-RD, ReVi-3C et ReVi-5C) combinant nos modèles basés CNN (AtRD, AtR3C et AtR5C) avec ceux basés Vit (ViRD,ViR3C et ViR5C). Nous avons mené des expériences distinctes pour évaluer les capacités de classification de la RD de chaque modèle. Nos résultats démontrent que nos

modèles et surtout l'architecture hybride proposée surpassent les modèles individuels et ceux de l'état de l'art sur l'ensemble des métriques d'évaluation.

Conclusion Générale et Perspectives

La présente thèse s'est attachée à répondre à un enjeu majeur en ophtalmologie moderne : l'amélioration de la détection et de la classification de la rétinopathie diabétique (RD) à l'aide des techniques de deep learning. Cette pathologie constitue l'une des principales causes de cécité évitable à l'échelle mondiale, et son diagnostic précoce reste une étape critique pour prévenir les complications irréversibles.

Dans ce contexte, nous avons conçu, implémenté et évalué plusieurs architectures, basées sur des réseaux convolutifs profonds (CNN), sur des Vision Transformers (ViTs), et sur une combinaison hybride des deux paradigmes.

Nos travaux ont abouti à trois contributions majeures.

1. Premièrement, nous avons proposé trois nouvelles familles de modèles adaptés à la détection et à la classification de la RD.
   - Les modèles AtRD, AtR3C et AtR5C, basés sur ResNet50 et optimisés par une recherche bayésienne des hyperparamètres, se sont distingués par leur robustesse dans l'extraction de caractéristiques locales, atteignant 99.22% d'exactitude en détection, 94.26% en classification en trois classes et 85.42% en classification en trois classes.
   - Les modèles ViRD, ViR3C et ViR5C, quant à eux, exploitent la capacité des ViTs à modéliser des relations globales et contextuelles, et affichent de trés bonnes performances avec 97.73% en détection, 92.97% et 87.33% en classification en 3 et 5 classes respectivement.
   - Enfin, les modèles hybride ReVi-RD, ReVi-3C et ReVi-5C, qui combinent les avantages complémentaires des CNN et des ViTs, ont permis d'obtenir des résultats remarquables, atteignant 99.55% de précision en détection , 98.26% et 96.21% en classification multi-classes 3 et 5 respectivement.
2. Deuxièmement, l'utilisation de la classification en 3 classes a permis de réduire la confusion entre stades intermédiaires et d'améliorer la fiabilité du diagnostic auto-

matisé, en particulier dans les phases initiales où l'intervention clinique est la plus bénéfique sachant.

3. Troisièmement, nous avons démontré, à travers des validations expérimentales sur la base de données APTOS 2019, que nos modèles atteignent des performances supérieures à l'état de l'art. Les modèles hybrides, en particulier, ont non seulement surpassé les approches CNN et ViT isolées, mais ils ont également permis une détection plus fine des premiers signes de RD, tels que les microanévrismes, souvent négligés par les architectures existantes. L'intégration de l'optimisation bayésienne a par ailleurs renforcé la généralisabilité et la robustesse de nos modèles, ouvrant la voie à une utilisation en conditions cliniques réelles.

Dans cette thèse, nous avons proposé une contribution méthodologique et appliquée, démontrant qu'une combinaison entre les modèles convolutionnels, transformeurs visuels et optimisation bayésienne peut non seulement repousser les limites actuelles de la détection automatique de la RD, mais aussi contribuer à un diagnostic précoce. Ces résultats ouvrent la voie à plusieurs perspectives futures, notamment :

1. Dans la continuité de ce travail, il serait intéressant d'associer le modèle à un *Generative Adversarial Network* (GAN) afin de générer automatiquement de nouvelles images pour l'augmentation du jeu de données. Nous avons par ailleurs proposé dans [7] une architecture qui pourrait être approfondie et évaluée sur des bases plus variées.
2. L'extension à des bases de données multi-sources permettrait de limiter le biais du modèle envers des populations spécifiques ou des conditions particulières d'acquisition, et de soutenir ainsi le développement d'outils de diagnostic applicables à l'échelle mondiale.
3. l'intégration des techniques d'IA explicables fournirait de la transparence dans la prise de décision, favorisant une plus grande confiance et adoption de nos modèles dans des contextes médicaux réels.
4. l'exploration d'architectures encore plus légère favoriserait leur déploiement sur des dispositifs mobiles.

Bibliographie

[1] Saif Hameed Abbood, Haza Nuzly Abdull Hamed, Mohd Shafry Mohd Rahim, Amjad Rehman, Tanzila Saba, and Saeed Ali Bahaj. Hybrid retinal image enhancement algorithm for diabetic retinopathy diagnostic using deep learning model. *IEEE Access*, 10 :73079–73086, 2022.

[2] Michael David Abràmoff, Yiyue Lou, Ali Erginay, Warren Clarida, Ryan Amelon, James C Folk, and Meindert Niemeijer. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative ophthalmology & visual science*, 57(13) :5200–5206, 2016.

[3] Kemal Adem, Mahmut Hekim, and Selim Demir. Detection of hemorrhage in retinal images using linear classifiers and iterative thresholding approaches based on firefly and particle swarm optimization algorithms. *Turkish Journal of Electrical Engineering and Computer Sciences*, 27(1) :499–515, 2019.

[4] Samira Ait Kaci Azzou, Djamila Boukredera, and Sifeddine Baouz. A hybrid cnn-transformer approach for precise three-class diabetic retinopathy classification. *Jordanian Journal of Computers and Information Technology (JJCIT)*, 11(03) :279–299, Sept 2025.

[5] Samira Ait Kaci Azzou, Djamila Boukredera, Sifeddine Baouz, and Toufik Azzi. Early diabetic retinopathy detection with vision transformers and optimized data preprocessing. In *African Conference on Research in Computer Science and Applied Mathematics*, pages 379–386. Springer, 2024.

[6] Samira Ait Kaci Azzou, Djamila Boukredera, Sifeddine Baouz, and Toufik Azzi. Early diabetic retinopathy detection with vision transformers and optimized data preprocessing. In *Research in Computer Science : 17th African Conference on Research in Computer Science and Applied Mathematics, CARI 2024, Bejaia, Algeria, November 24–26, 2024, Proceedings*, volume 2462, page 379. Springer Nature, 2025.

[7] Samira Ait Kaci Azzou, Djamila Boukredera, and Imen BENADJAOUD. Proposition d'une architecture gan pour le prétraitement et la classification des images de la rétinopathie diabétique. In *Colloque International MOAD'2022 (Méthodes et Outils d'Aide à la Décision) Université de Béjaia, 15 - 17 Novembre*, 2022.

[8] Samira Ait Kaci Azzou, Boukredera Djamila, Yacine Abiche, Akram Amokrane, and Achour Achroufene. Early detection and severity grading of diabetic retinopathy using fine-tuned deep learning models with automated hyperparameter optimization. *SN Computer Science*, 6(6) :1–19, 2025.

[9] Mohammad T Al-Antary and Yasmine Arafa. Multi-scale attention network for diabetic retinopathy classification. *IEEE Access*, 9 :54190–54200, 2021.

[10] Marco Alban and Tanner Gilligan. Automated detection of diabetic retinopathy using fluorescein angiography photographs. *Report of standford education*, 2016.

[11] Hanan S Alghamdi, Hongying Lilian Tang, Saad A Waheeb, and Tunde Peto. Automatic optic disc abnormality detection in fundus images : A deep learning approach. In *Proceedings of the ophthalmic medical image analysis international workshop*, volume 3. University of Iowa, 2016.

[12] Wejdan L Alyoubi, Maysoon F Abulkhair, and Wafaa M Shalash. Diabetic retinopathy fundus image classification and lesions localization system using deep learning. *Sensors*, 21(11) :3704, 2021.

[13] Wejdan L Alyoubi, Maysoon F Abulkhair, and Wafaa M Shalash. Diabetic retinopathy fundus image classification and lesions localization system using deep learning. *Sensors*, 21(11) :3704, 2021.

[14] Wejdan L Alyoubi, Wafaa M Shalash, and Maysoon F Abulkhair. Diabetic retinopathy detection through deep learning techniques : A review. *Informatics in Medicine Unlocked*, 20 :100377, 2020.

[15] Farrikh Alzami, Rama Arya Megantara, Ahmad Zainul Fanani, et al. Diabetic retinopathy grade classification based on fractal analysis and random forest. In *2019 International Seminar on Application for Technology of Information and Communication (iSemantic)*, pages 272–276. IEEE, 2019.

[16] Javeria Amin, Muhammad Sharif, Mussarat Yasmin, Hussam Ali, and Steven Lawrence Fernandes. A method for the detection and classification of diabetic retinopathy using structural predictors of bright lesions. *Journal of Computational Science*, 19 :153–164, 2017.

[17] Akram Amokrane, Yacine Abiche, Samira Ait Kaci Azzou, and Djamila Boukredera. Enhancing diabetic retinopathy detection using transfer learning. In *Colloque sur les Objets et Systèmes Connectés Learning» 4 eme edition du colloque international sur les objets et les systems connectés (COC'2023) Tunisie, 22-23 juin*, 2023.

[18] Prawej Ansari, Noushin Tabasumma, Nayla Nuren Snigdha, Nawfal Hasan Siam, Rachana V. N. R. S. Panduru, Shofiul Azam, J. M. A. Hannan, and Yasser H. A. Abdel-Wahab. Diabetic retinopathy : An overview on mechanisms, pathophysiology and pharmacotherapy. *Diabetology*, 3 :159–175, 2022.

[19] Alessandro Arrigo, Michel Teussink, Emanuela Aragona, Francesco Bandello, and Maurizio Battaglia Parodi. Multicolor imaging to detect different subtypes of retinal microaneurysms in diabetic retinopathy. *Eye*, 35(1) :277–281, 2021.

[20] Norah Asiri, Muhammad Hussain, Fadwa Al Adel, and Nazih Alzaidi. Deep learning based computer-aided diagnosis systems for diabetic retinopathy : A survey. *Artificial intelligence in medicine*, 99 :101701, 2019.

[21] American Optometric Association. Diabetic retinopathy, 2013. Accessed on juin 10, 2023.

[22] TR Athira and Jyothisha J Nair. Diabetic retinopathy grading from color fundus images : An autotuned deep learning approach. *Procedia Computer Science*, 218 :1055–1066, 2023.

[23] S Zulaikha Beevi. Multi-level severity classification for diabetic retinopathy based on hybrid optimization enabled deep learning. *Biomedical Signal Processing and Control*, 84 :104736, 2023.

[24] Francine Behar-Cohen, Emmanuelle Gelizé, Laurent Jonet, and Patricia Lassiaz. Anatomie de la rétine. *médecine/sciences*, 36(6-7) :594–599, 2020.

[25] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The journal of machine learning research*, 13(1) :281–305, 2012.

[26] Michael W Berry, Azlinah Mohamed, and Bee Wah Yap. *Supervised and unsupervised learning for data science*. Springer, 2019.

[27] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25 :197–227, 2016.

[28] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25 :197–227, 2016.

[29] Pratik Bijam and Smita Deshmukh. A review on detection of diabetic retinopathy using deep learning and transfer learning based strategies. *International Journal of Computer (IJC)*, 45(1) :164–175, 2023.

[30] RS Biyani and BM Patre. A clustering approach for exudates detection in screening of diabetic retinopathy. In *2016 International Conference on Signal and Information Processing (IConSIP)*, pages 1–5. IEEE, 2016.

[31] Jyostna Devi Bodapati. Stacked convolutional auto-encoder representations with spatial attention for efficient diabetic retinopathy diagnosis. *Multimedia Tools and Applications*, 81(22) :32033–32056, 2022.

[32] Jyostna Devi Bodapati, Veeranjaneyulu Naralasetti, Shaik Nagur Shareef, Saqib Hakak, Muhammad Bilal, Praveen Kumar Reddy Maddikunta, and Ohyun Jo. Blended multimodal deep convnet features for diabetic retinopathy severity prediction. *Electronics*, 9(6) :914, 2020.

[33] Jyostna Devi Bodapati, Nagur Shareef Shaik, and Veeranjaneyulu Naralasetti. Composite deep neural network with gated-attention mechanism for diabetic retinopathy severity classification. *Journal of Ambient Intelligence and Humanized Computing*, 12(10) :9825–9839, 2021.

[34] Philippe Bolon, Jean-Marc Chassery, Jean-Pierre Cocquerez, Didier Demigny, Christine Graffigne, Annick Montanvert, Sylvie Philipp, Rachid Zéboudj, Josiane Zerubia, and Henri Maître. Analyse d'images : filtrage et segmentation, 1995.

[35] Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. An analysis of deep neural network models for practical applications. *arXiv preprint arXiv :1605.07678*, 2016.

[36] Wen Cao, Nicholas Czarnek, Juan Shan, and Lin Li. Microaneurysm detection using principal component analysis and machine learning methods. *IEEE transactions on nanobioscience*, 17(3) :191–198, 2018.

[37] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[38] Mayank Arya Chandra and SS Bedi. Survey on svm and their application in image classification. *International Journal of Information Technology*, 13 :1–11, 2021.

[39] Nagesh Singh Chauhan. Optimization algorithms in neural networks. https://www.kdnuggets.com/2020/12/optimization-algorithms-neural-networks.html, 2020. Accessed : August 2023.

[40] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet : Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv :2102.04306*, 2021.

[41] Xue-Wen Chen and Xiaotong Lin. Big data deep learning : challenges and perspectives. *IEEE access*, 2 :514–525, 2014.

[42] 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.

[43] Amrita Roy Chowdhury, Tamojit Chatterjee, and Sreeparna Banerjee. A random forest classifier-based approach in the detection of abnormalities in the retina. *Medical & biological engineering & computing*, 57 :193–203, 2019.

[44] Jorge Cuadros et al. Eyepacs dataset. http://www.kaggle.com/c/diabetic-retinopathy-detection/data, 2009. Last accessed : May 17, 2023.

[45] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4) :303–314, 1989.

[46] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893, 2005.

[47] Etienne Decencière et al. Messidor2 dataset. http://www.adcis.net/en/Download-Third-Party/Messidor.html, 2014. Last accessed : May 17, 2023.

[48] Omar Dekhil, Ahmed Naglah, Mohamed Shaban, Mohammed Ghazal, Fatma Taher, and Ayman Elbaz. Deep learning based method for computer aided diagnosis of diabetic retinopathy. In *2019 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 1–4. IEEE, 2019.

[49] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*, 2018.

[50] Terrance DeVries and Graham W Taylor. Dataset augmentation in feature space. *arXiv preprint arXiv :1702.05538*, 2017.

[51] Rasha Ali Dihin, Ebtesam AlShemmary, and Waleed Al-Jawher. Diabetic retinopathy classification using swin transformer with multi wavelet. *Journal of Kufa for Mathematics and Computer*, 10(2) :167–172, 2023.

[52] 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.

[53] Noushin Eftekhari, Hamid-Reza Pourreza, Mojtaba Masoudi, Kamaledin Ghiasi-Shirazi, and Ehsan Saeedi. Microaneurysm detection in fundus images using a two-step convolutional neural network. *Biomedical engineering online*, 18 :1–16, 2019.

[54] Mohamed Elsharkawy, Ahmed Sharafeldeen, Fahmi Khalifa, Ahmed Soliman, Ahmed Elnakib, Mohammed Ghazal, Ashraf Sewelam, Aristomenis Thanos, Harpal S Sandhu, and Ayman El-Baz. A clinically explainable ai-based grading system for age-related macular degeneration using optical coherence tomography. *IEEE Journal of Biomedical and Health Informatics*, 28(4) :2079–2090, 2024.

[55] Mehdi Torabian Esfahani, Mahsa Ghaderi, and Raheleh Kafiyeh. Classification of diabetic and normal fundus images using new deep learning method. *Leonardo Electron. J. Pract. Technol*, 17(32) :233–248, 2018.

[56] Runze Fan, Yuhong Liu, and Rongfen Zhang. Multi-scale feature fusion with adaptive weighting for diabetic retinopathy severity classification. *Electronics*, 10(12) :1369, 2021.

[57] International Diabetes Federation. Idf diabetes atlas, 2021. Accessed on juin 10, 2023.

[58] Matthias Feurer and Frank Hutter. Hyperparameter optimization. In *Automated machine learning : Methods, systems, challenges*, pages 3–33. Springer International Publishing Cham, 2019.

[59] Donald S Fong, Lloyd P Aiello, Frederick L Ferris III, and Ronald Klein. Diabetic retinopathy. *Diabetes care*, 27(10), 2004.

[60] David A Forsyth and Jean Ponce. *Computer vision : a modern approach*. prentice hall professional technical reference, 2002.

[61] Luca Franceschi, Michele Donini, Valerio Perrone, Aaron Klein, Cédric Archambeau, Matthias Seeger, Massimiliano Pontil, and Paolo Frasconi. Hyperparameter optimization in machine learning. *arXiv preprint arXiv :2410.22854*, 2024.

[62] Kunihiko Fukushima. Neocognitron : A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4) :193–202, 1980.

[63] Akhilesh Kumar Gangwar and Vadlamani Ravi. Diabetic retinopathy detection using transfer learning and deep learning. In *Evolution in Computational Intelligence :*

Frontiers in Intelligent Computing : Theory and Applications (FICTA 2020), Volume 1, pages 679–689. Springer, 2021.

[64] James Kang Hao Goh, Carol Y Cheung, Shaun Sebastian Sim, Pok Chien Tan, Gavin Siew Wei Tan, and Tien Yin Wong. Retinal imaging techniques for diabetic retinopathy screening. *Journal of diabetes science and technology*, 10(2) :282–294, 2016.

[65] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[66] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[67] DR Study Research Group E.T.D.R.S.R(Early Treatment Diabetic Retinopathy Study Research Group). Grading diabetic retinopathy from stereoscopic color fundus photographs—an extension of the modified airlie house classification. *ETDRS report*, 10 :786–806, 1991.

[68] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern recognition*, 77 :354–377, 2018.

[69] Zongyun Gu, Yan Li, Zijian Wang, Junling Kan, Jianhua Shu, and Qing Wang. Classification of diabetic retinopathy severity in fundus images using the vision transformer and residual attention. *Computational Intelligence and Neuroscience*, 2023(1) :1305583, 2023.

[70] Yan Guex-Crosier and Francine Behar-Cohen. Ophtalmologie : Rétinopathie diabétique : nouvelles possibilités thérapeutiques. *Rev Med Suisse*, 101 :107, 2015.

[71] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, 316(22) :2402–2410, 2016.

[72] Pankaj Gupta, Chee Wah Tan, Carol Y Cheung, Charumathi Sabanayagam, Ching-Yu Cheng, Tien Yin Wong, and Dan Milea. Fgadr : Fine-grained annotated diabetic retinopathy dataset and analysis. *Medical Image Analysis*, 68 :101879, 2021.

[73] Pavel Hamet and Johanne Tremblay. Artificial intelligence in medicine. *metabolism*, 69 :S36–S40, 2017.

[74] Nikolaus Hansen. The cma evolution strategy : A tutorial. *arXiv preprint arXiv :1604.00772*, 2016.

[75] 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.

[76] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets : Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv :1704.04861*, 2017.

[77] Jingbo Hu, Huan Wang, Le Wang, and Ye Lu. Graph adversarial transfer learning for diabetic retinopathy classification. *IEEE Access*, 10 :119071–119083, 2022.

[78] 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.

[79] Yanbo Huang. Advances in artificial neural networks–methodological development and application. *Algorithms*, 2(3) :973–1007, 2009.

[80] 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.

[81] Shahzaib Iqbal, Tariq M Khan, Khuram Naveed, Syed S Naqvi, and Syed Junaid Nawaz. Recent trends and advances in fundus image analysis : A review. *Computers in Biology and Medicine*, 151 :106277, 2022.

[82] Uzair Ishtiaq, Sameem Abdul Kareem, Erma Rahayu Mohd Faizal Abdullah, Ghulam Mujtaba, Rashid Jahangir, and Hafiz Yasir Ghafoor. Diabetic retinopathy detection through artificial intelligent techniques : a review and open issues. *Multimedia Tools and Applications*, 79 :15209–15252, 2020.

[83] Md Robiul Islam, Lway Faisal Abdulrazak, Md Nahiduzzaman, Md Omaer Faruq Goni, Md Shamim Anower, Mominul Ahsan, Julfikar Haider, and Marcin Kowalski. Applying supervised contrastive learning for the detection of diabetic retinopathy and its severity levels from fundus images. *Computers in Biology and Medicine*, 146 :105602, 2022.

[84] Md Robiul Islam, Md Al Mehedi Hasan, and Abu Sayeed. Transfer learning based diabetic retinopathy detection with a novel preprocessed layer. In *2020 IEEE Region 10 Symposium (TENSYMP)*, pages 888–891. IEEE, 2020.

[85] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1) :1–54, 2019.

[86] Rohan Joseph. Grid search for model tuning. *Medium, Towards Data Science*, 2018.

[87] Ibrahem Kandel and Mauro Castelli. Transfer learning with convolutional neural networks for diabetic retinopathy image classification. a review. *Applied Sciences*, 10(6) :2021, 2020.

[88] Tejas Karkera, Chandranath Adak, Soumi Chattopadhyay, and Muhammad Saqib. Detecting severity of diabetic retinopathy from fundus images : A transformer network-based review. *Neurocomputing*, page 127991, 2024.

[89] Sagar Suresh Karki and Pradnya Kulkarni. Diabetic retinopathy classification using a combination of efficientnets. In *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)*, pages 68–72. IEEE, 2021.

[90] Sohier Dane Karthik, Maggie. Aptos 2019 blindness detection, 2019. Last accessed on juin 6, 2023.

[91] S Karthika, M Durgadevi, and T Yamuna Rani. Enhancing diabetic retinopathy diagnosis with resnet-50-based transfer learning : A promising approach. *Annals of Data Science*, 11(1) :1–24, 2024.

[92] Sara Hosseinzadeh Kassani, Peyman Hosseinzadeh Kassani, Reza Khazaeinezhad, Michal J Wesolowski, Kevin A Schneider, and Ralph Deters. Diabetic retinopathy classification using a modified xception architecture. In *2019 IEEE international symposium on signal processing and information technology (ISSPIT)*, pages 1–6. IEEE, 2019.

[93] Inam Ullah Khan, Mohaimenul Azam Khan Raiaan, Kaniz Fatema, Sami Azam, Rafi ur Rashid, Saddam Hossain Mukta, Mirjam Jonkman, and Friso De Boer. A computer-aided diagnostic system to identify diabetic retinopathy, utilizing a modified compact convolutional transformer and low-resolution images to reduce computation time. *Biomedicines*, 11(6) :1566, 2023.

[94] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision : A survey. *ACM computing surveys (CSUR)*, 54(10s) :1–41, 2022.

[95] Sotiris B Kotsiantis. Decision trees : a recent overview. *Artificial Intelligence Review*, 39 :261–283, 2013.

[96] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[97] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6) :84–90, 2017.

[98] Gaurav Kumar, Shraban Chatterjee, and Chiranjoy Chattopadhyay. Dristi : a hybrid deep neural network for diabetic retinopathy diagnosis. *Signal, Image and Video Processing*, 15(8) :1679–1686, 2021.

[99] Vasudevan Lakshminarayanan, Hoda Kheradfallah, Arya Sarkar, and Janarthanam Jothi Balaji. Automated detection and diagnosis of diabetic retinopathy : A comprehensive survey. *Journal of imaging*, 7(9) :165, 2021.

[100] Erik G Learned-Miller. Introduction to supervised learning. *I : Department of Computer Science, University of Massachusetts*, page 3, 2014.

[101] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11) :2278–2324, 1998.

[102] Jian Lian and Tianyu Liu. Lesion identification in fundus images via convolutional neural network-vision transformer. *Biomedical Signal Processing and Control*, 88 :105607, 2024.

[103] Songtao Liu, Xiaoyan Xu, Ming Ding, and Yiqiao Huang. Clinically applicable deep learning framework for reliable diabetic retinopathy diagnosis. *Nature Communications*, 11(1) :1–10, 2020.

[104] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2) :91–110, 2004.

[105] Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pages 2113–2122. PMLR, 2015.

[106] Romany F Mansour. Deep-learning-based automatic computer-aided diagnosis system for diabetic retinopathy. *Biomedical engineering letters*, 8 :41–57, 2018.

[107] Muhammad Mateen, Junhao Wen, Nasrullah Nasrullah, Song Sun, and Shaukat Hayat. Exudate detection for diabetic retinopathy using pretrained convolutional neural networks. *Complexity*, 2020(1) :5801870, 2020.

[108] John McCarthy, Marvin L Minsky, Nathaniel Rochester, and Claude E Shannon. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine*, 27(4) :12–12, 2006.

[109] Tom M Mitchell and Tom M Mitchell. *Machine learning*, volume 1. McGraw-hill New York, 1997.

[110] Jonas Močkus. On bayesian methods for seeking the extremum. In *IFIP Technical Conference on Optimization Techniques*, pages 400–404. Springer, 1974.

[111] N Jagan Mohan, R Murugan, Tripti Goel, and Parthapratim Roy. Vit-dr : Vision transformers in diabetic retinopathy grading using fundus images. In *2022 IEEE 10th region 10 humanitarian technology conference (R10-HTC)*, pages 167–172. IEEE, 2022.

[112] Sambit S Mondal, Nirupama Mandal, Krishna Kant Singh, Akansha Singh, and Ivan Izonin. Edldr : An ensemble deep learning technique for detection and classification of diabetic retinopathy. *Diagnostics*, 13(1) :124, 2022.

[113] Mohammad Amin Morid, Alireza Borjali, and Guilherme Del Fiol. A scoping review of transfer learning research on medical image analysis using imagenet. *Computers in biology and medicine*, 128 :104115, 2021.

[114] Muhammad Waqas Nadeem, Hock Guan Goh, Muzammil Hussain, Soung-Yue Liew, Ivan Andonovic, and Muhammad Adnan Khan. Deep learning for diabetic retinopathy analysis : a review, research challenges, and future directions. *Sensors*, 22(18) :6780, 2022.

[115] Dimple Nagpal, Surya Narayan Panda, Muthukumaran Malarvel, Priyadarshini A Pattanaik, and Mohammad Zubair Khan. A review of diabetic retinopathy : Datasets, approaches, evaluation metrics and future trends. *Journal of King Saud University-Computer and Information Sciences*, 34(9) :7138–7152, 2022.

[116] Syed Ali Gohar Naqvi, Muhammad Faisal Zafar, and Ihsan ul Haq. Referral system for hard exudates in eye fundus. *Computers in biology and medicine*, 64 :217–235, 2015.

[117] Waleed Nazih, Ahmad O Aseeri, Osama Youssef Atallah, and Shaker El-Sappagh. Vision transformer model for predicting the severity of diabetic retinopathy in fundus photography-based retina images. *IEEE Access*, 11 :117546–117561, 2023.

[118] Pradeep Nijalingappa and B Sandeep. Machine learning approach for the identification of diabetes retinopathy and its stages. In *2015 International conference on applied and theoretical computing and communication technology (iCATccT)*, pages 653–658. IEEE, 2015.

[119] Mehmet Evren Okur, Ioannis D Karantas, and Panoraia I Siafaka. Diabetes mellitus : a review on pathophysiology, current status of oral pathophysiology, current status of oral medications and future perspectives. *Acta Pharmaceutica Sciencia*, 55(1), 2017.

[120] Mohammed Oulhadj, Jamal Riffi, Khodriss Chaimae, Adnane Mohamed Mahraz, Bennis Ahmed, Ali Yahyaouy, Chraibi Fouad, Abdellaoui Meriem, Benatiya Andaloussi Idriss, and Hamid Tairi. Diabetic retinopathy prediction based on deep learning and deformable registration. *Multimedia Tools and Applications*, 81(20) :28709–28727, 2022.

[121] Mohammed Oulhadj, Jamal Riffi, Chaimae Khodriss, Adnane Mohamed Mahraz, Ahmed Bennis, Ali Yahyaouy, Fouad Chraibi, Meriem Abdellaoui, Idriss Benatiya Andsaloussi, and Hamid Tairi. Diabetic retinopathy prediction based on transfer learning and ensemble voting. In *International Conference on Digital Technologies and Applications*, pages 929–937. Springer, 2023.

[122] Mohammed Oulhadj, Jamal Riffi, Chaimae Khodriss, Adnane Mohamed Mahraz, Ali Yahyaouy, Meriem Abdellaoui, Idriss Benatiya Andaloussi, and Hamid Tairi. Diabetic retinopathy prediction based on vision transformer and modified capsule network. *Computers in Biology and Medicine*, 175 :108523, 2024.

[123] Sinno Jialin Pan. Transfer learning. *Learning*, 21 :1–2, 2020.

[124] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10) :1345–1359, 2009.

[125] Kamal Porwal et al. Idrid (indian diabetic retinopathy image dataset). https://idrid.grand-challenge.org/, 2018. Last accessed : May 17, 2023.

[126] David MW Powers. Evaluation : from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv :2010.16061*, 2020.

[127] Sarni Suhaila Rahim, Chrisina Jayne, Vasile Palade, and James Shuttleworth. Automatic detection of microaneurysms in colour fundus images for diabetic retinopathy screening. *Neural computing and applications*, 27 :1149–1164, 2016.

[128] Mihir Rao, Michelle Zhu, and Tianyang Wang. Conversion and implementation of state-of-the-art deep learning algorithms for the classification of diabetic retinopathy. *arXiv preprint arXiv :2010.11692*, 2020.

[129] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V. Le. Regularized evolution for image classifier architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019. arXiv :1802.01548.

[130] Ali M Reza. Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *Journal of VLSI signal processing systems for signal, image and video technology*, 38 :35–44, 2004.

[131] Jesus Rodriguez. Understanding hyperparameters optimization in deep learning models : concepts and tools. *Linkedin Pulse*, 2018.

[132] Sohini Roychowdhury, Dara D Koozekanani, and Keshab K Parhi. Dream : diabetic retinopathy analysis using machine learning. *IEEE journal of biomedical and health informatics*, 18(5) :1717–1728, 2013.

[133] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3) :211–252, 2015.

[134] Stuart J Russell and Peter Norvig. *Artificial intelligence : a modern approach*. pearson, 2016.

[135] Sameera V Mohd Sagheer and Sudhish N George. A review on medical image denoising algorithms. *Biomedical signal processing and control*, 61 :102036, 2020.

[136] Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3 :210–229, 1959.

[137] Elham Saraee, Mona Jalal, and Margrit Betke. Visual complexity analysis using deep intermediate-layer features. *Computer Vision and Image Understanding*, 195 :102949, 2020.

[138] Benedetta Savelli, Alessandro Bria, Mario Molinara, Claudio Marrocco, and Francesco Tortorella. A multi-context cnn ensemble for small lesion detection. *Artificial intelligence in medicine*, 103 :101749, 2020.

[139] Ganeshsree Selvachandran, Shio Gai Quek, Raveendran Paramesran, Weiping Ding, and Le Hoang Son. Developments in the detection of diabetic retinopathy : a state-of-the-art review of computer-aided diagnosis and machine learning methods. *Artificial intelligence review*, 56(2) :915–964, 2023.

[140] Fatma Shaheen, Brijesh Verma, and Md Asafuddoula. Impact of automatic feature extraction in deep learning architecture. In *2016 International conference on digital image computing : techniques and applications (DICTA)*, pages 1–8. IEEE, 2016.

[141] Nagur Shareef Shaik and Teja Krishna Cherukuri. Hinge attention network : A joint model for diabetic retinopathy severity grading. *Applied Intelligence*, 52(13) :15105–15121, 2022.

[142] Hossein Shakibania, Sina Raoufi, Behnam Pourafkham, Hassan Khotanlou, and Muharram Mansoorizadeh. Dual branch deep learning network for detection and stage grading of diabetic retinopathy. *Biomedical Signal Processing and Control*, 93 :106168, 2024.

[143] Shreya Shekar, Nitin Satpute, and Aditya Gupta. Review on diabetic retinopathy with deep learning methods. *Journal of Medical Imaging*, 8(6) :060901–060901, 2021.

[144] Dan Simon. *Evolutionary optimization algorithms*. John Wiley & Sons, 2013.

[145] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv :1409.1556*, 2014.

[146] Chanjira Sinthanayothin, James F Boyce, Tom H Williamson, Helen L Cook, Evelyn Mensah, Shantanu Lal, and David Usher. Automated detection of diabetic retinopathy on digital fundus images. *Diabetic medicine*, 19(2) :105–112, 2002.

[147] Dilip Singh Sisodia, Shruti Nair, and Pooja Khobragade. Diabetic retinal fundus images : Preprocessing and feature extraction for early detection of diabetic retinopathy. *Biomedical and Pharmacology Journal*, 10(2) :615–626, 2017.

[148] Ayoub Skouta, Abdelali Elmoufidi, Said Jai-Andaloussi, and Ouail Ouchetto. Deep learning for diabetic retinopathy assessments : a literature review. *Multimedia Tools and Applications*, 82(27) :41701–41766, 2023.

[149] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4) :427–437, 2009.

[150] Toufique Ahmed Soomro, Junbin Gao, Tariq Khan, Ahmad Fadzil M Hani, Mohammad AU Khan, and Manoranjan Paul. Computerised approaches for the detection of diabetic retinopathy using retinal fundus images : a survey. *Pattern Analysis and Applications*, 20 :927–961, 2017.

[151] Ruchir Srivastava, Lixin Duan, Damon WK Wong, Jiang Liu, and Tien Yin Wong. Detecting retinal microaneurysms and hemorrhages with robustness to the presence of blood vessels. *Computer methods and programs in biomedicine*, 138 :83–91, 2017.

[152] Kenneth O. Stanley and Risto Miikkulainen. Evolving neural networks through augmenting topologies. *Evolutionary Computation*, 10(2) :99–127, 2002.

[153] S Sudha, A Srinivasan, and T Gayathri Devi. Automatic detection of microaneurysms in diabetic retinopathy images using graph cut segmentation and svm classifier with pca. *International Journal of Pure and Applied Mathematics is a Mathematical Journal*, 119(15) :3365–3374, 2018.

[154] Ayaka Sugeno, Yasuyuki Ishikawa, Toshio Ohshima, and Rieko Muramatsu. Simple methods for the lesion detection and severity grading of diabetic retinopathy by image processing and transfer learning. *Computers in biology and medicine*, 137 :104795, 2021.

[155] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning : An introduction*, volume 1. MIT press Cambridge, 1998.

[156] 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.

[157] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis : Full training or fine tuning ? *IEEE transactions on medical imaging*, 35(5) :1299–1312, 2016.

[158] Srikanth Tammina. Transfer learning using vgg-16 with deep convolutional neural network for classifying images. *International Journal of Scientific and Research Publications (IJSRP)*, 9(10) :143–150, 2019.

[159] Mingxing Tan and Quoc Le. Efficientnet : Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[160] Daniel SW Ting, Carol Y Cheung, Quang Nguyen, Charumathi Sabanayagam, Gilbert Lim, Zhan Wei Lim, Gavin SW Tan, Yu Qiang Soh, Leopold Schmetterer, Ya Xing Wang, et al. Deep learning in estimating prevalence and systemic risk factors for diabetic retinopathy : a multi-ethnic study. *Npj Digital Medicine*, 2(1) :24, 2019.

[161] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends : algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[162] Alan M Turing. *Computing machinery and intelligence*. Springer, 2009.

[163] National Eye Institute. U.S. Diabetic retinopathy. *Department of Health and Human Services*, 2022. Accessed on juin 10, 2023, url=https ://www.nei.nih.gov/learn-about-eye-health/eye-conditions-anddiseases/diabetic-retinopathy.

[164] Mark JJP Van Grinsven, Bram van Ginneken, Carel B Hoyng, Thomas Theelen, and Clara I Sánchez. Fast convolutional neural network training using selective data sampling : Application to hemorrhage detection in color fundus images. *IEEE transactions on medical imaging*, 35(5) :1273–1284, 2016.

[165] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[166] Kritika Verma and Pradeep Kumar Singh. An insight to soft computing based defect prediction techniques in software. *International Journal of Modern Education and Computer Science*, 7(9) :52, 2015.

[167] Richa Vij and Sakshi Arora. A novel deep transfer learning based computerized diagnostic systems for multi-class imbalanced diabetic retinopathy severity classification. *Multimedia Tools and Applications*, 82(22) :34847–34884, 2023.

[168] Stela Vujosevic, Stephen J Aldington, Paolo Silva, Cristina Hernández, Peter Scanlon, Tunde Peto, and Rafael Simó. Screening for diabetic retinopathy : new perspectives and challenges. *The Lancet Diabetes & Endocrinology*, 8(4) :337–347, 2020.

[169] Juan Wang, Yujing Bai, and Bin Xia. Feasibility of diagnosing both severity and features of diabetic retinopathy in fundus photography. *IEEE access*, 7 :102589–102597, 2019.

[170] Shuangling Wang, Yilong Yin, Guibao Cao, Benzheng Wei, Yuanjie Zheng, and Gongping Yang. Hierarchical retinal blood vessel segmentation based on feature and ensemble learning. *Neurocomputing*, 149 :708–717, 2015.

[171] Wei Wang and Amy CY Lo. Diabetic retinopathy : pathophysiology and treatments. *International journal of molecular sciences*, 19(6) :1816, 2018.

[172] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath : Transformer-based self-supervised learning for histopathological image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021 : 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 186–195. Springer, 2021.

[173] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1) :1–40, 2016.

[174] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1) :1–40, 2016.

[175] Charles P Wilkinson, Frederick L Ferris III, Ronald E Klein, Paul P Lee, Carl David Agardh, Matthew Davis, Diana Dills, Anselm Kampik, R Pararajasegaram, Juan T Verdaguer, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, 110(9) :1677–1682, 2003.

[176] Jianfang Wu, Ruo Hu, Zhenghong Xiao, Jiaxu Chen, and Jingwei Liu. Vision transformer-based recognition of diabetic retinopathy grade. *Medical Physics*, 48(12) :7850–7863, 2021.

[177] Di Xiao, Shuang Yu, Janardhan Vignarajan, Dong An, Mei-Ling Tay-Kearney, and Yogi Kanagasingam. Retinal hemorrhage detection by rule-based and machine learning approach. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 660–663. IEEE, 2017.

[178] Yaoming Yang, Zhili Cai, Shuxia Qiu, and Peng Xu. A novel transformer model with multiple instance learning for diabetic retinopathy classification. *IEEE Access*, 2024.

[179] Xin Yao. Evolving artificial neural networks. *Proceedings of the IEEE*, 87(9) :1423–1447, 1999.

[180] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks ? *Advances in neural information processing systems*, 27, 2014.

[181] Shuang Yu, Kai Ma, Qi Bi, Cheng Bian, Munan Ning, Nanjun He, Yuexiang Li, Hanruo Liu, and Yefeng Zheng. Mil-vt : Multiple instance learning enhanced vision transformer for fundus image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021 : 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 45–54. Springer, 2021.

[182] Shuang Yu, Di Xiao, and Yogesan Kanagasingam. Exudate detection for diabetic retinopathy with convolutional neural networks. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1744–1747. IEEE, 2017.

[183] Tong Yu and Hong Zhu. Hyper-parameter optimization : A review of algorithms and applications. *arXiv preprint arXiv :2003.05689*, 2020.

[184] Jinghua Zhang, Chen Li, Yimin Yin, Jiawei Zhang, and Marcin Grzegorzek. Applications of artificial neural networks in microorganism image analysis : a comprehensive review from conventional multilayer perceptron to popular convolutional neural network and potential visual transformer. *Artificial Intelligence Review*, 56(2) :1013–1070, 2023.

[185] Zhaoyang Zhang, Fei Yin, Feng Liu, Yuhao Zhang, Ning Luo, and Wing Wong. Ddra : A large-scale database for diabetic retinopathy analysis. *arXiv preprint arXiv :1712.03917*, 2017.

[186] Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics : a survey. In *2020 IEEE symposium series on computational intelligence (SSCI)*, pages 737–744. IEEE, 2020.

[187] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

[188] Zhi-Hua Zhou. *Machine learning*. Springer Nature, 2021.

[189] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1) :43–76, 2020.

Résumé

La rétinopathie diabétique (RD) est une cause majeure de cécité évitable. Le dépistage précoce et la classification de sa sévérité sont essentiels, mais les approches traditionnelles demeurent limitées par la subjectivité et la disponibilité restreinte de données annotées. Dans ce contexte, cette thèse étudie l'apport du deep learning pour l'automatisation du diagnostic de la rétinopathie diabétique (RD). Nous avons d'abord évalué l'efficacité du transfert d'apprentissage à travers plusieurs architectures convolutionnelles (VGG16, ResNet50, InceptionV3, Xception), puis analysé le potentiel des Vision Transformers (ViTs), capables de modéliser les relations à longue distance entre lésions dispersées.
Nous avons réalisé deux contributions majeures. La première consiste en l'ajustement automatique des hyperparamètres par optimisation bayésienne, ce qui a considérablement amélioré les performances des modèles. La seconde porte sur le développement d'un modèle hybride CNN-Transformer, fusionnant extraction locale et attention globale. Ce modèle, automatiquement optimisé, a prouvé avoir de meilleures performances pour la détection et la classification multi-niveaux (2, 3 et 5 classes) de la RD, dépassant les méthodes traditionnelles et préparant le terrain pour une future application clinique.
En conclusion, cette recherche confirme la pertinence des approches hybrides en deep learning pour le dépistage de la RD et ouvre des perspectives vers des systèmes plus robustes, explicables et généralisables.

Mots clés : Rétinopathie Diabétique, Classification, Deep Learning, Transfert d'Apprentissage, Vision Transformers, Optimisation Bayésienne.

Abstract

Diabetic retinopathy (DR) is a leading cause of preventable blindness. Early screening and accurate severity classification are essential, yet traditional approaches remain limited by subjectivity and the restricted availability of annotated datasets. In this context, this thesis investigates the contribution of deep learning to the automation of DR diagnosis. We first assessed the effectiveness of transfer learning through several convolutional architectures (VGG16, ResNet50, InceptionV3, Xception), and then explored the potential of Vision Transformers (ViTs), which can model long-range dependencies between scattered retinal lesions.
This work makes two major contributions. The first is the automatic adjustment of hyperparameters using Bayesian optimization, which significantly improved model performance. The second is the development of a hybrid CNN–Transformer model, combining local feature extraction with global attention. This automatically optimized model demonstrated superior performance in DR detection and multi-level classification (2, 3, and 5 classes), outperforming traditional methods and paving the way for future clinical applications.
In conclusion, this research confirms the relevance of hybrid deep learning approaches for DR screening and opens perspectives toward the development of more robust, explainable, and generalizable diagnostic systems.

Keywords : Diabetic Retinopathy, Classification, Deep Learning, transfer Learning, Vision Transformers, Bayesian Optimization.

الملخص

تُعَدُّ اعتلالات الشبكية السكري (DR) من الأسباب الرئيسية للعمى الذي يمكن الوقاية منه. ويُعَدُّ الفحص المبكر والتصنيف الدقيق لدرجات شدته أمرًا بالغ الأهمية، غير أن الأساليب التقليدية ما زالت محدودة بسبب الطابع الذاتي للاختبارات وقلة توافر قواعد البيانات المعلَّمة. في هذا السياق، تبحث هذه الأطروحة في إسهام **التعلّم العميق** في أتمتة تشخيص اعتلال الشبكية السكري. لقد قمنا أولاً بتقييم فعالية **التعلّم بالنقل** من خلال عدة معماريات التفافية (VGG16, ResNet50, InceptionV3, Xception)، ثم استكشفنا إمكانات **محوّلات الرؤية (Vision Transformers – ViTs)** القادرة على نمذجة الترابطات بعيدة المدى بين الآفات الشبكية المتناثرة.

يقدّم هذا العمل إسهامين رئيسيين. يتمثل الأول في **الضبط التلقائي للمعاملات الفائقة** باستخدام **التحسين البايزي**، الأمر الذي حسّن بشكل ملحوظ من أداء النماذج. أما الإسهام الثاني فيتمثل في تطوير **نموذج هجين CNN–Transformer** يجمع بين الاستخراج المحلي للميزات وآلية الانتباه الشامل. وقد أثبت هذا النموذج، بعد تحسينه آليًا، تفوقًا ملحوظًا في مهام كشف اعتلال الشبكية السكري وتصنيفه على مستويات متعددة (2 و3 و5 أصناف)، متجاوزًا الطرق التقليدية وممهّدًا الطريق لتطبيقات سريرية مستقبلية.

ختامًا، تؤكد هذه الدراسة أهمية النهج الهجينة المعتمدة على التعلّم العميق في فحص اعتلال الشبكية السكري، كما تفتح آفاقًا نحو تطوير أنظمة تشخيصية أكثر قوة وقابلة للتفسير وذات قدرة أفضل على التعميم.

الكلمات المفتاحية: اعتلال الشبكية السكري، التصنيف، التعلّم العميق، التعلّم بالنقل، محوّلات الرؤية، التحسين البايزي.