République Algérienne Démocratique et Populaire
Ministère de l'Enseignement supérieur et de la Recherche scientifique

Université A. Mira-Béjaïa
Faculté des Sciences Exactes
Department D'informatique

Master Recherche
En
Informatique
Option
Systèmes d'information avancés (SIA)

Thème

Explicabilité des décisions de l'IA appliquée à la rétinopathie diabétique

Réalisé par:

Mr. ATMANI Yacine
Mr. MEHENNI Kussila

Devant le jury composé de :

Président :	Dr. SAAD Narimane	Université de Béjaïa.
Examinateur :	Dr. ACHROUFENE Achour	Université de Béjaïa.
Examinateur :	Dr. BOULAHROUZ Djamila	Université de Béjaïa.
Examinateur :	Dr. BOUCHEBBAH Fatah	Université de Béjaïa.
Encadrant :	Mme. S. Ait Kaci Azzou	Université de Béjaïa.

U. A/M Béjaïa, Juillet 2025.

Remerciements

Avant toute chose, il nous semble opportun de commencer ce mémoire en exprimant notre gratitude, tout d'abord à Dieu pour nous avoir accordé la force et le courage de mener à bien ce modeste travail.

Nos remerciements et notre gratitude vont à notre superviseur, Mme Samira AIT KACI AZZOU, qui nous a soutenu et guidée tout au long de cette expérience professionnelle avec patience et enthousiasme.

Nous tenons également à remercier les membres du jury d'avoir accepté d'examiner et d'évaluer notre travail.

À tous ceux qui ont contribué, directement ou indirectement, par leurs conseils, leurs encouragements ou leur amitié, à la réalisation de ce mémoire, on exprime notre profonde gratitude.

Pour leurs encouragements, leur soutien moral et la patience dont ils ont fait preuve tout au long de l'année, nous sommes sincèrement reconnaissants à tous les membres de nos familles.

Dédicaces

Nous dédions ce modeste travail à :
Nos chers parents, pour leur amour inconditionnel,

leur patience infinie et leur soutien indéfectible
tout au long de notre parcours.
À nos frères et sœurs, pour leur affection,

leurs encouragements et leur présence bienveillante.
À nos professeurs, pour nous avoir transmis le savoir

avec passion, exigence et bienveillance.
À nos amis, pour leur soutien constant,

leur écoute et leur présence dans les moments difficiles.
Et à toutes les personnes qui, de près ou de loin,

ont contribué à l'aboutissement de ce travail.

Y.A. & K.M.

Table des matières

Liste des figures

Liste des tableaux

Liste des acronymes

AAO Société Américaine d'Ophtalmologie (American Academy of Ophthalmology)

AMIR Anomalies Microvasculaires Intrarétiniennes

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

AtR5C Nom du modèle CNN

CNN Réseaux de Neurones Convolutifs (Convolutional Neural Networks)

DQN Deep Q-Network

DL Apprentissage profond (Deep Learning)

ETDRS Early Treatment Diabetic Retinopathy Study

Grad-CAM Gradient-weighted Class Activation Mapping

GPU Graphics Processing Unit

IA Intelligence artificielle

KNN K-Nearest Neighbors

LIME Local Interpretable Model-agnostic Explanations

LSTM Long Short-Term Memory

ML Apprentissage automatique (Machine Learning)

MLP Perceptron multicouche (Multi-Layer Perceptron)

MSA Multi-Head Self-Attention

OCT Tomographie par cohérence optique (Optical Coherence Tomography)

PCA Analyse en Composantes Principales (Principal Component Analysis)

PSEM Path-Sufficient Explanations Method

RNN Réseaux de Neurones Récurrents (Recurrent Neural Networks)

RD Rétinopathie diabétique

ReVi-5C Nom du modèle Hybride(CNN + ViT)

Score-CAM Score-Weighted Class Activation Mapping

SHAP SHapley Additive Explanations

S3VM Semi-supervised Support Vector Machines

SVM Support Vector Machines

ViR-5C Nom du modèle ViT

ViT Vision Transformer

VEGF Facteur de croissance de l'endothélium vasculaire (Vascular Endothelial Growth Factor)

VRAM Video Random Access Memory

XAI Intelligence artificielle explicable (Explainable Artificial Intelligence)

Introduction Générale

L'intelligence artificielle (IA) transforme profondément la pratique médicale contemporaine, notamment dans le domaine de l'imagerie ophtalmologique. La rétinopathie diabétique (RD), complication microvasculaire du diabète et première cause de cécité évitable chez l'adulte, illustre parfaitement ce potentiel : le dépistage précoce des lésions rétiniennes à l'aide de l'IA pourrait permettre de préserver la vision de millions de patients. Les modèles d'apprentissage profond ont déjà prouvé leur efficacité dans la classification des images du fond d'œil. Cependant, leur opacité demeure un frein majeur à la confiance des cliniciens et à leur déploiement à grande échelle.

L'intelligence artificielle explicable (XAI) a pour objectif de lever cet obstacle en rendant compréhensibles les décisions prises par les modèles. Comprendre pourquoi un modèle conclut à un certain niveau de sévérité de la RD est fondamental, tant pour évaluer la pertinence clinique de la décision que pour détecter d'éventuels biais ou anomalies. Pourtant, la diversité des méthodes XAI disponibles, fondées sur des principes variés (cartes de saillance, approximations locales, théorie des jeux, mécanismes d'attention, etc.) [5], complique leur comparaison. À ce jour, il n'existe pas de cadre d'évaluation universellement accepté.

Face à cette pluralité de méthodes XAI et à l'absence de critères d'évaluation normalisés, une question centrale émerge : comment choisir, implémenter et évaluer efficacement une méthode d'explicabilité pour la classification de la RD ?

Ce mémoire se concentre sur l'analyse comparative et l'évaluation rigoureuse de plusieurs méthodes XAI appliquées à la classification de la RD. Trois architectures profondes représentatives ont été étudiées : un réseau de neurones convolutifs (CNN) nommé AtR5C, un modèle Vision Transformer (ViR-5C) et une architecture hybride (ReVi-5C), tous fine-tunés sur le jeu de données APTOS [68]. Pour chacun de ces modèles, les techniques d'explicabilité les plus pertinentes ont été déployées :

- Grad-CAM, Score-CAM, LIME et SHAP pour le CNN AtR5C ;
- Attention Rollout, LIME et SHAP pour le ViT ViR-5C ;
- LIME et SHAP pour le modèle hybride ReVi-5C.

Afin d'assurer une comparaison objective de ces méthodes, des métriques d'explicabilité reconnues, telles que l'insertion, la suppression (deletion), la fidélité (faithfulness) et la robustesse, ont été employées [17, 54, 48]. Ces métriques permettent de quantifier respectivement la pertinence des régions mises en évidence, la cohérence des explications avec la prédiction, la fidélité globale au comportement du modèle, ainsi que la stabilité face à de légères perturbations de l'image.

Les résultats obtenus, issus d'une analyse qualitative approfondie, ont permis de mettre en évidence de nouveaux aspects concernant :

1. La capacité des cartes générées par chaque méthode à cibler les zones anatomiques pertinentes (macula, micro-anévrismes, exsudats) ;
2. La robustesse des explications aux variations d'acquisition ;
3. La compatibilité des explications produites avec les pratiques réelles de dépistage.

La démarche adoptée est reflétée par la structure de ce mémoire :

- Chapitre 1 : Introduction à la rétinopathie diabétique et à l'intelligence artificielle.
- Chapitre 2 : Présentation de l'état de l'art des méthodes d'intelligence artificielle explicable (XAI).
- Chapitre 3 : Conception et mise en œuvre du cadre expérimental.
- Conclusion : Synthèse des résultats obtenus et exploration des perspectives futures.

Chapitre 1

Rétinopathie Diabétique et Intelligence Artificielle

I Introduction

Ces dernières années, les avancées rapides de l'intelligence artificielle (IA) ont profondément transformé de nombreux secteurs, dont celui de la santé. L'IA permet aujourd'hui d'analyser de vastes volumes de données médicales, de détecter des motifs subtils et de proposer des prédictions utiles pour le diagnostic et la prise de décision clinique. L'un des domaines où ces technologies trouvent une application concrète est celui de la rétinopathie diabétique (RD), une complication fréquente du diabète, pouvant entraîner des troubles visuels sévères, voire la cécité.

Dans ce contexte, notre travail s'inscrit dans une volonté de mettre en lumière comment l'IA, et plus particulièrement les techniques de vision par ordinateur et d'apprentissage automatique, peuvent être utilisées pour améliorer le dépistage et le suivi de la RD. Ce premier chapitre a pour objectif d'introduire les concepts médicaux de base nécessaires à la compréhension de cette pathologie, ainsi que les fondements théoriques de l'IA et de ses sous-domaines.

Nous présentons notamment les principes du machine learning et du deep learning, les architectures classiques comme les réseaux de neurones convolutifs (CNN), ainsi que les modèles plus récents tels que les Vision Transformers (ViT). Nous abordons également l'apprentissage par transfert, une méthode précieuse dans le cadre médical, où les données annotées sont souvent limitées. Ce cadre théorique constitue la base sur laquelle s'appuieront les réflexions et expérimentations présentées dans les chapitres suivants.

II Rétinopathie Diabétique

La rétinopathie diabétique est une complication du diabète qui affecte les petits vaisseaux sanguins de la rétine, la membrane sensible à la lumière située à l'arrière de l'œil, responsable de la vision. Elle résulte principalement de l'hyperglycémie chronique, un facteur clé du diabète [1].Cette condition fragilise les capillaires rétiniens, entraînant des micro-anévrismes et des fuites de liquide [26].

Les vaisseaux rétiniens peuvent se déformer, et des néovaisseaux anormaux se forment, marquant une forme sévère de la maladie appelée rétinopathie proliférante. Si la rétine centrale est atteinte, on parle de maculopathie diabétique, qui entraîne un œdème dans la macula, réduisant l'acuité visuelle [38].

Les symptômes incluent une baisse de l'acuité visuelle, des « mouches volantes » et des hémorragies rétiniennes, mais la maladie peut être asymptomatique au début [37]. Si non traitée, la rétinopathie diabétique peut conduire à la cécité, notamment en raison de la maculopathie diabétique. Le dépistage annuel est essentiel pour prévenir cette complication, et le contrôle de la glycémie, ainsi que l'équilibrage de la pression artérielle, sont des mesures cruciales pour prévenir la progression de la maladie [82].

II.1 Physiopathologie de la rétinopathie diabétique

La rétinopathie diabétique résulte principalement d'une microangiopathie provoquée par l'hyperglycémie chronique, entraînant la disparition progressive des péricytes. Cela engendre trois conséquences majeures : la fragilité, l'hyperperméabilité et l'occlusion capillaire.

Ces altérations conduisent à une capillaropathie à double mécanisme [28], comme illustré à la figure 1.1.

- **Oedémateux :** la rupture de la barrière hémato-rétinienne permet le passage d'ions, protéines et lipides dans la rétine, générant un œdème (via l'osmose) et la formation d'exsudats durs en zone péri-oedémateuse. L'œdème aggrave l'hypoxie et altère la vision, surtout au niveau maculaire.
- **Ischémique :** l'obstruction des capillaires accentue l'hypoxie, renforcée par l'œdème, et déclenche une réponse compensatoire via la sécrétion de facteurs de croissance proangiogéniques (VEGF). Cela entraîne une prolifération anarchique de néo-vaisseaux.

Ces néo-vaisseaux, bien que destinés à compenser l'ischémie, sont fragiles et pathogènes, menant à trois complications principales :

- Hémorragie intra-vitréenne
- Décollement rétinien par traction
- Glaucome néo-vasculaire

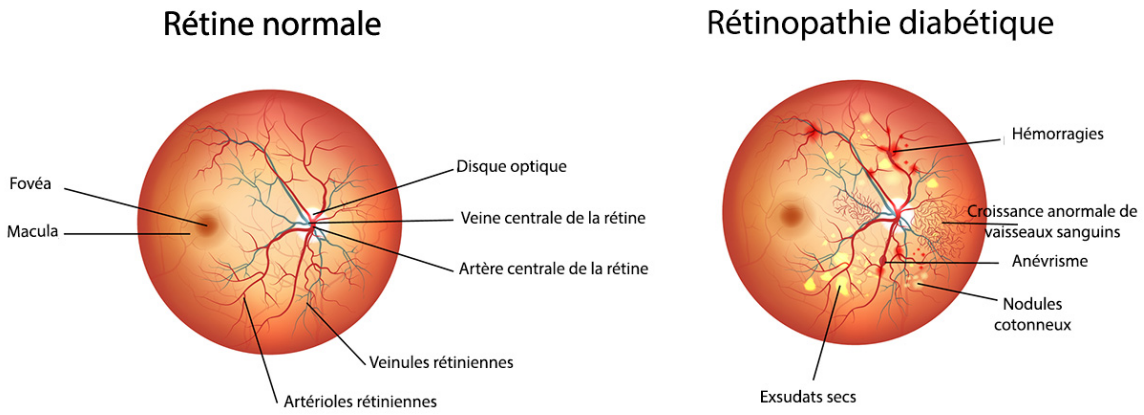


Figure 1.1: Retine normale vs Retine diabetique [75]

III Stades de la rétinopathie diabétique

La classification de la rétinopathie diabétique proposée par la Société Américaine d'Ophtalmologie (AAO) [79], dérivée de l'étude ETDRS (Early Treatment Diabetic Retinopathy Study), repose sur la localisation de l'œdème par rapport au centre de la macula. Cette évaluation, réalisée à partir de photographies du fond d'œil, permet d'estimer le risque visuel : plus l'œdème est proche du centre, plus la menace pour la vision est importante voir figure 1.2.

La classification est présentée dans le tableau 1.1 :

Table 1.1: Classification de la rétinopathie diabétique selon le degré de sévérité

Classe	Degré de sévérité	Caractéristiques	Symptômes
0	Pas de RD apparente	Aucun signe détecté au fond d'œil.	Aucune plainte ; vision normale.
1	RD non-proliférative légère	Présence de microanévrismes.	Généralement asymptomatique, vision normale.
2	RD non-proliférative modérée	Microanévrismes, hémorragies, exsudats.	Vision légèrement floue ou perturbée.
3	RD non-proliférative sévère	Nombreux microanévrismes, hémorragies, exsudats, AMIR.	Vision floue, perte de contraste, "mouches volantes".
4	RD proliférative	Néovaisseaux, hémorragies internes, risque de décollement.	Vision gravement altérée, risque de cécité.

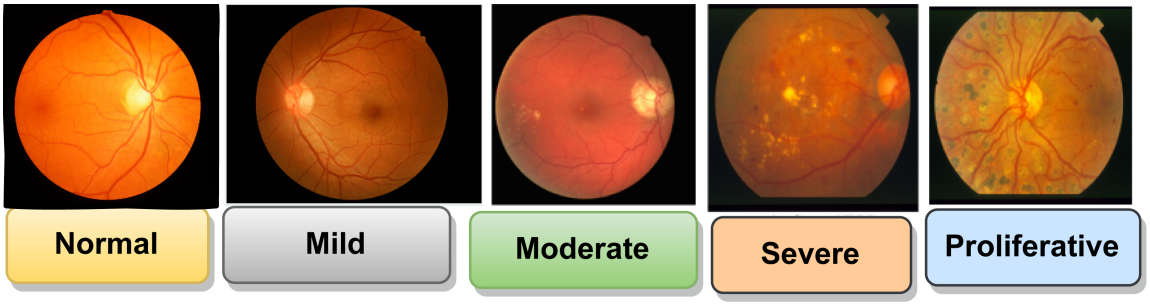


Figure 1.2: Stades de la rétinopathie diabétique [18]

Dans cette section nous présentons les différents symptômes de la RD ainsi que les différentes techniques de son diagnostic.

III.1 Symptômes communs de la rétinopathie diabétique

La rétinopathie diabétique est souvent asymptomatique dans ses premières étapes. Cependant, à mesure que la maladie progresse, les symptômes peuvent inclure [26] :

- Vision floue ou altérée.

- Perte de la vision centrale ou périphérique.
- Apparition de "mouches volantes" (myodésopsies).
- La cécité

Les symptômes peuvent se développer progressivement ou de manière brutale, selon la sévérité de la rétinopathie.

III.2 Méthodes de prévention

La prévention de la rétinopathie diabétique repose principalement sur un contrôle rigoureux de la glycémie, ainsi que sur la gestion de la pression artérielle et du cholestérol. Voici les principales mesures préventives [82] :

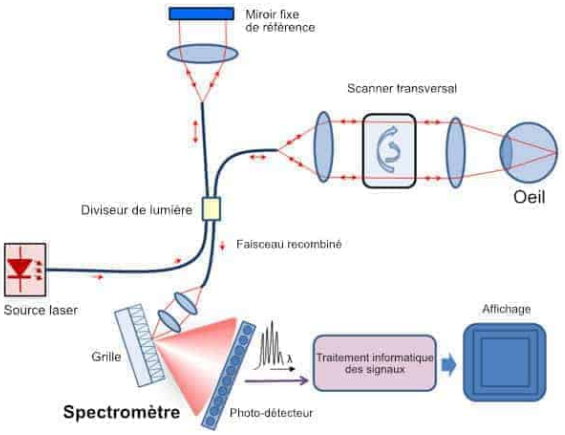
- **Suivi régulier de la glycémie :** Maintenir une glycémie stable pour éviter les fluctuations importantes qui endommagent les vaisseaux rétiniens.
- **Contrôle de la pression artérielle et du cholestérol :** Ces facteurs peuvent contribuer à la progression de la rétinopathie, c'est pourquoi leur gestion est cruciale.
- **Dépistage annuel :** Les personnes diabétiques doivent subir un examen de fond d'œil une fois par an, dès le diagnostic du diabète. Un suivi plus fréquent peut être recommandé si des anomalies sont détectées.

III.3 Techniques de diagnostic

Le diagnostic de la rétinopathie diabétique repose sur plusieurs examens ophtalmologiques, illustrés à la Figure 1.3 :

- **Examen du fond d'œil** : Cet examen permet de visualiser directement la rétine et de détecter des anomalies comme les micro-anévrismes[20], les hémorragies, et les néovaisseaux.
- **Tomographie par cohérence optique (OCT)** : Cette technique permet de visualiser les couches de la rétine et d'évaluer l'épaisseur de la macula pour détecter un œdème maculaire [30].
- **Angiographie fluorescéinique** : Injection d'un colorant pour visualiser les vaisseaux rétiniens et détecter des fuites, des occlusions ou des néovaisseaux dans la rétine [21].
- **Échographie en mode B** : Utilise des ondes ultrasonores pour examiner la rétine et détecter des anomalies comme des hémorragies ou des décollements, particulièrement utile en cas d'opacité du cristallin [19].

(a) Examen du fond d'œil [20]



(b) Tomographie par cohérence optique [30]

(c) Angiographie fluorescéinique [21]

(d) Échographie en mode B [19]

Figure 1.3: Techniques de diagnostic

III.4 Traitement

Le traitement de la rétinopathie diabétique dépend du type et du stade de la maladie. Dans les formes plus avancées, plusieurs options thérapeutiques peuvent être envisagées [26]:

- le traitement au laser (photocoagulation rétinienne) ;
- la thérapie par injections intravitréennes d'agents anti-VEGF (facteur de croissance de l'endothélium vasculaire) ;
- l'injection de corticostéroïdes ;
- la vitrectomie, en cas de complications sévères comme l'hémorragie intravitréenne ou le décollement de la rétine.

Face aux limites des méthodes traditionnelles de diagnostic de la rétinopathie diabétique, l'intelligence artificielle s'impose comme un outil innovant pour améliorer la détection précoce et le suivi de cette pathologie. Elle permet d'automatiser l'analyse des images médicales, de réduire les erreurs humaines et de renforcer la précision du diagnostic, même dans des contextes à ressources limitées.

IV Intelligence Artificielle

L'intelligence artificielle (IA) désigne un ensemble de techniques qui permettent aux machines de simuler des fonctions cognitives humaines, telles que l'apprentissage, la perception, la prise de décision et la résolution de problèmes. L'IA est principalement utilisée pour automatiser des tâches complexes et pour analyser de grandes quantités de données, ce qui permet d'identifier des motifs, des tendances et des anomalies qui échappent souvent à l'œil humain[60].

Les domaines de l'apprentissage automatique (machine learning) et de l'apprentissage profond (deep learning) sont des sous-ensembles de l'IA, tandis que l'intelligence artificielle explicable (XAI) représente une approche transversale visant à interpréter et expliquer les décisions prises par ces modèles. Comme le montre la figure 1.4, ces concepts s'articulent selon une structure imbriquée.

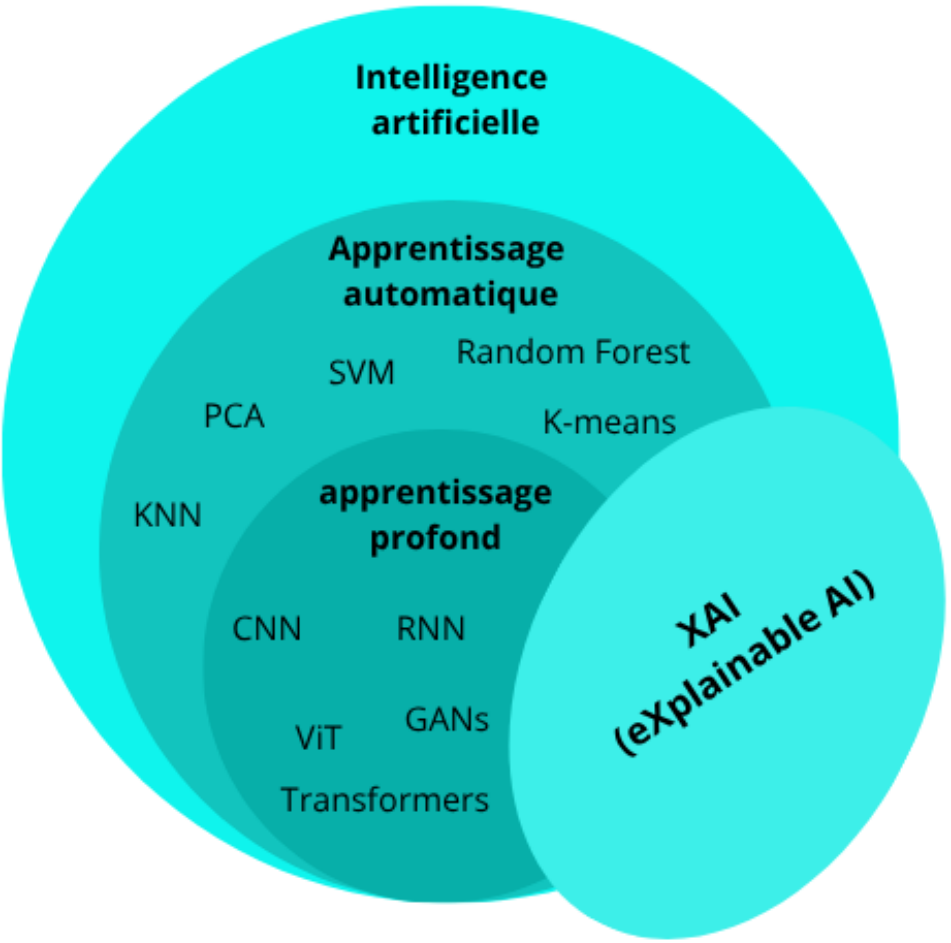


Figure 1.4: Illustration de la relation entre les concepts clés de l'IA, incluant l'apprentissage automatique, l'apprentissage profond et l'IA explicable (XAI)

V Apprentissage automatique

Le machine learning (ML) est un sous-ensemble de l'IA qui permet à une machine d'apprendre et d'améliorer ses performances sur une tâche sans avoir été explicitement programmée pour cela. Contrairement aux approches traditionnelles où un algorithme est écrit pour résoudre un problème, le ML permet à l'algorithme d'apprendre des données historiques et d'adapter ses prévisions ou décisions en conséquence [60].

V.1 Apprentissage Supervisé

L'apprentissage supervisé repose sur l'utilisation d'un ensemble de données étiquetées, dans lequel chaque exemple comporte une entrée et une sortie cible. L'algorithme apprend ainsi la relation entre les variables d'entrée et de sortie, afin de pouvoir prédire ou classer correctement de nouvelles données. La figure 1.5 illustre ce principe général, où le modèle est entraîné à partir d'exemples annotés pour généraliser à de nouveaux cas. Parmi les approches les plus couramment utilisées, nous pouvons citer :

- **Support Vector Machines (SVM) :** Cette méthode vise à trouver un hyperplan optimal séparant les différentes classes dans un espace à plusieurs dimensions [42]. En ophtalmologie, elle est souvent utilisée pour distinguer les stades de la RD à partir de caractéristiques extraites d'images rétiniennes.
- **Régression logistique :** Technique probabiliste permettant de prédire l'appartenance à une classe (par exemple, présence ou absence de RD sévère), en modélisant la relation entre des variables explicatives et une variable binaire ou multinomiale [29].
- **K-Nearest Neighbors (KNN) :** Algorithme non paramétrique qui classe une nouvelle observation en fonction des classes majoritaires parmi ses k plus proches voisins dans l'espace des caractéristiques [29].
- **Random Forests :** Méthode d'ensemble reposant sur une collection d'arbres de décision construits à partir de sous-échantillons aléatoires des données. Elle permet d'obtenir des prédictions robustes et précises, et est particulièrement adaptée à la classification d'images médicales, y compris celles de la RD [29].

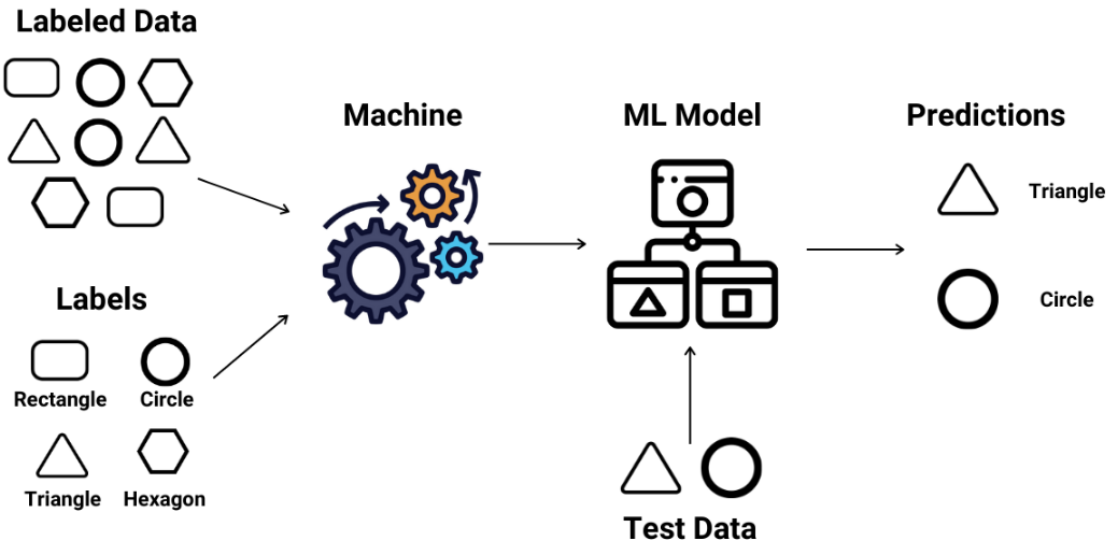


Figure 1.5: Illustration du principe de l'apprentissage supervisé, dans lequel le modèle apprend à partir d'exemples étiquetés afin de prédire ou classer de nouvelles instances [27].

V.2 Apprentissage Non supervisé

Dans l'apprentissage non supervisé, l'algorithme découvre des structures sous-jacentes dans les données sans utiliser d'étiquettes. Il peut regrouper des données similaires ou réduire la dimensionnalité pour extraire les caractéristiques principales. La figure 1.6 illustre ce principe, où l'analyse se base uniquement sur la structure intrinsèque des données. Parmi les approches les plus courantes, nous pouvons citer :

- **Clustering :** Utilisé pour regrouper des données similaires sans information préalable sur les catégories. Par exemple, l'algorithme K-means peut identifier des groupes naturels au sein d'un ensemble de données en se basant uniquement sur les caractéristiques observées [23].
- **Analyse en Composantes Principales (PCA) :** Méthode de réduction de dimensionnalité permettant de projeter les données dans un espace de moindre dimension tout en conservant l'essentiel de l'information. Elle facilite l'analyse, la visualisation et le prétraitement des données [36].
- **Autoencoders :** Type de réseau de neurones artificiels conçu pour apprendre des représentations compressées et efficaces des données, souvent utilisé pour la réduction de dimension, la détection d'anomalies ou la pré-initialisation de modèles profonds [41].

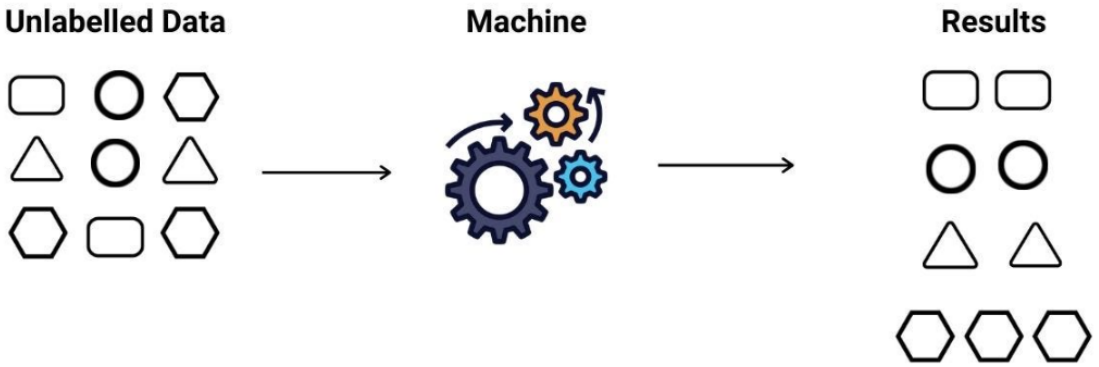


Figure 1.6: Schéma de l'apprentissage non supervisé [27]

V.3 Apprentissage Semi-supervisé

L'apprentissage semi-supervisé combine un petit nombre de données étiquetées avec une grande quantité de données non étiquetées afin d'améliorer la performance des modèles. Cette approche est particulièrement utile lorsque l'annotation manuelle des données est coûteuse et difficile à obtenir. La figure 1.7 illustre le principe général de ce type d'apprentissage.

- **Méthodes de propagation de labels :** Techniques visant à propager les étiquettes d'un petit ensemble annoté vers un grand ensemble non étiqueté, souvent à l'aide de graphes ou de modèles de similarité [70].
- **Semi-supervised Support Vector Machines (S3VM) :** Variante des SVM intégrant des données non étiquetées dans le processus d'optimisation de la frontière de décision [70].
- **Autoencoders semi-supervisés :** Réseaux de neurones combinant apprentissage non supervisé (via reconstruction) et supervision partielle pour apprendre des représentations utiles à la classification [39].

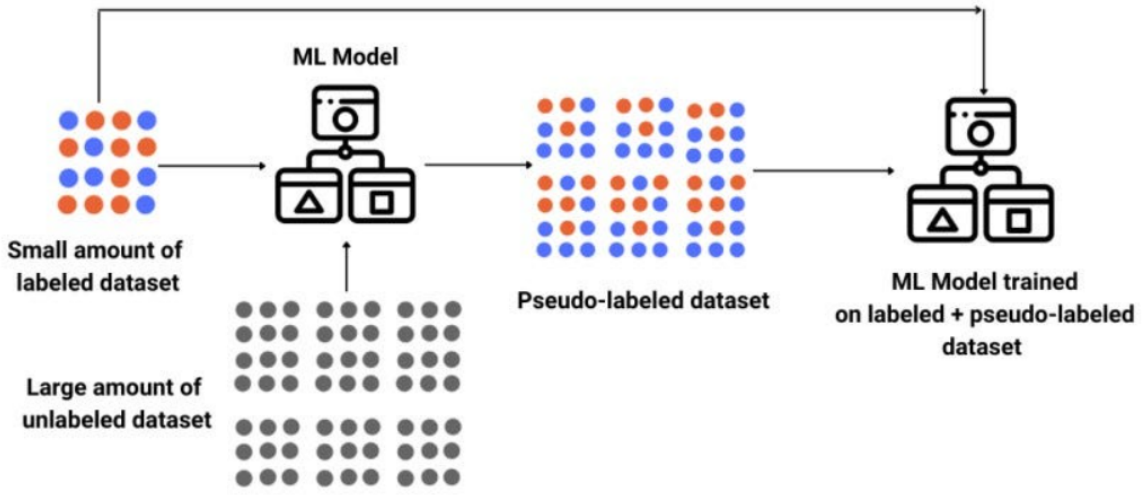


Figure 1.7: Illustration du processus d'apprentissage semi-supervisé, combinant données étiquetées et non étiquetées pour entraîner un modèle plus performant [27].

V.4 Apprentissage par Renforcement

L'apprentissage par renforcement repose sur l'interaction d'un agent avec un environnement. L'agent apprend à prendre des décisions séquentielles en recevant des récompenses (ou punitions) selon les actions entreprises [49]. L'objectif est de maximiser la récompense cumulative à long terme. Contrairement aux autres types d'apprentissage, il n'y a pas de supervision directe mais un système de feedback basé sur l'expérience. La figure 1.8 illustre le fonctionnement général de ce paradigme. Voici quelques algorithmes représentatifs :

- **Q-Learning :** Algorithme hors-politique qui utilise une table de valeurs (Q-table) pour estimer la qualité d'une action dans un état donné. Il repose sur la mise à jour itérative des valeurs Q selon l'équation de Bellman. Très utilisé dans des environnements simples où les états sont discrets [78].
- **Deep Q-Network (DQN) :** Extension du Q-Learning utilisant un réseau de neurones profond pour approximer la Q-table. Il est adapté aux environnements complexes à grands espaces d'états comme les jeux vidéo, où une table explicite serait impraticable [49].
- **Actor-Critic :** Méthode hybride combinant deux composants : l'*Actor*, qui apprend la politique (quelle action faire), et le *Critic*, qui estime la valeur de l'état ou de l'action. Cette architecture permet des mises à jour plus stables et plus rapides [40].
- **REINFORCE :** Méthode de gradient de politique (policy gradient) stochastique. L'agent collecte des trajectoires complètes et met à jour les paramètres de la politique proportionnellement à la récompense reçue. Simple à implémenter mais souvent instable ou lent à converger [80].

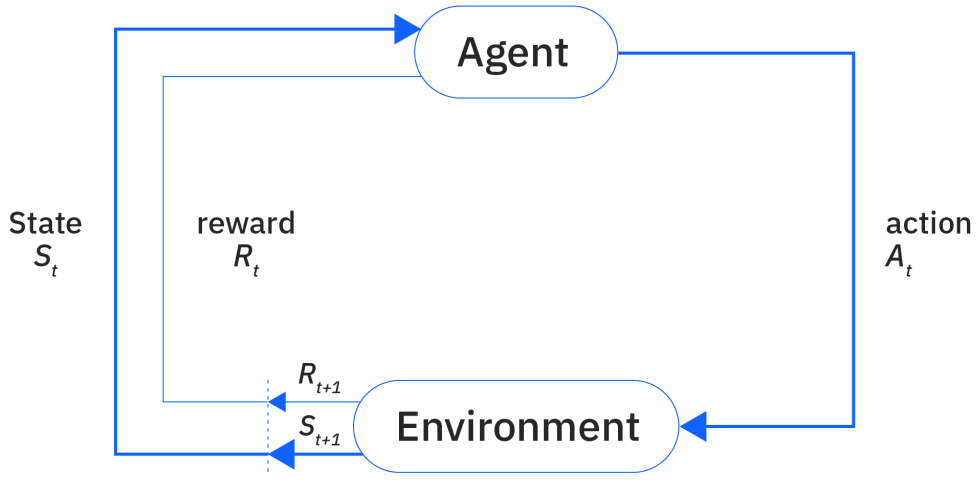


Figure 1.8: Illustration du fonctionnement de l'apprentissage par renforcement, où un agent interagit avec un environnement et ajuste sa stratégie en fonction des récompenses reçues [27].

VI Apprentissage Profond

Le deep learning est un sous-ensemble du machine learning qui utilise des réseaux de neurones artificiels à plusieurs couches (d'où le terme "profond") pour modéliser des représentations complexes de données. Ces réseaux peuvent apprendre des représentations hiérarchiques de données, permettant d'extraire des caractéristiques de plus en plus abstraites à mesure que les données passent par les différentes couches du réseau [60].

VI.1 Réseaux de Neurones Profonds

Les réseaux de neurones profonds sont constitués de plusieurs couches cachées situées entre la couche d'entrée et la couche de sortie. Chaque couche apprend à extraire des représentations de plus en plus abstraites et complexes des données d'entrée. Grâce à cette architecture hiérarchique, ces réseaux sont capables de modéliser des relations non linéaires complexes, ce qui les rend particulièrement performants dans des tâches telles que la classification, la régression ou la reconnaissance de formes. Ils sont couramment utilisés dans des domaines nécessitant l'apprentissage automatique de caractéristiques à différents niveaux d'abstraction (voir la Figure 1.9).

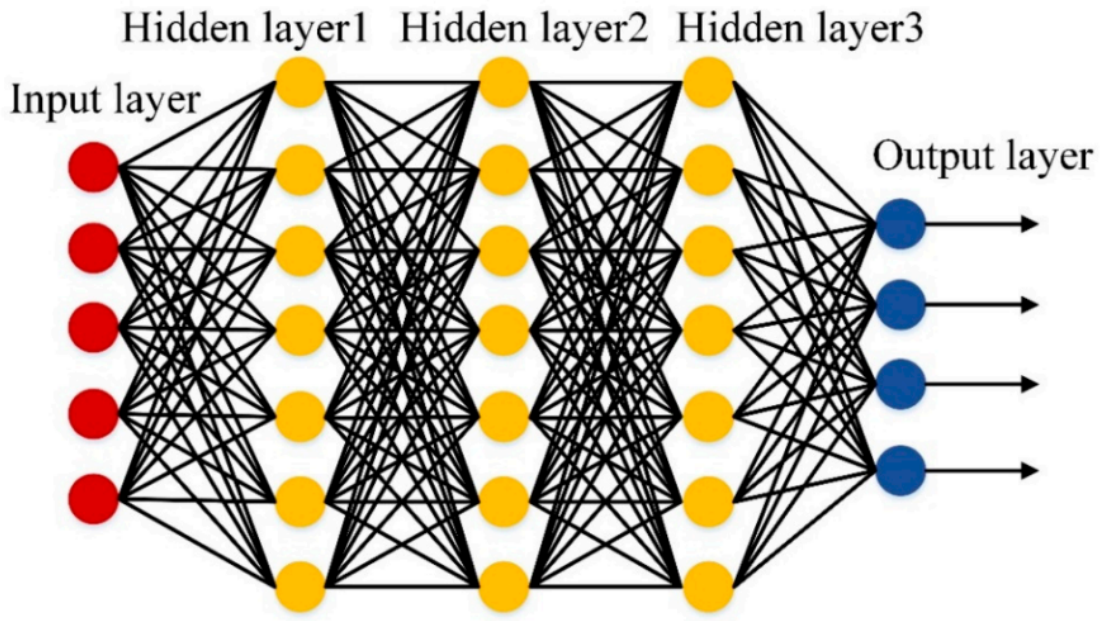


Figure 1.9: Représentation d'un réseau de neurones multicouches [52]

VI.2 Réseaux de Neurones Récurrents

Les réseaux de neurones récurrents (*Recurrent Neural Networks* -RNN) sont conçus pour traiter des données séquentielles ou temporelles. Contrairement aux réseaux de neurones classiques, les RNN intègrent des connexions récurrentes formant des boucles, ce qui leur permet de mémoriser et de transmettre l'information d'une étape à l'autre dans une séquence. Cette capacité les rend particulièrement adaptés aux tâches impliquant des séries temporelles ou des données séquentielles, telles que le traitement du langage naturel ou l'analyse de signaux audio.

Cependant, les RNN classiques présentent certaines limitations lorsqu'ils sont confrontés à de longues séquences, en raison du problème de disparition ou d'explosion du gradient lors de l'apprentissage. Pour pallier ces difficultés, l'architecture *Long Short-Term Memory* (LSTM) a été proposée. Les LSTM constituent une amélioration des RNN traditionnels, capable de mieux gérer les dépendances à long terme grâce à un mécanisme de portes (oubli, entrée et sortie) qui contrôle le flux d'information et permet de préserver les informations pertinentes sur de longues périodes.

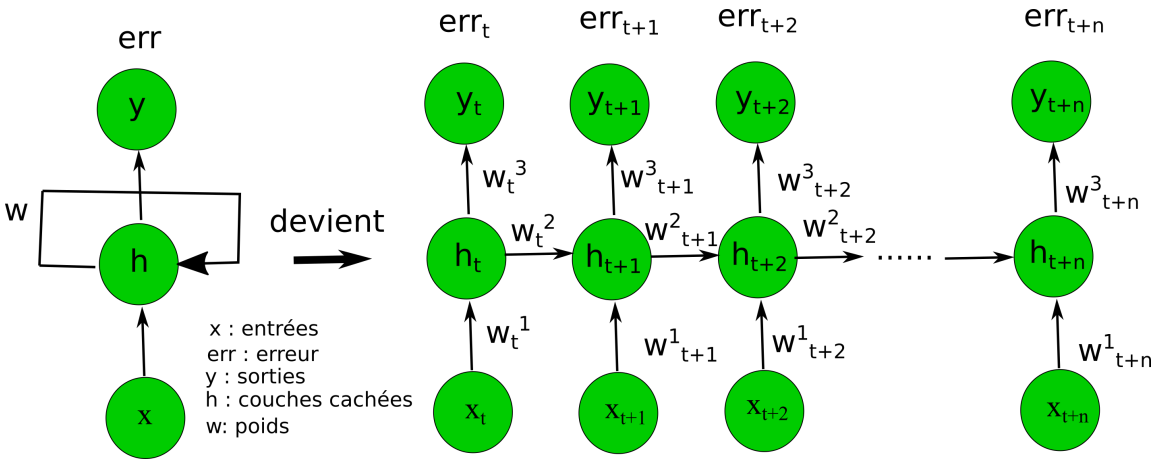


Figure 1.10: Fonctionnement d'un réseau neuronal récurrent[52]

VI.3 Réseaux de Neurones Convolutifs

Les réseaux de neurones convolutifs (CNN) sont des architectures particulièrement efficaces pour l'analyse d'images. Ces réseaux sont composés de couches de convolution, qui appliquent des filtres afin d'extraire des caractéristiques locales des données d'entrée, telles que les bords, les textures ou les formes. Ensuite, des couches de sous-échantillonnage (ou *pooling*) permettent de réduire la taille des représentations tout en conservant l'information essentielle. Les CNN sont largement utilisés dans le domaine de la vision par ordinateur ainsi que dans d'autres tâches d'analyse d'images. L'architecture générale d'un CNN est illustrée à la figure 1.11.

Les CNN reposent principalement sur trois types de couches fondamentales :

1. **Couche de convolution :** La convolution est une opération mathématique clé utilisée en traitement d'images et de signaux numériques. Elle permet d'extraire automatiquement des caractéristiques pertinentes à partir des images d'entrée en appliquant des filtres (ou noyaux) entraînables. Chaque filtre est représenté par une matrice de poids, appliquée sur des régions locales de l'image pour produire une carte de caractéristiques. Contrairement aux couches entièrement connectées classiques, les couches de convolution n'utilisent pas de connexions globales, mais se concentrent localement sur les motifs, ce qui rend le traitement plus efficace et moins coûteux en paramètres [15].

2. **Couche de pooling :** Placée généralement après une couche de convolution, la couche de pooling a pour objectif de réduire la dimensionnalité des cartes de caractéristiques, tout en préservant les informations essentielles. Les techniques les plus courantes sont le *max pooling* et le *average pooling*. Le pooling permet ainsi de limiter la complexité computationnelle et d'éviter une explosion du nombre de paramètres lors de l'empilement des couches [15].

3. **Couche entièrement connectée :** À la fin du processus d'extraction de caractéristiques, une ou plusieurs couches entièrement connectées (*Fully Connected layers*) sont ajoutées pour réaliser la classification finale. Chaque neurone de ces couches est connecté à toutes les sorties de la couche précédente, ce qui permet de combiner l'ensemble des informations extraites pour prédire la classe de l'échantillon. Le vecteur en sortie de la partie convolutionnelle (appelé parfois *code CNN*) est ainsi transformé en une décision finale [13].

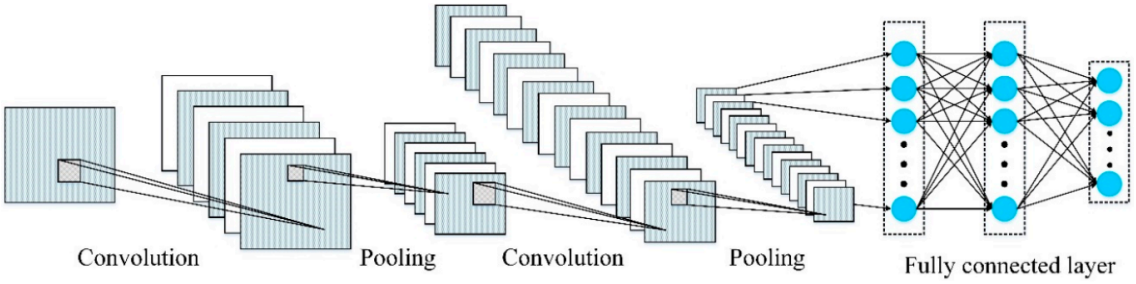


Figure 1.11: Architecture d'un réseau de neurones convolutif [52]

VI.4 Les Transformers

Les Transformers, introduits par Vaswani et al. (2017) [73], sont des architectures de deep learning basées sur un mécanisme d'attention. Initialement conçus pour le traitement du langage naturel (NLP), ils utilisent des blocs encodeurs et décodeurs dotés de modules d'auto-attention et d'attention multi-têtes, permettant de capturer efficacement les dépendances contextuelles.

Afin de mieux comprendre le fonctionnement interne des Transformers, il convient d'examiner en détail les principaux composants qui les constituent. La figure 1.12 illustre la structure complète d'un modèle Transformer.

VI.4.1 Mécanisme d'attention

L'attention simule le processus cognitif humain en se concentrant sur les informations pertinentes d'une entrée tout en négligeant celles jugées moins importantes. Une fonction d'attention associe un vecteur de requête à un ensemble de paires clé-valeur pour générer un vecteur de sortie. Le mécanisme d'Attention à produit scalaire mis à l'échelle (Scaled Dot-Product Attention), proposé par (Vaswani et al., 2017) [73], est formulé ainsi :

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- **Q** sert à poser des questions aux autres patches : « À qui dois-je faire attention ? »
- **K** sert à déterminer à quel point chaque patch est pertinent pour ces questions.
- **V** contient l'information que l'on va réellement utiliser pour construire la nouvelle représentation du patch.

Et d_k désigne la dimension des requêtes et des clés. Le facteur d'échelle $\frac{1}{\sqrt{d_k}}$ est introduit pour éviter la saturation de la fonction softmax, qui rendrait les gradients extrêmement faibles lorsque d_k est élevé [73].

VI.4.2 Attention Multi-têtes

L'attention multi-têtes consiste à appliquer simultanément plusieurs mécanismes d'attention afin de capturer diverses informations de l'entrée sous différents angles. Cette approche est particulièrement adaptée aux tâches de traitement du langage naturel, où il est crucial d'analyser les relations complexes entre les mots.

Concrètement, les matrices Q, K et V sont projetées linéairement vers différentes représentations avant d'être soumises à la fonction d'attention. Ce processus est répété h fois en parallèle [73]. Les résultats sont ensuite concaténés et projetés de nouveau dans un espace adapté:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

avec :

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

où W_i^Q, W_i^K, W_i^V et W^O sont des matrices de poids apprises spécifiques à chaque tête d'attention.

Lorsque $K = V$, on parle de self-attention, et dans le cas multi-têtes, de Multi-Head Self-Attention (MSA).

VI.4.3 Encodage positionnel

Le Transformer traite les entrées simultanément sans tenir compte de leur ordre. Pour intégrer l'information séquentielle [73], un encodage positionnel est ajouté à chaque entrée. Une méthode fréquente repose sur l'utilisation de fonctions sinusoïdales de différentes fréquences, définies par :

$$\text{PE}(\text{pos}, 2i) = \sin\left(\frac{\text{pos}}{10000^{\frac{2i}{d_{\text{model}}}}}\right), \quad \text{PE}(\text{pos}, 2i+1) = \cos\left(\frac{\text{pos}}{10000^{\frac{2i}{d_{\text{model}}}}}\right)$$

où `pos` représente la position dans la séquence et i l'indice de la dimension dans le vecteur d'embedding.

VI.4.4 Bloc MLP et Structure Globale du Transformer

Après le bloc de self-attention, les sorties sont transmises à un perceptron multicouche (MLP),[73, 10] constitué de deux couches entièrement connectées avec une fonction d'activation non linéaire, souvent ReLU. Formellement, pour une entrée x :

$$\text{MLP}(x) = (xW_1 + b_1)\,W_2 + b_2$$

De plus, une normalisation de couche (Layer Normalization, LN) est appliquée après les connexions résiduelles, selon la structure suivante :

$$x'_l = \text{LN}(x_l + \text{MSA}(x_l)), \quad x_{l+1} = \text{LN}(x'_l + \text{MLP}(x'_l))$$

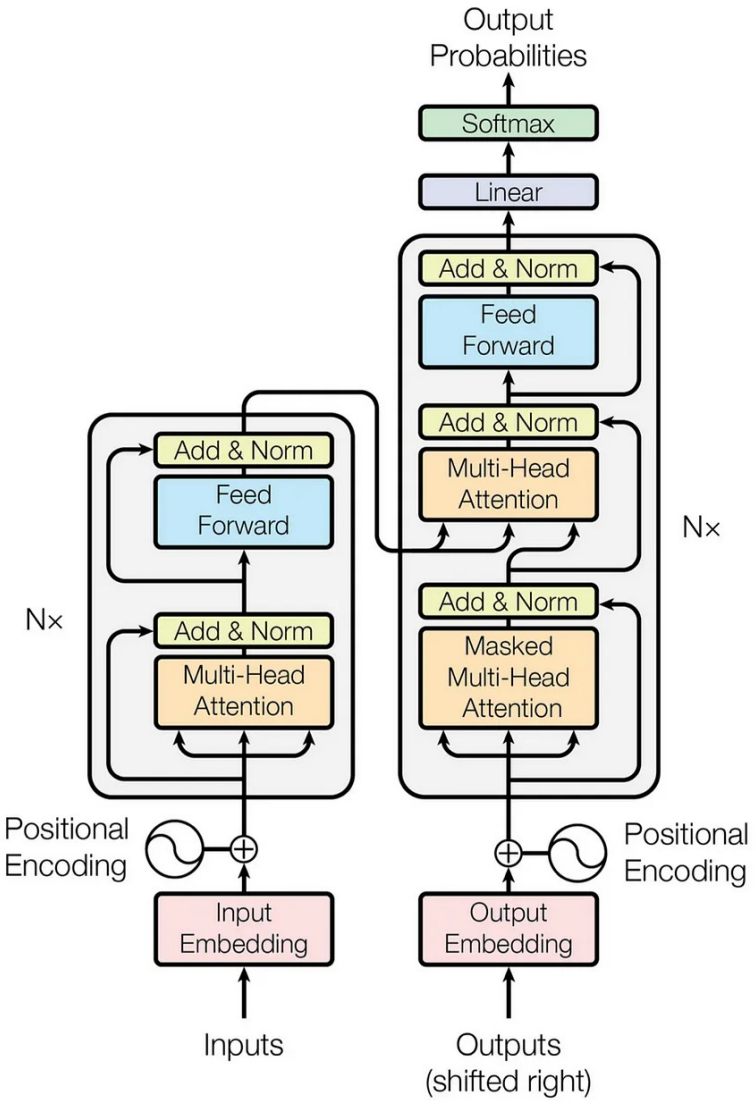


Figure 1.12: Architecture du modèle Transformer [73]

VI.5 Vision Transformers

Cherchant à reproduire le succès des Transformers en NLP, [25] ont proposé le *Vision Transformer* (ViT). L'image est divisée en petits patchs, chacun étant projeté en un vecteur d'une dimension spécifique, formant ainsi une séquence analogique aux mots dans le NLP. Contrairement aux CNNs qui traitent des régions locales via des filtres convolutifs, le ViT capture des relations globales grâce au mécanisme d'attention.

L'architecture du ViT repose uniquement sur l'encodeur Transformer. Elle se compose des étapes suivantes, comme illustré à la figure 1.13:

- **Découpage en patchs** : L'image d'entrée est de taille $x \in \mathbb{R}^{H \times W \times C}$, où H et W représentent respectivement la hauteur et la largeur de l'image, et C le nombre de canaux (par exemple, $C = 3$ pour une image RGB) [25]. Elle est divisée en N patchs non superposés de taille $P \times P$.

$$N = \frac{H \cdot W}{P^2}$$

Chaque patch est aplati en un vecteur de dimension $P^2 \cdot C$, puis projeté dans un espace de dimension D via une transformation linéaire W_p :

$$x_p^i \in \mathbb{R}^{P^2 \cdot C} \xrightarrow{W_p} z_0^i \in \mathbb{R}^D$$

Cela permet de transformer chaque patch image en une représentation vectorielle utilisable par le Transformer.

- **Encodage positionnel** : Contrairement aux CNNs, les Transformers sont invariants à l'ordre des tokens [25]. Pour donner un sens à la position de chaque patch dans l'image, un encodage positionnel appris E_{pos} est ajouté aux embeddings :

$$Z_0 = \{z_0^{[\text{CLS}]}, z_0^1, ..., z_0^N\} + E_{\text{pos}}$$

Ici, $z_0^{[\text{CLS}]}$ est un token spécial inséré au début de la séquence, servant à agréger l'information globale. L'encodage positionnel E_{pos} est de même dimension que les embeddings.

- **Token [CLS]** : Le vecteur spécial $z_0^{[\text{CLS}]}$ est initialisé comme un paramètre appris, et il est utilisé comme représentation finale pour la classification après passage à travers les couches Transformer [25] .

- **Blocs Transformer** : La séquence complète (contenant les patchs encodés + [CLS]) passe à travers L blocs Transformer [25] . Chaque bloc est composé :

  – D'une couche *Multi-Head Self-Attention (MSA)*, qui capture les dépendances globales entre les patches. Chaque tête d'attention utilise la formule :

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

  où :

  * **Q** sert à poser des questions aux autres patches : « À qui dois-je faire attention ? »
  * **K** sert à déterminer à quel point chaque patch est pertinent pour ces questions.
  * **V** contient l'information que l'on va réellement utiliser pour construire la nouvelle représentation du patch.
  * d_k est la dimension des clés (souvent égale à D/h où h est le nombre de têtes).
  * Le terme $\frac{QK^\top}{\sqrt{d_k}}$ permet de normaliser les produits scalaires pour éviter des gradients trop grands.

  – D'un perceptron multicouche (*MLP*) avec une fonction d'activation non linéaire GELU.

  – De mécanismes de normalisation de couche (*LayerNorm*) et de connexions résiduelles qui stabilisent l'apprentissage. Ces opérations sont décrites par les équations suivantes :

$$\hat{z}_\ell = \text{MSA}(\text{LN}(z_{\ell-1})) + z_{\ell-1}$$
$$z_\ell = \text{MLP}(\text{LN}(\hat{z}_\ell)) + \hat{z}_\ell$$

où $z_{\ell-1}$ est l'entrée du bloc ℓ, $\hat{z}_\ell$ est la sortie après l'attention et normalisation, et z_ℓ est la sortie finale du bloc après passage dans le MLP.

- **MLP Head** : À la sortie du dernier bloc Transformer (L), seul le token $z_L^{[\text{CLS}]}$ est conservé pour la tâche de classification [25]. Ce vecteur est envoyé à une couche dense (MLP) suivie d'une fonction softmax pour produire les probabilités des classes :

$$\hat{y} = \text{softmax}(W \cdot z_L^{[\text{CLS}]} + b)$$

où W et b sont les poids et biais de la couche linéaire finale, et $\hat{y}$ est un vecteur de probabilités représentant les scores des classes.

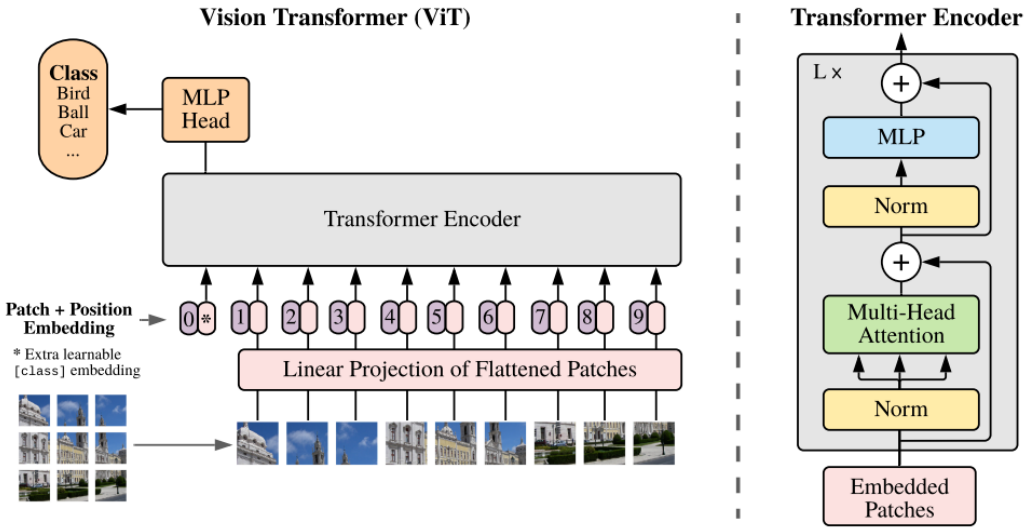


Figure 1.13: Architecture d'un Vision Transformer[25]

VII L'IA dans la Rétinopathie Diabétique

Apport du Deep Learning

L'apprentissage profond a significativement amélioré le diagnostic automatisé de la rétinopathie diabétique. Les réseaux de neurones convolutifs (CNN) sont largement utilisés pour analyser les images de fond d'œil, permettant de détecter des anomalies telles que les micro-anévrismes, les hémorragies ou les exsudats. Ces modèles peuvent classifier les images selon les stades de la maladie et parfois surpasser les ophtalmologistes en précision et rapidité [51].

Émergence des Vision Transformers

Plus récemment, les Vision Transformers (ViT) ont été proposés comme une alternative performante aux CNN. Contrairement aux CNN, les ViT utilisent des mécanismes d'attention pour modéliser les relations globales dans l'image. Cela leur permet de capturer des motifs complexes avec une précision souvent supérieure dans les tâches de classification des stades de la RD [51].

VIII Applications Cliniques de l'IA dans la RD

Diagnostic assisté par IA

Les systèmes basés sur l'IA sont capables d'identifier les signes précoces de la RD avec une grande fiabilité. En plus d'accélérer le processus de diagnostic, ils apportent une standardisation et réduisent les erreurs humaines, rendant le diagnostic plus cohérent.

Dépistage et Suivi Améliorés

L'IA permet d'optimiser les campagnes de dépistage de la RD, notamment en réduisant les taux de faux positifs et faux négatifs. Elle permet également un triage intelligent des patients, ce qui facilite le travail des ophtalmologistes et améliore la gestion des ressources médicales [32].

Rôle spécifique des ViT

Les ViT, en exploitant les dépendances globales entre les pixels d'image, améliorent la détection de stades complexes de la RD. Ils constituent aujourd'hui un outil de choix dans le développement de solutions IA de nouvelle génération pour l'analyse rétinienne.

IX Conclusion

Ce chapitre introductif nous a permis de poser les bases médicales et technologiques nécessaires à la compréhension de notre problématique. Nous avons d'abord rappelé les éléments essentiels relatifs à la rétinopathie diabétique, en expliquant son origine, ses manifestations cliniques, ainsi que l'importance d'un dépistage précoce et d'un suivi régulier pour en limiter les effets. Ensuite, nous avons présenté les fondements de l'intelligence artificielle, en particulier le *machine learning* et le *deep learning*, qui constituent aujourd'hui des approches majeures dans le traitement automatisé des données médicales.

Nous avons examiné les principales architectures de réseaux de neurones utilisées pour l'analyse d'images, notamment les CNN et les ViT, en soulignant leurs performances respectives dans les tâches de classification, notamment pour les images de fond d'œil utilisées dans la détection de la rétinopathie diabétique (RD).

Le chapitre suivant sera ainsi consacré à l'exploration des principales techniques explicatives, visant à rendre les prédictions des modèles d'IA plus intelligibles, interprétables et acceptables pour les professionnels de santé.

Chapitre 2

L'Explicabilité en Intelligence Artificielle

I Introduction

De nombreux modèles d'IA modernes, notamment les réseaux de neurones profonds, fonctionnent comme des « boîtes noires », c'est à dire qu'ils génèrent des résultats sans que leur raisonnement soit directement compréhensible par un humain [50]. Cette opacité devient problématique dans les domaines sensibles où il est indispensable d'expliquer et de justifier les décisions prises par un système automatisé.

II Définition de l'XAI

L'intelligence artificielle explicable, ou *XAI* (pour *eXplainable Artificial Intelligence*), regroupe un ensemble de méthodes et de techniques visant à rendre les décisions des modèles d'intelligence artificielle compréhensibles pour les utilisateurs humains. Contrairement aux modèles dits "boîte noire", dont le fonctionnement interne est difficilement interprétable, l'XAI cherche à clarifier le raisonnement du modèle, en répondant à des questions telles que : "pourquoi cette prédiction a-t-elle été faite ?" ou "quels éléments ont influencé la décision ?"

Selon Adadi et Berrada (2018), l'XAI désigne "tout ensemble de processus qui produit des descriptions compréhensibles de modèles et de prédictions complexes d'IA" [5].

II.1 Motivations et enjeux de l'XAI

L'intelligence artificielle explicable (XAI) joue un rôle essentiel dans l'acceptation et l'intégration de l'IA dans les domaines critiques. Tout d'abord, elle favorise la **transparence** des systèmes automatisés, ce qui renforce la **confiance** des utilisateurs finaux en rendant les décisions plus compréhensibles et traçables [83, 57]. Cette transparence contribue également à la **responsabilisation** des modèles d'IA, en facilitant la détection de biais potentiels et en permettant de respecter les exigences **éthiques et réglementaires**, notamment dans des secteurs sensibles comme la santé ou la finance [83].

En outre, l'XAI constitue un outil précieux pour le **débogage et l'amélioration des modèles**, elle aide les experts à mieux comprendre les erreurs, à affiner les algorithmes et à améliorer leurs performances globales [57]. Enfin, dans un contexte de régulation croissante, l'explicabilité facilite la **conformité aux cadres légaux**, en répondant aux exigences de

transparence, de traçabilité et de redevabilité imposées aux systèmes d'intelligence artificielle [83].

III Les Concepts Fondamentaux de l'XAI

Avant d'aborder les différentes approches méthodologiques de l'explicabilité, il est essentiel de clarifier les concepts clés qui structurent ce domaine et qui sont souvent employés de manière interchangeable dans la littérature.

III.1 Définitions et terminologie fondamentale

Plusieurs concepts fondamentaux permettent de structurer le champ de l'intelligence artificielle explicable. Bien que souvent utilisés de manière interchangeable, des distinctions subtiles existent entre explicabilité, interprétabilité, transparence et fidélité.

L'explicabilité désigne la capacité d'un système à fournir un raisonnement compréhensible expliquant ses décisions. Elle s'appuie sur des informations transparentes, contextualisées à la lumière des connaissances du domaine, dans le but d'offrir une compréhension significative à un utilisateur donné [69].

L'interprétabilité est souvent considérée comme une condition préalable à l'explicabilité. Elle renvoie à la facilité avec laquelle un humain peut comprendre les mécanismes internes d'un modèle, c'est-à-dire la relation entre les entrées et les sorties du système [22].

La transparence se rapporte à la clarté du fonctionnement interne d'un modèle. Un modèle est dit transparent si sa logique de décision peut être entièrement retracée et comprise sans recours à des techniques externes d'analyse [16].

La fidélité correspond à la capacité d'une explication à représenter de manière précise le comportement réel du modèle. Une explication fidèle ne doit pas simplifier ou altérer la logique du système, au risque de produire une compréhension erronée de ses décisions [85, 74].

La figure 2.1 illustre les relations conceptuelles entre l'explicabilité et ses notions connexes, notamment l'interprétabilité, la transparence et la fidélité.

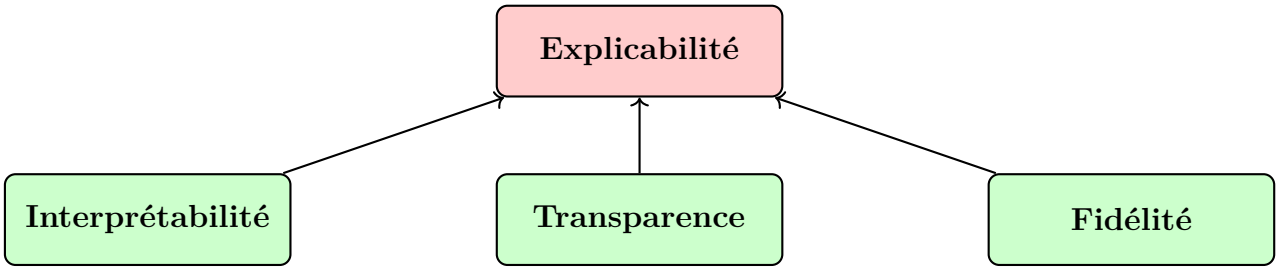


Figure 2.1: Relations entre les concepts fondamentaux en XAI

III.2 Le paradigme des boîtes noires et des boîtes blanches

Les modèles d'intelligence artificielle sont souvent classés en deux grandes catégories : les **modèles boîte noire** et les **modèles boîte blanche**.

Les modèles boîte noire, tels que les réseaux de neurones profonds, les Transformers ou les modèles d'ensemble, se distinguent par leurs performances élevées, mais leur fonctionnement interne reste largement opaque pour l'utilisateur[31].

À l'inverse, les modèles boîte blanche, comme les régressions linéaires ou les arbres de décision, offrent une transparence totale : leur logique est simple, traçable et compréhensible.

Dans les domaines sensibles (santé, droit, finance), la capacité à justifier une décision est indispensable. L'opacité des modèles boîte noire pose alors des défis éthiques et réglementaires.

Les approches XAI visent justement à **rendre les modèles boîte noire interprétables**, soit en privilégiant des modèles intrinsèquement explicables, soit en appliquant des méthodes d'explication a posteriori (post-hoc)[31].

La figure 2.2 illustre la distinction entre modèles boîte noire et boîte blanche, ainsi que le rôle des méthodes XAI pour rendre les prédictions des modèles complexes plus compréhensibles.

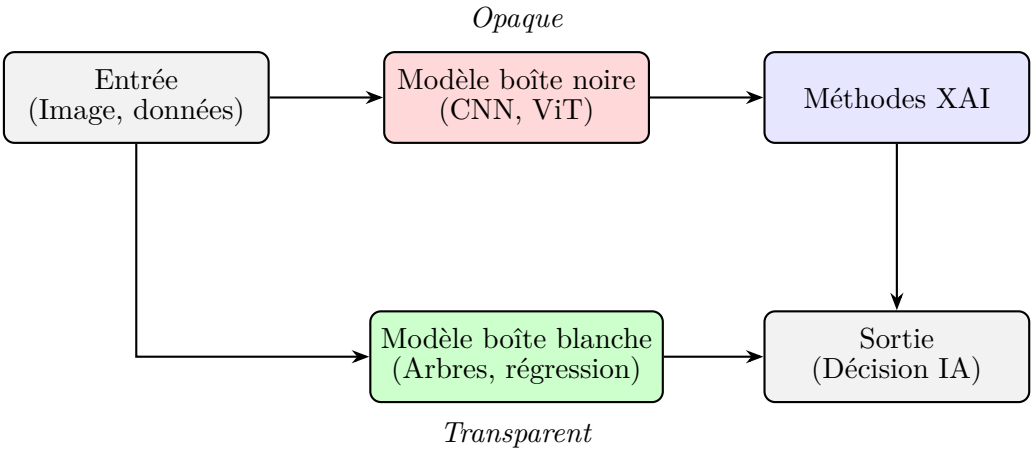


Figure 2.2: Comparaison entre les modèles boîte noire et blanche, et rôle de la XAI

III.3 Les différents niveaux d'explication

Les méthodes d'intelligence artificielle explicable (XAI) peuvent être classées selon le niveau auquel elles opèrent. On distingue principalement deux types : les explications globales, qui visent à interpréter le comportement général d'un modèle, et les explications locales, qui se concentrent sur les décisions prises pour des instances particulières.

III.3.1 Explication au niveau du modèle (Explications Globales)

Les explications globales ont pour objectif de décrire le comportement général d'un modèle d'IA. Elles permettent de comprendre de manière agrégée comment les différentes caractéristiques d'entrée influencent les décisions du modèle [6]. Ces explications sont particulièrement utiles pour évaluer la cohérence interne du modèle, identifier les sources d'erreurs telles que les faux positifs ou faux négatifs [55], et pour analyser la structure décisionnelle sous-jacente [61].

Dans des domaines critiques comme la santé, il a été montré que des explications globales à visée narrative, contextualisées sur des groupes de patients ou des profils types, peuvent répondre efficacement aux attentes des professionnels non spécialistes [66].

III.3.2 Explication au niveau de l'instance (Explications Locales)

Les explications locales, quant à elles, visent à justifier une décision particulière prise par le modèle, en identifiant les caractéristiques spécifiques de l'exemple d'entrée qui ont influencé la prédiction [6]. Elles sont essentielles pour l'analyse de cas individuels, notamment dans les situations où chaque prédiction a des implications importantes, comme en diagnostic médical.

Des recherches récentes montrent que les préférences des utilisateurs varient selon le contexte : certains privilégient des explications globales pour comprendre la logique générale du système, tandis que d'autres préfèrent des explications locales, plus opérationnelles [66].

La figure 2.3 synthétise la distinction entre les explications globales et locales en XAI, en mettant en évidence leurs caractéristiques et objectifs respectifs.

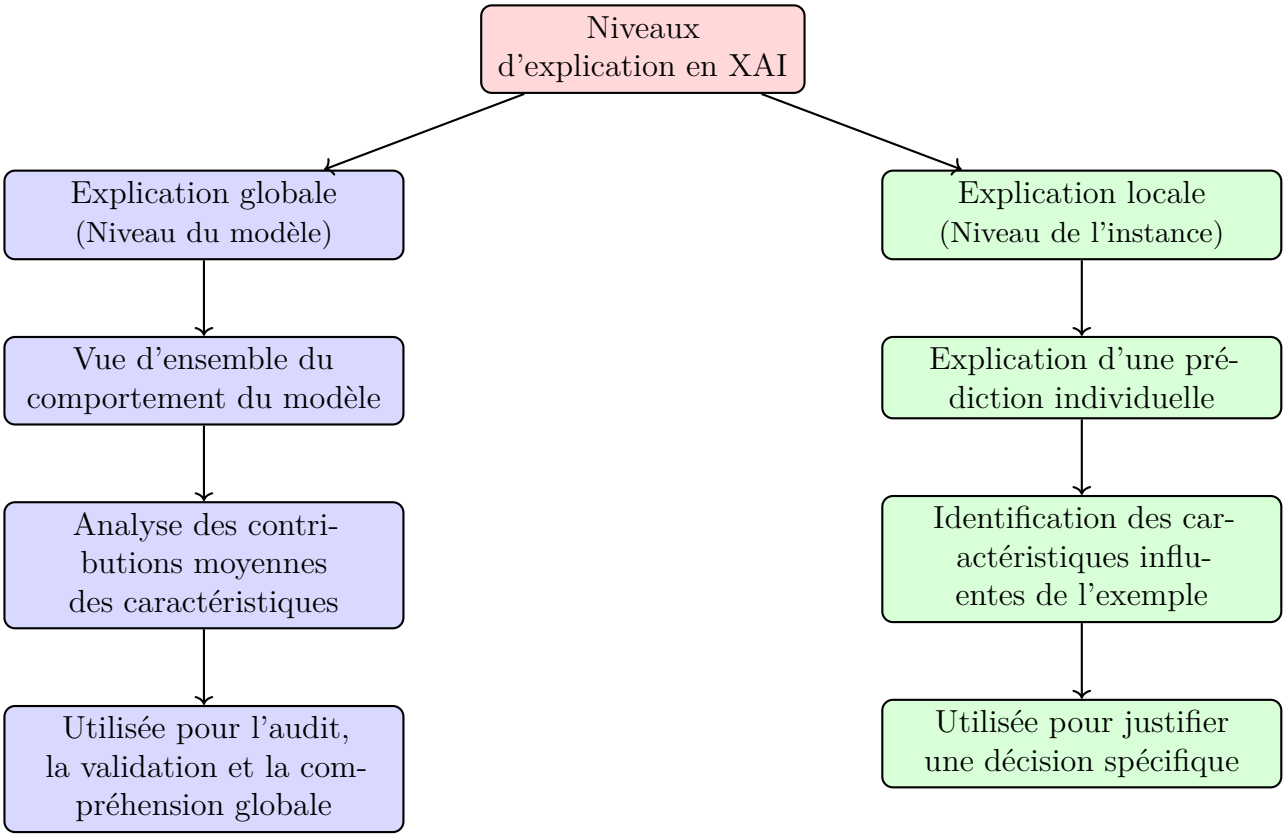


Figure 2.3: Les deux niveaux d'explication en XAI : globale vs locale

III.4 Propriétés des explications en XAI

Les méthodes d'intelligence artificielle explicable (XAI) doivent produire des explications qui soient à la fois informatives, utilisables et adaptées aux utilisateurs finaux. Plusieurs propriétés sont généralement considérées comme souhaitables pour évaluer la qualité d'une explication :

- **Compréhensibilité :** une explication est considérée comme compréhensible lorsqu'elle peut être facilement interprétée et assimilée par un utilisateur non expert. Cette caractéristique est essentielle pour favoriser l'acceptabilité des systèmes d'IA [33, 45].
- **Fidélité :** la fidélité mesure à quel point l'explication reflète fidèlement le comportement réel du modèle. Une explication fidèle ne déforme pas les décisions du modèle, ce qui évite des interprétations erronées [45].
- **Stabilité :** une explication est dite stable si de petites variations de l'entrée ne provoquent pas de changements majeurs dans l'explication. Le PSEM (Path-Sufficient Explanations Method) vise notamment à garantir cette propriété [45].

- **Suffisance :** une explication est suffisante si elle fournit assez d'éléments pour justifier une décision sans surcharge cognitive. Elle doit couvrir les éléments réellement déterminants de la prédiction [45].
- **Certitude :** cette propriété renvoie à la capacité de l'explication à indiquer le niveau de confiance ou d'incertitude associé à une prédiction. Cela aide à ajuster la confiance de l'utilisateur dans le système [77].
- **Compacité :** une explication compacte est brève, synthétique et exempte de redondances, ce qui facilite son assimilation rapide par l'utilisateur.

La figure 2.4 synthétise les principales propriétés attendues d'une bonne explication en XAI, en illustrant leur articulation autour de l'objectif central : rendre la décision du modèle intelligible et fiable pour l'utilisateur.

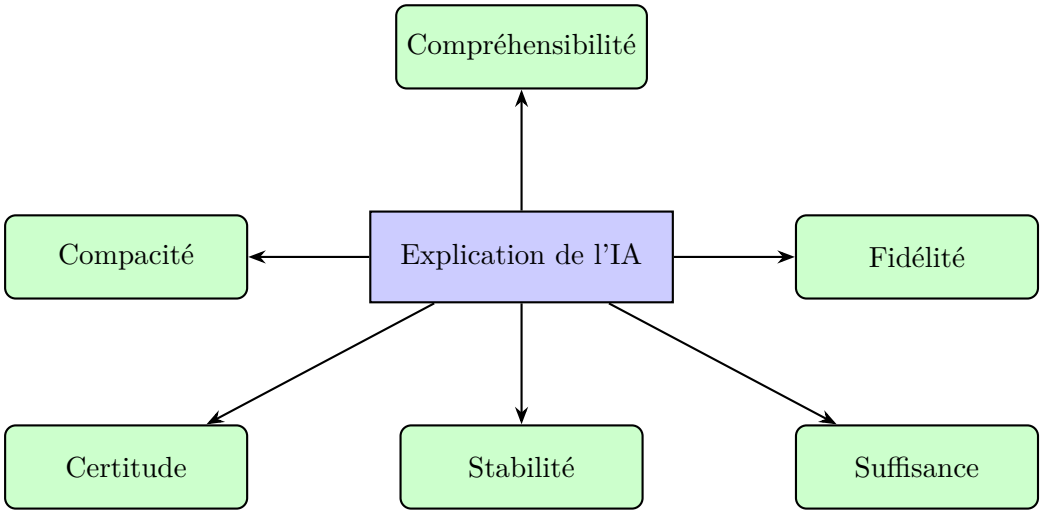


Figure 2.4: Propriétés souhaitables d'une explication en XAI

IV Méthodes et Techniques d'IA Explicative

Les méthodes d'explicabilité en intelligence artificielle (XAI) jouent un rôle fondamental dans les systèmes critiques, tels que les systèmes d'aide à la décision médicale. Elles visent à rendre les modèles plus transparents, compréhensibles et dignes de confiance pour les utilisateurs finaux. On distingue principalement deux grandes familles de méthodes : les approches intrinsèquement interprétables et les approches post-hoc.

IV.1 Méthodes intrinsèquement interprétables

Ces méthodes sont conçues pour être interprétables par nature, avec une structure simple qui permet de comprendre directement le fonctionnement et les décisions du modèle, sans recours à des techniques d'explication externes. Elle sont peu utilisées dans l'imagerie médicale, mais elles constituent une base conceptuelle utile.

IV.1.1 Modèles linéaires

Les modèles linéaires, comme la régression linéaire ou logistique, modélisent la relation entre les variables d'entrée $\mathbf{x} = (x_1, x_2, ..., x_n)$ et la sortie y sous une forme additive et pondérée. Cette simplicité structurelle permet une interprétation directe des poids du modèle, qui représentent l'impact de chaque variable sur la prédiction [34].

La prédiction est donnée par :

$$\hat{y} = \mathbf{w}^\top \mathbf{x} + b = \sum_{i=1}^{n} w_i x_i + b$$

où $\mathbf{w} = (w_1, w_2, ..., w_n)$ sont les coefficients du modèle, et b est le biais. Par exemple, dans une régression logistique, la probabilité d'appartenance à une classe est donnée par la fonction sigmoïde :

$$P(y = 1|\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b) = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x} + b)}}$$

L'interprétation est intuitive : un poids positif augmente la probabilité, un poids négatif la diminue, et la magnitude reflète l'importance relative.

IV.1.2 Arbres de décision

Les arbres de décision sont des structures hiérarchiques composées de nœuds internes correspondant à des tests conditionnels sur les variables d'entrée, et de feuilles associées à des prédictions [34]. À chaque nœud, la décision s'effectue par une règle simple de la forme :

$$\text{if } x_j \leq \theta \quad \Rightarrow \quad \text{branche gauche}, \quad \text{sinon} \quad \text{branche droite}$$

Les prédictions sont obtenues en suivant un chemin unique de la racine à une feuille, ce qui facilite une interprétation humaine du processus décisionnel en étapes successives, similaires à un raisonnement expert. La figure 2.5 illustre un exemple d'arbre de décision appliqué à la classification de la rétinopathie diabétique, en simulant un raisonnement médical hiérarchique.

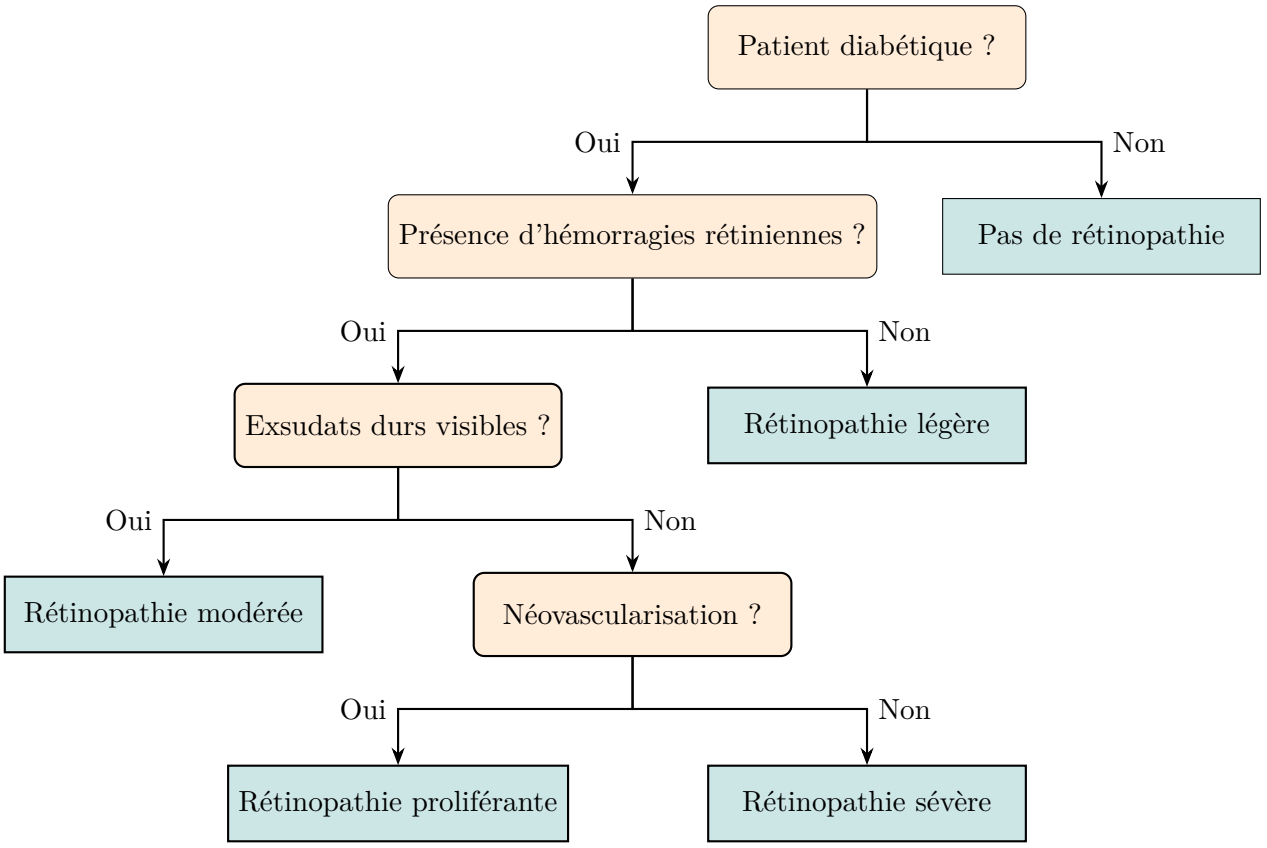


Figure 2.5: Exemple alternatif d'arbre de décision pour la classification de la rétinopathie diabétique

IV.1.3 Limitations des méthodes intrinsèquement interprétables

Les méthodes intrinsèquement interprétables, telles que la régression linéaire ou les arbres de décision peu profonds, offrent une transparence immédiate. Toutefois, elles présentent plusieurs limites lorsqu'elles sont appliquées à des contextes complexes :

- **Modélisation limitée :** Difficulté à capturer des relations non linéaires ou des interactions complexes entre les variables.
- **Performances réduites :** Moins précises que les modèles complexes dans les tâches à forte dimensionnalité.
- **Vulnérabilité au bruit :** Sensibilité accrue aux données déséquilibrées, bruitées ou incomplètes.
- **Faible capacité de généralisation :** Moins efficaces face à des données hétérogènes ou atypiques.
- **Explications simplistes :** Risque de masquer des comportements complexes non représentés par le modèle.

Ces limitations justifient le recours à des méthodes post-hoc, qui permettent d'expliquer le fonctionnement de modèles plus performants mais souvent opaques, tels que les réseaux neuronaux profonds ou les architectures à base de Transformers.

IV.2 Méthodes post-hoc agnostiques au modèle

Ces méthodes sont appliquées après entraînement, indépendamment de l'architecture, et cherchent à expliquer les décisions d'un modèle complexe.

IV.2.1 SHAP (SHapley Additive Explanations)

La méthode SHAP repose sur la théorie des jeux coopératifs. Elle attribue à chaque caractéristique d'entrée une contribution ϕ_i à la prédiction finale du modèle.[44]

L'idée est de représenter chaque prédiction comme une somme pondérée des contributions des variables :

$$f(x) = \phi_0 + \sum_{i=1}^{n} \phi_i$$

où :

- ϕ_0 est la valeur de base (souvent la moyenne des prédictions sur l'ensemble),
- ϕ_i est la valeur de Shapley attribuée à la variable x_i.

La valeur de Shapley ϕ_i est définie comme la contribution marginale moyenne de la variable x_i à toutes les coalitions $S \subseteq N \setminus \{i\}$, avec $N = \{1, \ldots, n\}$:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

Les propriétés fondamentales de SHAP sont :

- *Additivité* : les contributions se somment exactement à la prédiction,
- *Cohérence* : si une variable a un plus grand effet, sa valeur SHAP est plus élevée,
- *Exactitude locale* : la somme des ϕ_i correspond exactement à $f(x)$.

La figure 2.6 présente un exemple de visualisation SHAP appliquée à un modèle ViT, illustrant les contributions locales des différentes régions de l'image à la prédiction.

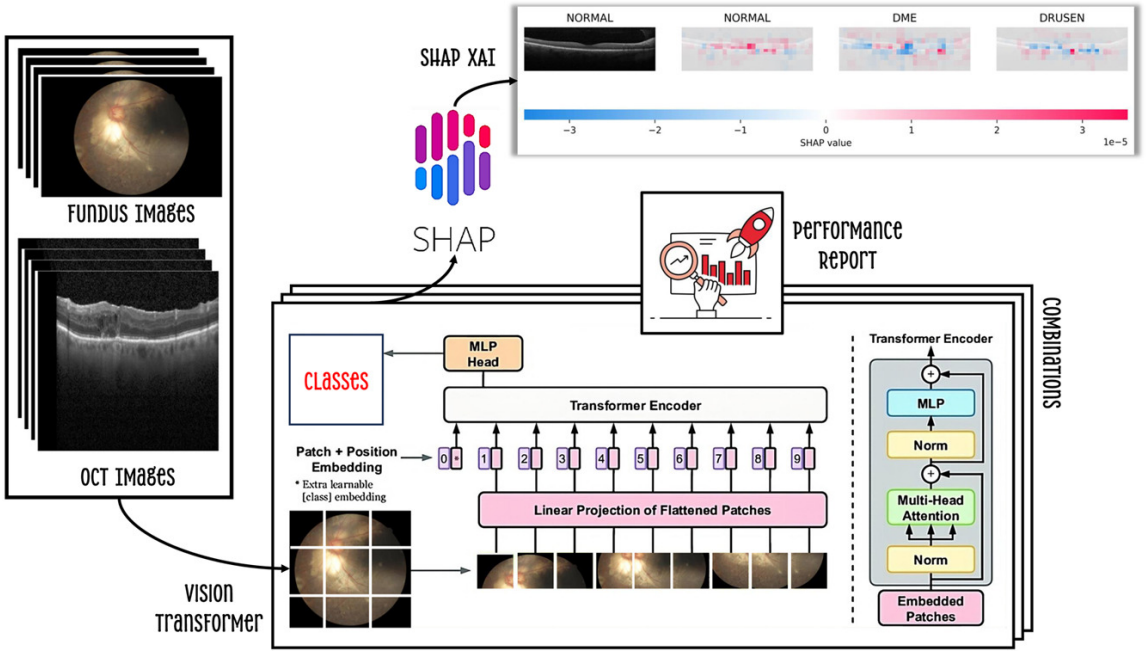


Figure 2.6: Exemple de visualisation SHAP appliquée à un modèle ViT[11]

IV.2.2 LIME (Local Interpretable Model-agnostic Explanations)

La méthode LIME fournit une explication locale d'un modèle boîte noire en le remplaçant localement autour d'une instance x par un modèle interprétable $g \in G$ (souvent linéaire ou arborescent).[59]

L'objectif est de minimiser une fonction de perte pondérée, combinant fidélité au modèle complexe et simplicité du modèle explicatif :

$$\hat{g} = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

où :

- $\mathcal{L}(f, g, \pi_x)$ mesure l'écart entre f et g autour de x, pondéré par π_x,
- $\pi_x(z)$ est une mesure de proximité entre z et x,
- $\Omega(g)$ est une pénalisation de la complexité du modèle explicatif (ex. : nombre de variables utilisées).

Ainsi, LIME génère une approximation locale fidèle de f, offrant une interprétation compréhensible pour une prédiction particulière. La figure 2.7 montre un exemple de visualisation LIME appliquée à une prédiction individuelle, mettant en évidence les variables les plus influentes localement.

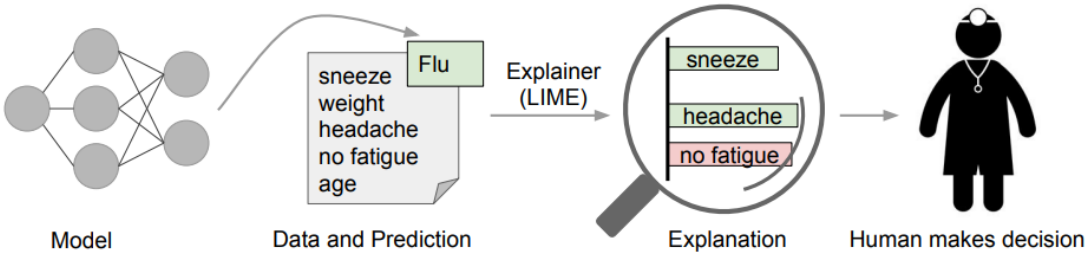


Figure 2.7: Explication d'une prédiction avec LIME[59]

IV.3 Méthodes post-hoc spécifiques aux architectures de modèles

Après avoir exploré les méthodes agnostiques au modèle, nous nous intéressons désormais aux approches post-hoc spécifiquement conçues pour certaines architectures, notamment les réseaux neuronaux convolutifs (CNN) et les Transformers (ViT).

IV.3.1 Méthodes spécifiques aux réseaux neuronaux convolutifs (CNN)

Ces méthodes exploitent la structure interne du modèle (notamment les réseaux de neurones) pour produire des explications plus détaillées sur le fonctionnement interne et l'origine des prédictions. Elles sont particulièrement utilisées dans les architectures profondes (CNN).

IV.3.1.1 Grad-CAM (Gradient-weighted Class Activation Mapping)

Grad-CAM est une méthode post-hoc d'explicabilité visuelle spécifiquement conçue pour les réseaux de neurones convolutionnels (CNN). Elle génère des cartes de chaleur qui mettent en évidence les régions de l'image ayant le plus influencé la prédiction d'une classe cible c.

Le principe repose sur l'analyse des activations $A^k \in \mathbb{R}^{H \times W}$ d'une couche convolutionnelle choisie, ainsi que des gradients associés $\frac{\partial y^c}{\partial A^k}$, où y^c désigne la sortie (logit) du modèle pour la classe c [63].

Les poids α_k^c associés à chaque carte A^k sont calculés par une moyenne spatiale des gradients :

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad \text{avec } Z = H \times W$$

La carte Grad-CAM finale est ensuite obtenue par combinaison linéaire pondérée des cartes d'activation, suivie d'un ReLU :

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$$

- A^k : k-ième carte d'activation de la couche sélectionnée,
- α_k^c : importance de A^k pour la classe c,
- ReLU : permet de ne conserver que les contributions positives à la prédiction.

Grad-CAM est largement utilisé pour localiser les zones discriminantes d'une image dans des tâches de classification ou de détection, tout en restant applicable à de nombreuses architectures CNN (ResNet, VGG, etc.). La figure 2.8 illustre un exemple de visualisation Grad-CAM, révélant les régions de l'image qui ont contribué le plus fortement à la prédiction d'une classe cible.

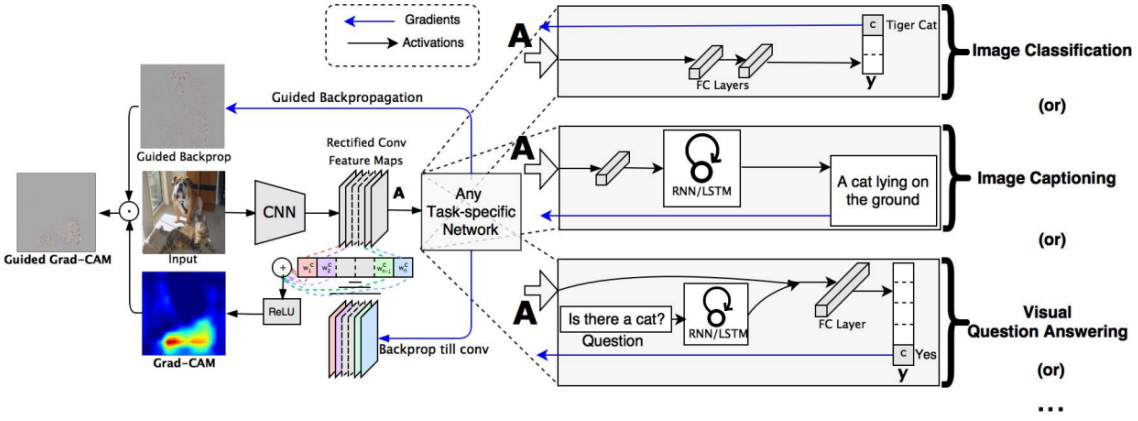


Figure 2.8: Exemple de visualisation Grad-CAM mettant en évidence les régions activées liées à une classe cible [63].

IV.3.1.2 Score-CAM (Score-Weighted Class Activation Mapping)

Score-CAM est une amélioration de la famille des méthodes CAM. Contrairement à Grad-CAM, elle ne dépend pas des gradients rétropropagés. Elle évalue l'impact de chaque carte d'activation uniquement à partir de passes avant (forward pass), ce qui la rend plus robuste et plus fidèle visuellement [76].

Le processus commence par normaliser chaque carte d'activation A_l^k par la fonction suivante :

$$s(A_l^k) = \frac{A_l^k - \min(A_l^k)}{\max(A_l^k) - \min(A_l^k)}$$

On redimensionne ensuite A_l^k à la taille de l'entrée pour obtenir $H_l^k = s(\mathrm{Up}(A_l^k))$, et on masque l'image d'entrée X avec cette carte :

$$X \circ H_l^k$$

Ce masque est passé dans le modèle pour évaluer sa contribution via le score de classe :

$$\alpha_k^c = f(X \circ H_l^k) - f(X_b)$$

où :

- $f(\cdot)$ est la sortie du modèle (logit ou softmax),
- X_b est une baseline (image de fond ou bruit),
- $X \circ H_l^k$ est le produit élément-wise entre l'entrée et la carte activée.

La carte Score-CAM est ensuite construite comme :

$$L^c_{\text{Score-CAM}} = \text{ReLU}\left(\sum_k \alpha^c_k A^k_l\right)$$

Score-CAM fournit des visualisations souvent plus stables et fidèles, au prix d'un coût de calcul plus élevé (car plusieurs passes avant sont nécessaires par carte). La figure 2.9 montre un exemple de visualisation obtenue avec Score-CAM, mettant en évidence les zones discriminantes sans recours aux gradients rétropropagés.

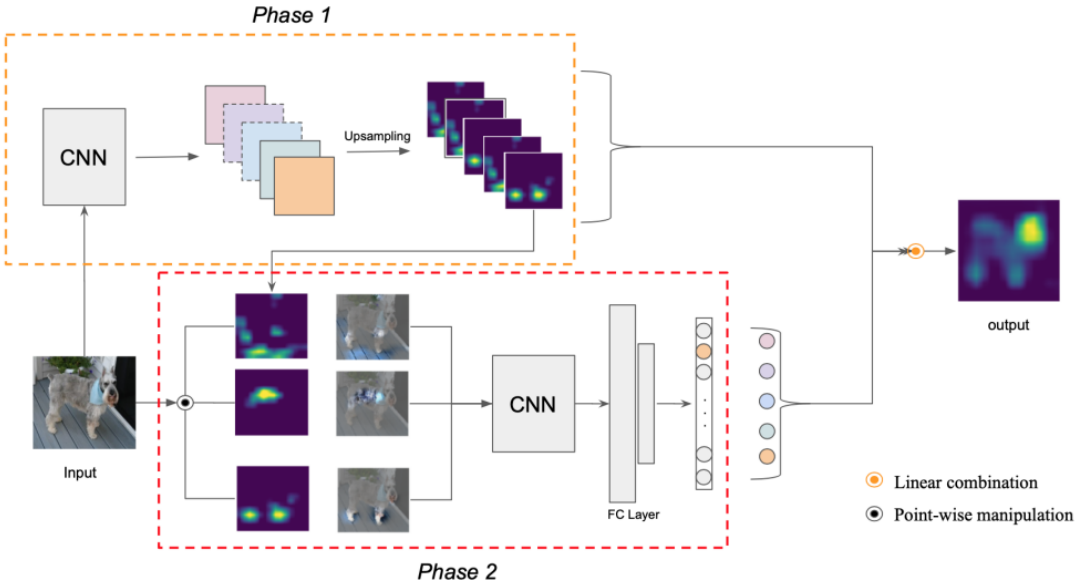


Figure 2.9: Exemple de visualisation Score-CAM illustrant l'impact visuel de chaque activation sans rétropropagation [76].

IV.3.2 Méthodes spécifiques aux Transformers (ViT)

Les modèles de type Transformer utilisent des mécanismes d'attention qui permettent d'apprendre des dépendances à long terme dans des séquences. Les méthodes d'explicabilité exploitent principalement les poids d'attention pour générer des explications.

IV.3.2.1 Attention Rollout

Attention Rollout est une méthode post-hoc conçue pour interpréter les Vision Transformers (ViT) en retraçant le flux d'information entre les couches. Contrairement aux cartes d'attention classiques, elle cumule les matrices d'attention de toutes les couches pour estimer la contribution globale des tokens d'entrée à la sortie du modèle [4].

La méthode repose sur une propagation récursive des poids d'attention :

$$\tilde{A}^{(l)} = \begin{cases} A^{(0)} & \text{si } l = 0 \\ A^{(l)} \cdot \tilde{A}^{(l-1)} & \text{sinon} \end{cases}$$

Pour mieux refléter l'architecture réelle des Transformers, les connexions résiduelles sont prises en compte via une interpolation :

$$A'^{(l)} = \alpha \cdot A^{(l)} + (1 - \alpha) \cdot I$$

où $\alpha \in [0, 1]$ est un facteur d'équilibrage (typiquement $\alpha = 0{,}5$). Cela permet d'agréger l'attention tout en respectant la structure du modèle.

Afin d'analyser et d'interpréter les décisions prises par les modèles d'apprentissage automatique appliqués à la rétinopathie diabétique, plusieurs méthodes d'explicabilité ont été sélectionnées. Le tableau suivant 2.1 présente un résumé des principales approches utilisées dans ce mémoire, en précisant leur principe et les types de modèles auxquels elles s'appliquent.

Méthode	Modèle ciblé	Principe
LIME	Tous	Apprend un modèle linéaire localement fidèle autour d'un exemple cible en générant des perturbations locales.
SHAP	Tous	Attribue à chaque caractéristique une contribution moyenne fondée sur la théorie des valeurs de Shapley (théorie des jeux coopératifs).
Score-CAM	CNN	Génère des cartes de chaleur en pondérant les activations par les scores du modèle, sans recourir aux gradients.
Grad-CAM	CNN	Utilise les gradients d'une couche convolutive pour identifier les régions les plus influentes dans une image.
Attention Rollout	ViT	Agrège les matrices d'attention de toutes les couches pour visualiser les dépendances globales entre les pixels.

Table 2.1: Résumé des principales méthodes d'explicabilité utilisées

V Évaluation de l'XAI

L'évaluation des méthodes d'intelligence artificielle explicable peut être abordée selon deux axes complémentaires : d'une part, une évaluation centrée sur l'humain, qui s'intéresse à la perception et à l'utilité des explications pour les utilisateurs ; d'autre part, une évaluation fonctionnelle, qui repose sur des critères techniques et des métriques quantitatives visant à mesurer la qualité intrinsèque des explications.

V.1 Évaluation centrée sur l'humain

Cette approche repose sur la perception et la compréhension des utilisateurs humains face aux explications fournies par les modèles d'IA. Elle implique :

- Des études utilisateur (tests d'utilisabilité, questionnaires, interviews) pour évaluer la satisfaction, la confiance, la compréhension ou encore la prise de décision.
- Elle reste essentielle pour valider si les explications sont réellement utiles aux humains dans des contextes concrets [43].

V.2 Évaluation fonctionnelle

L'évaluation des méthodes d'IA explicable est essentielle pour garantir leur utilité dans des domaines sensibles comme la médecine. Plusieurs métriques ont été proposées dans la

littérature. Dans ce travail, nous avons sélectionné celles les plus adaptées à notre objectif d'analyse des méthodes explicatives appliquées à la classification de la rétinopathie diabétique.

V.2.1 Faithfulness

La fidélité mesure à quel point une explication reflète réellement le comportement du modèle. Elle repose sur la corrélation entre les scores d'importance attribués par la méthode explicative et l'impact effectif de la suppression des caractéristiques correspondantes sur la sortie du modèle [17].

$$\text{Faithfulness} = \text{corr}_{s\subseteq[d],|s|=k}\left(\sum_{i\in s} g(f,x)_i,\ \ f(x) - f(x_{[s=\bar{s}]})\right) \tag{2.1}$$

Explication des termes :

- $f(x)$: prédiction du modèle sur l'entrée d'origine.
- $x_{[s=\bar{s}]}$: entrée modifiée où les caractéristiques du sous-ensemble s sont supprimées ou masquées.
- $g(f,x)_i$: importance attribuée à la caractéristique i par la méthode explicative g.
- $\sum_{i\in s} g(f,x)_i$: importance totale du sous-ensemble s selon g.
- $f(x) - f(x_{[s=\bar{s}]})$: variation de la sortie du modèle causée par la suppression des caractéristiques de s.
- La corrélation (souvent de Pearson) entre les deux quantités est utilisée pour estimer la fidélité.

Intervalle : $[-1, 1]$
Interprétation : Plus la valeur est proche de 1, plus l'explication est fidèle au comportement du modèle.

V.2.2 Insertion

La métrique d'insertion évalue dans quelle mesure la sortie du modèle augmente lorsque l'on introduit progressivement les régions jugées importantes par une méthode d'explication. Elle permet ainsi d'estimer la pertinence des zones identifiées comme influentes [54].

$$\text{Insertion} = \frac{1}{N}\int_0^N f(x_t)\,dt$$

Explication des termes :

- x_t : version partielle de l'entrée contenant uniquement les t pixels les plus importants selon la carte d'explicabilité.
- $f(x_t)$: prédiction du modèle sur cette entrée partielle.
- L'intégrale calcule l'aire sous la courbe représentant l'évolution de $f(x_t)$ en fonction de t.

- La normalisation par N permet la comparaison entre différentes méthodes explicatives ou images.

Intervalle : $[0, 1]$
Interprétation : Une valeur plus élevée indique que les régions insérées sont effectivement informatives pour le modèle.

V.2.3 Deletion

La métrique mesure la diminution de la prédiction du modèle lorsqu'on supprime progressivement les régions jugées importantes par la méthode explicative [54]. Elle vise à évaluer dans quelle mesure les zones identifiées sont réellement déterminantes pour la décision du modèle.

$$\text{Deletion} = \frac{1}{N} \int_0^N f(x_t)\, dt$$

Explication des termes :

- x_t : version de l'image où les t pixels les plus importants (selon l'explication) ont été masqués ou supprimés.
- $f(x_t)$: prédiction du modèle sur cette version dégradée.
- La courbe trace l'évolution de la prédiction en fonction du nombre de pixels supprimés.
- Une baisse rapide de la courbe (donc une aire plus faible) indique que les régions supprimées avaient une forte influence sur la décision.

Intervalle : $[0, 1]$
Interprétation : Plus la valeur est faible, plus l'explication est considérée comme pertinente.

V.2.4 Robustesse

La métrique Robustesse mesure la stabilité d'une explication lorsque l'entrée subit une légère perturbation [48].

$$\text{Robustesse} = 1 - \frac{\|\text{expl}_1 - \text{expl}_2\|}{\|\text{expl}_1\| + \|\text{expl}_2\| + \epsilon}$$

Explication des termes :

- expl_1, expl_2 : cartes de saillance obtenues avant et après une perturbation mineure de l'entrée.

Intervalle : $[0, 1]$
Interprétation : Plus la valeur est proche de 1, plus l'explication est stable.

V.3 Les défis majeurs de l'XAI

L'intelligence artificielle explicable (XAI) vise à rendre les décisions des systèmes d'IA compréhensibles pour les humains. Toutefois, plusieurs défis importants freinent son adoption dans des domaines critiques tels que la santé, la finance ou les véhicules autonomes.

- **Compromis entre interprétabilité et performance**
  Les modèles les plus performants, comme les réseaux neuronaux profonds, sont souvent opaques, tandis que les modèles transparents manquent de puissance prédictive. Ce compromis rend leur utilisation complexe dans des contextes sensibles [14].

- **Fidélité des explications**
  Certaines méthodes XAI produisent des explications qui ne reflètent qu'approximativement le raisonnement réel du modèle, ce qui peut induire les utilisateurs en erreur [67].

- **Scalabilité aux architectures complexes**
  Certaines techniques XAI fonctionnent bien sur des modèles simples, mais deviennent inefficaces ou coûteuses en temps de calcul sur des architectures complexes ou de grandes bases de données [67].

- **Absence de causalité dans les explications**
  La plupart des méthodes se contentent de mettre en évidence des corrélations entre variables d'entrée et prédictions, sans garantir un lien de causalité. Cela limite la robustesse et la fiabilité des décisions expliquées [14].

- **Biais dans les explications**
  Dans des applications médicales comme la détection de la rétinopathie diabétique, des biais peuvent apparaître si les données d'entraînement sont peu représentatives (âge, origine ethnique, qualité des images, etc.). Ces biais compromettent la généralisation des explications à l'ensemble des patients [24].

VI État de l'art

Dans cette partie, nous présentons une vue d'ensemble des travaux existants portant sur la classification de la rétinopathie diabétique (RD), en mettant un accent particulier sur les approches d'intelligence artificielle explicable. En effet, le développement récent de techniques telles que SHAP, LIME et Grad-CAM a significativement renforcé l'interprétabilité des modèles de deep learning dans le domaine médical (voir Tableau 2.2), et plus spécifiquement dans le cadre de la RD, comme le montre le Tableau 2.3.

Pour dépasser les limites des modèles CNN traditionnels, des architectures hybrides combinant CNN et Transformers ont été récemment proposées. **Sampath et Khan (2025)**[62] ont ainsi développé *EfficientViT*, une architecture dotée de mécanismes d'attention croisée et intégrant Grad-CAM++ guidé par attention. Leur approche a permis de générer des cartes de chaleur fortement alignées avec les annotations cliniques, avec une amélioration notable de l'IoU par rapport à Grad-CAM standard (63 % contre 51 %). Malgré ses performances prometteuses, cette architecture reste complexe à entraîner et nécessite des ressources computationnelles importantes.

Dans la même lignée, le modèle *VR-FuseNet* proposé par **Refat et al. (2025)**[58] constitue une contribution majeure. Il fusionne des données issues de cinq bases publiques

(APTOS, DDR, IDRiD, Messidor2, Retino) et applique plusieurs méthodes d'explicabilité post-hoc (Grad-CAM, Score-CAM, Layer-CAM) pour produire des visualisations riches et interprétables. Toutefois, ce modèle présente certaines limites, notamment le déséquilibre des classes, un temps d'entraînement élevé, et l'absence d'intégration des Transformers pour la capture des dépendances globales.

Par ailleurs, **Zhang et al. (2025)**[84] ont réalisé une étude comparative approfondie entre les modèles CNN, ViT et les architectures hybrides, en utilisant Grad-CAM et Attention Rollout comme outils d'interprétabilité. Leur analyse met en évidence les avantages des architectures modernes telles que SwinV2-Tiny, LeViT-256 et CvT-13, tout en soulignant certaines limites, notamment la taille restreinte des jeux de données, l'équilibrage artificiel des classes, ainsi que l'absence de validation externe.

Dans une autre étude, **Herrero-Tudela et al. (2024)**[35] ont proposé une approche basée sur SHAP pour identifier les régions pathologiques dans les images de fond d'œil. Cette méthode a été appliquée à une large gamme de modèles CNN (DenseNet121, InceptionV3, MobileNet, VGG-19, Xception, ResNet-50) sur plusieurs jeux de données publics (APTOS, EyePACS, DDR, IDRiD, SUSTech-SYSU). Les résultats montrent une utilité clinique importante, avec des performances prédictives élevées. Toutefois, les auteurs soulignent l'importance de la qualité des images pour éviter les erreurs de diagnostic.

LIME a également été utilisée dans de nombreuses études, notamment celle de **Shahzad et al. (2024)**[64], qui visait à rendre les modèles CNN plus interprétables pour favoriser leur adoption en milieu clinique. De leur côté, **Vasireddi et al. (2024)**[72] ont combiné LIME avec plusieurs architectures (ResNet50, DenseNet121, MobileNetV2) et démontré que l'intégration d'outils explicatifs n'impacte pas négativement les performances de classification multi-classes. Néanmoins, ces approches souffrent de certaines limitations, telles que la taille réduite des jeux de données et la charge computationnelle importante.

Enfin, Grad-CAM reste l'une des méthodes les plus répandues pour la visualisation des activations neuronales. Les travaux de **Alavee et al. (2024)**[8] ont appliqué cette méthode à différentes architectures CNN classiques et validé la capacité de généralisation des modèles à travers des validations croisées sur les bases Messidor2 et IDRiD.

Table 2.2: Résumé des approches XAI dans différents contextes cliniques

Article	**Cas Clinique**	**Méthode XAI utilisée**	**Modèle utilisé**
Ainhoa Osa-Sanchez et al[53], 2024	Dégénérescence Maculaire liée à l'âge	SHAP, LIME	MLP, CNN, Transformer
Meera Radhakrishnan et al[56], 2024	Diagnostic du Cancer de l'Ovaire	Integrated Gradients, Saliency Maps, Grad-CAM, DeepLIFT	ResNeXt50, VGG19, MobileNetV2, ResNet18, Xception, EfficientNetB0, InceptionV3
Rafique Ahmed, Ali Shariq Imran[7], 2024	Ostéoarthrite du genou	Grad-CAM	VGG-16, VGG-19, ResNet-50, ResNet-101, EfficientNetB7
Maria Nancy A et K. Sathyarajasekaran[2], 2024	Segmentation des tumeurs cérébrales (IRM volumétriques)	Grad-CAM	SwinVNETR

Table 2.2 – suite

Article	Cas Clinique	Méthode XAI utilisée	Modèle utilisé
Mitchell D. Woodbright et al[81], 2024	Troubles neurologiques (épilepsie, tumeurs cérébrales, Alzheimer)	Grad-CAM	VGG16, VGG19, ResNet50, DenseNet121, Xception, DTCWT + CVANN-2, SVM, RF, ART-Explain
S. M. Mahim et al[46], 2024	Alzheimer	LIME, SHAP, Attention Maps	ResNet50, MobileNetV2, VGG19, Xception, InceptionV3, DenseNet121, VGG16, ViT-GRU
Lin Zou et al[86], 2023	Pneumonie Communautaire grave et Infections respiratoires au COVID-19	Grad-CAM, Grad-CAM++, SHAP, LIME, Saliency, Ensemble XAI	InceptionV3, VGG16
Ruey-Kai Sheu et al[65], 2023	Pneumonie	Grad-CAM, SHAP	DenseNet-121, ResNet152v2
V. A. Ashwath et al[9], 2023	Lésions cutanées, Pathologies pulmonaires	Grad-CAM, Saliency Maps	DenseNet-121, InceptionV3, Xception, ResNet-50, CNN personnalisé
Dara Varam et al[71], 2023	Classification d'images endoscopiques (maladies gastro-intestinales)	Grad-CAM, Grad-CAM++, Layer-CAM, LIME, SHAP	InceptionV3, EfficientNetV2, VGG-16/VGG-19, MobileNetV3Large, ResNet152v2, ViT
Mafalda Malafaia et al[47], 2022	Cancer du poumon	Saliency Maps, Integrated Gradients, LRP, DeepLIFT	CNN personnalisé

Table 2.3: Résumé comparatif des approches XAI appliquées à la rétinopathie diabétique

Article	**Année**	**Méthode**	**Résultats**	**Dataset**	**Modèle**	**Avantages**	**Inconvénients**
EfficientViT: A Hybrid CNN-Transformer Framework With Cross-Attention Fusion For Clinically Interpretable Diabetic Retinopathy Grading **Mahalakshmi Sampath, Mohammad Akram Khan**[62]	2025	Grad-CAM++ guidé par attention	Amélioration de l'IoU (63% vs 51%) ; alignement avec annotations cliniques	APTOS 2019	EfficientNetV2 + ViT (EfficientViT)	Interprétabilité renforcée ; fusion efficace CNN + ViT ; performances élevées	Complexité élevée ; besoin de grandes bases non annotées ; validation clinique encore limitée
VR-FuseNet: A Fusion of Heterogeneous Fundus Data and Explainable Deep Network for Diabetic Retinopathy Classification **Shamim Rahim Refat et al.**[58]	2025	Grad-CAM, Grad-CAM++, Score-CAM, Faster Score-CAM, Layer-CAM	Explications visuelles variées ; modèle performant et interprétable	APTOS 2019, DDR, IDRiD, Messidor 2, Retino	VGG16, VGG19, ResNet50V2, MobileNetV2, Xception, VR-FuseNet (proposé)	Bonne fusion des caractéristiques ; transparence accrue ; adoption clinique facilitée	Complexité élevée ; déséquilibre de classes ; absence de ViTs ; non-prise en compte de données cliniques

Suite page suivante

Table 2.3 – suite

Article	Année	Méthode	Résultats	Dataset	Modèle	Avantages	Inconvénients
Interpretable Deep Learning for Diabetic Retinopathy: A Comparative Study of CNN, ViT, and Hybrid Architectures **Weijie Zhang et al.**[84]	2025	Grad-CAM, Attention Rollout	Comparaison approfondie ; visualisations cohérentes	EyePACS, APTOS-2019	ResNet50, EfficientNet-B0, CvT-13, ViT-Small, DINOv2, SwinV2-Tiny, LeViT-256	Évaluation de plusieurs architectures ; bonne base de comparaison ; techniques explicables avancées	Dataset restreint ; équilibrage artificiel ; faible nombre d'époques ; manque de validation externe ; interprétabilité partielle
An explainable deep-learning model reveals clinical clues in diabetic retinopathy through SHAP **María Herrero-Tudela et al.**[35]	2024	SHAP	Visualisation des régions pertinentes via SHAP	APTOS 2019, EyePACS, DDR, IDRiD, SUSTech-SYSU	DenseNet121, InceptionV3, MobileNet, VGG-19, Xception, ResNet-50	Haute utilité clinique, prédiction fiable, réduction des faux positifs/négatifs, soulagement du personnel médical	Ne tient pas compte de la qualité des images ; nécessité d'introduire des Vision Transformers et autres approches XAI
Developing a Transparent Diagnosis Model for Diabetic Retinopathy Using Explainable AI **Tariq Shahzad et al.**[64]	2024	LIME	Visualisation interprétable via LIME	Non précisé	CNN	Renforce la confiance clinique ; rend la boîte noire interprétable	Petit dataset, complexité computationnelle élevée, absence d'évaluation en temps réel

Suite page suivante

Table 2.3 – suite

Article	Année	Méthode	Résultats	Dataset	Modèle	Avantages	Inconvénients
DR-XAI: Explainable Deep Learning Model for Accurate Diabetic Retinopathy Severity Assessment **Hemanth Kumar Vasireddi et al.**[72]	2024	LIME	Classification multi-classes précise et interprétable	MESSIDOR	ResNet50, DenseNet121, E-DenseNet, MobileNetV2, Méthode proposée	Bonne précision sans perte de performance ; modèle interprétable	Jeu de données limité ; complexité élevée, surtout en contexte clinique
Enhancing Early Detection of Diabetic Retinopathy Through the Integration of Deep Learning Models and Explainable Artificial Intelligence **Kazi Ahnaf Alavee et al.**[8]	2024	Grad-CAM	Visualisation des lésions ; validation multi-datasets	APTOS 2019, IDRiD, Messidor2	DenseNet121, Xception, ResNet50, VGG16/19, InceptionV3, CNN	Bonne généralisation ; intégration efficace de XAI	Absence de validation en clinique réelle

VII Conclusion

Ce chapitre a présenté une vue d'ensemble complète des fondements, méthodes et enjeux liés à l'intelligence artificielle explicable (XAI). Nous avons d'abord défini les concepts clés tels que l'explicabilité, l'interprétabilité, la transparence et la fidélité, tout en explorant les différents niveaux d'explication, globale et locale, ainsi que les propriétés souhaitables d'une bonne explication. Ensuite, les principales techniques XAI ont été classées entre méthodes intrinsèquement interprétables, méthodes post-hoc agnostiques, et méthodes spécifiques aux modèles. Chacune a été illustrée par des exemples concrets, accompagnés de leurs avantages et limites. Les métriques d'évaluation de la qualité des explications ont également été abordées, distinguant les approches centrées sur l'humain et celles fondées sur des critères quantitatifs. Enfin, nous avons analysé les défis actuels de l'XAI, allant du compromis entre interprétabilité et performance à la présence de biais dans les explications, sans oublier son intégration dans les processus de développement des systèmes d'IA.

L'importance de l'XAI apparaît ainsi comme cruciale pour développer une IA de confiance, transparente, équitable et éthique, notamment dans les domaines sensibles comme la santé ou la justice. Alors que l'IA devient omniprésente, fournir des explications claires et utiles est d'une nécessité cruciale.

Le prochain chapitre de ce mémoire approfondira l'application de ces concepts dans un contexte particulier : la classification de la rétinopathie diabétique. Nous verrons comment les techniques XAI peuvent contribuer à la transparence et à la fiabilité des décisions médicales automatisées, tout en répondant aux exigences des professionnels de santé.

Chapitre 3

Conception & Réalisation

I Introduction

Dans ce chapitre, nous présentons l'application des techniques d'explicabilité (XAI) à trois modèles de classification d'images médicales utilisés pour la détection de la rétinopathie diabétique : AtR5C (CNN), ViR-5C (ViT) et ReVi-5C (modèle hybride) combinant ces deux architectures [3, 12], ont déjà démontré de bonnes performances en termes de précision. Cependant, la nature critique du diagnostic médical impose une meilleure compréhension des décisions prises par ces modèles. Ainsi, notre objectif principal est d'appliquer diverses techniques d'explicabilité afin de fournir des interprétations visuelles et quantitatives des prédictions, améliorant ainsi la confiance et la transparence du système. Après une brève présentation des modèles, nous détaillerons les méthodes XAI utilisées, avant d'analyser les résultats visuels et quantitatifs obtenus.

II Méthodologie

La Figure 3.1 illustre une vue d'ensemble structurée de notre méthodologie. Celle-ci s'articule autour des étapes suivantes : une présentation synthétique des trois modèles utilisés, l'application des différentes méthodes d'explicabilité, puis une évaluation à la fois quantitative et qualitative des cartes générées. Le processus global d'explicabilité mis en œuvre dans cette étude est décrit dans l'algorithme 1.

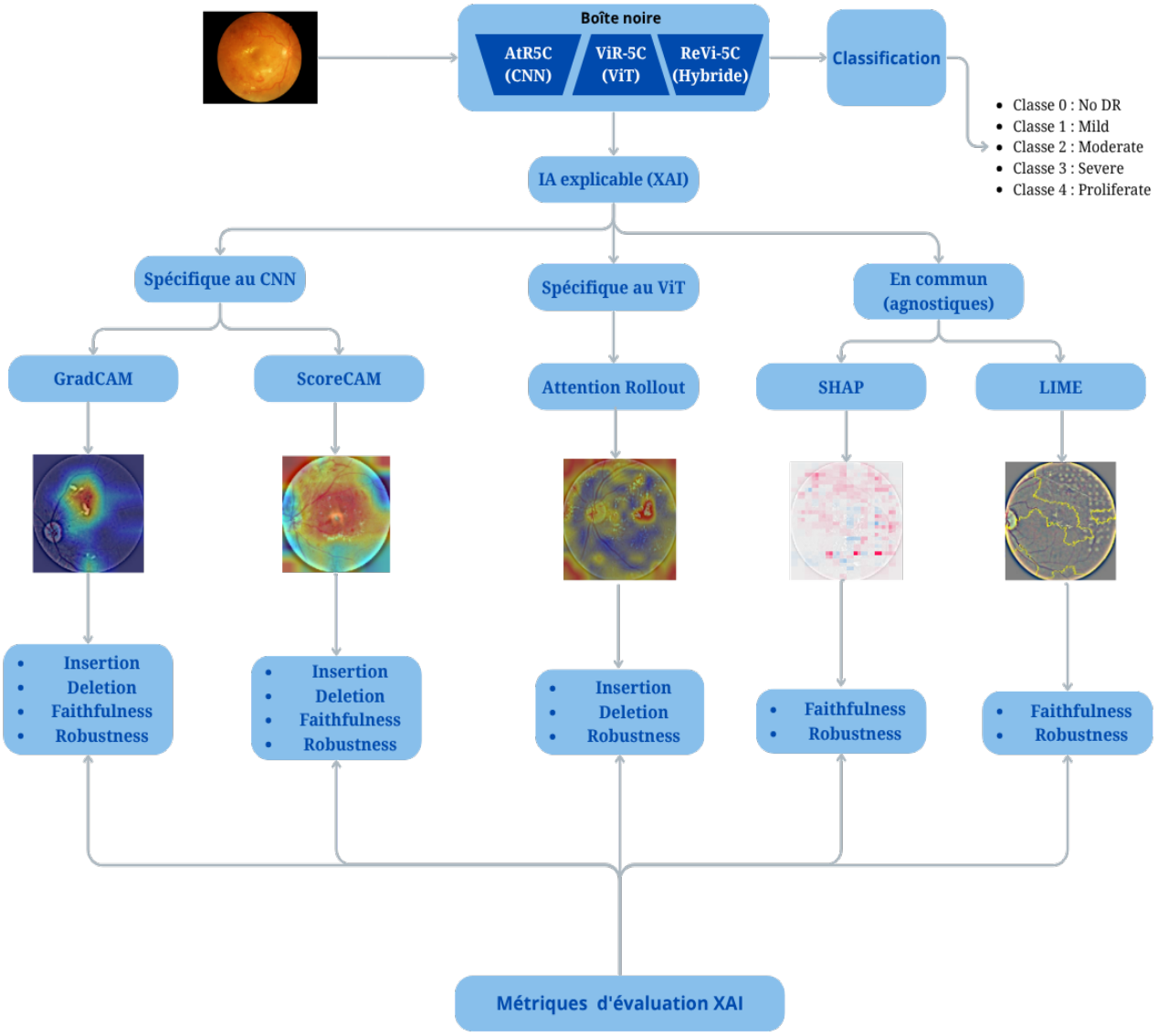


Figure 3.1: Démarche d'explicabilité des modèles boîtes noires de classification d'images rétiniennes et analyse des performances XAI

Algorithm 1: Pipeline d'explicabilité pour la classification de la rétinopathie diabétique

Input: Image du fond d'œil $x \in \mathbb{R}^{H \times W \times C}$
Output: Classe prédite y et carte explicative $E(x)$
Choisir un modèle $f \in \{\text{AtR5C}, \text{ViR-5C}, \text{ReVi-5C}\}$;
Charger le modèle pré-entraîné f ;
Calculer la prédiction : $y \leftarrow f(x)$;
if $f = \textit{AtR5C}$ **then**
    $M \leftarrow \{\text{Grad-CAM}, \text{Score-CAM}, \text{SHAP}, \text{LIME}\}$;
else
    if $f = \textit{ViR-5C}$ **then**
        $M \leftarrow \{\text{Attention Rollout}, \text{SHAP}, \text{LIME}\}$;
    else
        $M \leftarrow \{\text{SHAP}, \text{LIME}\}$;
foreach $m \in M$ **do**
    Calculer la carte explicative $E_m(x)$ avec la méthode m ;
    Visualiser $E_m(x)$;
    foreach $\mu \in \{\textit{faithfulness}, \textit{insertion}, \textit{deletion}, \textit{robustness}\}$ **do**
        Évaluer $E_m(x)$ selon la métrique μ ;
return $(y, E(x))$

III Modèles utilisés

Dans ce qui suit, nous décrivons les trois modèles principaux utilisés dans le cadre de ce mémoire.

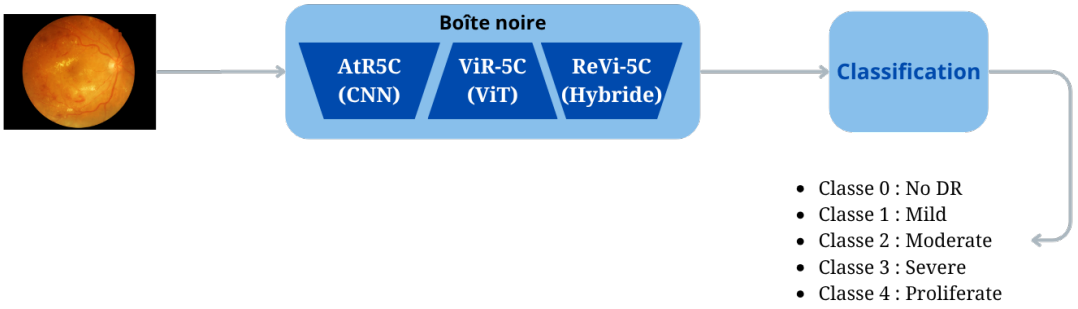


Figure 3.2: représentation des modèles utilisés

III.1 Modèle AtR5C

Dans notre travail, nous nous appuyons sur le modèle **AtR5C** tel qu'adapté par Abiche Yacine et Amokrane Akram (2023) dans leur mémoire dédié à la classification de la rétinopathie diabétique [3].

Le modèle AtR5C a été pré-entraîné sur le dataset ImageNet, puis *fine-tuné* sur des images de fond d'œil issues du jeu de données APTOS 2019, classées selon cinq niveaux de gravité de la rétinopathie diabétique. Pour adapter le modèle à cette tâche spécifique, la dernière couche entièrement connectée (prévue initialement pour 1000 classes) a été remplacée par une couche adaptée aux 5 classes du problème.

Ils ont également intégré diverses stratégies d'optimisation, incluant une fonction de perte pondérée pour corriger le déséquilibre des classes, des techniques de régularisation, ainsi qu'un ajustement précis des hyperparamètres en utilisant une optimisation bayesienne. Grâce à ces choix méthodologiques, le modèle a atteint une précision de **85,42 %**.

III.2 Modèle ViR-5C

Dans le mémoire de Baouz Sif Eddine (2024), l'architecture Vision Transformer (ViT), désignée par la série **ViR**, est utilisée pour la classification de la rétinopathie diabétique [12]. Le modèle transforme une image d'entrée de taille 224 × 224 en 256 patchs de 16 × 16 pixels, chacun étant linéarisé puis enrichi d'un *embedding* de position. Ces séquences sont ensuite traitées par une succession de blocs Transformer comprenant des mécanismes de self-attention multi-tête suivis de couches MLP. La tête MLP finale est spécifiquement adaptée à la tâche de classification.

Dans la configuration à 5 classes (**ViR-5C**), le modèle atteint une précision de **87,33 %** sur le dataset APTOS 2019. Ce modèle se distingue par sa capacité à capturer des relations globales entre les régions de l'image, grâce à l'attention, mais montre certaines limites en matière de sensibilité locale.

III.3 Modèle ReVi-5C

Toujours selon Baouz Sif Eddine (2024), une architecture hybride, nommée **ReVi**, combine les avantages des CNN et des Transformers. Elle commence par un ResNet50 pré-entraîné (avec des poids AtR5C) pour extraire des caractéristiques locales via des couches convolutives. Ces représentations sont ensuite transmises à un module ViT, chargé de modéliser les dépendances globales. La classification finale est effectuée via une tête dense [12].

Dans la version à 5 classes (**ReVi-5C**), cette architecture atteint une précision de **96,21 %** avec un score F1 de **95,80 %** dépassant significativement la performance du ViT seul. Cette supériorité s'explique par la complémentarité des informations extraites localement et globalement.

L'ensemble des expériences repose sur le dataset APTOS 2019, prétraité (recadrage circulaire, redimensionnement, normalisation, et augmentation de données). L'apprentissage est optimisé via un *fine-tuning* rigoureux avec un *batch size* de 32 ou 64, un taux d'apprentissage de 2×10^{-5} et un *weight decay* de 0,01.

Le tableau 3.1 résume les performances obtenues par les trois modèles analysés dans ce mémoire (AtR5C, ViR-5C et ReVi-5C), en termes de précision, rappel, F1-score et exactitude. Il met en évidence la supériorité du modèle hybride ReVi-5C, qui combine efficacement les avantages des CNN et des Transformers.

Table 3.1: Comparaison des performances

Métriques	**AtR5C**	**ViR-5C**	**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 %**

IV Application des Méthodes d'Explicabilité (XAI)

Cette section présente en détail l'implémentation des cinq méthodes d'explicabilité post-hoc retenues dans le cadre de ce travail : Grad-CAM, Score-CAM, Attention Rollout, SHAP et LIME. Pour chacune de ces méthodes, nous décrivons le principe algorithmique ainsi que les résultats obtenus après application sur les modèles étudiés. Le tableau 3.2 offre une synthèse des différentes techniques utilisées, en précisant les types de modèles pour lesquels elles sont applicables.

Table 3.2: Méthodes XAI adaptées aux différentes architectures

Méthode XAI	**AtR5C**	**ViR-5C**	**ReVi-5C**	**Pourquoi c'est adapté**
Grad-CAM	Oui	Non	Non	Utilise les couches convolutionnelles pour générer des cartes d'activation. Fonctionne uniquement avec des CNN.
Score-CAM	Oui	Non	Non	Repose sur les cartes de caractéristiques convolutives, donc fonctionne uniquement avec des CNN.
Attention Rollout	Non	Oui	Non	Exploite les poids d'attention, donc spécifique aux architectures de type Transformer comme ViT.
SHAP	Oui	Oui	Oui	Génère des versions masquées de l'image pour estimer la contribution de chaque région à la prédiction. Indépendant de l'architecture.
LIME	Oui	Oui	Oui	Méthode agnostique au modèle qui identifie les zones importantes via des perturbations locales. Compatible avec tous les types de modèles.

IV.1 Matériel et Environnement

Les expérimentations ont été réalisées sur **Kaggle Notebooks** avec GPU `P100`, en utilisant principalement **TensorFlow/Keras** et **PyTorch**. Le tableau 3.3 présente les détails de l'environnement logiciel et matériel.

Table 3.3: Résumé du matériel et de l'environnement logiciel

Composant	Détails
Plateforme	Kaggle Notebooks
GPU	NVIDIA P100 (16 Go VRAM)
Langage de programmation	Python 3.10
Frameworks	TensorFlow 2.17.1, Keras 3.5.0 PyTorch 2.2.2 (Score-CAM, SHAP), 2.1.2 (LIME, SHAP)
Bibliothèques complémentaires	Transformers 4.41.1 (HuggingFace), SHAP 0.44.1, LIME 0.2.0.1, NumPy, Pandas, OpenCV, Matplotlib, Pillow
Modèles utilisés	AtR5C, ViR-5C, ReVi-5C

IV.2 Adéquation entre les méthodes XAI et les métriques d'évaluation

Chaque méthode d'explicabilité (XAI) repose sur des fondements algorithmiques spécifiques qui conditionnent la nature des explications qu'elle produit. De ce fait, toutes les métriques ne sont pas systématiquement applicables à toutes les méthodes. Voici une justification détaillée, méthode par méthode :

- **Grad-CAM et Score-CAM**
  Ces méthodes visuelles, spécifiques aux réseaux de neurones convolutifs (CNN), produisent des cartes de saillance spatiales directement exploitables pixel par pixel. Elles sont compatibles avec les métriques *insertion* et *deletion*, qui mesurent l'impact de la suppression ou de l'ajout progressif des zones jugées importantes. Elles permettent également d'évaluer la *fidélité* par corrélation entre l'importance estimée et l'effet sur la prédiction, ainsi que la *robustesse*, en mesurant la stabilité des cartes face à de petites perturbations.

- **Attention Rollout**
  Méthode conçue pour les Transformers, elle repose sur l'agrégation des matrices d'attention à travers les couches du modèle. Bien qu'elle produise des cartes spatiales compatibles avec *insertion* et *deletion*, elle ne fournit pas de scores d'importance individuels par caractéristique. De ce fait, la métrique de *fidélité* (telle que définie par la corrélation entre importance et variation de sortie) n'est pas applicable. En revanche, la *robustesse* peut être évaluée par la stabilité des cartes d'attention.

- **SHAP**
  Basée sur la théorie des jeux coopératifs, cette méthode attribue à chaque caractéristique une valeur de Shapley représentant sa contribution marginale. Elle est parfaitement adaptée à la mesure de la *fidélité*, puisque les contributions se somment à la prédiction. Toutefois, les métriques *insertion* et *deletion* sont inadaptées ici, car SHAP ne génère pas de heatmaps continues mais des attributions discrètes. La *robustesse* est applicable, puisqu'il est possible de comparer les valeurs SHAP avant et après perturbation.

- **LIME**
  Méthode d'explication locale, LIME génère des perturbations autour d'une instance et

ajuste un modèle linéaire local. Elle permet une évaluation directe de la *fidélité*, puisque le modèle explicatif est optimisé pour reproduire localement la sortie du modèle complexe. Comme SHAP, LIME ne produit pas de cartes continues, rendant *insertion* et *deletion* inapplicables. En revanche, sa *robustesse* peut être mesurée efficacement.

IV.3 Méthodes XAI appliquées au modèle AtR5C

Les méthodes Grad-Cam et Score-Cam ne sont applicables qu'aux modèles CNN.

IV.3.1 Grad-CAM

L'implémentation de Grad-CAM a été réalisée avec `TensorFlow 2.17.1` et `Keras 3.5.0`. Le principe repose sur la génération d'une carte de chaleur à partir des activations de la dernière couche convolutionnelle. Voici le détail des étapes du fonctionnement algorithmique :

- **Propagation directe** : une image d'entrée normalisée (dimensions 224×224) est transmise au sous-modèle. On récupère à la fois la sortie de la couche convolutionnelle ciblée (`conv5_block3_out`) et la prédiction du modèle.
- **Calcul des gradients** : à l'aide de `GradientTape`, on calcule les gradients de la probabilité prédite (correspondant à la classe cible) par rapport aux activations de la couche convolutionnelle. Ces gradients reflètent la sensibilité de chaque neurone à la sortie du modèle.
- **Pondération des activations** : une moyenne spatiale est effectuée sur les gradients pour chaque canal afin d'obtenir un poids d'importance global pour chaque filtre.
- **Génération de la carte de chaleur** : les poids calculés sont multipliés canal par canal avec les activations correspondantes. Le résultat est ensuite agrégé et passé à une fonction `ReLU` pour ne conserver que les zones à impact positif.
- **Visualisation finale** : la carte Grad-CAM peut être superposée à l'image originale afin d'en faciliter l'interprétation visuelle par les experts médicaux.

Algorithme Grad-CAM

L'algorithme suivant illustre le pipeline exact d'implémentation de Grad-CAM appliqué au modèle AtR5C utilisé dans ce travail (algorithm 2).

Algorithm 2: Fonctionnement de Grad-CAM pour AtR5C

Input: Image d'entrée prétraitée x, modèle entraîné f, couche convolutionnelle cible L
Output: Carte de saillance Grad-CAM $M(x)$

Étape 1 : Propagation avant
Effectuer une passe avant $f'(x)$
Obtenir :
- $A^L(x)$: cartes d'activations de la couche L
- $f(x)$: prédiction du modèle (vecteur de probabilités)

Étape 2 : Calcul des gradients
Utiliser `GradientTape` pour calculer :
$\nabla_{A^L(x)} f_c(x)$ où $c = \arg\max f(x)$

Étape 3 : Moyennisation des gradients
for *chaque canal* k *dans* $A^L(x)$ **do**
  Calculer le poids $\alpha_k = \frac{1}{Z} \sum_{i,j} \nabla_{A^L_{i,j,k}} f_c(x)$

Étape 4 : Agrégation pondérée
$M(x) = \text{ReLU}\left(\sum_k \alpha_k A^L_k(x)\right)$

Étape 5s : Visualisation
Superposer $M(x)$ sur l'image x comme carte de chaleur

Résultats visuels obtenus

La figure 3.3 illustre une carte de chaleur Grad-CAM générée pour une image test. Les zones colorées indiquent les régions fortement contributives à la décision du modèle.

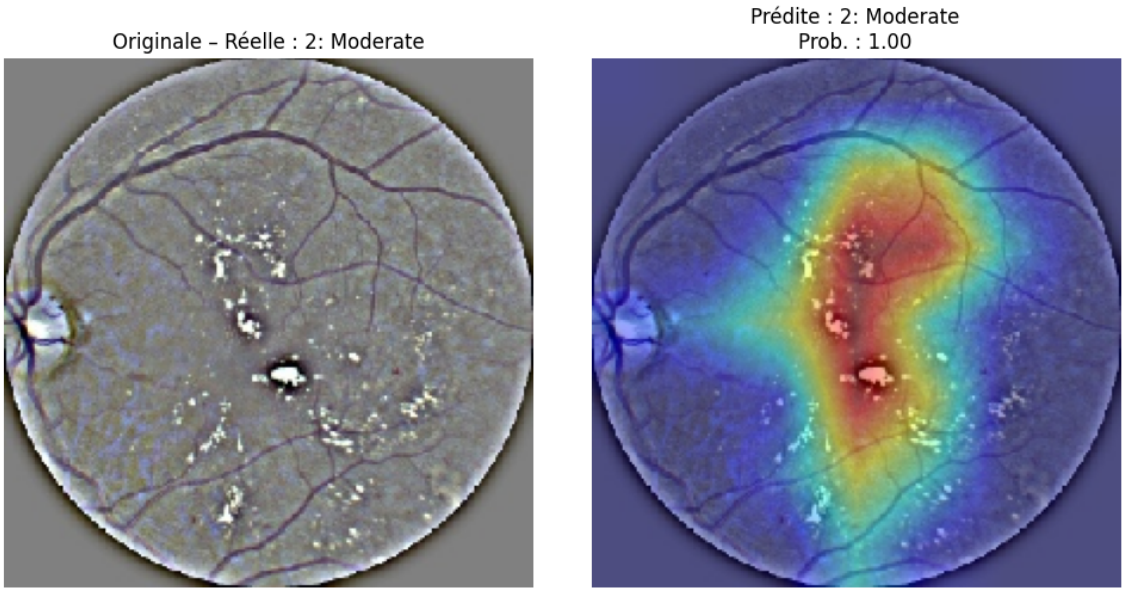


Figure 3.3: Exemple de visualisation Grad-CAM pour la classe 2 (Moderate)

L'explication visuelle générée par Grad-CAM montre que le modèle **AtR5C** base sa prédiction de rétinopathie « modérée » (classe 2) sur une zone centrale marquée par un amas d'exsudats. Cette région, fortement activée, apparaît en rouge/orange sur la carte de chaleur superposée à l'image originale.

Toutefois, certaines lésions périphériques comme la fovéa sombre ou les micro-anévrismes ne sont pas mises en évidence, ce qui suggère que le modèle peut ignorer des signes cliniquement importants. Cette analyse confirme que les exsudats majeurs jouent un rôle central dans la décision du CNN, illustrant ainsi l'utilité de Grad-CAM pour visualiser les critères déterminants de classification.

Évaluation Quantitative

Les métriques d'évaluation de la qualité explicative de Grad-CAM ont été calculées sur 500 images du jeu de test (100 images par classe). Les résultats sont présentés dans le tableau 3.4.

Table 3.4: Métriques d'explicabilité pour Grad-CAM sur AtR5C

Métrique	**AtR5C**
Insertion	0.827504
Deletion	0.620674
Faithfulness	1.000000
Robustness	0.385266
Temps total d'exécution	1h 22min 51s

Les résultats métriques obtenus pour Grad-CAM montrent une bonne capacité à identifier les régions pertinentes (insertion : 0,83), mais une suppression partielle de l'information importante (deletion : 0,62), indiquant une carte encore trop large. La fidélité (1,00) semble parfaite, mais ce score est à relativiser car il reflète surtout une adéquation technique avec la métrique utilisée. Enfin, la robustesse reste faible (0,39), ce qui suggère une forte sensibilité aux perturbations, un point critique dans le contexte de l'imagerie médicale.

IV.3.2 Score-CAM

L'implémentation de Score-CAM a été réalisée avec `TensorFlow 2.17.1` et `Keras 3.5.0`. La méthode repose sur une analyse directe de l'impact des activations de la couche convolutionnelle finale sur la prédiction finale. Les étapes principales sont détaillées ci-dessous:

- **Chargement de l'image** : chaque image est redimensionnée à 224×224 pixels et normalisée dans l'intervalle [0, 1].
- **Prédiction initiale** : le modèle **AtR5C** effectue une première prédiction sur l'image pour identifier la classe cible.
- **Extraction des activations** : un sous-modèle est construit pour extraire les cartes d'activation de la dernière couche convolutionnelle (`conv5_block3_out`).
- **Génération de cartes pondérées** : pour chaque canal d'activation, une carte d'importance est redimensionnée, normalisée et utilisée pour pondérer l'image d'entrée. Cette nouvelle image est ensuite repassée dans le modèle pour mesurer sa contribution à la prédiction cible.

- **Agrégation des scores** : les scores obtenus sont utilisés pour pondérer les cartes d'activation correspondantes. La somme pondérée donne la carte Score-CAM.
- **Visualisation** : la carte finale est superposée à l'image originale sous forme de carte thermique.

Algorithme Score-CAM

L'algorithme suivant présente le pipeline de Score-CAM tel qu'il a été appliqué dans ce travail (algorithm 3).

Algorithm 3: Fonctionnement de Score-CAM pour AtR5C

Input: Image d'entrée x, modèle entraîné f, couche convolutionnelle cible L
Output: Carte Score-CAM $M(x)$
Étape 1 : Prédiction initiale
Obtenir $f(x)$ et $c = \arg\max f(x)$
Étape 2 : Extraction des activations
Construire un modèle f' tel que $f'(x) = A^L(x)$
où $A^L(x)$ est la sortie de la couche L
Étape 3 : Génération de heatmaps
for *chaque canal* k *de* $A^L(x)$ **do**
    Redimensionner $A_k^L(x)$ à la taille de x
    Normaliser la carte
    Multiplier x pixel à pixel avec la carte normalisée $\rightarrow x_k$
    Calculer $s_k = f_c(x_k)$
    $M(x) \leftarrow M(x) + s_k \cdot A_k^L(x)$
Étape 4 : Post-traitement
$M(x) \leftarrow \text{ReLU}(M(x))$
Normaliser $M(x)$ entre [0,1]
Redimensionner à la taille d'entrée
Étape 5 : Visualisation
Superposer $M(x)$ sur l'image originale

Résultats visuels

La figure 3.4 montre une visualisation générée par Score-CAM. Les zones colorées révèlent les régions les plus importantes détectées par le modèle pour effectuer sa classification.

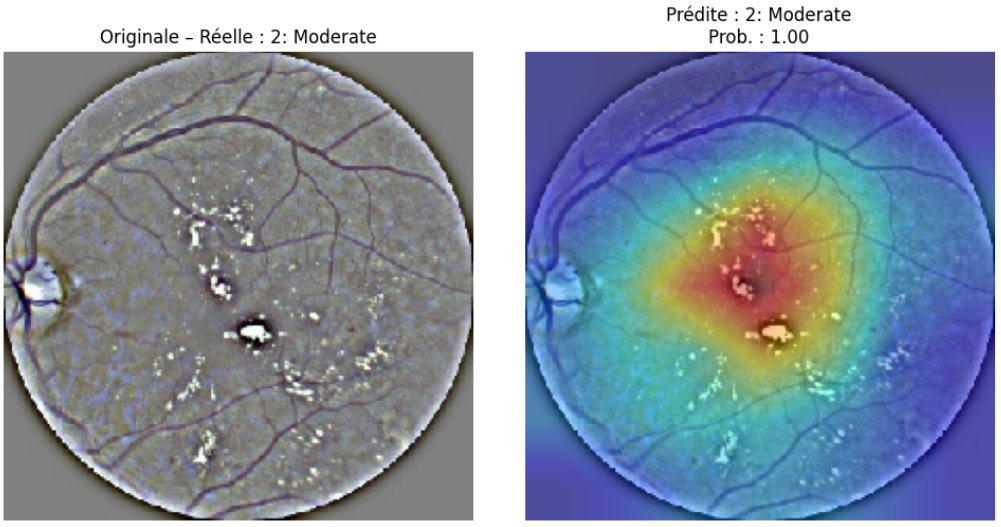


Figure 3.4: Carte Score-CAM générée pour une image de la classe 2

Score-CAM met en évidence une forte activation autour de l'exsudat central, avec un halo s'étendant vers les vaisseaux adjacents. Le modèle AtR5C prédit une rétinopathie « modérée » (classe 2, probabilité 1,00) en s'appuyant principalement sur cette lésion.

Comparée à Grad-CAM, la carte Score-CAM offre des contours plus nets et un champ d'attention légèrement plus large, incluant des structures périphériques. Toutefois, certaines lésions comme la fovéa sombre ou les micro-anévrismes restent peu activées.

Évaluation Quantitative

Les métriques utilisées sont identiques à celles appliquées pour Grad-CAM, permettant ainsi une comparaison cohérente. Les résultats sont présentés dans le tableau 3.5.

Table 3.5: Métriques d'explicabilité pour Score-CAM sur AtR5C

Métrique	**AtR5C**
Insertion	0.781137
Deletion	0.553961
Faithfulness	0.906650
Robustness	0.742029
Temps total d'exécution	4h 47min 42s

Les résultats obtenus pour Score-CAM confirment sa bonne capacité à localiser les régions critiques (insertion : 0,78), avec une montée rapide de la probabilité correcte lors de la réintroduction progressive des pixels importants. Le score de deletion (0,55) reste moyen, ce qui indique que la suppression des zones importantes affecte la prédiction, mais que la carte conserve une certaine largeur. La fidélité est excellente (0,91), témoignant d'une bonne cohérence entre l'importance attribuée aux pixels et leur impact réel sur la sortie, sans effet de saturation artificielle. Enfin, la robustesse (0,74) est bonne : les cartes de chaleur varient peu

face à de légères perturbations, montrant une stabilité supérieure aux méthodes basées sur les gradients, ce qui est un atout pour les applications en imagerie médicale.

IV.4 Méthodes XAI appliquées au modèle ViR-5C

Rollout est une technique qui n'est utilisée que par les ViTs

IV.4.1 Attention Rollout

L'implémentation a été effectuée en `PyTorch 2.2.2` et appliquée sur un modèle **ViR-5C** fine-tuné sur le dataset APTOS 2019. La procédure inclut l'extraction des attentions multi-têtes, leur agrégation, et leur propagation couche par couche pour générer une carte de chaleur finale.

- **Extraction des matrices d'attention** : lors de l'inférence, les matrices d'attention de chaque couche sont récupérées à l'aide du paramètre `output_attentions=True`.
- **Fusion des têtes** : pour chaque couche, les attentions sont moyennées sur les différentes têtes.
- **Ajout des connexions résiduelles** : une matrice identité est ajoutée à chaque couche pour simuler les connexions résiduelles.
- **Propagation cumulée** : les matrices sont multipliées les unes avec les autres (produit matriciel) afin de propager l'attention depuis la couche d'entrée.
- **Projection spatiale** : le vecteur d'attention associé au token `[CLS]` est extrait et redimensionné pour obtenir une carte 14×14, interpolée ensuite à la taille 224×224.
- **Colorisation et superposition** : la carte est normalisée, colorisée selon une palette personnalisée, puis superposée à l'image d'origine avec une transparence $\alpha = 0.4$.

Algorithme Attention Rollout

L'algorithme suivant décrit le fonctionnement de la méthode Attention Rollout pour le modèle ViR-5C (algorithm 4).

Algorithm 4: Fonctionnement de l'Attention Rollout pour ViR-5C

Input: Image d'entrée x, modèle ViT f entraîné, couches d'attention A_l
Output: Carte de saillance Attention Rollout $M(x)$
Étape 1 : Inférence avec extraction des attentions
Obtenir $(\hat{y}, \{A_1, A_2, ..., A_L\}) = f(x)$ avec `output_attentions=True`
Étape 2 : Propagation de l'attention
$R \leftarrow I$ (matrice identité)
for *chaque couche* $l = 1$ *à* L **do**
  $A'_l \leftarrow$ moyenne(A_l) sur les têtes
  $A'_l \leftarrow A'_l + I$
  $A'_l \leftarrow$ normaliser par ligne
  $R \leftarrow A'_l \cdot R$
Étape 3 : Projection spatiale
Extraire $R[\texttt{CLS}, 1:]$ et reformer une carte 14×14
Interp. bilinéaire à 224×224 et normalisation $[0, 1]$
Étape 4 : Affichage
Coloriser en RGB personnalisé et superposer à x

Résultats visuels obtenus

La figure 3.5 montre une visualisation générée par Attention Rollout pour une image test. On observe que l'attention est concentrée sur les zones typiquement affectées par la rétinopathie diabétique.

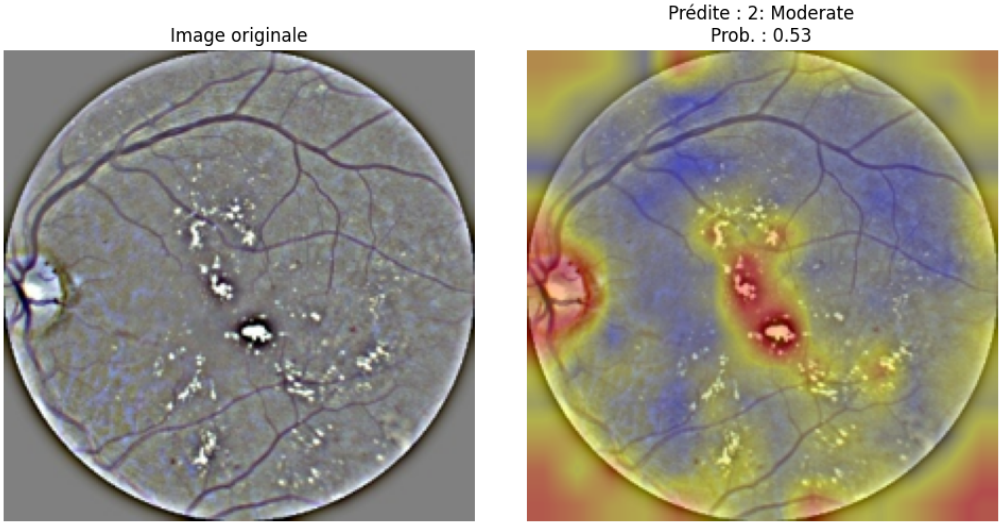


Figure 3.5: Visualisation Attention Rollout

Attention Rollout sur le modèle ViT (ViR-5C)

La carte générée montre une forte attention sur l'exsudat central et, de façon notable, sur le disque optique, malgré son absence de caractère pathologique. Le modèle prédit une rétinopathie « modérée » (classe 2, probabilité 0,53).

Des activations secondaires apparaissent autour des vaisseaux proches, mais la fovéa sombre et les micro-anévrismes sont peu pris en compte, révélant une sensibilité partielle aux petites lésions cliniques.

Évaluation Quantitative

Sur un échantillon de 500 images, plusieurs métriques d'évaluation ont été calculées afin de quantifier la pertinence des explications générées par Attention Rollout, présentées dans le tableau 3.6.

Table 3.6: Métriques d'explicabilité pour Attention Rollout sur ViR-5C

Métrique	**ViR-5C**
Insertion	0.346490
Deletion	0.381975
Robustness	0.724510
Temps total d'exécution	0h 12min 7s

Les résultats métriques obtenus pour Attention Rollout révèlent une performance contrastée. Le score d'insertion est faible (0,35), ce qui traduit une récupération lente de l'information utile au fur et à mesure de la réintroduction des pixels. Cela suggère une dispersion excessive de l'attention, qui peine à cibler efficacement les lésions spécifiques comme les micro-anévrismes. En revanche, la métrique de deletion (0,38) est relativement bonne : la suppression des zones jugées importantes entraîne une chute notable de la probabilité, ce qui indique une localisation plutôt précise des régions clés, même si certaines zones pertinentes échappent encore à la carte. Enfin, la robustesse est excellente (0,72), témoignant d'une forte stabilité des cartes face aux perturbations légères (bruit, recadrage), probablement grâce à l'agrégation des informations d'attention sur plusieurs couches du modèle ViR-5C.

IV.5 Méthodes appliquées aux modèles AtR5C, ViR-5C et ReVi-5C

Dans cette section nous presenterons les méthodes de XIA qui peuvent être appliquées aux 3 modèles pour une pouvoir effectuer une comparaison.

IV.5.1 SHAP (SHapley Additive exPlanations)

Nous avons implémenté SHAP pour trois architectures : **AtR5C**, **ViR-5C** et le modèle hybride **ReVi-5C**. L'objectif était d'analyser la sensibilité des prédictions aux variations locales des pixels, à travers des valeurs de Shapley calculées à l'aide de `DeepExplainer` pour AtR5C et le modèle hybride, et `ImageExplainer` pour ViT.

- **Préparation du modèle** : chargement du modèle entraîné (AtR5C, ViR-5C ou ReVi-5C) avec ses poids, mis en mode évaluation.
- **Prétraitement des images** : redimensionnement à 224×224, normalisation et transformation en tenseurs compatibles avec l'entrée du modèle.
- **Sélection des échantillons de fond** : un ensemble réduit d'images de fond est utilisé pour estimer l'effet marginal moyen.
- **Calcul des valeurs de SHAP** : l'explainer calcule les contributions de chaque pixel (ou patch dans le cas de ViR-5C) pour une image donnée.

- **Visualisation** : génération de cartes de chaleur superposées à l'image d'entrée montrant les zones ayant une contribution positive (rouge) ou négative (bleu) à la prédiction.

Algorithme SHAP

L'algorithme suivant résume l'application de SHAP sur les architectures utilisées (algorithm 5).

Algorithm 5: Fonctionnement de SHAP pour les modèles d'images

Input: Image d'entrée x, modèle entraîné f, échantillons de fond B
Output: Carte de valeurs SHAP $S(x)$
Étape 1 : Prétraitement
Redimensionner et normaliser x
Construire B avec n images de fond
Étape 2 : Initialiser l'explainer
if *modèle est CNN (AtR5C, ReVi-5C)* **then**
  Explainer $\leftarrow$ `DeepExplainer`(f, B)
else if *modèle est ViR-5C* **then**
  Explainer $\leftarrow$ `ImageExplainer`(f, B)
Étape 3 : Calcul des SHAP values
$S(x) \leftarrow$ `explainer.shap_values`(x)
Étape 4 : Visualisation
Afficher $S(x)$ sous forme de carte de chaleur colorisée

Résultats visuels

La figure 3.6 montre les cartes SHAP obtenues pour les trois architectures.

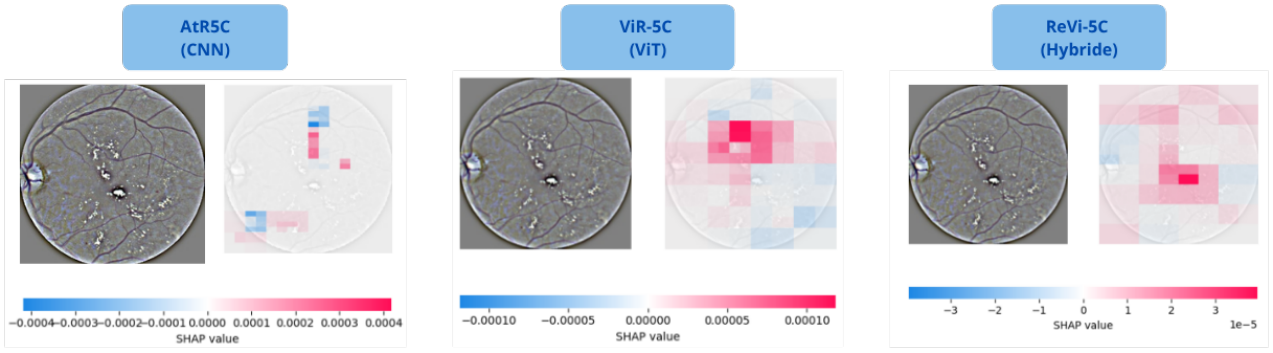


Figure 3.6: Visualisation des valeurs SHAP sur AtR5C, ViR-5C et ReVi-5C

Interprétation des couleurs : Les zones en rouge foncé indiquent une forte contribution positive (poussant la prédiction vers une rétinopathie sévère), tandis que celles en bleu foncé traduisent une contribution négative (associée à une faible sévérité). Les teintes claires correspondent à un impact faible ou neutre.

CNN (*AtR5C*) : La carte SHAP met en évidence plusieurs petites zones rouges dispersées, notamment autour des exsudats et des vaisseaux, indiquant que le CNN prend en compte diverses régions localisées sans se focaliser exclusivement sur une seule zone dominante.

ViT (*ViR-5C*) : Une grande zone rouge centrée sur l'exsudat principal domine la carte, avec peu d'autres activations. Le ViT semble focaliser son attention sur cette seule lésion, au détriment des anomalies périphériques.

Modèle hybride (*ReVi-5C*) : L'activation est plus compacte que pour le CNN, mais moins concentrée que pour le ViT. Elle englobe l'exsudat principal sans intégrer les structures fines environnantes, traduisant une stratégie intermédiaire visant à préserver l'essentiel tout en réduisant le bruit.

Évaluation Quantitative

L'efficacité explicative de SHAP a été mesurée via deux métriques sur 500 images de test. Les résultats sont reportés dans le tableau 3.7.

Table 3.7: Comparaison des métriques d'explicabilité

Métrique	**AtR5C**	**ViR-5C**	**ReVi-5C**
Faithfulness	**1.000000**	**1.000000**	**1.000000**
Robustness	0.484171	0.515607	0.566952
Temps total d'exécution	3h 14min 37s	0h 43min 40s	4h 15min 22s

Les scores de fidélité obtenus avec SHAP sont parfaits pour tous les modèles testés, avec une valeur de **1,00**. Cette fidélité maximale est une conséquence directe de la construction mathématique de SHAP : fondée sur la théorie des jeux, la somme des contributions des pixels correspond exactement à la sortie du modèle. Ce résultat garantit une corrélation théorique parfaite entre importance attribuée et prédiction, mais cela ne signifie pas nécessairement que la carte localise avec précision les lésions. Il convient donc de compléter cette analyse par une évaluation qualitative, notamment visuelle.

En ce qui concerne la robustesse, SHAP présente des scores moyens : **0,48** pour AtR5C, **0,52** pour ViR-5C, et **0,57** pour le modèle hybride. Ces valeurs indiquent une sensibilité notable aux perturbations de l'image (bruit, flou, etc.). Toutefois, on observe une amélioration progressive entre les architectures : ViR-5C est légèrement plus stable que AtR5C, et le modèle hybride (ReVi-5C) obtient la meilleure robustesse. Cela suggère que la combinaison de deux types d'architectures permet de produire des cartes SHAP plus cohérentes même lorsque les images varient légèrement.

Enfin, le principal inconvénient de SHAP reste son **temps d'exécution**. Cette méthode est particulièrement coûteuse en calcul, car elle repose sur la génération et l'évaluation de milliers de versions perturbées de l'image. Dans ce contexte, le modèle ViR-5C est le plus rapide, avec un temps moyen d'environ **0,7 seconde par image**, grâce à sa structure légère et à une implémentation efficace. À l'inverse, le modèle ReVi-5C est le plus lent, avec un temps d'exécution dépassant les **4 heures** pour le traitement du jeu de test, en raison de la double passe nécessaire (AtR5C puis ViR-5C).

IV.5.2 LIME (Local Interpretable Model-agnostic Explanations

Nous avons appliqué LIME sur les trois architectures : **AtR5C**, **ViR-5C** et le modèle hybride **ReVi-5C**, à l'aide de la bibliothèque `lime` adaptée à la classification d'images. LIME

a permis de mettre en évidence les superpixels les plus influents dans la prédiction finale pour une image donnée.

- **Chargement du modèle** : le modèle (AtR5C, ViR-5C ou ReVi-5C) est chargé avec ses poids et mis en mode évaluation.
- **Prétraitement de l'image** : redimensionnement à 224×224, normalisation et transformation vers le format d'entrée du modèle.
- **Segmentation en superpixels** : division de l'image en zones homogènes via `quickshift` ou `slic`.
- **Génération de perturbations** : des versions modifiées de l'image sont générées en masquant aléatoirement certains superpixels.
- **Prédiction des images perturbées** : le modèle prédit la classe pour chaque image perturbée.
- **Ajustement d'un modèle local** : un modèle linéaire est entraîné pour approximer localement la décision du modèle initial.
- **Visualisation** : surlignement des superpixels ayant une influence positive ou négative sur la prédiction finale.

Algorithme LIME

L'algorithme algorithm 6 résume l'approche LIME utilisée.

Algorithm 6: Fonctionnement de LIME pour les modèles d'images

Input: Image d'entrée x, modèle entraîné f
Output: Carte LIME $L(x)$ mettant en évidence les superpixels influents
Étape 1 : Segmentation de l'image
Segmenter x en k superpixels $\{s_1, s_2, \ldots, s_k\}$
Étape 2 : Génération de perturbations
Créer n versions $x^{(i)}$ de l'image avec superpixels masqués
Étape 3 : Prédictions
Pour chaque $x^{(i)}$, calculer $f(x^{(i)})$
Étape 4 : Modèle local
Ajuster un modèle linéaire g pour approximer f localement
Étape 5 : Visualisation
Extraire les coefficients de g et afficher les superpixels les plus influents

IV.5.2.1 Résultats visuels

La figure 3.7 présente les visualisations LIME générées pour chaque architecture.

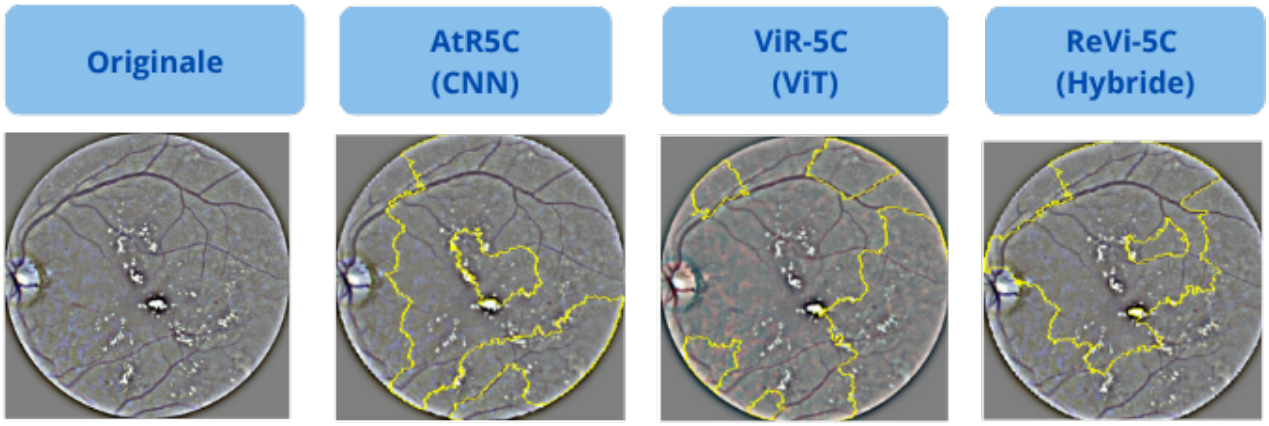


Figure 3.7: Visualisation des explications LIME sur AtR5C, ViR-5C et ReVi-5C

CNN (*AtR5C*)

Les contours détectés par LIME entourent principalement l'amas central d'exsudats et les vaisseaux les plus visibles. Le CNN s'appuie donc sur des anomalies bien marquées, en se concentrant sur les lésions majeures.

ViT (*ViR-5C*)

LIME met en évidence des zones plus petites et fragmentées, souvent localisées autour de micro-anévrismes et de courts segments vasculaires. Le ViT semble capter des détails fins et très localisés.

Modèle hybride (*ReVi-5C*)

Les zones entourées sont de taille intermédiaire, centrées sur l'exsudat principal tout en négligeant certaines lésions périphériques. Le modèle hybride semble combiner la focalisation du CNN sur les anomalies dominantes et la finesse d'analyse du ViT.

Évaluation Quantitative

Les performances explicatives de LIME ont été mesurées selon ces deux métriques comme montré dans le tableau 3.8.

Table 3.8: Comparaison des métriques d'explicabilité selon l'architecture

Métrique	AtR5C	ViR-5C	ReVi-5C
Faithfulness	**1.000000**	**1.000000**	**1.000000**
Robustness	0.905470	0.911545	0.905470
Temps total d'exécution	3h 32min 54s	2h 45min 42s	4h 14min 26s

LIME atteint un score parfait de fidélité (**1,00**) sur tous les modèles évalués. Ce résultat s'explique par le principe même de LIME, qui consiste à ajuster une régression linéaire locale sur des échantillons perturbés de l'image, de manière à reconstruire exactement la sortie du modèle dans cette zone locale. La corrélation avec la prédiction est donc structurellement maximale.

Concernant la robustesse, LIME présente des résultats remarquables avec des scores compris entre **0,905** et **0,912** selon l'architecture utilisée, ce qui montre une excellente stabilité

des explications face aux perturbations légères de l'image. Le modèle ViR-5C obtient le meilleur score (**0,912**), probablement grâce à sa représentation en patches, qui amortit les effets du bruit pixel à pixel.

Cependant, cette qualité d'explication a un coût en termes de **temps d'exécution**. Chaque carte LIME repose sur la génération et l'analyse de **500 à 1000 perturbations** de l'image originale. Le modèle ViT s'en sort le mieux en raison de sa structure séquentielle directe, tandis que le modèle hybride ReVi-5C est le plus lent, du fait du double passage nécessaire à chaque perturbation.

IV.6 Analyse Visuelle Comparative par Stade de Rétinopathie

Cette section propose une analyse visuelle comparative des principales méthodes d'explicabilité (XAI) appliquées aux trois modèles étudiés **AtR5C**, **ViR-5C** et **ReVi-5C** pour chaque stade de rétinopathie diabétique (de 0 à 4). L'objectif est d'évaluer, à travers les figures présentées 3.8 3.9 3.10, si les cartes générées permettent de localiser correctement les zones pathologiques cliniquement reconnues (macula, microanévrismes, exsudats, néovaisseaux).

IV.6.1 Modèle ViR-5C

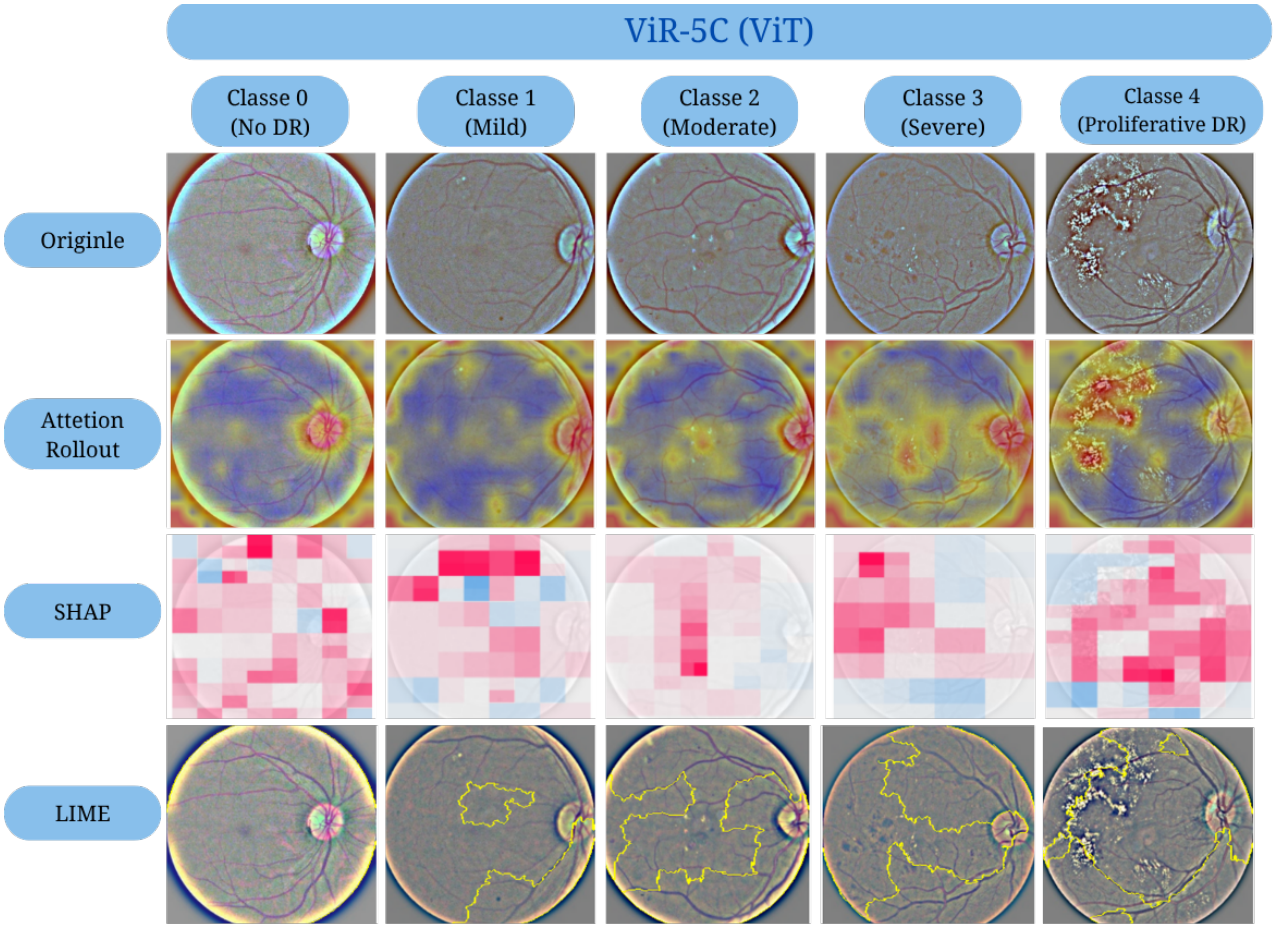


Figure 3.8: Comparaison des explications XAI sur le modèle ViR-5C

Classe 0 (absence de rétinopathie) : Les trois méthodes convergent vers une absence d'activation significative. Attention Rollout montre un léger halo périphérique, traduisant

l'absence de zones suspectes. SHAP présente des tuiles dispersées sans structure cohérente, signalant l'absence d'impact localisé. LIME n'affiche aucun contour, confirmant l'absence de pathologie reconnue par le modèle.

Classe 1 (stade léger) : Attention Rollout révèle une activation centrée sur la macula, traduisant la détection des premiers signes, comme les microanévrismes. SHAP affiche quelques blocs roses dans la partie supérieure de la rétine, mais leur lisibilité est réduite par un bruit coloré. LIME produit un petit contour polygonal autour de la macula, mettant en évidence une région critique bien localisée.

Classe 2 (modérée) : La carte Attention Rollout devient plus diffuse, couvrant la macula et les vaisseaux avoisinants. SHAP dessine une colonne centrale de tuiles rosées alignée sur la macula, montrant une contribution notable mais difficile à interpréter à cause de la granularité. LIME élargit ses contours pour englober des structures vasculaires autour de la macula, ce qui reflète l'extension des lésions.

Classe 3 (sévère) : Attention Rollout affiche deux foyers distincts autour de la macula, signe d'une attention répartie sur plusieurs lésions. SHAP montre une carte instable, mêlant tuiles positives et négatives, ce qui rend l'interprétation plus délicate. LIME trace plusieurs polygones sur les exsudats, offrant une segmentation lisible, bien que légèrement englobante.

Classe 4 (proliférative) : Attention Rollout met en évidence une activation intense couvrant la macula et la périphérie, cohérente avec la sévérité des lésions prolifératives. SHAP génère une carte globalement rose, traduisant une forte influence des zones atteintes, mais sans distinction nette entre foyers. LIME entoure la quasi-totalité de la rétine avec un large contour, indiquant une détection généralisée, mais manquant de finesse dans la segmentation des foyers multiples.

IV.6.2 Modèle AtR5C

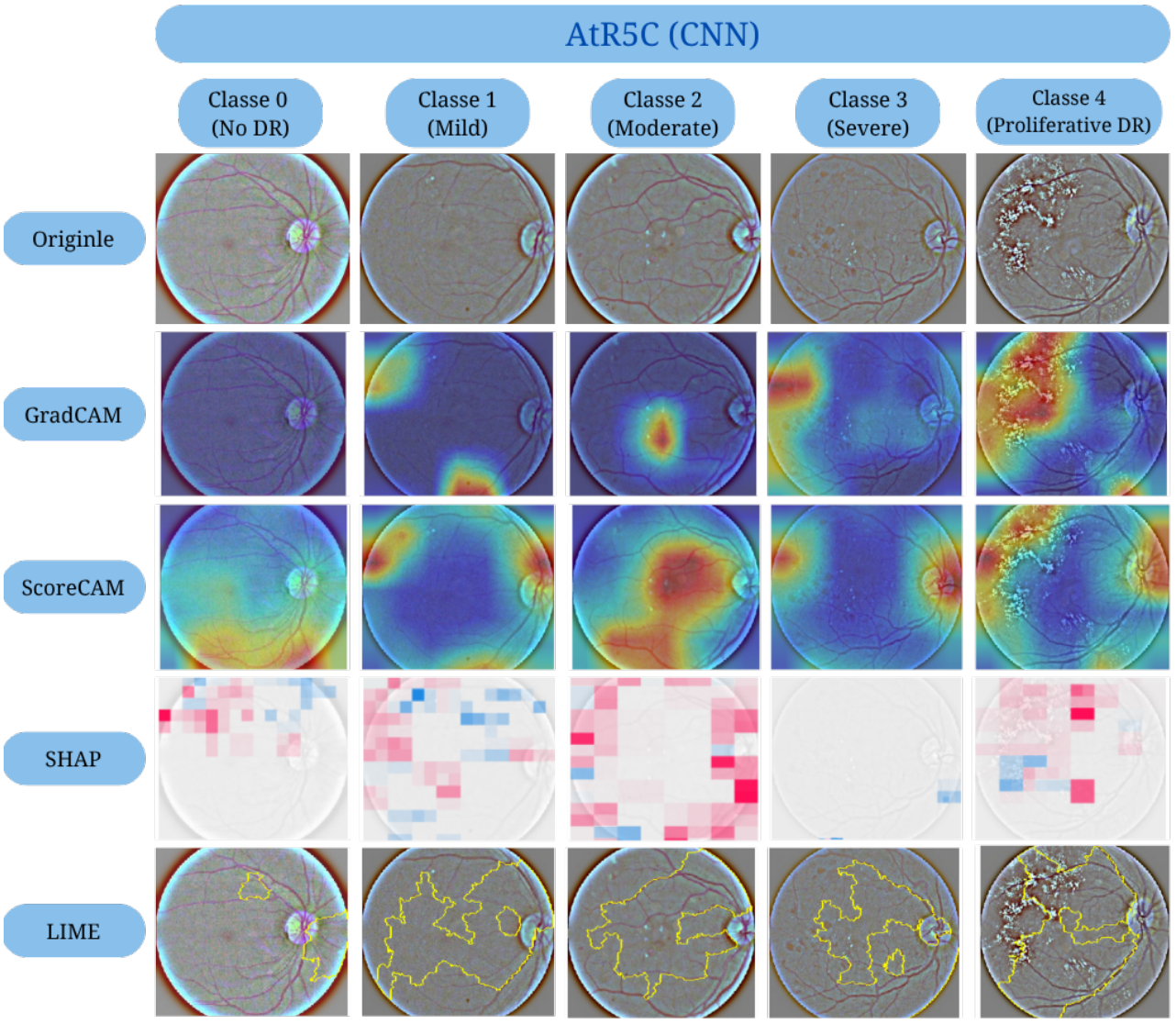


Figure 3.9: Comparaison des explications XAI sur le modèle AtR5C

Classe 0 (absence de rétinopathie) : Les quatre méthodes convergent vers une absence notable d'activation. Grad-CAM produit des cartes quasi vierges, confirmant l'absence d'anomalie détectée. Score-CAM montre un léger halo périphérique peu informatif. SHAP révèle principalement des zones neutres sans contribution marquée. LIME n'affiche aucun contour explicatif, traduisant bien la normalité des images.

Classe 1 (stade léger) : Grad-CAM identifie efficacement les microanévrismes à travers des zones rouges bien localisées. Score-CAM produit des activations plus diffuses autour de la macula, avec une précision réduite. SHAP indique quelques pixels rouges correspondant aux microanévrismes, mais noyés dans du bruit. LIME cerne localement certains microanévrismes avec de petits contours jaunes, traduisant une bonne capacité de focalisation locale.

Classe 2 (modérée) : Grad-CAM surligne la macula et les exsudats avec une intensité croissante. Score-CAM couvre largement les vaisseaux et régions pathologiques, mais de

façon imprécise. SHAP marque quelques zones contributives mais reste affecté par une forte granularité. LIME trace des formes polygonales sur les exsudats et la macula, permettant une lecture claire, bien que parfois trop englobante.

Classe 3 (sévère) : Grad-CAM reste pertinent, focalisant sur les lésions visibles autour de la région maculaire. Score-CAM active une grande partie de la rétine, donnant une impression de suractivation. SHAP devient moins expressif, les zones rouges étant peu marquées, indiquant une difficulté à expliquer les prédictions dans les cas sévères. LIME montre des contours étendus autour des exsudats, avec une interprétation plus lisible que SHAP.

Classe 4 (proliférative) : Grad-CAM identifie correctement les régions prolifératives, bien que la différenciation des foyers reste perfectible. Score-CAM recouvre presque toute la rétine en rouge intense, illustrant la sévérité mais sans finesse. SHAP génère quelques tuiles rouges isolées, insuffisantes pour représenter la complexité des lésions. LIME affiche une large zone délimitée, cohérente avec l'étendue de la pathologie, mais manquant de précision sur les localisations fines.

IV.6.3 Modèle ReVi-5C

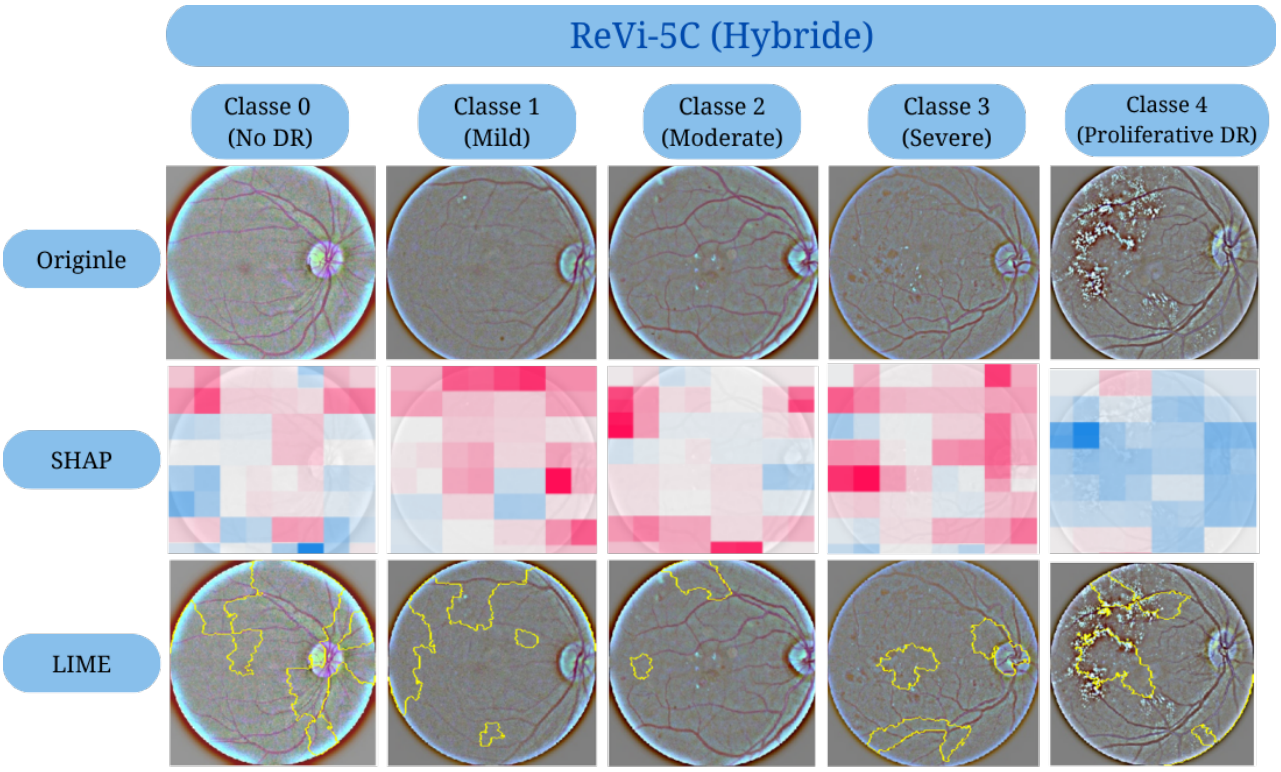


Figure 3.10: Comparaison des explications XAI sur le modèle ReVi-5C

Classe 0 (absence de rétinopathie) : SHAP affiche des tuiles très pâles, avec quelques blocs bleus en périphérie, traduisant une absence d'activation significative, conforme à la normalité. LIME génère un contour jaune très fin en périphérie, probablement lié à du bruit, sans lien avec une lésion réelle.

Classe 1 (stade léger) : SHAP montre des dominantes rosées autour de la macula, correspondant à la détection des microanévrismes, malgré une certaine présence de bruit visuel. LIME cerne efficacement la macula avec un périmètre arrondi, traduisant une bonne localisation des régions critiques.

Classe 2 (modérée) : SHAP produit une colonne verticale de tuiles rouges centrée sur la macula, bien alignée avec les exsudats modérés ; le reste de l'image demeure neutre. En revanche, LIME génère un contour décalé sur une petite zone latérale, ne couvrant pas les principales anomalies, ce qui révèle un déficit de détection des zones réellement pathologiques.

Classe 3 (sévère) : Les tuiles rouges SHAP entourent les foyers d'exsudats, mais leur interprétation est affaiblie par la co-présence de zones bleues. LIME offre une explication plus claire, avec des contours bien positionnés englobant les foyers majeurs, ce qui renforce sa pertinence à ce stade.

Classe 4 (proliférative) : SHAP présente une carte paradoxalement dominée par des tuiles bleues, ce qui rend l'interprétation moins intuitive du point de vue médical, malgré la sévérité du stade. À l'inverse, LIME génère plusieurs contours précis qui longent les néovaisseaux et les exsudats prolifératifs, permettant une bonne lecture des zones critiques.

V Discussion et Comparaison

Cette section présente une discussion des résultats obtenus, suivie d'une comparaison des méthodes d'explicabilité et des modèles évalués. Elle permet de mettre en lumière les forces, les limites et les différences entre les approches utilisées.

V.1 Interprétation des Résultats

L'analyse combinée des résultats quantitatifs (métriques explicatives) et qualitatifs (visualisations) appliqués aux modèles **AtR5C** (CNN), **ViR-5C** (ViT) et **ReVi-5C** (Hybride) révèle des comportements distincts, tant sur le plan de la précision explicative que sur celui de la lisibilité visuelle. Sur la base des métriques d'explicabilité, **Grad-CAM** sur AtR5C excelle en insertion avec un score de **0,83**, indiquant une excellente capacité à identifier les zones réellement discriminantes. **Score-CAM** suit de près (**0,78**), tandis qu'**Attention Rollout** sur ViR-5C montre une insertion plus faible (**0,35**), malgré sa rapidité d'exécution (**12 minutes**), la meilleure de toutes les méthodes. Sur la métrique de suppression (deletion), ViR-5C se démarque avec Attention Rollout (**0,38**), approchant le seuil optimal, tandis que Grad-CAM (**0,62**) et Score-CAM (**0,55**) sont moins performants. En termes de **fidélité**, **SHAP**, **LIME** et **Grad-CAM** atteignent **1,00**. Score-CAM obtient un excellent score réaliste (**0,91**). La **robustesse** révèle d'autres contrastes : LIME sur ViR-5C atteint **0,91**, preuve de stabilité, tandis que Grad-CAM chute à **0,39**, signalant une forte sensibilité aux perturbations. Enfin, sur le critère de **temps d'exécution**, **Score-CAM** (4 h 47 min 42 s), LIME et SHAP (jusqu'à **4 h 15 min**) se révèlent les plus lents, ce qui limite leur déploiement sur de larges jeux de données.

L'évaluation visuelle confirme ces tendances. Pour la **classe 0**, tous les modèles identifient correctement l'absence de pathologie, avec des cartes neutres et non activées. En **classe 1**, ViR-5C et ReVi-5C détectent efficacement les microanévrismes via Attention Rollout et LIME,

tandis qu'AtR5C les localise bien avec Grad-CAM. À partir de la **classe 2**, ViR-5C offre une attention diffuse mais cohérente, tandis qu'AtR5C devient moins précis avec Score-CAM. ReVi-5C montre ici un déficit localisé avec LIME. En **classe 3**, ViR-5C capte plusieurs foyers, AtR5C reste lisible avec Grad-CAM, mais Score-CAM tend à suractiver. ReVi-5C, avec LIME, reste robuste, même si SHAP présente du bruit. En **classe 4** (prolifération), ViR-5C couvre intensément les lésions, parfois trop largement. AtR5C est précis via Grad-CAM, tandis que Score-CAM perd en finesse. ReVi-5C combine bien SHAP et LIME, mais SHAP peut générer des cartes contre-intuitives.

Ainsi, ViR-5C se distingue par sa **rapidité** et son attention structurée à large échelle (via Attention Rollout), mais il manque parfois de précision fine. AtR5C, notamment avec Grad-CAM, est très performant pour la détection locale, mais moins robuste. Le modèle hybride ReVi-5C combine les avantages des deux : une bonne localisation via LIME et une vision globale via SHAP, au prix d'un temps de traitement plus élevé. LIME se révèle comme la méthode la plus constante et robuste pour les trois modèles, même si elle est lente et peut légèrement sur-segmenter. SHAP reste fidèle théoriquement, mais bruité visuellement, surtout dans les stades avancés. Score-CAM émerge comme un **compromis intéressant** entre fidélité et robustesse, mais son temps d'exécution élevé (4 h 47 min 42 s) constitue une limite significative.

V.2 Comparaison des méthodes

Le tableau 3.9 présente un classement global des méthodes d'explicabilité selon leurs points forts et faibles, en tenant compte des critères principaux : fidélité, robustesse, précision visuelle et temps d'exécution.

Méthode	Points forts	Points faibles
Score-CAM	Bonne fidélité, robuste, insertion efficace	Localisation moyenne, temps d'exécution très élevé
LIME	Excellente robustesse	Très coûteuse en temps
Attention Rollout	Rapide, bonne robustesse	Moyenne précision, cartes diffuses
Grad-CAM	Meilleure insertion	Faible robustesse
SHAP	Bonne fidélité	Temps long, instabilité sur certains cas

Table 3.9: Résumé comparatif des méthodes

Recommandations selon les cas d'usage :

- **Score-CAM avec AtR5C :** offre un excellent compromis entre fidélité, insertion et robustesse visuelle, mais son temps d'exécution élevé le réserve à des analyses ciblées ou à de petits jeux de données, nécessitant des ressources de calcul importantes.
- **LIME avec ViR-5C :** choix pertinent dans les contextes nécessitant une forte robustesse, notamment pour des analyses reproductibles.
- **Attention Rollout avec ViR-5C :** méthode rapide, convenant bien aux applications en temps réel, malgré une fidélité légèrement inférieure.

Le choix de la méthode d'explicabilité doit ainsi être adapté aux priorités du contexte clinique : besoin de précision, de stabilité, de lisibilité ou encore de rapidité.

V.3 Comparaison des Modèles

Le tableau 3.10 résume les points forts et faibles des trois architectures testées, en tenant compte de leur précision, robustesse et complexité d'exécution.

Modèle	Points forts	Points faibles
ReVi-5C	Meilleure robustesse	Temps d'exécution très élevé
ViR-5C	Vitesse, bonne robustesse	Moins précis sur les petits détails
AtR5C	Bonne fidélité	Moins robuste que ViR-5C

Table 3.10: Résumé comparatif des modèles

Discussion : Chaque architecture présente des avantages spécifiques selon les objectifs recherchés :

- **AtR5C** : offre une forte capacité d'identification (insertion élevée), idéal pour la localisation précise des lésions. Cependant, ses explications manquent de stabilité face aux perturbations (robustesse faible).
- **ViR-5C** : excelle en robustesse et en rapidité, notamment avec LIME et Attention Rollout. Sa représentation globale par patchs assure une bonne tolérance aux transformations, mais certaines méthodes y produisent des cartes floues ou peu contrastées.
- **ReVi-5C** : combine les avantages des approches AtR5C et ViR-5C, ce qui renforce la qualité des explications (fidélité + robustesse). Toutefois, son coût de calcul est très élevé, ce qui limite son usage en pratique.

V.4 Analyse explicative des erreurs de classification des modèles

La figure 3.11 montre sur quelle base les modèles AtR5C (CNN), ViR-5C (ViT) et ReVi-5C (hybride), utilisant les méthodes LIME et SHAP, ont effectué leurs prédictions. Elle présente les principales fautes de classification sur les stades *Mild* (1) et *Moderate* (2), suivie d'une interprétation des causes possibles de ces erreurs.

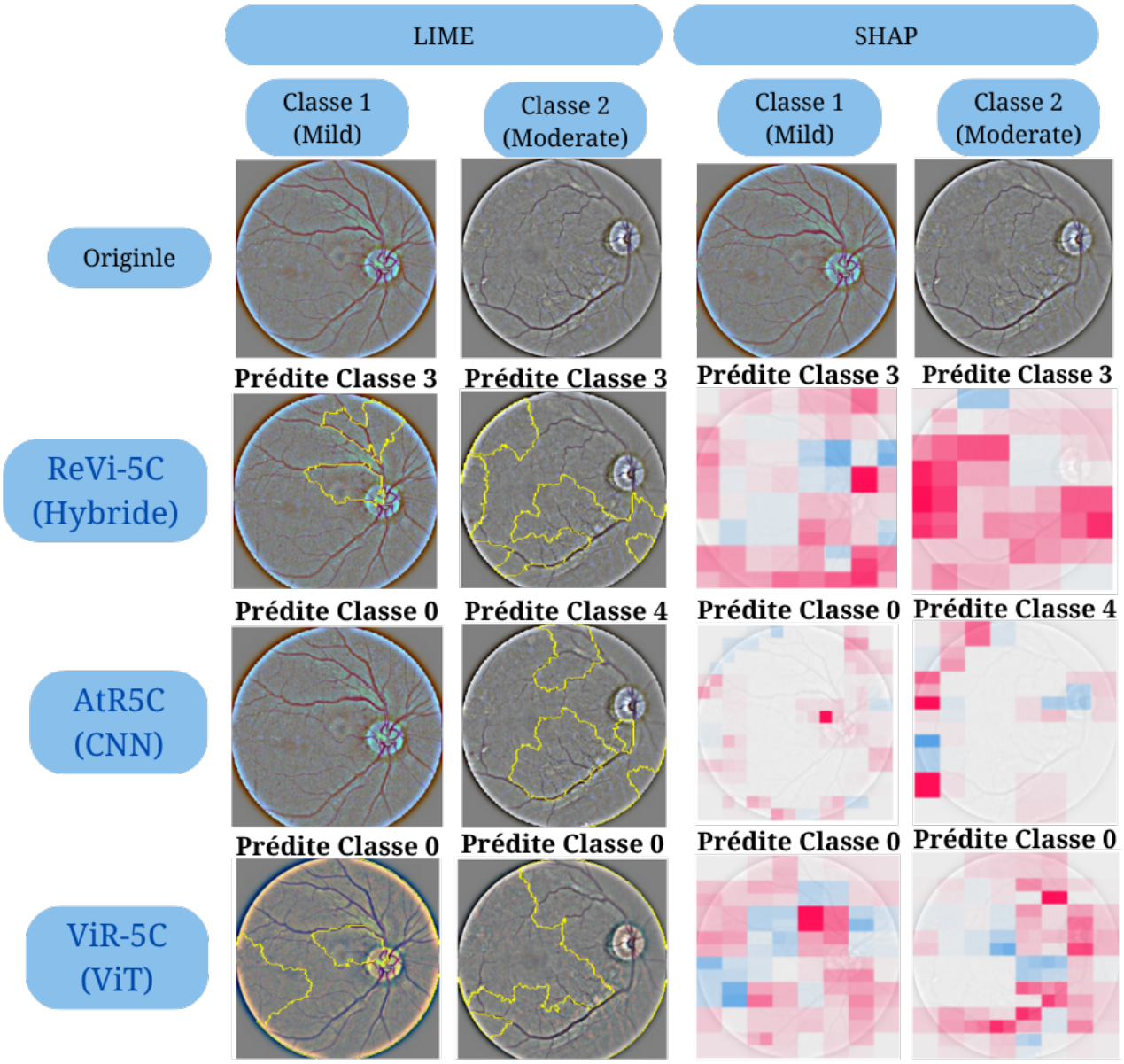


Figure 3.11: Visualisation des erreurs de classification expliquées par LIME et SHAP sur les stades Mild et Moderate

Nous avons souhaité illustrer le comportement des modèles dans des cas de mauvaise classification. L'exemple présenté dans la Figure 3.11 montre que, pour les stades *Mild* (1) et *Moderate* (2), les trois modèles AtR5C (CNN), ViR-5C (ViT) et ReVi-5C (hybride) rencontrent des difficultés à localiser et exploiter correctement les lésions centrales (microanévrismes ou exsudats).

- **CNN (AtR5C)** : a tendance à classer ces cas en *No DR* ou à surestimer en *Proliferative DR*, en raison d'une mauvaise détection ou d'une absence d'intégration des lésions dans les explications LIME et SHAP.
- **ViT (ViR-5C)** : classe fréquemment ces cas en *No DR*, bien qu'il détecte parfois un signal faible au niveau de la macula.
- **Hybride (ReVi-5C)** : tend à les classer en *Sévère* au lieu de *Modéré*, car il explore davantage de régions, ce qui peut amplifier l'impact de lésions périphériques ou diffuses.

Ces erreurs peuvent être dues à une mauvaise annotation initiale ou à la présence de bruit ou de lésions subtiles non visibles à l'œil nu. Il est toutefois préférable de surclasser un cas douteux plutôt que de le sous-estimer, notamment pour éviter un faux négatif dans le diagnostic de la rétinopathie diabétique. De ce point de vue, le modèle hybride ReVi-5C reste le plus prudent et peut être considéré comme performant dans ce type de cas.

VI Conclusion

Ce travail de comparaison des méthodes d'explicabilité appliquées à la classification de la rétinopathie diabétique montre qu'aucune solution n'est parfaite : chaque méthode XAI présente des forces mais aussi des limites qu'il faut prendre en compte selon le contexte médical.

Score-CAM ressort comme la méthode la plus équilibrée : elle offre une bonne fidélité et une robustesse correcte, mais son temps d'exécution élevé constitue une limite majeure. LIME, quant à elle, se distingue par une stabilité exceptionnelle, ce qui la rend très fiable pour l'analyse clinique, mais elle est très lente à exécuter. Attention Rollout se distingue principalement par sa rapidité d'exécution, ce qui est idéal pour une utilisation en temps réel, mais ses cartes sont plus floues et moins précises. Grad-CAM donne d'excellents résultats sur certaines métriques comme l'insertion, mais sa faible robustesse limite son intérêt. Enfin, SHAP propose une fidélité parfaite, mais cela résulte surtout d'un effet de construction ; en pratique, ses cartes restent sensibles aux perturbations et très coûteuses à générer.

Du point de vue des architectures, le modèle ViR-5C se révèle rapide et relativement robuste, ce qui le rend adapté aux situations cliniques dynamiques. Le AtR5C reste fiable et bien compris, avec de bons résultats en insertion et fidélité. Le modèle ReVi-5C atteint la meilleure robustesse, mais au prix d'un temps d'exécution élevé.

En résumé, le choix de la méthode XAI et du modèle dépendra toujours des besoins spécifiques du terrain : rapidité, lisibilité, robustesse ou précision.

Conclusion Générale

Ce mémoire a mis en évidence la possibilité de quantifier l'explicabilité des modèles d'apprentissage profond appliqués au dépistage de la rétinopathie diabétique, ainsi que de comparer de manière rigoureuse différentes techniques d'intelligence artificielle explicable (XAI), dans l'optique d'orienter leur utilisation en contexte clinique.

Les principaux enseignements issus de ce travail peuvent être résumés comme suit :

1. Complémentarité des approches : aucune méthode d'explicabilité ne s'impose comme étant supérieure dans tous les contextes. Grad-CAM permet une localisation intuitive des régions importantes à un coût de calcul modéré, tandis que Score-CAM offre une localisation plus précise mais à un coût computationnel élevé. SHAP fournit des explications fines, parfois sensibles au bruit. LIME se révèle pertinent pour des interprétations locales ponctuelles, et Attention Rollout offre une visualisation claire de l'attention dans les architectures de type Vision Transformer (ViT).

2. Pertinence du choix des métriques : la métrique faithfulness permet de s'assurer que l'explication suit véritablement le raisonnement interne du modèle. Les métriques insertion et deletion sont adaptées aux situations où la précision médicale est prioritaire, tandis que robustness est à privilégier dans des environnements cliniques hétérogènes, nécessitant une stabilité sur différents supports.

3. Intérêt du cadre expérimental proposé : la combinaison d'évaluations quantitatives et qualitatives a permis de valider la pertinence des cartes d'explication produites, tout en mettant en lumière certains cas où un ajustement du modèle est nécessaire.

Les contributions majeures de notre travail sont les suivantes :

1. La réalisation d'une revue critique des principales techniques d'explicabilité (XAI) appliquées à la classification de la rétinopathie diabétique ;

2. La conception d'un protocole d'évaluation rigoureux permettant d'analyser les explications générées par trois architectures de classification de la rétinopathie diabétique : un modèle basé sur les CNN, un modèle Vision Transformer (ViT), et une architecture hybride CNN-ViT ;

3. Une étude comparative approfondie des performances et des explications produites par différentes méthodes XAI appliquées à ces trois modèles ;

4. Une visualisation interprétable des zones discriminantes activées par les modèles, offrant ainsi une ouverture de la « boîte noire » et une meilleure compréhension du processus décisionnel de chaque architecture.

Perspectives : Afin de renforcer à la fois l'utilité clinique et la qualité technique des méthodes explicatives étudiées, plusieurs pistes d'amélioration sont proposées :

1. **Réduction de l'erreur de classification** : Realiser une etude statistique des images mal classées, et agir soit sur les données ou sur les poids de décision.
2. **Impliquer les cliniciens dans la conception** : organiser des sessions de validation avec des ophtalmologues pour confronter les cartes générées à leur expertise.
3. **Variété de prétraitements d'images** : tester d'autres méthodes de filtrage et normalisation adaptative pour renforcer la robustesse aux variations d'acquisition.
4. **Occlusion du disque optique** : masquer artificiellement le disque optique pendant l'entraînement pour forcer le modèle à se concentrer sur les vraies lésions.
5. **Entraînement avec dropout spatial** : masquer aléatoirement des blocs entiers de la carte de caractéristiques pour diversifier l'attention et obliger le réseau à explorer d'autres régions et améliorer la robustesse, favorisant une meilleure généralisation.

Pour conclure, ce travail contribue à rapprocher l'intelligence artificielle des réalités et besoins du domaine médical. Il démontre qu'une évaluation rigoureuse de l'explicabilité ne relève pas d'une simple exigence académique, mais constitue une condition essentielle pour une adoption responsable, transparente et éthique des systèmes d'IA. Le développement de cadres d'évaluation tels que celui proposé ici ouvre la voie à une médecine plus fiable et résolument centrée sur le patient.

BIBLIOGRAPHIE

[1] Assurance Maladie (Ameli.fr). *Les complications du diabète au niveau des yeux*. https://www.ameli.fr/assure/sante/themes/diabete-adulte/diabete-symptomes-evolution/complications-yeux-diabete. Consulté le 8 avril 2025. 2025.

[2] M.N. A and K. Sathyarajasekaran. "SwinVNETR: Swin V-net Transformer with Non-local Block for Volumetric MRI Brain Tumor Segmentation". In: *Automatika* 65.4 (2024), pp. 1350–1363.

[3] Y. Abiche and A. Amokrane. "Diabetic Retinopathy classification using transfer learning and GAN". Mémoire de master. Béjaïa, Algérie: Université A. Mira-Béjaïa, 2023.

[4] Samira Abnar and Willem Zuidema. "Quantifying Attention Flow in Transformers". In: *ACL*. 2020.

[5] A. Adadi and M. Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6 (2018), pp. 52138–52160.

[6] J. Aechtner et al. "Comparing User Perception of Explanations Developed with XAI Methods". In: *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. Padua, Italy: IEEE, 2022, pp. 1–7.

[7] R. Ahmed and A.S. Imran. "Knee Osteoarthritis Analysis Using Deep Learning and XAI on X-Rays". In: *IEEE Access* 12 (2024), pp. 68870–68879.

[8] Kazi Ahnaf Alavee et al. "Enhancing Early Detection of Diabetic Retinopathy Through the Integration of Deep Learning Models and Explainable Artificial Intelligence". In: *IEEE Access* 12 (2024), pp. 73950–73964. DOI: 10.1109/ACCESS.2024.3405570.

[9] V. Ashwath, S. O. K., and R. Benitez. "TS-CNN: A Three-Tier Self-Interpretable CNN for Multi-Region Medical Image Classification". In: *IEEE Access* (2023).

[10] J.L. Ba, J.R. Kiros, and G.E. Hinton. *Layer Normalization*. https://arxiv.org/abs/1607.06450. arXiv:1607.06450. 2016.

[11] H. Balaha et al. "Advancing Eye Disease Detection: A Comprehensive Study on Computer-Aided Diagnosis with Vision Transformers and SHAP Explainability Techniques". In: *Biocybernetics and Biomedical Engineering* 45 (2025), pp. 23–33.

[12] S. E. Baouz. "Diagnostique automatique de la rétinopathie diabétique en exploitant les architectures basées sur le transfer learning". Mémoire de master. Université A. Mira-Béjaïa, 2024.

[13] Shaik Sulaiman Basha et al. "Impact of Fully Connected Layers on Performance of Convolutional Neural Networks for Image Classification". In: *Neurocomputing* 378 (2019).

[14] M. Belghachi. "A Review on Explainable Artificial Intelligence Methods, Applications, and Challenges". In: *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)* (2023), pp. 1007–1024.

[15] Mohamed Benali. *Mémoire de Master sur les réseaux de neurones*. https://bucket.theses-algerie.com/files/repositories-dz/1490651004040837.pdf. Université Mohamed Larbi Ben M'hidi - Oum El bouaghi, consulté le 12 mai 2025. 2021.

[16] O. Benchekroun et al. *The Need for Standardized Explainability*. 2020. arXiv: arXiv:2009.09288.

[17] Umang Bhatt, Adrian Weller, and José M. F. Moura. "Evaluating and Aggregating Feature-based Model Explanations". In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*. 2020, pp. 3016–3022. DOI: 10.24963/ijcai.2020/417. URL: https://doi.org/10.24963/ijcai.2020/417.

[18] D. Bhulakshmi and D. S. Rajput. "A systematic review on diabetic retinopathy detection and classification based on deep learning techniques using fundus images". In: *PeerJ Computer Science* (2024). Accessed: 2025-04-09.

[19] Centre d'Ophtalmologie Jean Jaurès. *Échographie oculaire à Toulouse – Plateau technique*. https://www.centreophtalmologiejeanjaures.fr/centre-ophtalmologie-jean-jaures-toulouse/plateau-technique-ophtalmologie/echographie-oculaire-toulouse.html. Accessed: 2025-04-09. 2025.

[20] Centre Neuchâtelois de Vision et d'Ophtalmologie (CNVO). *Les examens ophtalmologiques*. https://www.cnvo.ch/les-examens-ophtalmologiques/. Accessed: 2025-04-09. 2025.

[21] Centre Ophtalmologique Spécialisé de Sourdun (COSS). *Angiographie – Plateau technique*. https://www.coss-ophtalmologie.paris/le-centre/plateau-technique/angiographie/. Accessed: 2025-04-09. 2025.

[22] M.-A. Clinciu and H. Hastie. "A Survey of Explainable AI Terminology". In: *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*. Tokyo, Japan: Association for Computational Linguistics, 2019, pp. 8–13.

[23] N. Dhanachandra, K. Manglem, and Y.J. Chanu. "Image Segmentation Using K-means Clustering Algorithm and Subtractive Clustering Algorithm". In: *Procedia Comput. Sci.* 54 (2015), pp. 764–771.

[24] John Doe. *Exploring Explainable Artificial Intelligence Technologies: Approaches, Challenges, and Applications*. https://www.researchgate.net/publication/381293565_Exploring_Explainable_Artificial_Intelligence_Technologies_Approaches_Challenges_and_Applications. Accessed: 2025-05-18. 2024.

[25] A. Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. https://arxiv.org/abs/2010.11929. arXiv:2010.11929. 2021.

[26] Groupe Elsan. *Rétinopathie diabétique*. https://www.elsan.care/fr/pathologie-et-traitement/maladies-des-yeux/retinopathie-diabetique. Consulté le 8 avril 2025. 2025.

[27] EnjoyAlgorithms. *Supervised, Unsupervised and Semi-supervised Learning [Electronic resource]*. https://www.enjoyalgorithms.com/blogs/supervised-unsupervised-and-semisupervised-learning. Accessed: 2025-05-19. 2025.

[28] P. Fischer-Ghanassia, É. Ghanassia, and M.-C. Baraut. *Endocrinologie, diabétologie, nutrition*. 9e. Paris: Éditions Vernazobres-Grego, 2017.

[29] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. 3rd. O'Reilly Media, 2022. ISBN: 9781098125974.

[30] Glaucome.Tech. *La tomographie en cohérence optique (OCT)*. https://glaucome.tech/index.php/la-tomographie-en-coherence-optique/. Accessed: 2025-04-09. 2025.

[31] Riccardo Guidotti et al. "A survey of methods for explaining black box models". In: *ACM Computing Surveys (CSUR)* 51.5 (2018), pp. 1–42.

[32] V. Gulshan et al. "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs". In: *JAMA* 316.22 (2016), p. 2402.

[33] S.U. Hamida et al. "Exploring the Landscape of Explainable Artificial Intelligence (XAI): A Systematic Review of Techniques and Applications". In: *Big Data and Cognitive Computing* 8.11 (2024), p. 149.

[34] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. Springer Science & Business Media, 2009. ISBN: 9780387848570.

[35] María Herrero-Tudela et al. "An explainable deep-learning model reveals clinical clues in diabetic retinopathy through SHAP". In: *Biomedical Signal Processing and Control* 102 (2025), p. 107328. DOI: 10.1016/j.bspc.2024.107328.

[36] Adam Hoover and Michael Goldbaum. *Automated localization of the optic disc, fovea and retinal blood vessels from digital color fundus images*. https://www.researchgate.net/publication/12885182_Automated_localization_of_the_optic_disc_fovea_and_retinal_blood_vessels_from_digital_color_fundus_images. Consulté le 12 mai 2025. 2001.

[37] National Eye Institute. *Diabetic Retinopathy*. https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/diabetic-retinopathy. Consulté le 8 avril 2025. 2025.

[38] Centre d'Ophtalmologie Jean Jaurès. *Diagnostic et traitement de la rétinopathie diabétique à Toulouse*. https://www.centreophtalmologiejeanjaures.fr/pathologies-ophtalmologiques/pathologies-retine-et-vitre/traitement-retinopathie-diabetique-toulouse.html. Consulté le 8 avril 2025. 2025.

[39] D.P. Kingma et al. *Semi-Supervised Learning with Deep Generative Models*. https://arxiv.org/abs/1406.5298. arXiv:1406.5298. 2014.

[40] Vijay R Konda and John N Tsitsiklis. "Actor-critic algorithms". In: *Advances in neural information processing systems*. 2000, pp. 1008–1014.

[41] K.B. Lebaka and S.P. Sithungu. "Anomaly Detection in Retinal Imagery Using Variational Autoencoders". In: *Proceedings of the 2024 7th International Conference on Computational Intelligence and Intelligent Systems*. Nagoya, Japan: ACM, 2024, pp. 53–58.

[42] C.-I. Lin et al. "Nanopodia - Thin, Fragile Membrane Projections with Roles in Cell Movement and Intercellular Interactions". In: *J. Vis. Exp. JoVE* 86 (2014), p. 51320.

[43] Pedro Lopes et al. "XAI Systems Evaluation: A Review of Human and Computer-Centred Methods". In: *Applied Sciences* 12.19 (2022), p. 9423. DOI: 10.3390/app12199423.

[44] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[45] R. Luss and A. Dhurandhar. *When Stability meets Sufficiency: Informative Explanations that do not Overwhelm*.

[46] S.M. Mahim et al. “Unlocking the Potential of XAI for Improved Alzheimer’s Disease Detection and Classification Using a ViT-GRU Model”. In: *IEEE Access* 12 (2024), pp. 8390–8412.

[47] M. Malafaia et al. “Robustness Analysis of Deep Learning-Based Lung Cancer Classification Using Explainable Methods”. In: *IEEE Access* 10 (2022), pp. 112731–112741.

[48] Melkamu Abay Mersha et al. “A Unified Framework with Novel Metrics for Evaluating the Effectiveness of XAI Techniques in LLMs”. In: *arXiv preprint arXiv:2503.05050* (Mar. 2025). URL: https://arxiv.org/abs/2503.05050.

[49] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al. “Human-level control through deep reinforcement learning”. In: *Nature* 518.7540 (2015), pp. 529–533.

[50] G. Montavon, W. Samek, and K.-R. Müller. “Methods for interpreting and understanding deep neural networks”. In: *Digital Signal Processing* 73 (2018), pp. 1–15.

[51] M.W. Nadeem et al. “Deep Learning for Diabetic Retinopathy Analysis: A Review, Research Challenges, and Future Directions”. In: *Sensors* 22.18 (2022), p. 6780.

[52] Saeed Nosratabadi et al. “Data Science in Economics: Comprehensive Review of Advanced Machine Learning and Deep Learning Methods”. In: *Mathematics* 8.10 (2020). DOI: 10.3390/math8101799. URL: https://www.mdpi.com/2227-7390/8/10/1799.

[53] A. Osa Sanchez et al. “Explainable AI-Based Approach for Age-Related Macular Degeneration (AMD) Detection via Fundus Imaging”. In: *IEEE Access* (2024).

[54] Vitali Petsiuk, Abir Das, and Kate Saenko. “RISE: Randomized Input Sampling for Explanation of Black-box Models”. In: *arXiv preprint arXiv:1806.07421* (2018). URL: https://arxiv.org/abs/1806.07421.

[55] M. Radensky et al. *Exploring The Role of Local and Global Explanations in Recommender Systems*. 2021. arXiv: 2109.13301.

[56] M. Radhakrishnan et al. “Advancing Ovarian Cancer Diagnosis Through Deep Learning and eXplainable AI: A Multiclassification Approach”. In: *IEEE Access* (2024).

[57] N. Rane, S. Choudhary, and J. Rane. “Explainable Artificial Intelligence (XAI) Approaches for Transparency and Accountability in Financial Decision-Making”. In: *SSRN Electronic Journal* (2023). DOI: 10.2139/ssrn.4390847.

[58] Shamim Rahim Refat et al. “VR-FuseNet: A Fusion of Heterogeneous Fundus Data and Explainable Deep Network for Diabetic Retinopathy Classification”. In: *arXiv preprint arXiv:2504.21464* (2025). URL: https://arxiv.org/abs/2504.21464.

[59] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “"Why should I trust you?": Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144.

[60] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 4th. Pearson, 2020. ISBN: 978-0134610993.

[61] R. Saleem et al. “Explaining deep neural networks: A survey on the global interpretation methods”. In: *Neurocomputing* 513 (2022), pp. 165–180.

[62] Mahalakshmi Sampath and Mohammad Akram Khan. “EfficientViT: A Hybrid CNN-Transformer Framework With Cross-Attention Fusion For Clinically Interpretable Diabetic Retinopathy Grading”. In: *International Journal of Creative Research Thoughts (IJCRT)* 13.4 (2025). DOI: 10.1729/Journal.44896. URL: https://www.ijcrt.org/papers/IJCRT2504630.pdf.

[63] Ramprasaath R Selvaraju et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision (ICCV)*. 2017, pp. 618–626. URL: https://doi.org/10.1109/ICCV.2017.74.

[64] Tariq Shahzad et al. “Developing a Transparent Diagnosis Model for Diabetic Retinopathy Using Explainable AI”. In: *IEEE Access* 12 (2024), pp. 149700–149709. DOI: 10.1109/ACCESS.2024.3475550.

[65] R.-K. Sheu et al. “Interpretable Classification of Pneumonia Infection using eXplainable AI (XAI-ICP)”. In: *IEEE Access* (2023).

[66] A. Sivaprasad et al. “Evaluation of Human-Understandability of Global Model Explanations Using Decision Tree”. In: *Artificial Intelligence. ECAI 2023 International Workshops*. Ed. by S. Nowaczyk et al. Vol. 1947. Cham: Springer Nature Switzerland, 2024, pp. 43–65.

[67] Jane Smith and John Doe. “A Survey on Explainable AI: Techniques and Challenges”. In: *International Journal of Innovations in Engineering Research and Technology* (2023).

[68] Dane Sohier and Maggie Karthik. *APTOS 2019 Blindness Detection*. Last accessed on June 6, 2023. 2019. URL: https://www.kaggle.com/competitions/aptos2019-blindness-detection.

[69] K. Sokol and P. Flach. *Explainability Is in the Mind of the Beholder: Establishing the Foundations of Explainable Artificial Intelligence*. 2022. arXiv: 2112.14466 [cs.AI].

[70] Carnegie Mellon University. *CMU-CALD-02-107*. https://www.cs.cmu.edu/~zhuxj/pub/CMU-CALD-02-107.pdf. Consulté le 12 mai 2025. 2002.

[71] D. Varam et al. “Wireless Capsule Endoscopy Image Classification: An Explainable AI Approach”. In: *IEEE Access* 11 (2023), pp. 105262–105280.

[72] Hemanth Kumar Vasireddi, K. Suganya Devi, and G. N. V. Raja Reddy. “DR-XAI: Explainable Deep Learning Model for Accurate Diabetic Retinopathy Severity Assessment”. In: *Arabian Journal for Science and Engineering* 49 (2024), pp. 12899–12917. DOI: 10.1007/s13369-024-08836-7.

[73] A. Vaswani et al. *Attention Is All You Need*. https://arxiv.org/abs/1706.03762. arXiv:1706.03762. 2023.

[74] M. Velmurugan et al. *Developing a Fidelity Evaluation Approach for Interpretable Machine Learning*. 2021. arXiv: 2106.08492.

[75] Visiopôle du Beaujolais. *Rétinopathie diabétique | Visiopôle du Beaujolais | Villefranche-sur-Saône [Electronic resource]*. https://www.visiopoledubeaujolais.com/retine/retinopathie-diabetique/. Accessed: 2025-04-09. 2025.

[76] Haofan Wang et al. “Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks”. In: *arXiv preprint arXiv:1910.01279* (2020).

[77] X. Wang and M. Yin. “Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making”. In: *26th International Conference on Intelligent User Interfaces*. College Station, TX, USA: ACM, 2021, pp. 318–328.

[78] Christopher JCH Watkins and Peter Dayan. “Q-learning”. In: *Machine learning* 8.3-4 (1992), pp. 279–292.

[79] C.P. Wilkinson, F.L. Ferris III, R.E. Klein, et al. “Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales”. In: *Ophthalmology* (2003).

[80] Ronald J Williams. “Simple statistical gradient-following algorithms for connectionist reinforcement learning”. In: *Machine learning* 8.3-4 (1992), pp. 229–256.

[81] M.D. Woodbright et al. “Toward Transparent AI for Neurological Disorders: A Feature Extraction and Relevance Analysis Framework”. In: *IEEE Access* 12 (2024), pp. 37731–37743.

[82] World Health Organization. *Diabetic retinopathy screening: a short guide. Increase effectiveness, maximize benefits and minimize harm*. https://iris.who.int/bitstream/handle/10665/336660/9789289055321-eng.pdf. Accessed: 2025-05-19.

[83] Rasiklal B. Yadav. “The Ethics of Understanding: Exploring Moral Implications of Explainable AI”. In: *International Journal of Science and Research (IJSR)* 13.6 (2024), pp. 1–7.

[84] Weijie Zhang, Veronika Belcheva, and Tatiana Ermakova. “Interpretable Deep Learning for Diabetic Retinopathy: A Comparative Study of CNN, ViT, and Hybrid Architectures”. In: *Computers* 14 (2025). DOI: 10.3390/computers14050187. URL: https://www.mdpi.com/2073-431X/14/5/187.

[85] X. Zheng et al. *F-Fidelity: A Robust Framework for Faithfulness Evaluation of Explainable AI*. 2025. arXiv: 2410.02970.

[86] L. Zou et al. “Ensemble Image Explainable AI (XAI) Algorithm for Severe Community-Acquired Pneumonia and COVID-19 Respiratory Infections”. In: *IEEE Transactions on Artificial Intelligence* (2022).

ABSTRACT

This thesis demonstrates how explainable AI (XAI) can make deep learning models used for diabetic retinopathy screening more transparent. Three fine-tuned models AtR5C (CNN-based), ViR-5C (Vision Transformer), and ReVi-5C (hybrid) are evaluated on the APTOS dataset. A set of widely used XAI methods (Grad-CAM, Score-CAM, LIME, SHAP, and Attention Rollout) are compared using four established metrics: insertion, deletion, faithfulness, and robustness. The results show that no single method is perfect; rather, they are complementary. Simple recommendations are proposed to help clinicians choose the appropriate explanation and assess its reliability a crucial requirement for the responsible use of AI in ophthalmology.

Keywords: Diabetic Retinopathy, Explainable AI, Convolutional Neural Network, Vision Transformer, Hybrid Model, APTOS Dataset, Deep Learning.

RÉSUMÉ

Ce mémoire montre comment l'explicabilité (XAI) peut rendre les réseaux profonds utilisés pour le dépistage de la rétinopathie diabétique plus transparents. Trois modèles fine-tunés AtR5C (CNN), ViR-5C (Vision Transformer) et ReVi-5C (hybride) sont testés sur la base APTOS. Un ensemble de méthodes XAI répandues (Grad-CAM, Score-CAM, LIME, SHAP et Attention Rollout) sont comparées à l'aide de quatre métriques reconnues : insertion, deletion, faithfulness et robustness. Les résultats montrent qu'aucune méthode n'est parfaite ; elles se complètent. Des recommandations simples aident les cliniciens à choisir la bonne explication et à juger sa fiabilité, condition essentielle pour un usage responsable de l'IA en ophtalmologie.

Mots-clés : Rétinopathie Diabétique, Intelligence Artificielle Explicable, Réseau de Neurones Convolutifs, Vision Transformer, Modèle Hybride, Base APTOS, Apprentissage Profond.