



République Algérienne Démocratique Populaire
Ministère de l'Enseignement Supérieur et de la
Recherche Scientifique
Université Abou Bekr Belkaïd de Tlemcen –
Faculté de Technologie
Département de Génie Biomédical

Mémoire

présenté en vue de l'obtention du diplôme de
Master en Génie Biomédical
Spécialité : Informatique Médicale et Hospitalière

Présenté par : Hasna Mrabent

Modèles de Fondation Auto-Supervisés pour la Prédiction du Risque Cardiovasculaire à partir d'Images Rétiniennes

Soutenu publiquement, le 22/06/2026, devant le jury composé de :

M. TALEB Tarik	Univ. Tlemcen	Président
Mme HAMZA CHERIF Souad	Univ. Tlemcen	Encadrante
M. LAZOUNI Amine	Univ. Tlemcen	Co-encadrant
M. EL HABIB DAHO Mostafa	UBO – LaTIM, Brest	Co-encadrant
Mme BELAIDI Asma	Univ. Tlemcen	Examinatrice

Organisme d'accueil : Université de Bretagne Occidentale – UBO
Année universitaire 2025–2026

Dédicace

À mes chers parents,

*Pour votre amour sans limite, vos sacrifices silencieux
et votre confiance absolue en moi.*

*Ce travail est aussi le vôtre, reflet de tout
ce que vous avez semé avec tant de cœur.*

*À ma petite sœur Lina et mes frères Karim et Islem,
qui m'ont accompagnée même de loin.*

À mes encadrants,

*M. Mostafa El Habib Daho, Mme Hamza Cherif Souad
et M. Lazouni Amine,*

Vous m'avez ouvert une porte que je n'oublierai jamais.

*Grâce à vous, j'ai appris à travailler, à chercher
et surtout, à croire en ce que je fais.*

Ce travail vous doit énormément.

À toute l'équipe du LaTIM,

Pour l'accueil chaleureux et votre bienveillance.

À mon amie Meryem,

*Tu m'as guidée, soutenue, et tu es devenue
bien plus qu'une amie , tu es une sœur.*

Remerciements

Avant toute chose, je rends grâce à Dieu Tout-Puissant pour m'avoir accordé la santé, la force et la persévérance nécessaires à l'accomplissement de ce travail.

Ce mémoire est le résultat d'un parcours fait de rencontres, d'apprentissages et de belles opportunités. Je tiens à exprimer ma profonde gratitude à toutes les personnes qui ont contribué, de près ou de loin, à sa réalisation.

Je souhaite adresser mes remerciements les plus sincères à mes encadrants, **Mme Hamza Cherif Souad**, **M. Lazouni Amine** et **M. Mostafa El Habib Daho**, pour leur accompagnement tout au long de ce projet.

À **Mme Hamza Cherif Souad**, j'exprime une reconnaissance particulière pour la confiance qu'elle m'a accordée et pour avoir cru en mon potentiel dès le début. Son engagement, son implication et son soutien constant ont été essentiels pour obtenir cette opportunité de stage et pour mon parcours. Je lui suis profondément reconnaissante pour sa bienveillance et pour avoir ouvert des portes importantes pour mon avenir.

À **M. Lazouni Amine**, je tiens à adresser mes sincères remerciements pour son aide précieuse et concrète tout au long de cette expérience. Son soutien, sa disponibilité et son accompagnement dans les étapes importantes de ce projet ont beaucoup compté pour sa réussite. Je lui suis également reconnaissante pour la confiance qu'il m'a accordée.

À **M. Mostafa El Habib Daho**, je remercie chaleureusement pour son accueil et sa grande bienveillance. Sa gentillesse, son accompagnement et ses conseils ont contribué au bon déroulement de ce travail. Je lui suis également reconnaissante pour l'opportunité qui m'a été donnée et pour son soutien tout au long du projet.

Je remercie également l'ensemble de l'équipe du **LaTIM** pour son accueil, sa disponibilité et la qualité des échanges qui ont rendu cette expérience très enrichissante.

J'adresse mes sincères remerciements aux **membres du jury** pour le temps qu'ils consacrent à l'évaluation de ce travail.

Une pensée toute particulière va à **Meryem**, une amie précieuse qui a été présente tout au long de cette aventure. Sa présence et son soutien.

Enfin, j'exprime ma profonde gratitude à ma famille, en particulier à **mes parents** ainsi qu'à **Lina**, **Karim** et **Islem**, pour leur amour, leur confiance et leur soutien sans faille. Leur présence a été ma plus grande force tout au long de ce parcours.

Résumé

Les maladies cardiovasculaires constituent la première cause de mortalité dans le monde, avec environ 17,9 millions de décès par an. La rétine, seul tissu permettant l’observation non invasive de la microcirculation systémique, offre ainsi une fenêtre privilégiée pour une évaluation précoce du risque cardiovasculaire.

Ce mémoire propose une approche complète, fondée sur l’apprentissage auto-supervisé, pour la prédiction du risque cardiovasculaire à partir d’images rétinienne multi-modalités (*OCTA* : Plexelite et Cirrus ; *CFP* : Clarus et Optos).

Nous construisons six modèles de fondation **VA-MAE** (*Vessel-Aware Masked Autoencoder*), entraînés séparément par appareil sur 83 243 images. L’architecture **VA-MAE** étend le **MAE** standard via un décodeur à trois têtes supervisant l’intensité, la *vesselness* et le squelette vasculaire, couplé à un masquage *vessel-aware* et à un curriculum de masquage progressif. Les images *OCTA* sont représentées en RGB multi-plexus (superficiel / profond / full retina). En complément, 30 modèles de diffusion conditionnels (SD 1.5, SD 2.0, SDXL) sont affinés par **LoRA** pour générer des images synthétiques conditionnées par des métadonnées cliniques ; SD 1.5 obtient les meilleures performances (FID 147.2, Inception Score 1.37).

Les modèles sont évalués sur deux benchmarks cardiovasculaires. Sur la cohorte **RASTA** (491 patients, score CHA2DS2-VASc), une fusion clinique par *cross-attention* atteint une AUC de 0.968 et une précision balancée de 87.5% en classification à trois classes, surpassant la baseline supervisée EfficientNetV2-B3 (AUC 0.868, BAcc 63.3%). Sur le challenge international **VAMPIRE OCTA-CVD** (MICCAI 2025 ; 1 659 patients ; 12 équipes), le backbone MAE-PLEX6×6 se classe 4^e avec une macro AUC de 0.801 ± 0.008 .

Ces résultats montrent que le pré-entraînement auto-supervisé spécifique à la modalité et à l’appareil, combiné à une fusion clinique multimodale, constitue une approche robuste et généralisable pour la stratification du risque cardiovasculaire à partir d’images rétinienne.

Mots-clés : Apprentissage auto-supervisé, MAE, ViT, imagerie rétinienne, OCTA, CFP, risque cardiovasculaire, modèles de fondation, modèles de diffusion, LoRA, fusion multimodale, CHA2DS2-VASc.

Abstract

Cardiovascular diseases are the leading cause of death worldwide, accounting for approximately 17.9 million deaths per year. The retina, as the only tissue allowing non-invasive observation of systemic microcirculation, offers a privileged window for early cardiovascular risk assessment.

This thesis proposes a comprehensive approach based on self-supervised learning for cardiovascular risk prediction from multi-modal retinal images (*OCTA* : Plexelite and Cirrus ; *CFP* : Clarus and Optos).

We build six **VA-MAE** (*Vessel-Aware Masked Autoencoder*) foundation models, trained separately per device on 83 243 images. The **VA-MAE** architecture extends the standard **MAE** with a three-head decoder supervising intensity, vesselness and vascular skeleton, combined with vessel-aware masking and a progressive masking curriculum. OCTA images are represented as multi-plexus RGB inputs (superficial / deep / full retina). In addition, 30 conditional diffusion models (SD 1.5, SD 2.0, SDXL) are fine-tuned via **LoRA** to generate synthetic images conditioned on clinical metadata ; SD 1.5 achieves the best performance (FID 147.2, Inception Score 1.37).

The models are evaluated on two cardiovascular benchmarks. On the **RASTA** cohort (491 patients, CHA2DS2-VASc score), cross-attention clinical fusion achieves an AUC of 0.968 and a balanced accuracy of 87.5% on the three-class task, outperforming the supervised EfficientNetV2-B3 baseline (AUC 0.868, BAcc 63.3%). On the international **VAMPIRE OCTA-CVD** challenge (MICCAI 2025 ; 1 659 patients ; 12 teams), the MAE-PLEX6×6 backbone ranks 4th with a macro AUC of 0.801 ± 0.008 .

These results demonstrate that modality- and device-specific self-supervised pretraining, combined with multimodal clinical fusion, provides a robust and generalizable approach for cardiovascular risk stratification from retinal images.

Keywords : Self-supervised learning, MAE, ViT, retinal imaging, OCTA, CFP, cardiovascular risk, foundation models, diffusion models, LoRA, multimodal fusion, CHA2DS2-VASc.

Table des matières

Résumé	4
Abstract	5
Table des figures	10
Liste des tableaux	11
Liste des acronymes	12
Introduction générale	13
I État de l'Art	16
1 Contexte médical et technologies d'imagerie	17
1.1 Introduction	17
1.2 Pathologies rétinienne et risque cardiovasculaire	17
1.2.1 La rétine comme fenêtre sur le système vasculaire	17
1.2.2 Risque cardiovasculaire : définition et enjeux	18
1.2.3 Liens entre altérations rétinienne et maladies cardiovasculaires	18
1.3 Technologies d'imagerie ophtalmologique	19
1.3.1 OCT-Angiographie (OCTA)	19
1.3.2 Photographie du fond d'oeil couleur (CFP)	21
1.3.3 Comparaison des modalités	23
1.4 Conclusion	24
2 Apprentissage auto-supervisé pour l'imagerie médicale	25
2.1 Introduction	25
2.2 Principes de l'apprentissage auto-supervisé	25
2.2.1 Définition et motivation	25
2.2.2 Taxonomie des méthodes SSL	26
2.3 Masked Autoencoders (MAE)	27

2.3.1	Architecture : encodeur ViT et décodeur léger	27
2.3.2	Stratégie de masquage	27
2.3.3	Fonction de reconstruction	28
2.3.4	Propriétés des représentations	28
2.4	Vision Transformers (ViT)	28
2.4.1	Architecture générale	28
2.4.2	ViT-B/16 : configuration	29
2.4.3	Avantages pour l'imagerie médicale	30
2.5	Modèles de fondation en imagerie médicale	30
2.5.1	Définition et intérêt	30
2.5.2	Enjeux du pré-entraînement spécifique par machine	31
2.6	Conclusion	31
3	Modèles de Diffusion pour la Génération d'Images Médicales	32
3.1	Introduction	32
3.2	Principes des modèles de diffusion	32
3.2.1	Processus de diffusion avant (<i>forward process</i>)	32
3.2.2	Processus de débruitage inverse (<i>reverse process</i>)	33
3.2.3	<i>Denoising Diffusion Probabilistic Models</i> (DDPM)	33
3.3	Fine-tuning efficace avec LoRA	34
3.3.1	Principe	34
3.3.2	Avantages computationnels	35
3.3.3	Application aux modèles de diffusion	35
3.4	Applications en imagerie rétinienne	35
3.4.1	Augmentation de données	35
3.4.2	Synthèse d'images pathologiques	36
3.5	Conclusion	36
4	Prédiction du Risque Cardiovasculaire par Deep Learning	38
4.1	Introduction	38
4.2	Revue des approches existantes	38
4.2.1	Approches unimodales OCTA	38
4.2.2	Approches multimodales	38
4.2.3	Transfer learning et SSL	39
4.3	Datasets pour la prédiction du risque cardiovasculaire	39
4.3.1	Dataset RASTA	39
4.3.2	Dataset VAMPIRE OCTA-CVD	40
4.3.3	Comparaison et complémentarité	41
4.4	Métriques d'évaluation	42
4.4.1	AUC-ROC, BAcc et QWK	42

4.4.2	Gestion du déséquilibre de classes	42
4.5	Conclusion	43
5	Synthèse de la Littérature sur les Modèles de Fondation SSL et la Prédiction du Risque Cardiovasculaire	44
5.1	Introduction	44
5.2	Comparaison des approches de modèles de fondation SSL en ophtalmologie et prédiction du risque cardiovasculaire	44
5.3	Limites de la littérature	45
5.4	Conclusion	46
II	Contributions	47
6	Construction des Modèles de Fondation SSL	48
6.1	Introduction	48
6.2	Vue d'ensemble des modèles construits	48
6.3	Données d'entraînement	49
6.3.1	Corpus OCTA Plexelite	49
6.3.2	Corpus OCTA Cirrus	49
6.3.3	Corpus CFP grand champ	49
6.3.4	Prétraitements	49
6.4	Architecture VA-MAE	51
6.4.1	Encodeur ViT	51
6.4.2	Décodeur à trois têtes	51
6.4.3	Masquage vessel-aware	52
6.4.4	Curriculum de masquage progressif	53
6.4.5	Protocole d'optimisation	53
6.5	Résultats du pré-entraînement	53
6.5.1	Convergence et perte de reconstruction	53
6.5.2	Visualisations par attention rollout	54
6.5.3	Impact du curriculum de masquage progressif	54
6.5.4	Modèles CFP grand champ	55
6.6	Modèles de diffusion conditionnels pour la génération d'images rétinienne	55
6.6.1	Framework et approche	55
6.6.2	Organisation des modèles entraînés	55
6.6.3	Préparation des données et structure des prompts	56
6.6.4	Résultats qualitatifs	57
6.7	Conclusion	58

7	Prédiction du Risque Cardiovasculaire : Expérimentations	59
7.1	Introduction	59
7.2	Protocole expérimental	59
7.2.1	Stratégie de validation	59
7.2.2	Stratégie de fine-tuning à partir des modèles de fondation OCTA	60
7.2.3	Architecture de classification	61
7.2.4	Gestion du déséquilibre de classes	62
7.3	Baselines de comparaison	62
7.3.1	Baselines cliniques : SVM et XGBoost	62
7.3.2	Baseline supervisée : EfficientNetV2-B3	63
7.3.3	Courbes ROC des baselines cliniques	63
7.4	Résultats sur RASTA	65
7.4.1	Performances globales	65
7.4.2	Comparaison : pré-entraînement Plexelite vs Cirrus	66
7.4.3	Comparaison SSL vs EfficientNetV2-B3	66
7.5	Résultats sur VAMPIRE OCTA-CVD	67
7.5.1	Architecture de fine-tuning OCTA-CVD	67
7.5.2	Performances globales	67
7.5.3	Comparaison des modèles	67
7.5.4	Analyse des erreurs par sous-tâche	68
7.6	Discussion transversale	69
7.6.1	Apport du pré-entraînement spécifique à la machine	69
7.6.2	Transfert de connaissance OCTA vers le risque cardiovasculaire	69
7.6.3	Limites et biais identifiés	70
7.7	Interprétabilité du modèle	70
7.7.1	Introduction	70
7.7.2	Attention rollout sur les images OCTA	71
7.7.3	Conclusion	72
7.8	Conclusion	73
	Conclusion générale	74
	A Détails d'implémentation	76
	B Résultats complémentaires	79

Table des figures

1.1	Image OCTA — EviRED	20
1.2	Image OCTA — Cirrus 6000	21
1.3	Image CFP — Optos	22
1.4	Image CFP — Clarus 700	23
1.5	Champs d’acquisition — comparaison	24
2.1	Architecture MAE	27
2.2	Architecture Vision Transformer (ViT)	29
3.1	Processus DDPM — chaîne de Markov	33
3.2	Principe de LoRA	34
4.1	Structure du dataset RASTA	41
6.1	Rehaussement de contraste (CLAHE)	50
6.2	Ajustement colorimétrique — Clarus	50
6.3	Schéma VA-MAE — OCTA	52
6.4	Pipeline VA-MAE	52
6.5	Images synthétiques CFP — Clarus	57
7.1	Validation croisée RASTA (5-fold)	60
7.2	Pipeline de fine-tuning — RASTA	61
7.3	Courbes ROC — tâche binaire	64
7.4	Courbes ROC — tâche à trois classes	65
7.5	Cartes d’attention rollout — tâche binaire	71
7.6	Cartes d’attention rollout — 3 classes	72
B.1	Matrice de confusion — MAE-CIRRUS6×6	79
B.2	Matrice de confusion — MAE-PLEX6×6	80
B.3	Incertitude du modèle — RASTA	82
B.4	Comparaison image réelle vs image synthétique — full retina (SDXL + LoRA)	83
B.5	Comparaison image réelle vs image synthétique — plexus profond (SD 2.0 + LoRA)	83

Liste des tableaux

1.1	Modalités d'imagerie — synthèse	23
2.1	Paramètres ViT-B/16	29
4.1	Variables cliniques — RASTA	40
4.2	Score CHA ₂ DS ₂ -VASc	40
5.1	Modèles de fondation SSL — comparaison	45
6.1	Modèles de fondation — vue d'ensemble	48
6.2	Curriculum de masquage	53
6.3	Perte de reconstruction finale après 200 époques de pré-entraînement.	54
6.4	Modèles de diffusion — organisation	56
6.5	<code>metadata.csv</code> — champs	56
6.6	Métriques de qualité générative (FID et Inception Score) pour les trois architectures de diffusion fine-tunées par LoRA sur le corpus Clarus 700.	58
7.1	Hyperparamètres d'entraînement	62
7.2	Baselines cliniques — performances	63
7.3	Performances RASTA — 3 classes	65
7.4	Performances RASTA — binaire	66
7.5	Ablation des canaux OCTA	66
7.6	Classement VAMPIRE OCTA-CVD	68
7.7	AUC par sous-tâche VAMPIRE OCTA-CVD	68
A.1	Bibliothèques Python	76
A.2	Paramètres d'entraînement VA-MAE	77
A.3	Hyperparamètres LoRA	77
A.4	Hyperparamètres fine-tuning VAMPIRE	78
B.1	Rappel par classe — MAE-PLEX6×6	80
B.2	Curriculum de masquage — ablation	81
B.3	AUC macro par fold — VAMPIRE	81

Liste des acronymes

SSL	Self-Supervised Learning — Apprentissage Auto-Supervisé
MAE	Masked Autoencoder — Autoencodeur Masqué
VA-MAE	Vessel-Aware Masked Autoencoder
OCTA	Optical Coherence Tomography Angiography — Angiographie par Tomographie en Cohérence Optique
CFP	Colour Fundus Photography — Photographie du Fond d’Œil Couleur
CVD	Cardiovascular Disease — Maladie Cardiovasculaire
ViT	Vision Transformer
DDPM	Denoising Diffusion Probabilistic Model
AUC	Area Under the Curve — Aire sous la Courbe ROC
ROC	Receiver Operating Characteristic
QWK	Quadratic Weighted Kappa — Coefficient Kappa Pondéré Quadratique
BAcc	Balanced Accuracy — Précision Balancée
LoRA	Low-Rank Adaptation
SWA	Stochastic Weight Averaging
LLRD	Layer-wise Learning Rate Decay
FID	Fréchet Inception Distance
SSIM	Structural Similarity Index Measure
MSE	Mean Squared Error — Erreur Quadratique Moyenne
BCE	Binary Cross-Entropy — Entropie Croisée Binaire
MLP	Multi-Layer Perceptron
MHSA	Multi-Head Self-Attention — Attention Multi-Têtes
RASTA	Retinal OCT-Angiography and cardiovascular STatus
CRAE	Central Retinal Arteriolar Equivalent
CRVE	Central Retinal Venular Equivalent

Introduction générale

Contexte général

Les maladies cardiovasculaires constituent la première cause de mortalité mondiale, responsables de 17,9 millions de décès annuels selon l'Organisation mondiale de la santé. L'hypertension artérielle, le diabète et l'athérosclérose partagent un substrat vasculaire commun dont les manifestations sont directement accessibles à travers le fond d'œil : la rétine est le seul tissu du corps humain où la microcirculation peut être observée de façon non invasive, sans injection de produit de contraste ni irradiation. Des modifications microvasculaires rétinienues, mesurables plusieurs années avant la première manifestation clinique, constituent des biomarqueurs précoces et répétables du risque cardiovasculaire.

Quatre modalités d'imagerie rétinienne sont exploitées dans ce travail. La **photographie du fond d'œil couleur** (*Colour Fundus Photography*, CFP) produite par le **Clarus 700** (Zeiss, 133°) offre un champ large en lumière polychromatique. L'**Optos California** étend ce champ à 200° en *ultra-wide field*. L'**OCT-angiographie** (*Optical Coherence Tomography Angiography*, OCTA) réalisée par le **Plexelite 9000** (Zeiss, champs 6×6 mm et 15×15 mm) fournit des cartes vasculaires de la perfusion capillaire organisées en plexus anatomiques distincts. Le **Cirrus HD-OCT 6000** (Zeiss, champs 3×3 mm et 6×6 mm) complète l'acquisition OCTA en routine clinique.

Problématique

L'exploitation de ces modalités pour la prédiction du risque cardiovasculaire se heurte à deux obstacles majeurs. La **rareté des annotations** constitue le premier : constituer des bases de données rétinienues labellisées pour le risque cardiovasculaire exige un suivi longitudinal structuré, disponible seulement dans quelques cohortes dédiées. L'**hétérogénéité instrumentale** constitue le second : les images produites par la Plexelite et la Cirrus présentent des résolutions, des profils de bruit et des contrastes fondamentalement différents, si bien qu'un modèle entraîné sur un mélange non contrôlé confond les variabilités instrumentales avec les signaux biologiques.

Ces deux contraintes rendent inadéquates les approches supervisées classiques. Une stratégie qui extrait des représentations utiles à partir de données non étiquetées, tout en

traitant séparément chaque équipement, est donc nécessaire.

Objectifs

Ce mémoire propose de répondre à ces deux difficultés par une approche en deux temps. Dans un premier temps, six **modèles de fondation auto-supervisés** sont construits, un par modalité et par machine, par apprentissage d'autoencodeur masqué. Pour les équipements OCTA, une extension originale est proposée : le **VA-MAE** (*Vessel-Aware Masked Autoencoder*) [2], qui intègre explicitement la structure vasculaire rétinienne dans la fonction de reconstruction via un décodeur à trois têtes parallèles (reconstruction d'intensité, carte de vesselness, carte de squelette) et un curriculum de masquage progressif. Dans un second temps, les représentations extraites sont adaptées par *fine-tuning* supervisé à deux benchmarks : la cohorte **RASTA** [4] (491 patients, classification en trois classes du score CHA₂DS₂-VASc) et le challenge **VAMPIRE MICCAI 2025** (1 659 patients, prédiction multi-label de cinq facteurs de risque, 12 équipes participantes) [39].

Organisation du mémoire

Le mémoire s'organise en deux parties. La Partie I couvre l'état de l'art en cinq chapitres ; la Partie II présente les contributions originales en deux chapitres.

Chapitre 1 : *Contexte médical et technologies d'imagerie* : liens physiologiques entre la microcirculation rétinienne et le risque cardiovasculaire systémique, description des quatre équipements d'imagerie (Clarus, Optos, Plexelite, Cirrus) et comparaison de leurs caractéristiques.

Chapitre 2 : *Apprentissage auto-supervisé pour l'imagerie médicale* : principes du SSL, familles de méthodes contrastives et génératives, architecture MAE et ViT-B/16, enjeux du pré-entraînement spécifique par machine.

Chapitre 3 : *Modèles de diffusion pour la génération d'images médicales* : processus de diffusion DDPM, fine-tuning efficace avec LoRA, applications à l'augmentation de données rétiniennes.

Chapitre 4 : *Prédiction du risque cardiovasculaire par deep learning* : revue des approches existantes, description des datasets RASTA et VAMPIRE OCTA-CVD, métriques d'évaluation adaptées au déséquilibre de classes.

Chapitre 5 : *Synthèse de la littérature* : tableau comparatif des travaux existants, identification des lacunes adressées par ce mémoire.

Chapitre 6 : *Construction des modèles de fondation SSL* : corpus d'entraînement, architecture VA-MAE, curriculum de masquage progressif, résultats du pré-entraînement, exploration diffusion sur EVIRED.

Chapitre 7 : *Prédiction du risque cardiovasculaire : expérimentations* : protocole de fine-tuning à trois phases, module de fusion clinique par cross-attention, résultats sur RASTA et sur le challenge VAMPIRE MICCAI 2025.

Première partie

État de l'Art

Chapitre 1

Contexte médical et technologies d'imagerie

1.1 Introduction

La rétine occupe une position singulière parmi les tissus de l'organisme : ses vaisseaux sont les seuls directement accessibles à l'observation optique non invasive, ce qui en fait un site privilégié pour évaluer l'état de la microcirculation systémique. Comprendre les liens quantifiés entre les altérations vasculaires rétiniennes et le risque cardiovasculaire est indispensable pour motiver l'utilisation de l'imagerie ophtalmologique à des fins de stratification du risque [23]. Ce chapitre décrit ces liens, puis présente les quatre équipements d'imagerie rétinienne exploités dans ce travail.

1.2 Pathologies rétiniennes et risque cardiovasculaire

1.2.1 La rétine comme fenêtre sur le système vasculaire

La **rétine** est un tissu neural hautement vascularisé, accessible de manière non invasive par imagerie ophtalmologique, dont la microcirculation reflète fidèlement l'état du système vasculaire systémique. Sa vascularisation repose sur deux systèmes distincts : la **circulation rétinienne**, issue de l'artère centrale de la rétine, branche de l'artère ophtalmique, irrigue les couches internes de la rétine ; la **circulation choroïdienne**, issue des artères ciliaires postérieures, irrigue les couches externes, notamment les photorécepteurs et l'épithélium pigmentaire.

La circulation rétinienne constitue un territoire microvasculaire directement observable, dont les altérations morphologiques, fonctionnelles et structurelles sont corrélées à des pathologies systémiques telles que l'hypertension artérielle, le diabète, les dyslipidémies et les maladies cardiovasculaires. Des cohortes de grande envergure, notamment la **UK Biobank** (plus de 500 000 participants) [10] et l'**ARIC Study** (*Atherosclerosis Risk in*

Communities) [41], ont établi des associations quantitatives entre les paramètres rétiens et le risque cardiovasculaire incident, conférant à la rétine le statut de **biomarqueur systémique** à fort potentiel prédictif.

1.2.2 Risque cardiovasculaire : définition et enjeux

Les **maladies cardiovasculaires** (*cardiovascular diseases*, CVD) constituent la première cause de mortalité mondiale, responsables de 17,9 millions de décès annuels selon l'Organisation mondiale de la santé, soit environ 32 % de la mortalité globale [42]. Elles regroupent les cardiopathies ischémiques, les accidents vasculaires cérébraux, les artériopathies périphériques et l'insuffisance cardiaque.

La quantification du risque cardiovasculaire individuel repose sur des **scores de risque** validés à l'échelle populationnelle. Le **score de Framingham** (*Framingham Risk Score*) [40] estime la probabilité d'événement cardiovasculaire majeur à dix ans en intégrant l'âge, le sexe, la pression artérielle systolique, le taux de cholestérol total et LDL, le statut tabagique et la présence d'un diabète. Le **SCORE2** (*Systematic COronary Risk Estimation 2*) [34], développé par la Société européenne de cardiologie, est calibré sur des cohortes européennes contemporaines et stratifie le risque en quatre catégories : faible, modéré, élevé et très élevé. Ces outils présentent des limites reconnues, notamment une sous-estimation du risque chez les sujets jeunes, les femmes et les populations non caucasiennes, justifiant le développement de biomarqueurs complémentaires issus de l'imagerie rétinienne.

1.2.3 Liens entre altérations rétiniennes et maladies cardiovasculaires

Les modifications morphologiques et microvasculaires de la rétine constituent des indicateurs précoces et quantifiables des pathologies cardiovasculaires systémiques ; plusieurs paramètres ont fait l'objet d'études cliniques à grande échelle.

Calibre vasculaire et rapport AV/AN Le **calibre vasculaire rétinien** est évalué par deux indices standardisés : le **CRAE** (*Central Retinal Arteriolar Equivalent*), qui synthétise le diamètre moyen des artéioles, et le **CRVE** (*Central Retinal Venular Equivalent*), qui synthétise celui des veinules. Le **rapport artério-veineux (rapport AV/AN)** est défini par le quotient CRAE/CRVE ; une valeur inférieure à 0,67 signe un rétrécissement artériolaire généralisé, associé à l'hypertension artérielle chronique et au risque d'AVC. L'étude ARIC a démontré qu'un rapport AV/AN abaissé prédit indépendamment les événements coronariens incidents après ajustement sur les facteurs de risque classiques [41]. La mesure de ces indices est réalisée à partir de photographies du fond d'oeil par des algorithmes semi-automatiques ou, plus récemment, par des méthodes d'apprentissage profond [29]. Des descripteurs de géométrie vasculaire (tortuosité, angles de bifurcation,

densité de branchement) ont également été associés à l'incidence et à la progression de la rétinopathie diabétique [9].

Rétinopathie hypertensive La **rétinopathie hypertensive** désigne l'ensemble des modifications rétinienne induites par l'hypertension artérielle ; elle est classifiée selon la grille de **Keith-Wagener-Barker** [22] en quatre stades de sévérité croissante. Le stade I correspond à un rétrécissement artériolaire modéré et à une augmentation du reflet lumineux artériolaire ; le stade II associe un rétrécissement artériolaire plus marqué à un signe de croisement artério-veineux (*signe de Salus-Gunn*) ; le stade III ajoute des exsudats cotonneux, des hémorragies en flammèches et des nodules durs ; le stade IV correspond au tableau précédent avec un oedème papillaire. Les stades III et IV sont associés à un risque de mortalité cardiovasculaire multiplié par deux à cinq par rapport aux sujets normo-tendus appariés.

Densité capillaire en OCTA L'**OCT-angiographie (OCTA)** permet la quantification non invasive de la **densité capillaire rétinienne** dans plusieurs plexus vasculaires : le plexus superficiel, le plexus profond, la choriocapillaire et la couche des capillaires radiaux péripapillaires. Deux phénomènes sont particulièrement associés au risque cardiovasculaire : la **raréfaction périfovéale**, définie par la réduction de la densité capillaire dans la zone annulaire entourant la fovéa, observée chez les patients hypertendus et diabétiques ; la **raréfaction péripapillaire**, soit la diminution de la densité des capillaires radiaux entourant la papille, corrélée à la pression de perfusion oculaire et à la pression artérielle systémique [28]. Des études récentes ont montré que la densité capillaire en OCTA est un prédicteur indépendant de la rigidité aortique et de l'hypertrophie ventriculaire gauche, deux marqueurs établis de risque cardiovasculaire élevé [14].

Ces altérations microvasculaires, bien que cliniquement établies, nécessitent des technologies d'imagerie adaptées pour être mesurées de façon reproductible à grande échelle ; la section suivante présente les quatre équipements exploités dans ce travail.

1.3 Technologies d'imagerie ophtalmologique

Quatre modalités d'imagerie sont utilisées dans le cadre de ce travail, couvrant deux grandes familles technologiques : l'OCT-angiographie et la photographie du fond d'oeil couleur.

1.3.1 OCT-Angiographie (OCTA)

L'**OCT-Angiographie (OCTA)** est une technique d'imagerie fonctionnelle dérivée de la **tomographie par cohérence optique** (*Optical Coherence Tomography*, OCT) ; elle

détecte le flux sanguin par analyse des variations de signal entre acquisitions successives en un même point rétinien, sans injection de produit de contraste. Le contraste vasculaire est obtenu par des algorithmes de **décorrélation de speckle** ou de calcul de **variance d'amplitude**, permettant la génération de cartes angiographiques en face pour chaque couche rétinienne segmentée. La résolution axiale, de l'ordre de 5 à 7 μm , et la résolution latérale, de l'ordre de 15 à 20 μm , permettent la visualisation des capillaires de petit calibre inaccessibles à l'angiographie fluorescéinique conventionnelle.

Plexelite 9000 (Zeiss) Le **Plex Elite 9000** (Carl Zeiss Meditec) est un système OCTA de recherche à **swept-source** (*SS-OCTA*), opérant à une longueur d'onde centrale de 1 060 nm, avec une vitesse d'acquisition de 100 000 A-scans par seconde. Ses caractéristiques distinctives sont : un champ d'acquisition maximal de 15×15 mm sans mosaïque, une pénétration choroïdienne supérieure grâce à la longueur d'onde dans le proche infrarouge, et une segmentation automatique en six couches vasculaires. La résolution axiale est de 6,3 μm dans les tissus ; la résolution latérale, de 20 μm . Ce dispositif est adapté à l'analyse de la choriocapillaire et des plexus profonds.

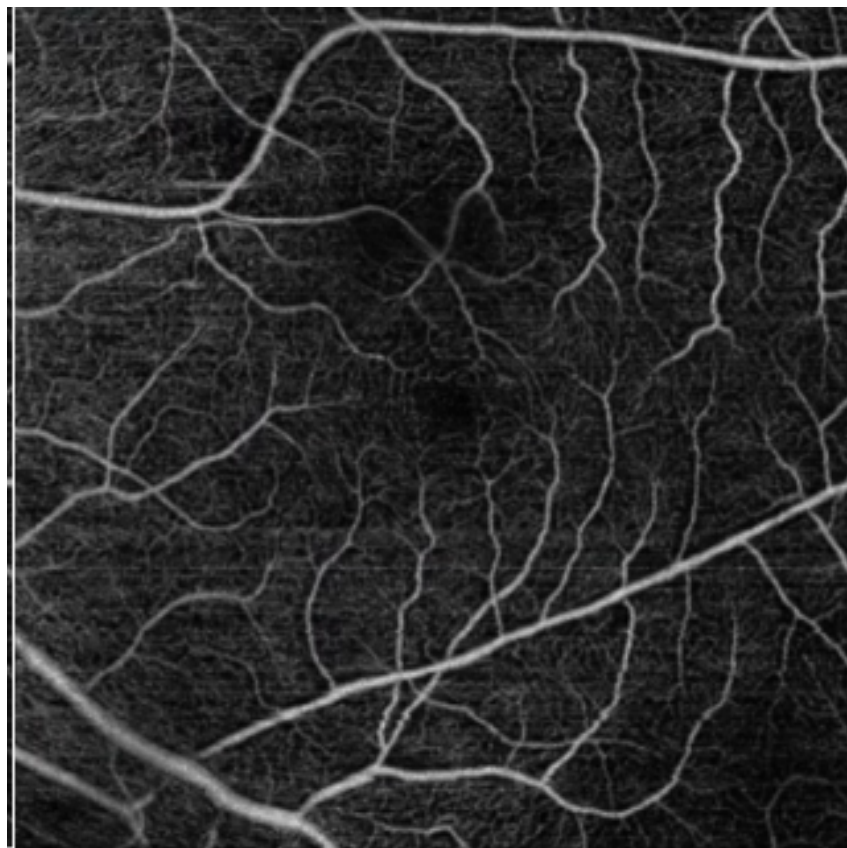


Figure 1.1 — Image rétinienne OCTA — Dataset EviRED (Plex Elite 9000).

Cirrus HD-OCT 6000 (Zeiss) Le **Cirrus HD-OCT 6000** (Carl Zeiss Meditec) est un système OCTA clinique à **spectral-domain** (*SD-OCTA*), opérant à 840 nm, avec une

vitesse d'acquisition de 100 000 A-scans par seconde. Son champ standard est de 6×6 mm ou 3×3 mm, centré sur la macula ou la papille ; la résolution axiale est de $5 \mu\text{m}$ dans les tissus ; la résolution latérale, de $15 \mu\text{m}$. Il intègre le module **AngioPlex** avec correction des artéfacts de mouvement par *FastTrac*. Sa diffusion clinique large en fait un dispositif de référence pour les études épidémiologiques multicentriques.

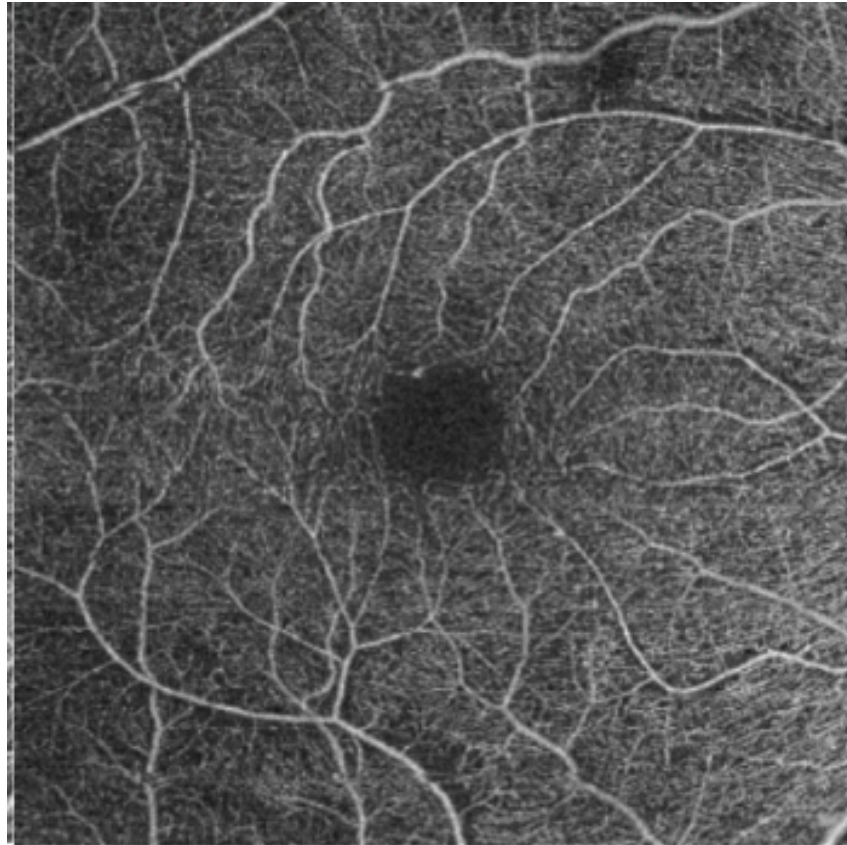


Figure 1.2 — Image rétinienne OCTA — Cirrus HD-OCT 6000 (Zeiss).

1.3.2 Photographie du fond d'oeil couleur (CFP)

La **photographie du fond d'oeil couleur** (*Color Fundus Photography*, CFP) est la modalité d'imagerie rétinienne la plus répandue en pratique clinique et en recherche épidémiologique ; elle documente la morphologie des vaisseaux rétiniens, de la papille et de la macula en lumière visible. Les progrès récents en optique grand champ et ultra-grand champ ont étendu le territoire rétinien accessible à l'imagerie sans mydriase systématique.

Optos (ultra-grand champ, 200°) L'Optos (Optos plc, modèles Daytona et California) est un système de **photographie ultra-grand champ** exploitant un principe de **balayage laser confocal** (*scanning laser ophthalmoscopy*, SLO) à deux longueurs d'onde simultanées (532 nm et 633 nm) ; il permet l'acquisition d'un champ de 200°, couvrant environ 80 % de la surface rétinienne totale, sans dilatation pupillaire dans la majorité des cas. La

résolution au centre du champ est de $14\mu\text{m}$; elle se dégrade vers la périphérie en raison des aberrations optiques périphériques. Cette modalité est particulièrement adaptée à la détection des lésions rétinienne périphériques, inaccessibles aux systèmes à champ standard.

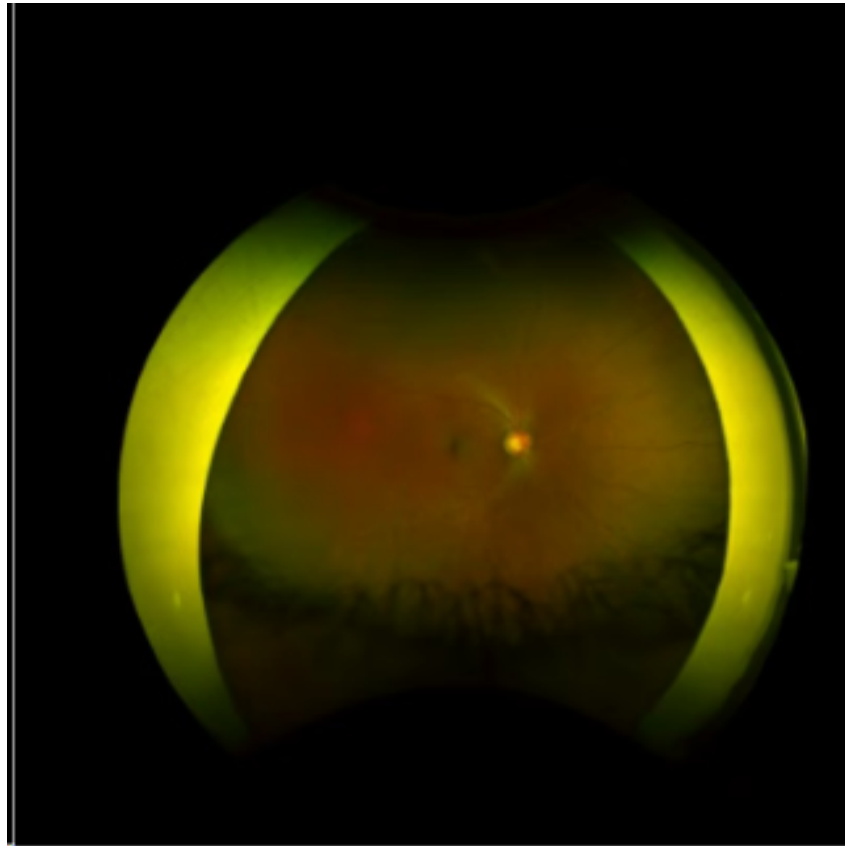


Figure 1.3 – Image du fond d’oeil couleur ultra-grand champ (CFP) — Optos.

Clarus 700 (Zeiss, 133°) Le **Clarus 700** (Carl Zeiss Meditec) est un système de photographie grand champ par **illumination à fond clair** (*true color fundus imaging*), produisant des images couleur à 133° de champ avec une résolution de 14,6 mégapixels. Il utilise une source lumineuse LED à spectre large, restituant une colorimétrie proche de la fundoscopie directe, et propose des modes d’acquisition en lumière verte, rouge et infrarouge pour la différenciation des couches rétinienne. L’absence de principe SLO confère des images à haute fidélité colorimétrique, avantageuses pour l’analyse des exsudats et des hémorragies.



Figure 1.4 – Image du fond d’œil couleur grand champ (CFP) — Clarus 700 (Zeiss).

1.3.3 Comparaison des modalités

La figure 1.5 illustre les champs d’acquisition relatifs des quatre modalités sur un schéma rétinien normalisé. Le tableau 1.1 synthétise les caractéristiques techniques et cliniques correspondantes.

Table 1.1 – Comparaison des quatre modalités d’imagerie ophtalmologique utilisées.

Critère	Plex Elite 9000	Cirrus 6000	Optos	Clarus 700
Type	SS-OCTA	SD-OCTA	SLO couleur	CFP couleur
Longueur d’onde	1 060 nm	840 nm	532/633 nm	LED spectre large
Champ maximal	12×12 mm	6×6 mm	200°	133°
Résolution axiale	6,3 μ m	5 μ m	N/A	N/A
Résolution latérale	20 μ m	15 μ m	14 μ m (centre)	~5 μ m
Flux sanguin	Oui	Oui	Non	Non
Sans dilatation	Oui	Partiel	Oui	Oui
Usage principal	Recherche	Clinique/Recherche	Dépistage périphérique	Clinique/Cohorte

Les modalités OCTA fournissent une information microvasculaire quantitative en trois dimensions, inaccessible à la CFP ; en contrepartie, leur champ d’acquisition est restreint à la région maculaire ou péripapillaire, et leur coût limite leur déploiement dans les grandes cohortes épidémiologiques. La CFP grand champ, notamment avec l’Optos, offre une couverture rétinienne large permettant la détection de lésions périphériques associées au diabète et à l’hypertension ; le Clarus 700, avec sa colorimétrie fidèle et sa haute résolution, est adapté à l’analyse automatisée des structures du pôle postérieur. La complémentarité de ces modalités motive l’approche multimodale adoptée dans ce travail.

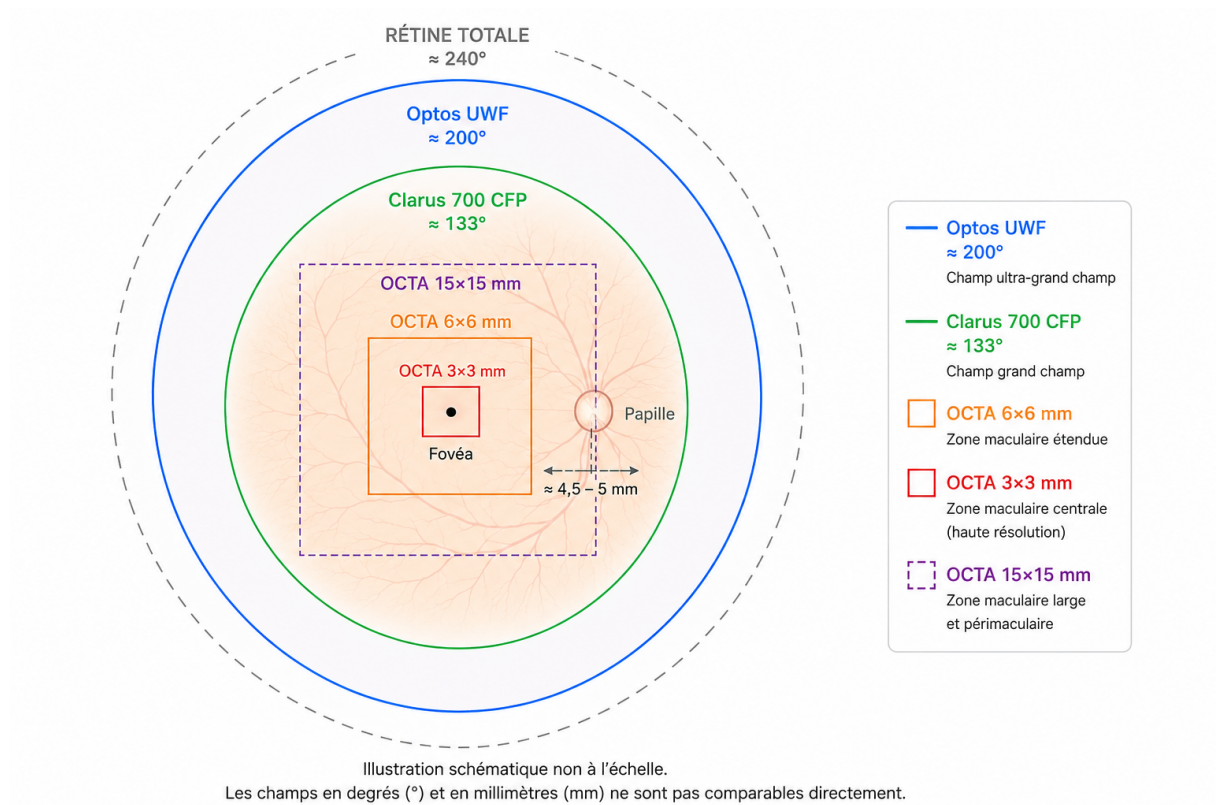


Figure 1.5 – Comparaison des champs d'acquisition des quatre modalités d'imagerie rétinienne (vue schématique, vue en face). L'Optos couvre 80 % de la surface rétinienne ; le Clarus 700 couvre un champ de 133° ; les systèmes OCTA sont limités à des fenêtres maculaires de 6×6 mm ou 3×3 mm.

La complémentarité de ces modalités, tant en termes de résolution spatiale que de couverture du champ rétinien, motive l'approche multi-machines adoptée dans ce travail. Le chapitre suivant présente le cadre algorithmique qui permet d'en exploiter la richesse sans annotation manuelle.

1.4 Conclusion

La rétine offre un accès direct, non invasif et quantifiable à la microcirculation systémique ; ses altérations morphologiques et microvasculaires constituent des biomarqueurs précoces et indépendants du risque cardiovasculaire, validés sur des cohortes de plusieurs centaines de milliers de sujets. Les paramètres rétiens établis, tels que le rapport AV/AN, la classification de Keith-Wagener-Barker et la densité capillaire en OCTA, reflètent des processus physiopathologiques communs aux maladies cardiovasculaires : rigidité artérielle, inflammation vasculaire et raréfaction microvasculaire. Les quatre modalités d'imagerie décrites fournissent des informations complémentaires à différentes échelles spatiales et fonctionnelles, constituant la base technique sur laquelle s'appuient les architectures de *foundation models* développées dans les chapitres suivants.

Chapitre 2

Apprentissage auto-supervisé pour l'imagerie médicale

2.1 Introduction

L'annotation de données médicales exige l'intervention de spécialistes cliniques, ce qui rend la constitution de grands corpus étiquetés coûteuse, lente et difficile à standardiser. En imagerie rétinienne, des milliers d'examens sont acquis chaque année sans jamais être annotés. L'**apprentissage auto-supervisé** (*Self-Supervised Learning*, SSL) répond à cette contrainte en apprenant des représentations généralisables directement depuis des données non étiquetées, par résolution de tâches prétextes exploitant le contenu intrinsèque des images. Ce chapitre présente les familles de méthodes SSL, l'architecture des *Masked Autoencoders* (MAE) et le cadre des *Vision Transformers* (ViT) qui constituent le socle de ce travail.

2.2 Principes de l'apprentissage auto-supervisé

2.2.1 Définition et motivation

L'**apprentissage auto-supervisé** (*Self-Supervised Learning*, SSL) est un paradigme d'entraînement dans lequel le signal de supervision est extrait directement des données non étiquetées, sans recours à des annotations humaines. Le modèle résout une **tâche prétexte** (*pretext task*) conçue pour forcer l'apprentissage de représentations sémantiquement riches, transférables ensuite à des tâches cibles par *fine-tuning* ou par simple sonde linéaire.

Trois raisons motivent le recours au SSL en imagerie médicale. Premièrement, l'annotation de données médicales exige l'intervention de spécialistes cliniques, ce qui rend la constitution de grands corpus étiquetés coûteuse et lente. Deuxièmement, les images non étiquetées sont disponibles en quantité bien supérieure aux images annotées, en particulier en rétinographie où chaque appareil accumule des milliers d'examens de routine. Troisième-

mement, les représentations apprises par SSL sur des données de domaine présentent un meilleur transfert que les représentations apprises sur ImageNet, car elles capturent les structures morphologiques propres à l'imagerie concernée.

2.2.2 Taxonomie des méthodes SSL

Les méthodes SSL se répartissent en trois familles selon la nature de la tâche prétexte et le mécanisme de régularisation employé.

Méthodes contrastives (*contrastive methods*). Elles apprennent à rapprocher dans l'espace des représentations deux vues augmentées d'une même image (*positive pair*) tout en éloignant les représentations d'images différentes (*negative pairs*). **SimCLR** utilise un encodeur partagé suivi d'une tête de projection et minimise la perte NT-Xent sur des mini-batches de grande taille [8]. **MoCo** (*Momentum Contrast*) introduit une file de représentations maintenue par un encodeur à momentum, ce qui découple la taille effective de la banque de négatifs de la taille du batch. Ces méthodes produisent des représentations discriminantes mais requièrent une conception soignée des augmentations pour éviter les solutions triviales.

Méthodes génératives (*generative methods*). Elles formulent la tâche prétexte comme une reconstruction partielle ou bruitée de l'entrée. L'encodeur est entraîné à produire une représentation latente suffisamment informative pour que le décodeur reconstruise les portions masquées ou dégradées. Les **Masked Autoencoders (MAE)** constituent l'instance la plus étudiée de cette famille dans le contexte des *Vision Transformers*; leur architecture est détaillée à la section 2.3.

Méthodes par distillation et auto-distillation (*distillation-based methods*). Elles suppriment le besoin de paires négatives explicites en utilisant deux réseaux, un réseau étudiant et un réseau enseignant, dont les représentations sont alignées. **DINO** exploite une architecture *teacher-student* où l'enseignant est mis à jour par moyenne exponentielle des poids de l'étudiant; l'alignement porte sur les distributions de sorties d'une tête de type *softmax* avec centrage [6]. **Barlow Twins** adopte une approche différente en minimisant la redondance entre les composantes du vecteur de représentation via une perte basée sur la matrice de corrélation croisée, sans réseau enseignant séparé.

Parmi ces familles, les autoencodeurs masqués constituent l'approche retenue dans ce travail en raison de leur adéquation aux images haute résolution et de leur compatibilité naturelle avec l'architecture ViT décrite à la section suivante.

2.3 Masked Autoencoders (MAE)

2.3.1 Architecture : encodeur ViT et décodeur léger

L'architecture **MAE** (*Masked Autoencoder*), introduite par He et al. (2022) [16], est délibérément asymétrique : un encodeur lourd traite un sous-ensemble de patches visibles, tandis qu'un décodeur léger reconstruit l'image complète à partir des représentations latentes et de tokens de masque. Cette asymétrie réduit le coût de calcul de l'encodeur d'un facteur correspondant au taux de masquage, rendant le pré-entraînement sur de larges corpus pratiquement réalisable.

L'encodeur est un **Vision Transformer** (*ViT*) standard ; il reçoit uniquement les patches non masqués, auxquels des encodages positionnels sont ajoutés avant les couches d'attention. Le décodeur, plus petit (profondeur et dimension réduites), prend en entrée la concaténation des représentations latentes des patches visibles et d'un ensemble de tokens apprenables représentant les positions masquées. La reconstruction s'effectue en espace pixel sur les patches masqués uniquement. Dans le contexte de l'OCTA, des variantes *vessel-aware* ont été proposées afin de contraindre l'auto-encodeur à préserver les structures vasculaires fines lors de la reconstruction [2].

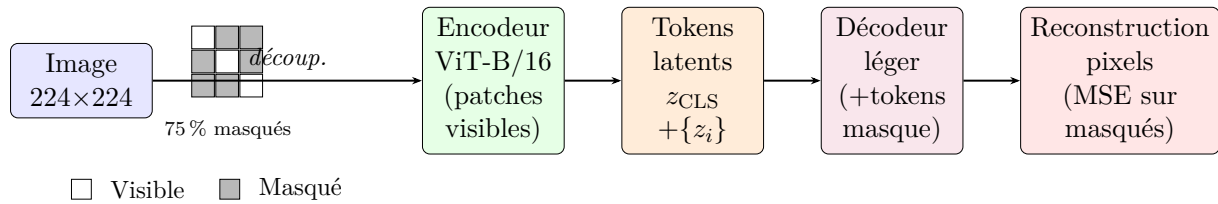


Figure 2.1 – Architecture MAE : l'encodeur ViT traite uniquement les 25 % de patches visibles ; le décodeur léger reconstruit les patches masqués à partir des tokens latents et de tokens de masque apprenables.

2.3.2 Stratégie de masquage

Le MAE applique un masquage aléatoire par patch avec un taux de 75 % comme valeur optimale. Ce taux élevé est justifié par deux arguments complémentaires. D'une part, un taux faible rend la tâche de reconstruction triviale car les patches voisins fournissent une information redondante suffisante pour interpoler les zones masquées sans apprendre de structure globale. D'autre part, un taux de 75 % force l'encodeur à construire des représentations intégrant le contexte large de l'image, ce qui conduit à des caractéristiques transférables plutôt qu'à des descripteurs locaux de bas niveau. Des taux supérieurs à 80 % dégradent la qualité de reconstruction sans apporter de bénéfice supplémentaire en transfert.

2.3.3 Fonction de reconstruction

La fonction de perte du MAE est l'erreur quadratique moyenne (*Mean Squared Error*, MSE) calculée exclusivement sur les patches masqués :

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|\hat{x}_i - x_i\|^2 \quad (2.1)$$

où $\mathcal{M}$ désigne l'ensemble des indices de patches masqués, x_i le patch original normalisé et $\hat{x}_i$ la reconstruction produite par le décodeur. La normalisation des patches cibles par leur moyenne et écart-type locaux améliore la qualité des représentations apprises en renforçant les gradients sur les régions à faible contraste, caractéristiques fréquentes en imagerie médicale.

2.3.4 Propriétés des représentations

Les représentations produites par un MAE pré-entraîné présentent une structure spatiale fine, car la tâche de reconstruction exige une modélisation précise de la géométrie locale et des relations entre patches distants. Cette propriété est particulièrement avantageuse pour les tâches de segmentation et de détection de lésions rétinienne, où la localisation spatiale est déterminante.

RETFound [45], développé par Zhou et al. (2023), illustre cette capacité : un ViT-Large pré-entraîné par MAE sur 1,6 million de rétinographies non étiquetées atteint des performances supérieures à l'état de l'art sur onze tâches cliniques de rétinographie et de tomographie par cohérence optique après *fine-tuning* sur des données étiquetées limitées. Sur la tâche de prédiction du risque cardiovasculaire à partir du fond d'œil, RETFound surpasse les modèles supervisés entraînés sur l'intégralité des données étiquetées disponibles, démontrant l'efficacité du pré-entraînement de domaine par rapport au transfert depuis ImageNet.

L'architecture ViT qui sert d'encodeur à ces modèles est décrite en détail dans la section suivante, avant d'aborder les enjeux spécifiques du pré-entraînement par machine.

2.4 Vision Transformers (ViT)

2.4.1 Architecture générale

Le **Vision Transformer** (*ViT*), introduit par Dosovitskiy et al. (2021) [12], adapte l'architecture *Transformer* originellement conçue pour le traitement du langage naturel au traitement des images en découpant l'image en patches de taille fixe. Chaque patch de dimensions $P \times P$ pixels est linéairement projeté vers un vecteur de dimension D , formant ainsi une séquence de tokens à laquelle est ajouté un token de classification [CLS] appris.

Un encodage positionnel appris ou sinusoïdal est additionné aux embeddings de patches afin de préserver l'information de position spatiale, absente du mécanisme d'attention par défaut. La séquence est ensuite traitée par L blocs *Transformer* identiques [38], chacun composé d'une couche d'**attention multi-têtes** (*Multi-Head Self-Attention*, MHSA) et d'un réseau *feed-forward* à deux couches avec normalisation de couche (*LayerNorm*) et connexions résiduelles.

L'attention multi-têtes est définie par :

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \quad (2.2)$$

où Q , K et V sont les matrices de requêtes, clés et valeurs issues de projections linéaires de la séquence d'entrée, et $d_k = D/h$ est la dimension par tête pour h têtes d'attention. La figure 2.2 illustre le pipeline de traitement d'une image par un ViT.

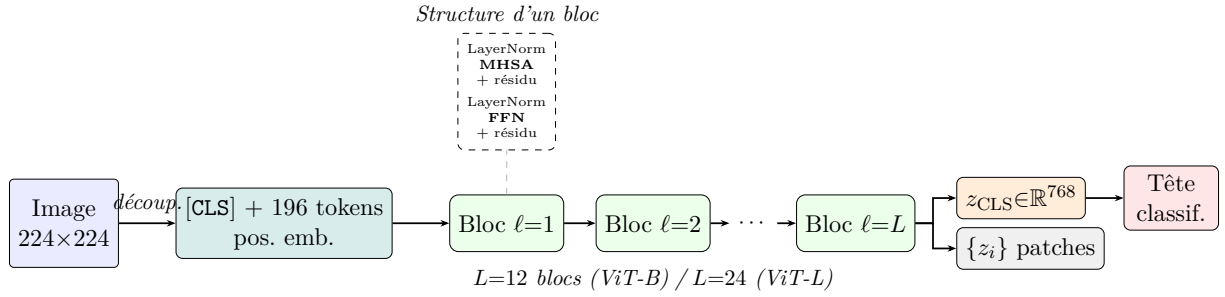


Figure 2.2 – Architecture Vision Transformer (ViT) : l'image est découpée en patches 16×16 projetés en tokens avec embedding de position ; le token [CLS] agrège la représentation globale après L blocs Transformer identiques, chacun composé d'une normalisation, d'une attention multi-têtes (MHSA) et d'un réseau *feed-forward* (FFN) avec connexions résiduelles.

2.4.2 ViT-B/16 : configuration

La variante **ViT-B/16** (*ViT-Base*, *patch size 16*) constitue le point de référence standard pour les expériences de pré-entraînement dans ce travail. Ses paramètres de configuration sont récapitulés dans le tableau 2.1.

Table 2.1 – Paramètres de configuration du modèle **ViT-B/16**.

Paramètre	Valeur
Nombre de blocs Transformer (L)	12
Dimension des embeddings (D)	768
Nombre de têtes d'attention (h)	12
Dimension par tête (d_k)	64
Dimension du réseau feed-forward	3 072
Taille des patches (P)	16×16 pixels
Nombre total de paramètres	$\approx 86 \times 10^6$

Pour une image d'entrée de 224×224 pixels, ViT-B/16 génère une séquence de 196 patches auxquels s'ajoute le token [CLS], soit 197 tokens traités par les 12 blocs d'attention. La représentation finale du token [CLS] est utilisée comme vecteur de caractéristiques globales pour les tâches de classification.

2.4.3 Avantages pour l'imagerie médicale

Les ViT présentent plusieurs propriétés favorables à l'imagerie médicale par rapport aux réseaux convolutifs (*Convolutional Neural Networks*, CNN). Le champ réceptif de l'attention est global dès la première couche, permettant la modélisation de dépendances longue distance entre structures anatomiques éloignées dans l'image, par exemple la relation entre la papille optique et la macula en rétinographie. Les cartes d'attention (*attention maps*) offrent une forme d'interprétabilité directe indiquant les régions de l'image ayant contribué à la décision, ce qui facilite la validation clinique. Enfin, les ViT s'adaptent facilement à des résolutions variables par interpolation des encodages positionnels, propriété utile lorsque les appareils d'imagerie produisent des images de dimensions hétérogènes.

2.5 Modèles de fondation en imagerie médicale

2.5.1 Définition et intérêt

Un **modèle de fondation** (*foundation model*) est un modèle de grande capacité pré-entraîné sur un corpus massif et hétérogène, conçu pour être adapté à de nombreuses tâches en aval avec un minimum de données étiquetées supplémentaires. Le terme, introduit par Bommasani et al. (2021) [5], recouvre les grands modèles de langage ainsi que leurs équivalents visuels et multimodaux. En ophtalmologie, des modèles visuels/multimodaux comme **VisionFM** visent à apprendre des représentations transférables sur un large spectre de tâches et de modalités [30].

En imagerie médicale, l'intérêt des modèles de fondation est triple : ils réduisent la dépendance aux annotations en exploitant de larges corpus non étiquetés ; ils centralisent les représentations apprises dans un seul modèle réutilisable au lieu de multiplier les entraînements spécialisés ; ils améliorent la généralisation aux distributions hors domaine grâce à la diversité des données de pré-entraînement. Le transfert d'un modèle de fondation nécessite souvent seulement quelques centaines d'exemples étiquetés pour atteindre des performances comparables à celles d'un modèle supervisé entraîné sur des milliers d'exemples.

2.5.2 Enjeux du pré-entraînement spécifique par machine

Une limite des modèles de fondation génériques en rétinographie est leur insensibilité aux variations instrumentales entre appareils d'acquisition. Deux rétinographes de gammes différentes, le **Plexelite** (Carl Zeiss Meditec, imagerie *wide-field* à 133° de champ) et le **Cirrus HD-OCT** (Carl Zeiss Meditec, imagerie structurale spectrale), produisent des images aux caractéristiques radiométriques, géométriques et de résolution distinctes. Un modèle pré-entraîné sur un mélange hétérogène d'images provenant de sources multiples peut apprendre des représentations bruitées par ces différences instrumentales, introduisant un biais confondant lors du *fine-tuning*.

Le pré-entraînement spécifique par machine (*per-device pretraining*) consiste à entraîner ou à adapter le modèle de fondation sur un corpus homogène acquis avec le même appareil que celui utilisé en déploiement. Cette stratégie présente deux avantages : les représentations latentes capturent les structures morphologiques de l'organe sans être perturbées par des artefacts propres à d'autres modalités ; le *fine-tuning* sur la tâche cible converge plus rapidement car le domaine source et le domaine cible sont identiques. Dans le cadre de ce travail, le pré-entraînement est conduit séparément sur les corpus Plexelite et Cirrus afin de disposer de représentations cohérentes avec les données de validation clinique correspondantes.

Ce cadre théorique étant posé, le chapitre suivant introduit les modèles de diffusion, une famille complémentaire de modèles génératifs utilisée dans ce travail pour la synthèse d'images rétinienne pathologiques.

2.6 Conclusion

L'apprentissage auto-supervisé par masquage, incarné par le MAE appliqué aux Vision Transformers, fournit un cadre général pour extraire des représentations riches à partir de données non étiquetées ; les méthodes contrastives et par distillation offrent des alternatives adaptées selon les contraintes computationnelles et la disponibilité des données. La configuration ViT-B/16, avec ses 86 millions de paramètres et son attention globale, constitue un encodeur de référence pour les tâches rétinienne. L'adoption du paradigme des modèles de fondation avec pré-entraînement spécifique par machine répond directement aux contraintes de l'imagerie clinique : hétérogénéité instrumentale, rareté des annotations et exigence de transférabilité vers des tâches de prédiction du risque cardiovasculaire.

Chapitre 3

Modèles de Diffusion pour la Génération d'Images Médicales

3.1 Introduction

Les **modèles de diffusion** constituent une famille de modèles génératifs dont le principe repose sur deux processus markoviens opposés : un processus de dégradation progressive qui transforme une image réelle en bruit gaussien, et un processus de débruitage appris qui reconstruit une image à partir de ce bruit. Leur capacité à modéliser des distributions de haute dimension avec une fidélité statistique supérieure aux *Generative Adversarial Networks* (GANs) les rend particulièrement adaptés à la synthèse d'images médicales, où la diversité et le réalisme des exemples générés conditionnent directement l'utilité des données augmentées. Ce chapitre présente les fondements mathématiques des DDPM, la stratégie d'adaptation efficace par LoRA, et les applications en imagerie rétinienne.

3.2 Principes des modèles de diffusion

3.2.1 Processus de diffusion avant (*forward process*)

Le **processus avant** définit une chaîne de Markov qui dégrade progressivement une image $\mathbf{x}_0$ issue de la distribution des données $q(\mathbf{x}_0)$ en ajoutant du bruit gaussien à chaque étape $t \in \{1, \dots, T\}$. La distribution de transition est définie par :

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}) \quad (3.1)$$

où $\{\beta_t\}_{t=1}^T$ est un calendrier de bruit (*noise schedule*) prédéfini, avec $\beta_t \in (0, 1)$ croissant en t . La structure markovienne permet d'exprimer $q(\mathbf{x}_t | \mathbf{x}_0)$ en forme close grâce à la reparamétrisation. En posant $\alpha_t = 1 - \beta_t$ et $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$, on obtient :

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I}) \quad (3.2)$$

ce qui permet d'échantillonner $\mathbf{x}_t$ directement à partir de $\mathbf{x}_0$ sans dérouler toute la chaîne : $\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}$, avec $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Pour T suffisamment grand (typiquement $T = 1000$), $\mathbf{x}_T$ converge vers un bruit gaussien isotrope $\mathcal{N}(\mathbf{0}, \mathbf{I})$. La figure 3.1 schématise les deux processus opposés.

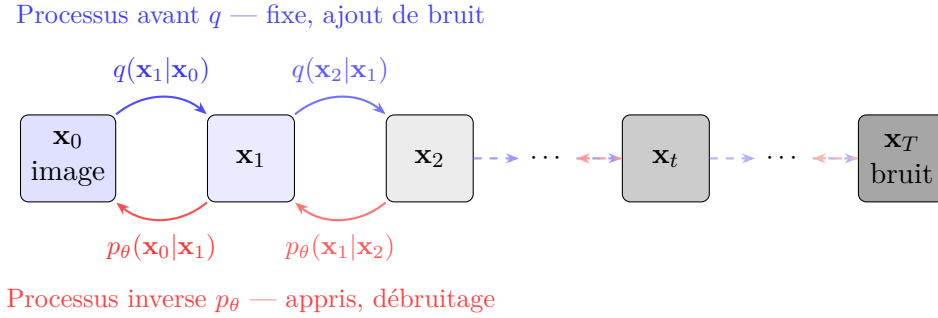


Figure 3.1 — Chaîne de Markov du processus DDPM : le processus avant q dégrade progressivement l'image $\mathbf{x}_0$ en bruit gaussien $\mathbf{x}_T$; le processus inverse p_θ reconstruit une image propre par débruitage itératif appris.

3.2.2 Processus de débruitage inverse (*reverse process*)

Le **processus inverse** apprend à inverser la dégradation en estimant la distribution $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$ pour chaque étape t . Cette distribution est paramétrée par un réseau de neurones θ comme une gaussienne :

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \quad (3.3)$$

La moyenne $\boldsymbol{\mu}_\theta$ est exprimée en fonction de la prédiction du bruit $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$:

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) \quad (3.4)$$

Le réseau $\boldsymbol{\epsilon}_\theta$ adopte une architecture **U-Net** avec des connexions de saut (*skip connections*) entre les blocs d'encodage et de décodage ; l'étape temporelle t est encodée par un plongement sinusoïdal (*sinusoidal embedding*) injecté dans chaque bloc résiduel. Des blocs d'attention (*self-attention*) multi-têtes sont intercalés aux résolutions intermédiaires pour capturer les dépendances longue portée dans l'image.

3.2.3 Denoising Diffusion Probabilistic Models (DDPM)

Les **DDPM**, introduits par Ho et al. (2020) [18], simplifient l'objectif d'entraînement en fixant $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t) = \sigma_t^2 \mathbf{I}$ et en réduisant la borne inférieure variationnelle (*ELBO*) à une

perte de prédiction du bruit. La fonction de perte s'écrit :

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{t, \mathbf{x}_0, \epsilon} \left[\left\| \epsilon - \epsilon_\theta \left(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t \right) \right\|^2 \right] \quad (3.5)$$

où $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ et t est échantillonné uniformément dans $\{1, \dots, T\}$. Cette formulation réduit l'entraînement à une régression sur le bruit, stabilisant considérablement l'optimisation par rapport aux objectifs adversariaux des GANs. Les **Latent Diffusion Models** (LDM) [32] étendent ce cadre en appliquant le processus de diffusion dans l'espace latent d'un auto-encodeur variationnel, réduisant le coût computationnel d'un facteur quadratique par rapport à la résolution spatiale.

La puissance générative des DDPM s'accompagne cependant d'un coût computationnel élevé pour l'adaptation à de nouveaux domaines ; la section suivante présente LoRA, la stratégie de fine-tuning efficace retenue pour contourner cette contrainte.

3.3 Fine-tuning efficace avec LoRA

3.3.1 Principe

LoRA (*Low-Rank Adaptation*) [20] est une méthode de fine-tuning à paramètres réduits qui postule que les adaptations nécessaires pour spécialiser un modèle pré-entraîné résident dans un sous-espace de faible rang. Pour une couche linéaire de matrice de poids $\mathbf{W} \in \mathbb{R}^{d \times k}$, LoRA gèle $\mathbf{W}$ et apprend une décomposition de faible rang additive :

$$\mathbf{W}' = \mathbf{W} + \mathbf{B}\mathbf{A} \quad (3.6)$$

où $\mathbf{B} \in \mathbb{R}^{d \times r}$, $\mathbf{A} \in \mathbb{R}^{r \times k}$, et $r \ll \min(d, k)$ est le rang de l'adaptation. À l'initialisation, $\mathbf{B} = \mathbf{0}$ et $\mathbf{A}$ est tiré d'une distribution gaussienne, garantissant que $\mathbf{W}' = \mathbf{W}$ au début de l'entraînement ; seuls $\mathbf{A}$ et $\mathbf{B}$ sont mis à jour par rétropropagation.

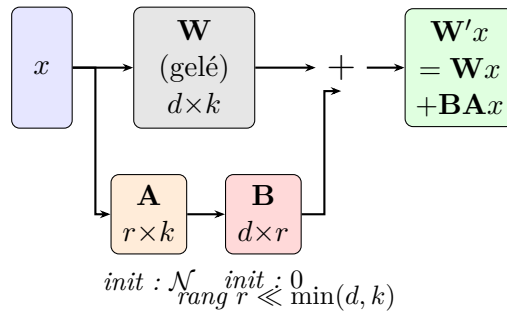


Figure 3.2 – Principe de LoRA : la matrice $\mathbf{W}$ est gelée ; seules les matrices $\mathbf{A}$ (rang r) et $\mathbf{B}$ (rang r) sont entraînées, réduisant les paramètres entraînaibles de $d \times k$ à $r \times (d + k)$.

3.3.2 Avantages computationnels

Sans LoRA, l’adaptation complète d’une couche exige la mise à jour de $d \times k$ paramètres ; avec LoRA de rang r , le nombre de paramètres entraînaibles se réduit à $r \times (d + k)$. Pour $r = 4$ et $d = k = 768$ (dimensions typiques des projections d’attention dans les U-Nets de grande taille), le ratio de réduction est :

$$\frac{d \times k}{r \times (d + k)} = \frac{768 \times 768}{4 \times (768 + 768)} = \frac{589\,824}{6\,144} \approx 96 \quad (3.7)$$

soit une réduction d’environ 96 fois le nombre de paramètres entraînaibles. Cette compression permet le fine-tuning sur des jeux de données médicaux de taille limitée sans risque de sur-apprentissage, et autorise la coexistence de plusieurs adaptateurs spécialisés chargés dynamiquement sans dupliquer les poids de base.

3.3.3 Application aux modèles de diffusion

Dans le contexte des modèles de diffusion, LoRA s’applique aux **projections QKV** (*Query, Key, Value*) des blocs d’auto-attention du U-Net. Les couches de convolution et les blocs résiduels restent gelés ; seules les matrices d’attention sont adaptées via les décompositions de faible rang. Pour un modèle de diffusion conditionnel guidé par des étiquettes pathologiques, chaque classe peut disposer d’un adaptateur LoRA distinct : le modèle de base capture la distribution générale des images rétiniennes, tandis que les adaptateurs encodent les variations morphologiques spécifiques à chaque pathologie. Lors de l’inférence, l’adaptateur correspondant à la pathologie souhaitée est activé, permettant la génération contrôlée d’images avec des caractéristiques ciblées.

Les propriétés de LoRA décrites ici seront exploitées concrètement au chapitre 6, lors du fine-tuning du modèle de diffusion sur le corpus EVIRED. La section suivante présente les applications directes de cette approche en imagerie rétinienne.

3.4 Applications en imagerie rétinienne

3.4.1 Augmentation de données

L’**augmentation par modèles de diffusion** dépasse les transformations géométriques et photométriques classiques en générant des variantes sémantiquement plausibles. Pour l’imagerie rétinienne, le processus consiste à entraîner un modèle de diffusion conditionnel sur un corpus d’images de fond d’œil annotées, puis à échantillonner de nouvelles images sous différentes conditions pathologiques ou démographiques. Les images synthétiques sont intégrées à l’ensemble d’entraînement d’un modèle aval, augmentant sa capacité de généralisation sur les classes sous-représentées. La condition de génération peut encoder

le grade de rétinopathie diabétique, le stade de dégénérescence maculaire liée à l’âge, ou des paramètres cardiovasculaires comme l’indice artério-veineux ; le modèle apprend ainsi à générer des images respectant les contraintes cliniques associées à chaque condition. L’efficacité de l’augmentation est mesurée par l’amélioration des métriques de classification sur le jeu de test réel après incorporation des données synthétiques.

3.4.2 Synthèse d’images pathologiques

La **synthèse d’images pathologiques** vise à produire des images réalistes présentant des altérations morphologiques spécifiques, adressant le déséquilibre sévère entre cas sains et cas pathologiques dans les bases de données rétinienne. Les **LDM** (*Latent Diffusion Models*) sont particulièrement adaptés à cette tâche : le processus de diffusion opère dans l’espace latent compressé de l’auto-encodeur, réduisant le coût d’échantillonnage tout en préservant les détails fins nécessaires à la crédibilité clinique des images générées.

La qualité des images synthétiques est évaluée selon deux métriques complémentaires. Le **FID** (*Fréchet Inception Distance*) [17] mesure la distance entre la distribution des caractéristiques des images réelles et synthétiques dans l’espace d’un réseau Inception pré-entraîné ; un FID faible indique une forte similarité distributionnelle. Le **SSIM** (*Structural Similarity Index Measure*) évalue la préservation des structures locales, de la luminance et du contraste entre une image de référence et son équivalent synthétique ; il est compris entre 0 et 1, une valeur proche de 1 indiquant une haute fidélité structurelle. Ces deux métriques capturent des aspects orthogonaux de la qualité : le FID évalue la diversité et le réalisme global de la distribution, tandis que le SSIM mesure la cohérence locale des structures anatomiques individuelles. Pour les images rétinienne, la préservation des structures vasculaires, des dimensions de la papille optique et de la zone maculaire constitue un critère de validation clinique indispensable au-delà des métriques quantitatives.

Ces travaux de génération s’inscrivent en complément du pipeline principal de ce mémoire. Le chapitre suivant aborde directement la tâche cible : la prédiction du risque cardiovasculaire par apprentissage profond à partir des images rétinienne.

3.5 Conclusion

Les modèles de diffusion offrent un cadre mathématiquement rigoureux pour la génération d’images médicales de haute qualité, reposant sur l’alternance d’un processus de dégradation analytiquement tractable et d’un processus de débruitage appris. La reparamétrisation en $\bar{\alpha}_t$ rend l’entraînement DDPM stable et scalable ; l’architecture U-Net avec attention fournit la capacité représentationnelle nécessaire à la modélisation des structures anatomiques rétinienne. L’adaptation par LoRA rend le fine-tuning accessible avec des données médicales limitées, réduisant le nombre de paramètres entraînaibles d’un facteur de

l’ordre de 96 pour des dimensions d’attention de 768. Les résultats expérimentaux obtenus sur la base EviRed avec ce cadre de génération sont détaillés à la section [6.6](#).

Chapitre 4

Prédiction du Risque Cardiovasculaire par Deep Learning

4.1 Introduction

L'imagerie **OCTA** (*Optical Coherence Tomography Angiography*) permet de visualiser sans injection de colorant la microvasculature rétinienne, dont les altérations reflètent fidèlement l'état du système vasculaire systémique. Les modifications de la densité capillaire, de la tortuosité artériolaire et de la zone avasculaire centrale constituent des biomarqueurs quantifiables du risque cardiovasculaire (*cardiovascular risk*, RCV). L'apprentissage profond appliqué à ces images ouvre la voie à une stratification automatique et non invasive du risque, complémentaire des scores cliniques classiques.

4.2 Revue des approches existantes

4.2.1 Approches unimodales OCTA

Mammo et al. (2016) [27] ont établi la première corrélation systématique entre les paramètres **OCTA** et les marqueurs de l'hypertension artérielle (*HTA*) : la densité du plexus capillaire superficiel diminue significativement chez les patients hypertendus, tandis que la zone avasculaire centrale s'élargit en proportion de la sévérité de l'atteinte vasculaire. Cette étude unimodale, fondée sur des mesures manuelles de segmentation, a démontré la faisabilité du diagnostic rétinien pour l'évaluation du risque systémique, mais ses métriques reposaient exclusivement sur des corrélations bivariées sans modèle prédictif entraînable.

4.2.2 Approches multimodales

Poplin et al. (2018) ont entraîné un réseau de neurones convolutif sur des photographies du fond d'œil (*Color Fundus Photography*, CFP) issues de la cohorte *UK Biobank* et d'autres

sources cliniques, obtenant une aire sous la courbe ROC (*AUC-ROC*) de 0,70 pour la prédiction d'événements cardiovasculaires majeurs. Ce résultat, obtenu sans variables biologiques explicites, a mis en évidence la capacité du réseau à extraire des représentations vasculaires latentes directement depuis l'image de fond d'œil. Des biomarqueurs rétiens appris par deep learning ont ensuite été validés spécifiquement sur *UK Biobank*, par exemple **Reti-CVD** [10]. La cohorte *UK Biobank*, qui associe imagerie rétinienne, données génomiques et suivis longitudinaux sur plusieurs centaines de milliers de participants, reste la référence en termes d'envergure pour l'évaluation des approches multimodales ; elle souffre néanmoins d'une représentation inégale des populations à haut risque et d'une résolution rétinienne inférieure à celle de l'OCTA.

4.2.3 Transfer learning et SSL

RETFound [45] est un modèle de fondation (*Foundation Model*) pré-entraîné par **apprentissage auto-supervisé** (*Self-Supervised Learning*, SSL) sur plus de un million d'images rétiennes CFP et OCT. Lorsqu'il est affiné sur le jeu de données **VAMPIRE OCTA-CVD**, il atteint une *AUC-ROC* de 0,802 sur la tâche de prédiction du risque cardiovasculaire, surpassant les architectures supervisées entraînées de zéro sur le même corpus. Sur la cohorte **RASTA**, Germanese et al. [15] ont proposé une approche supervisée (EfficientNetV2-B3) pour prédire le score $\text{CHA}_2\text{DS}_2\text{-VASc}$ à partir d'images SS-OCTA. Des revues récentes synthétisent ces approches et leurs limites en imagerie rétinienne pour le risque cardiovasculaire [23]. Aucun modèle de fondation SSL entraîné spécifiquement sur des images OCTA n'a encore été proposé pour la prédiction du RCV ; les modèles disponibles restent transférés depuis des modalités CFP ou OCT structurel.

Ces résultats confirment la pertinence de l'approche SSL pour la prédiction du risque cardiovasculaire, tout en soulignant l'absence de modèles spécialisés sur les images OCTA. Les jeux de données qui serviront à valider les contributions de ce travail sont décrits dans la section suivante.

4.3 Datasets pour la prédiction du risque cardiovasculaire

4.3.1 Dataset RASTA

Le jeu de données **RASTA** (*Retinal OCT-Angiography and cardiovascular STatus*) a été constitué au **CHU de Dijon** à partir de 491 patients consécutifs, imagés avec le dispositif **Plexelite** en mode $6 \times 6 \text{ mm}$ centré sur la fovéa [4]. La stratification du risque thrombo-embolique repose sur le score $\text{CHA}_2\text{DS}_2\text{-VASc}$ [25], dont les composantes sont détaillées dans le tableau 4.2. Trois classes ont été définies : **Faible** (204 images),

Intermédiaire (392 images) et **Élevé** (1409 images), reflétant le déséquilibre naturel de la population clinique. Les 491 patients sont répartis en **330** pour l’entraînement et la validation (validation croisée stratifiée à 5 plis, **seed** = 42 : 264 patients par fold d’entraînement et 66 par fold de validation) et **83** pour le test isolé final. Onze variables cliniques accompagnent chaque patient (Tab. 4.1).

Table 4.1 – Description des 11 variables cliniques utilisées dans les expériences multimodales sur la cohorte RASTA.

Variable	Type	Composante CHA2DS2-VASc
Âge	Continue	✓
Sexe	Binaire	✓
Insuffisance cardiaque (CHF)	Binaire	✓
Hypertension	Binaire	✓
Diabète	Binaire	✓
Antécédent d’AVC/AIT (Stroke/TIA)	Binaire	✓
Maladie vasculaire	Binaire	✓
IMC (BMI)	Continue	×
Apnée du sommeil (OSA)	Binaire	×
Tabagisme	Binaire	×
Dyslipidémie	Binaire	×

Les sept premières variables (✓) sont des composantes directes du score CHA2DS2-VASc [25] ; les quatre dernières (×) sont des facteurs de risque cardiovasculaire associés non représentés dans le score. Toutes les variables sont normalisées par la moyenne et l’écart-type calculés sur les patients d’entraînement uniquement, évitant toute fuite d’information depuis les ensembles de validation et de test.

Table 4.2 – Composantes du score **CHA₂DS₂-VASc** et points associés.

Lettre	Facteur de risque	Points
C	Insuffisance cardiaque congestive (<i>Congestive heart failure</i>)	1
H	Hypertension artérielle (<i>Hypertension</i>)	1
A ₂	Âge ≥ 75 ans (<i>Age</i>)	2
D	Diabète (<i>Diabetes mellitus</i>)	1
S ₂	Antécédent d’AVC ou AIT (<i>Stroke/TIA</i>)	2
V	Maladie vasculaire (<i>Vascular disease</i>)	1
A	Âge 65–74 ans (<i>Age 65–74</i>)	1
Sc	Sexe féminin (<i>Sex category female</i>)	1
Score maximum		9

4.3.2 Dataset VAMPIRE OCTA-CVD

Le jeu de données **VAMPIRE OCTA-CVD** regroupe **1659 patients** issus de plusieurs centres européens, chacun annoté selon cinq étiquettes binaires indépendantes :

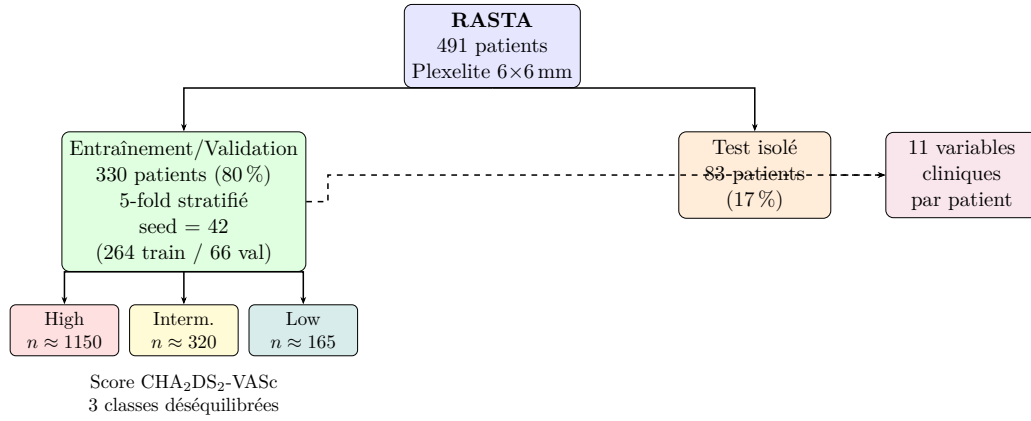


Figure 4.1 – Structure du jeu de données RASTA : 491 patients répartis en 330 pour l’entraînement et la validation (validation croisée stratifiée à 5 plis, **seed** = 42 : 264 patients par fold d’entraînement et 66 par fold de validation) et 83 pour le test isolé final ; les trois classes CHA₂DS₂-VASc (Low, Intermediate, High) présentent un fort déséquilibre naturel ; 11 variables cliniques accompagnent chaque patient.

HighRisk (risque cardiovasculaire global élevé), **HighGlucose** (glycémie élevée), **HighTC** (cholestérol total élevé), **HighTG** (triglycérides élevés) et **Hypertension**. Seules deux variables démographiques sont fournies aux participants : l’âge et le sexe ; aucune autre covariable clinique n’est disponible, ce qui impose aux modèles d’extraire l’intégralité de l’information prédictive depuis l’image OCTA. L’évaluation suit un protocole de **validation croisée à 5 plis** (*5-fold cross-validation*) standardisé, et la compétition associée a mobilisé **12 équipes** internationales. Ce protocole multi-labels, multi-sites et à faible supervision clinique constitue le cadre de référence le plus exigeant disponible pour la prédiction du RCV par OCTA [39]. L’acronyme *VAMPIRE* a été historiquement utilisé en analyse d’images rétinienne pour désigner des cadres de validation et de comparaison d’algorithmes [37].

4.3.3 Comparaison et complémentarité

Les deux jeux de données sont complémentaires selon trois axes. Premièrement, la granularité du score de risque diffère : RASTA fournit un score ordinal à trois niveaux issu d’un outil clinique validé (CHA₂DS₂-VASc), tandis que VAMPIRE OCTA-CVD propose cinq cibles binaires indépendantes couvrant des facteurs de risque distincts. Deuxièmement, la disponibilité des variables cliniques est asymétrique ; RASTA offre onze covariables exploitables pour une fusion multimodale, alors que VAMPIRE OCTA-CVD se limite à l’âge et au sexe. Troisièmement, la taille des corpus diffère substantiellement : 491 patients pour RASTA contre 1659 pour VAMPIRE OCTA-CVD, avec des protocoles d’acquisition distincts (Plexelite 6 × 6 mm pour RASTA, matériel hétérogène multi-sites pour VAMPIRE). L’utilisation conjointe des deux corpus permet d’évaluer la généralisabilité des représentations apprises.

La complémentarité de ces deux corpus permet d’évaluer les modèles dans deux régimes

distincts : supervision clinique riche pour RASTA, évaluation comparative standardisée pour VAMPIRE. Les métriques adaptées à ces deux régimes sont présentées dans la section suivante.

4.4 Métriques d'évaluation

4.4.1 AUC-ROC, BAcc et QWK

L'**aire sous la courbe ROC** (*Area Under the ROC Curve*, **AUC-ROC**) mesure la capacité discriminante d'un classifieur indépendamment du seuil de décision ; elle vaut 0,5 pour un classifieur aléatoire et 1,0 pour une séparation parfaite. Pour les tâches multi-classes à distribution déséquilibrée, la **précision équilibrée** (*Balanced Accuracy*, **BAcc**) calcule la moyenne arithmétique des rappels par classe, pénalisant les modèles qui ignorent les classes minoritaires. Le **coefficient Kappa pondéré quadratique** (*Quadratic Weighted Kappa*, **QWK**) est adapté aux sorties ordinales : il pénalise les erreurs proportionnellement au carré de leur distance ordinale, distinguant ainsi une confusion entre classes adjacentes d'une erreur de classification extrême. Ces trois métriques sont rapportées conjointement pour fournir une évaluation complète de la performance en contexte clinique déséquilibré.

4.4.2 Gestion du déséquilibre de classes

Le déséquilibre observé dans RASTA (rapport 1409 :392 :204 entre les classes Élevé, Intermédiaire et Faible) et dans VAMPIRE OCTA-CVD nécessite des stratégies d'atténuation explicites à trois niveaux.

- **WeightedRandomSampler** : chaque mini-lot est constitué en tirant les exemples avec une probabilité inversement proportionnelle à la fréquence de leur classe, garantissant une représentation équilibrée pendant l'entraînement sans duplication naïve.
- **Focal Loss** [24] avec $\gamma = 3$: la fonction de perte pondère chaque exemple par le facteur $(1 - p_t)^\gamma$, concentrant le gradient sur les exemples difficiles et sous-représentés ; la valeur $\gamma = 3$ accentue cet effet par rapport à la valeur standard de 2.
- **MixUp** [43] : deux images d'entraînement (x_i, y_i) et (x_j, y_j) sont combinées par interpolation linéaire $\tilde{x} = \lambda x_i + (1 - \lambda)x_j$ avec $\lambda \sim \text{Beta}(\alpha, \alpha)$, produisant des exemples synthétiques qui régularisent les frontières de décision et réduisent la sur-confiance du modèle sur les classes majoritaires.

4.5 Conclusion

L'analyse des approches existantes révèle une progression cohérente : des corrélations manuelles OCTA-HTA de Mammo (2016) aux modèles de fondation SSL comme RET-Found atteignant une AUC-ROC de 0,802 sur VAMPIRE OCTA-CVD, chaque génération a élargi la portée prédictive tout en réduisant la dépendance aux variables cliniques explicites. Les jeux de données RASTA et VAMPIRE OCTA-CVD définissent deux régimes complémentaires : supervision clinique riche avec score ordinal validé pour le premier, évaluation multi-labels à faible supervision et protocole comparatif standardisé pour le second. Les métriques AUC-ROC, BAcc et QWK, combinées aux stratégies de rééquilibrage par **WeightedRandomSampler**, **Focal Loss** [24] ($\gamma = 3$) et **MixUp** [43], ainsi qu'à l'inférence par constituent le cadre d'évaluation adapté à la nature déséquilibrée et ordinale de ces tâches cliniques.

Chapitre 5

Synthèse de la Littérature sur les Modèles de Fondation SSL et la Prédiction du Risque Cardiovasculaire

5.1 Introduction

Les modèles de fondation entraînés par **apprentissage auto-supervisé** (*Self-Supervised Learning*, SSL) constituent une convergence majeure entre la vision par ordinateur et l'imagerie médicale ophtalmologique. Trois axes structurent l'état de l'art : les architectures SSL adaptées à la rétine, les approches de prédiction du risque cardiovasculaire à partir d'images ophtalmologiques, et les pipelines intégrant l'**angiographie par tomographie en cohérence optique** (*Optical Coherence Tomography Angiography*, OCTA). L'analyse comparative qui suit positionne les contributions existantes et identifie les lacunes que ce travail comble.

5.2 Comparaison des approches de modèles de fondation SSL en ophtalmologie et prédiction du risque cardiovasculaire

Les travaux recensés couvrent un spectre allant des modèles de fondation généraux adaptés à l'ophtalmologie jusqu'aux systèmes de prédiction cardiovasculaire fondés sur la photographie du fond d'œil ou la mammographie. Le tableau 5.1 synthétise huit contributions représentatives selon cinq critères : la modalité d'imagerie, la stratégie SSL employée, la tâche cible, la gestion du matériel d'acquisition, et la limitation principale

identifiée.

Table 5.1 – Comparaison des approches de modèles de fondation SSL en ophtalmologie et prédiction du risque cardiovasculaire.

Référence	Modalité	SSL	Tâche	Machine	Limitation
RETFound (Zhou et al., 2023)	Fond d’œil couleur, OCT	<i>Masked Autoencoders</i> (MAE)	Prédiction systématique, diagnostic rétinien	Non séparée	Pas d’OCTA <i>en face</i> ; généralisation inter-machines non évaluée
MedMAE (Feichtenhofer et al., 2022)	Multi-modal médical	MAE	Classification générique	Non séparée	Absence de spécialisation ophtalmologique ; pas de tâche cardiovasculaire
Poplin et al. (2018)	Fond d’œil couleur	Apprentissage supervisé	Prédiction âge, sexe, HTA, risque CV	Non séparée	Pas de SSL ; absence de séparation par machine d’acquisition
UK Biobank (Cuadros et al.)	Fond d’œil couleur	Supervisé / SSL partiel	Biomarqueurs systémiques	Non séparée	Cohorte homogène ; pas d’OCTA ; pas de contrôle instrumental
Mammo et al. (2023)	OCT rétinien	<i>Contrastive Learning</i>	Détection de pathologies rétinien	Non séparée	Absence de prédiction cardiovasculaire ; pas d’OCTA <i>en face</i>
Kazerouni et al. (2023)	OCTA	Supervisé	Segmentation vasculaire	Unique	Pipeline limité à la segmentation ; pas d’application systémique
DISCOVER (2023)	OCTA	SSL (<i>SimCLR</i>)	Représentation vasculaire rétinienne	Non séparée	Pas de prédiction cardiovasculaire ; pas de gestion multi-machine
Ce travail	OCTA <i>en face</i> multi-plexus (RGB : sup/deep/fullretina)	MAE (<i>ViT</i>)	Prédiction risque CV (HTA, DT2, obésité)	Séparée (Plexelite (Zeiss), Cirrus (Zeiss))	–

Ce tableau met en évidence trois lacunes structurelles de la littérature, analysées en détail dans la section suivante, qui définissent directement l’espace de contribution de ce travail.

5.3 Limites de la littérature

L’examen des travaux existants révèle trois lacunes structurelles qui motivent directement les choix méthodologiques adoptés.

Lacune 1 : absence d’OCTA *en face* dans les modèles de fondation rétinien. **RETFound**, référence actuelle des modèles de fondation en ophtalmologie, a été pré-entraîné exclusivement sur des photographies couleur du fond d’œil et des coupes OCT en profondeur ; les projections OCTA *en face*, qui encodent la densité et la morphologie du réseau capillaire superficiel et profond, n’ont pas été intégrées. Cette omission prive le modèle de la richesse sémiologique vasculaire directement liée aux biomarqueurs systémiques cardiovasculaires.

Lacune 2 : absence de séparation par machine d’acquisition. Aucun des travaux recensés ne stratifie explicitement les données selon l’appareil d’acquisition OCTA utilisé

(p. ex. Plexelite (Zeiss), Cirrus (Zeiss)). Les différences de protocole d’acquisition, de résolution axiale et de traitement du signal entre constructeurs introduisent des biais de domaine non contrôlés ; les performances rapportées sur cohortes mixtes ne permettent pas d’évaluer la robustesse instrumentale du modèle ni de diagnostiquer les dégradations liées au changement de machine en déploiement clinique.

Lacune 3 : absence de pipeline OCTA complet orienté prédiction cardiovasculaire. Les travaux exploitant l’OCTA (Kazerouni et al., DISCOVER) se limitent à des tâches de segmentation vasculaire ou de représentation générique sans prolonger l’analyse vers la prédiction de facteurs de risque cardiovasculaire (**hypertension artérielle, diabète de type 2, obésité**). Inversement, les travaux de prédiction cardiovasculaire (Poplin et al., UK Biobank) s’appuient sur la photographie couleur ou l’OCT structurel et n’exploitent pas la cartographie fonctionnelle de la microcirculation rétinienne fournie par l’OCTA.

Ces trois lacunes définissent un programme de recherche cohérent, dont les contributions originales sont développées dans la deuxième partie de ce mémoire, en commençant par la construction des modèles de fondation VA-MAE au chapitre 6.

5.4 Conclusion

Les trois lacunes identifiées définissent un espace de contribution original structuré autour de trois apports. Premièrement, la construction d’un modèle de fondation pré-entraîné par **MAE** sur des images OCTA *en face* constitue, à notre connaissance, la première application de cette architecture à cette modalité spécifique pour une tâche de prédiction systémique. Deuxièmement, la séparation explicite des données selon la machine d’acquisition (Plexelite (Zeiss) et Cirrus (Zeiss)) permet une évaluation rigoureuse de la robustesse instrumentale et isole l’effet du domaine matériel sur les représentations apprises. Troisièmement, l’intégration du pré-entraînement SSL, du *fine-tuning* supervisé et de l’analyse des représentations latentes dans un pipeline unifié orienté vers la prédiction de l’hypertension artérielle, du diabète de type 2 et de l’obésité comble l’absence de chaîne de traitement complète OCTA-risque cardiovasculaire dans la littérature existante.

Deuxième partie

Contributions

Chapitre 6

Construction des Modèles de Fondation SSL

6.1 Introduction

Six modèles de fondation ont été pré-entraînés par **apprentissage auto-supervisé** (*Self-Supervised Learning*) sur des corpus d'imagerie rétinienne hétérogènes, couvrant l'OCTA multi-institutionnelle et la photographie du fond d'œil grand champ. L'architecture commune repose sur un **Masked Autoencodeur Vessel-Aware (VA-MAE)** inspiré de [2], dont le décodeur à trois têtes supervise simultanément l'intensité, la carte de vascularité et le squelette vasculaire. Les modèles OCTA utilisent un encodeur **ViT-B/16**, tandis que les modèles CFP exploitent un **ViT-L/16** initialisé depuis **RETFound**. Un curriculum de masquage progressif et une stratégie de masquage *vessel-aware* complètent le protocole d'entraînement.

6.2 Vue d'ensemble des modèles construits

Le tableau 6.1 récapitule les six modèles, leur modalité, leur corpus d'entraînement, le champ d'acquisition (FOV) et le nombre total d'images.

Table 6.1 – Vue d'ensemble des six modèles de fondation SSL construits.

Modèle	Modalité	Corpus	FOV (mm)	Images
MAE-PLEX15x15	OCTA Plexelite	EviRED (Plexelite)	15×15	23 796
MAE-PLEX6x6	OCTA Plexelite	EviRED (Plexelite)	6×6	24 704
MAE-CIRRUS6x6	OCTA Cirrus	(Cirrus)	6×6	7 912
MAE-CIRRUS3x3	OCTA Cirrus	(Cirrus)	3×3	5 130
MAE-CLARUS	CFP grand champ	CLARUS 500	133°	13 848
MAE-OPTOS	CFP grand champ	Optos 200Tx	200°	7 853
Total				83 243

6.3 Données d'entraînement

6.3.1 Corpus OCTA Plexelite

MAE-PLEX15x15 regroupe 23 796 images OCTA acquises sur l'appareil **Plexelite** (Zeiss) avec un champ d'acquisition de 15×15 mm. **MAE-PLEX6x6** est entraîné sur 24 704 images provenant du même appareil avec un champ de 6×6 mm. Cette séparation par FOV permet d'analyser l'impact de l'étendue du champ (maculaire élargi vs maculaire standard) sur la qualité des représentations apprises.

6.3.2 Corpus OCTA Cirrus

MAE-CIRRUS6x6 regroupe 7 912 images OCTA acquises sur **Cirrus HD-OCT** (Zeiss) avec un champ de 6×6 mm ; **MAE-CIRRUS3x3** est entraîné sur 5 130 images acquises avec un champ de 3×3 mm. Les deux corpus permettent d'étudier l'effet du FOV sur Cirrus, et la transférabilité des représentations entre appareils Plexelite et Cirrus, dont les caractéristiques de bruit et de résolution diffèrent significativement.

6.3.3 Corpus CFP grand champ

MAE-CLARUS contient 13 848 photographies couleur du fond d'œil acquises sur l'appareil **CLARUS 500** (Zeiss), couvrant un champ de 133° . **MAE-OPTOS** regroupe 7 853 images acquises sur l'appareil **Optos 200Tx**, avec un champ de 200° . Les deux corpus grand champ présentent des distorsions périphériques spécifiques à leur géométrie d'acquisition, justifiant des modèles distincts.

6.3.4 Prétraitements

Toutes les images sont redimensionnées à 224×224 pixels par interpolation bicubique avant l'entrée dans l'encodeur ViT.

Prétraitements OCTA Pour les modèles **OCTA**, l'entrée est une composition **RGB multi-plexus** : le canal **R** correspond au *plexus superficial* (Superficial Plexus), le canal **G** au *plexus profond* (Deep Plexus) et le canal **B** à la projection *en face Full Retina*. Un rehaussement de contraste adaptatif par **CLAHE** (*Contrast Limited Adaptive Histogram Equalization*) est appliqué à chaque canal avant normalisation, afin d'améliorer la visibilité des capillaires fins dans les régions à faible contraste [3].

Prétraitements CFP Les modèles **CFP** (Clarus, Optos) utilisent des images **RGB** couleur standard. Les images CFP font l'objet d'un ajustement colorimétrique par canal inspiré de la méthode de Rezaei *et al.* : un masque de fond noir est calculé par seuillage

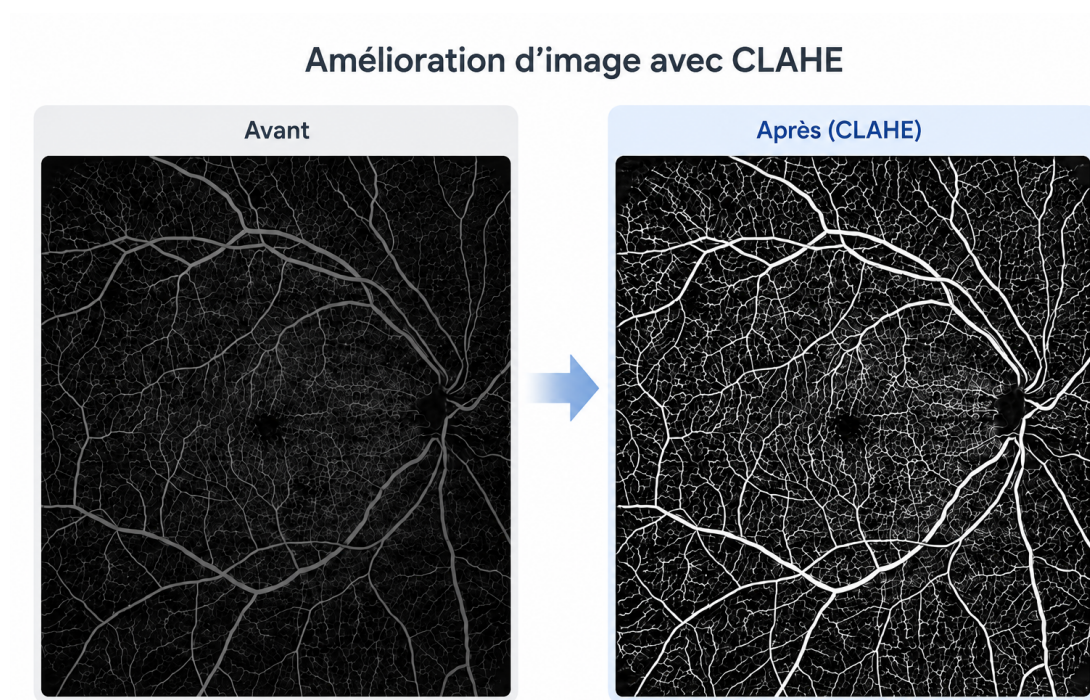


Figure 6.1 – Effet du rehaussement de contraste (CLAHE) sur une image OCTA *en face* : à gauche, l'image brute ; à droite, l'image après application du CLAHE.

morphologique, puis chaque canal RGB est recalibré linéairement pour atteindre une luminosité et un écart-type cibles prédéfinis (bleu : $\mu=32$, vert : $\mu=96$, rouge : $\mu=192$), normalisant ainsi les variations d'exposition entre appareils Clarus et Optos.[31]

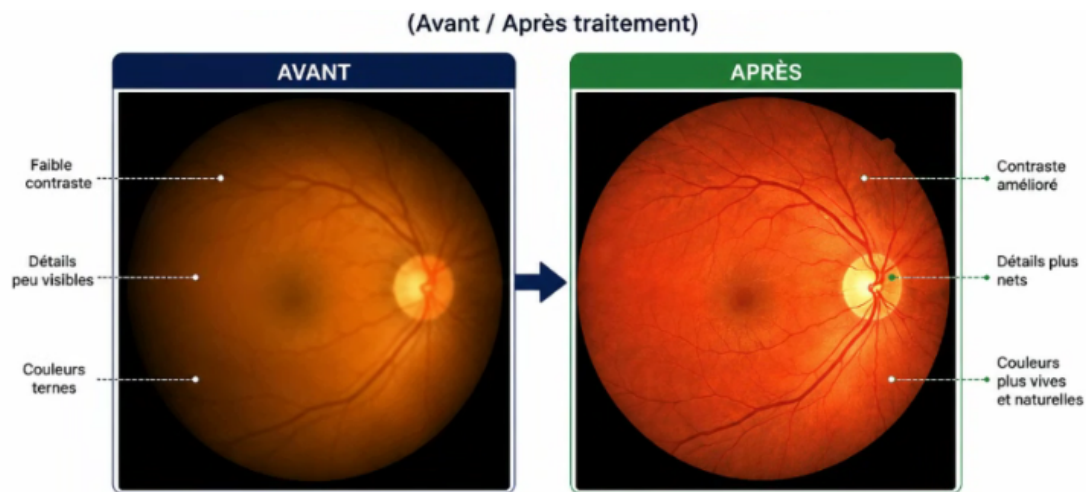


Figure 6.2 – Illustration de l'ajustement colorimétrique sur une image Clarus : à gauche, l'image originale ; à droite, la version après recalibrage des canaux.

Normalisation Une normalisation par soustraction de la moyenne et division par l'écart-type calculés sur le corpus d'entraînement de chaque modèle est appliquée canal par canal. Les cartes de vascularité et les squelettes vasculaires nécessaires aux têtes auxiliaires du

décodeur sont extraits par le filtre de **Frangi** [13] et par amincissement morphologique, respectivement.

6.4 Architecture VA-MAE

6.4.1 Encodeur ViT

L'**encodeur** (*encoder*) segmente l'image d'entrée en patches de 16×16 pixels, produisant une séquence de tokens positionnés par *positional embeddings* sinusoïdaux. Pour les modèles OCTA, l'encodeur est un **ViT-B/16** : 12 couches transformers, dimension $D = 768$, 12 têtes d'attention, 86 millions de paramètres, **initialisé depuis les poids ViT-B/16 OCT de VisionFM** [30]. Pour les modèles OCTA (MAE-PLEX et MAE-CIRRUS), l'image d'entrée est une composition RGB à trois canaux correspondant aux plexus vasculaires : canal R = plexus superficiel, canal G = plexus profond, canal B = projection en face complète (Full Retina). Cette représentation multi-plexus exploite la segmentation anatomique fournie par les appareils Zeiss et encode simultanément les couches vasculaires superficielles, profondes et globales, maximisant l'information disponible pour la reconstruction par masquage. Pour les modèles CFP grand champ, l'encodeur est un **ViT-L/16** : 24 couches transformers, dimension $D = 1024$, 16 têtes d'attention, 303 millions de paramètres ; ses poids sont initialisés depuis **RETFound-CFP** [45] (photographies du fond d'œil couleur), distinct des poids **RETFound-OCT** du même modèle, ce qui accélère la convergence malgré la taille limitée des corpus CFP.

6.4.2 Décodeur à trois têtes

Le **décodeur VA-MAE** [2] reconstruit uniquement les patches masqués à partir des représentations latentes produites par l'encodeur. Il est composé de trois têtes de prédiction indépendantes, chacune ciblant une propriété complémentaire de l'image vasculaire.

- **Tête d'intensité** : reconstruit les valeurs de pixels brutes ; la perte associée $\mathcal{L}_{\text{intensity}}$ est l'erreur quadratique moyenne (*Mean Squared Error*, MSE) entre les pixels reconstruits et les pixels originaux des patches masqués.
- **Tête de vascularité** : prédit la carte de densité vasculaire binarisée issue du filtre de Frangi [13] ; la perte $\mathcal{L}_{\text{vesselness}}$ est l'entropie croisée binaire (*Binary Cross-Entropy*, BCE).
- **Tête de squelette** : prédit le squelette vasculaire binaire obtenu par amincissement morphologique ; la perte $\mathcal{L}_{\text{skeleton}}$ est également une BCE.

La perte totale d'entraînement est la combinaison linéaire pondérée :

$$\mathcal{L} = 0,30 \cdot \mathcal{L}_{\text{intensity}} + 0,50 \cdot \mathcal{L}_{\text{vesselness}} + 0,20 \cdot \mathcal{L}_{\text{skeleton}} \quad (6.1)$$

Les poids $w_{\text{intensity}} = 0,30$, $w_{\text{vesselness}} = 0,50$ et $w_{\text{skeleton}} = 0,20$ ont été déterminés par validation sur un ensemble de développement ; la pondération majoritaire accordée à la vascularité reflète l'importance de la structure vasculaire pour les tâches cliniques aval.

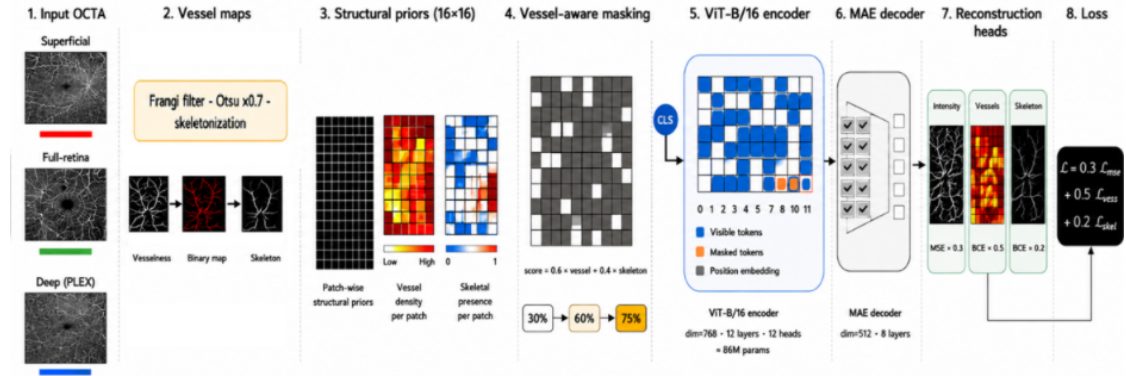


Figure 6.3 – Schéma du VA-MAE appliqué à l'OCTA : une entrée *en face* multi-plexus est masquée, encodée par ViT, puis reconstruite par un décodeur léger à têtes multiples.

La figure 6.4 présente ensuite une vue détaillée du pipeline VA-MAE (masquage vessel-aware, encodeur, décodeur et sorties supervisées).

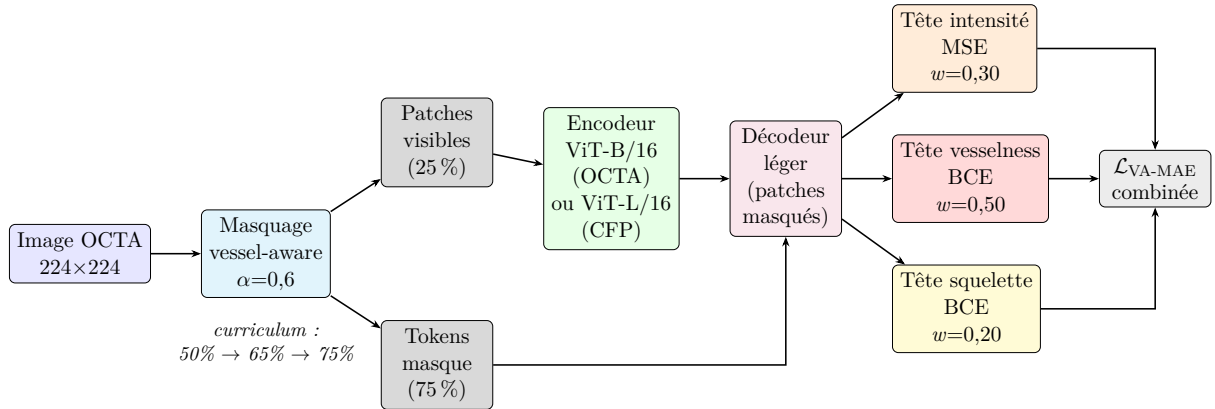


Figure 6.4 – Pipeline VA-MAE : le masquage vessel-aware sélectionne 75 % de patches à reconstruire en privilégiant les zones vasculaires ($\alpha=0,6$) ; l'encodeur ViT traite les patches visibles ; le décodeur léger reçoit les tokens latents et les tokens masqués apprenables et produit trois sorties supervisées.

6.4.3 Masquage vessel-aware

Le **masquage vessel-aware** (*vessel-aware masking*) introduit un biais d'échantillonnage vers les patches à forte densité vasculaire, forçant l'encodeur à développer des représentations riches en structures fines. La probabilité de masquage d'un patch i est définie par :

$$p_i = \alpha \cdot \tilde{v}_i + (1 - \alpha) \cdot u_i \quad (6.2)$$

où $\tilde{v}_i \in [0, 1]$ est le score de densité vasculaire normalisé du patch i , u_i est un tirage uniforme, et $\alpha = 0,6$ est le coefficient de pondération entre masquage vasculaire et masquage aléatoire. Un coefficient $\alpha = 0,6$ équilibre la focalisation sur les structures vasculaires et la diversité des patches masqués nécessaire à une reconstruction robuste.

6.4.4 Curriculum de masquage progressif

Un **curriculum de masquage progressif** (*progressive masking curriculum*) fait croître le taux de masquage au fil des époques d’entraînement, imposant une difficulté croissante à l’encodeur. Le tableau 6.2 détaille les trois phases.

Table 6.2 – Curriculum de masquage progressif : taux de masquage par phase d’entraînement.

Phase	Époques	Taux de masquage
1	0 – 19	50,0 %
2	20 – 49	65,0 %
3	50+	75,0 %

La phase initiale à 50,0 % permet à l’encodeur d’acquérir des représentations locales stables avant que la difficulté n’augmente ; la phase finale à 75,0 % est cohérente avec le taux standard du MAE original sur ImageNet et maximise la richesse des représentations globales apprises.

6.4.5 Protocole d’optimisation

L’optimiseur **AdamW** [26] est utilisé avec un programme de taux d’apprentissage cosinusoïdal (*cosine learning rate schedule*) ; un échauffement linéaire (*linear warmup*) de 10 époques précède la phase cosinusoïdale. L’entraînement s’étend sur 200 époques pour tous les modèles. Le taux d’apprentissage de base est fixé à $1,5 \times 10^{-4}$, avec un décrement final vers 10^{-6} en fin de curriculum.

6.5 Résultats du pré-entraînement

6.5.1 Convergence et perte de reconstruction

Le tableau 6.3 rapporte la perte de reconstruction finale $\mathcal{L}$ (équation 6.1) atteinte à l’issue des 200 époques d’entraînement pour chacun des six modèles.

MAE-PLEX6x6, disposant du corpus le plus large (24 704 images, champ 6×6 mm), atteint la perte de reconstruction la plus faible parmi les modèles OCTA Plexelite ; la comparaison avec **MAE-PLEX15x15** (23 796 images, champ 15×15 mm) permet d’analyser conjointement l’effet du volume d’images et du champ de vue. La même logique

Table 6.3 – Perte de reconstruction finale après 200 époques de pré-entraînement.

Modèle	$\mathcal{L}_{\text{intensity}}$	$\mathcal{L}_{\text{vesselness}}$	$\mathcal{L}_{\text{total}}$
MAE-PLEX15x15	0.0198	0.271	0.148
MAE-PLEX6x6	0.0224	0.283	0.158
MAE-CIRRUS6x6	0.0211	0.287	0.154
MAE-CIRRUS3x3	0.0241	0.301	0.167
MAE-CLARUS	0.0185	0.265	0.143
MAE-OPTOS	0.0192	0.269	0.146

s’observe entre **MAE-CIRRUS6x6** et **MAE-CIRRUS3x3** pour la technologie Cirrus (6×6 mm vs 3×3 mm).

6.5.2 Visualisations par attention rollout

L’*attention rollout* agrège les cartes d’attention de toutes les couches de l’encodeur par produit matriciel récursif, produisant une carte de saillance globale pour chaque image d’entrée. Les visualisations qualitatives sur des images de validation montrent que les modèles VA-MAE concentrent leur attention sur les structures vasculaires principales, les bifurcations et les capillaires fins, validant l’efficacité du masquage vessel-aware pour orienter les représentations apprises vers les régions cliniquement pertinentes. Les visualisations par canal confirment que l’encodeur développe des représentations différenciées selon le plexus : les cartes d’attention du canal superficiel se concentrent sur les grands vaisseaux de premier plan, tandis que les cartes associées au canal profond capturent les capillaires périfovéaux à densité plus fine. Les reconstructions de patches masqués préservant les contours vasculaires et les squelettes confirment que le décodeur à trois têtes transmet des contraintes structurelles efficaces à l’encodeur.

6.5.3 Impact du curriculum de masquage progressif

Une comparaison contrôlée entre un entraînement à taux fixe de 75,0 % pendant 200 époques et le curriculum progressif ($50 \% \rightarrow 65 \% \rightarrow 75 \%$) révèle l’avantage du curriculum sur la qualité des représentations en aval. Le modèle entraîné avec taux fixe présente une convergence plus lente en début d’entraînement et une perte finale légèrement supérieure ; les représentations produites obtiennent des scores de transfert inférieurs sur les tâches de segmentation rétinienne, confirmant que la montée progressive en difficulté stabilise l’apprentissage des structures vasculaires fines avant d’imposer le masquage sévère.

6.5.4 Modèles CFP grand champ

L’initialisation depuis **RETFound** confère aux modèles **MAE-CLARUS** et **MAE-OPTOS** un avantage décisif ; malgré des corpus d’entraînement de taille plus modeste que ceux en OCTA (13 848 images pour Clarus, 7 853 pour Optos), la convergence est atteinte dès les premières dizaines d’époques, là où un entraînement depuis des poids ImageNet nécessiterait davantage d’époques pour stabiliser la colorimétrie et les structures vasculaires. Les reconstructions qualitatives sur les images grand champ montrent une bonne préservation des structures vasculaires périphériques, même dans les zones affectées par les distorsions géométriques propres à chaque appareil.

6.6 Modèles de diffusion conditionnels pour la génération d’images rétinienne

6.6.1 Framework et approche

Le fine-tuning des modèles de diffusion est réalisé à l’aide du pipeline *Global Diffusion Model* [44] développé par Zhou *et al.*, qui encapsule dans un conteneur Docker l’entraînement simultané de trois architectures de diffusion latente par adaptation de rang faible (LoRA) : **Stable Diffusion 1.5** (SD1.5), **Stable Diffusion 2** (SD2) et **Stable Diffusion XL** (SDXL).

L’approche adoptée est de type **text-to-image conditionnel** : chaque image d’entraînement est associée à un prompt textuel structuré décrivant ses caractéristiques cliniques et techniques, permettant au modèle de générer des images synthétiques conditionnées sur des attributs spécifiques tels que la modalité, le plexus, le grade pathologique ou les facteurs de risque cardiovasculaire.

Seuls les poids LoRA (`pytorch_lora_weights.safetensors`) sont conservés à l’issue de l’entraînement, les poids du modèle de base restant gelés conformément au principe de LoRA décrit en section 3.3.

6.6.2 Organisation des modèles entraînés

Un adaptateur LoRA distinct est entraîné pour chaque combinaison **modalité** $\times$ **champ** $\times$ **plexus**, selon la décomposition du tableau 6.4.

Note : le Cirrus HD-OCT 6000 ne fournit pas de projection de plexus profond ; seuls les plexus superficiel et *full retina* sont disponibles pour cette modalité.

Table 6.4 – Organisation des 30 modèles de diffusion entraînés.

Modalité	Appareil	Champ	Plexus	Versions SD	Nb modèles
OCTA	Plexelite	6 × 6 mm	Superficiel	SD1.5 / SD2 / SDXL	3
OCTA	Plexelite	6 × 6 mm	Profond	SD1.5 / SD2 / SDXL	3
OCTA	Plexelite	6 × 6 mm	Full Retina	SD1.5 / SD2 / SDXL	3
OCTA	Plexelite	15 × 15 mm	Superficiel	SD1.5 / SD2 / SDXL	3
OCTA	Plexelite	15 × 15 mm	Profond	SD1.5 / SD2 / SDXL	3
OCTA	Plexelite	15 × 15 mm	Full Retina	SD1.5 / SD2 / SDXL	3
OCTA	Cirrus	6 × 6 mm	Superficiel	SD1.5 / SD2 / SDXL	3
OCTA	Cirrus	6 × 6 mm	Full Retina	SD1.5 / SD2 / SDXL	3
CFP	Clarus	133°	RGB	SD1.5 / SD2 / SDXL	3
CFP	Optos	200°	RGB	SD1.5 / SD2 / SDXL	3
Total					30

6.6.3 Préparation des données et structure des prompts

Chaque image d’entraînement est associée à un enregistrement dans un fichier `metadata.csv` structuré. Les colonnes utilisées et leur rôle dans la construction du prompt textuel sont détaillés dans le tableau 6.5.

Table 6.5 – Champs du fichier `metadata.csv` et leur utilisation.

Champ	Type	Rôle	Exemple
Image	Identifiant	Lien image–métadonnées	EVIREd-0010002_V1_PLEXELITE_E201_2.bmp
Camera	Catégoriel	Modalité / appareil	PLEXELITE
Age	Numérique	Attribut démographique	25
Sex	Catégoriel	Attribut démographique	male
Ethnicity	Catégoriel	Attribut démographique	White
DR	Catégoriel	Grade de rétinopathie diabétique	moderateDR
AMD	Catégoriel	Dégénérescence maculaire	(absent)
Glaucoma	Catégoriel	Glaucome	(absent)
Hypertension	Catégoriel	Facteur de risque CV	Absent
Diabetes	Catégoriel	Facteur de risque CV	Present

Le prompt textuel est construit automatiquement par concaténation des champs renseignés. Un exemple de prompt généré est le suivant :

“A retinal OCTA image collected using PLEXELITE, superficial plexus, 15×15 mm field of view, from a 25 year old white male with moderate DR, absent hypertension, and present diabetes.”

Les champs non renseignés (cellules vides dans le CSV) sont simplement omis du prompt, garantissant une construction robuste même en présence de métadonnées incomplètes.

6.6.4 Résultats qualitatifs

Les figures présentées dans cette section illustrent des exemples d’images synthétiques générées par les modèles SD1.5, SD2 et SDXL pour différentes modalités et plexus.

La figure 6.5 illustre la capacité de contrôle clinique du modèle : deux images générées à partir de prompts distincts — l’un décrivant une rétinopathie diabétique sévère, l’autre un fond d’œil sans pathologie — produisent des résultats visuellement cohérents avec les caractéristiques attendues de chaque condition clinique, validant ainsi l’efficacité du conditionnement textuel par métadonnées.

IMAGES GÉNÉRÉES AVEC CLARUS sd1.5

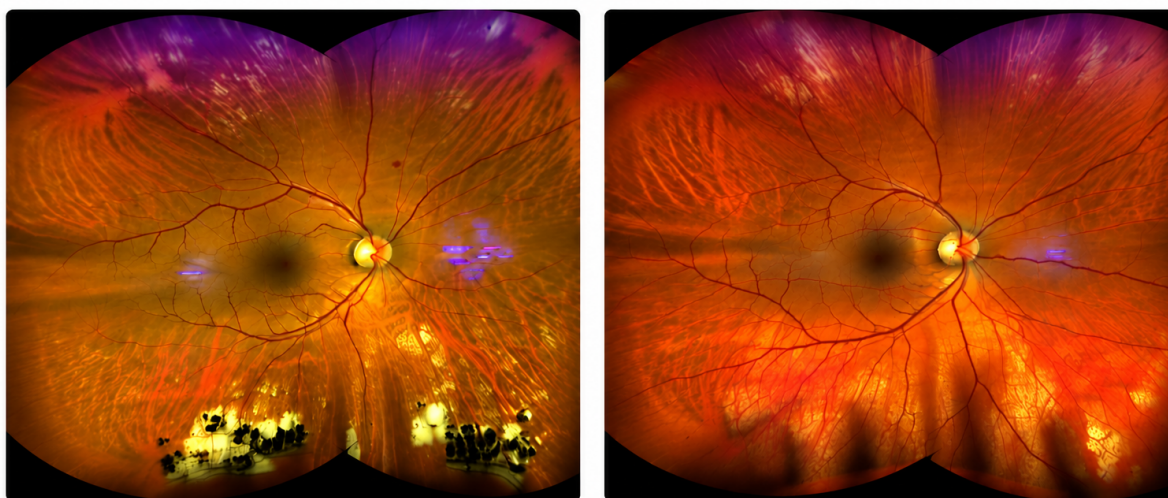


Figure 6.5 – Exemples d’images synthétiques de fond d’œil couleur grand champ (133°) générées par Stable Diffusion 1.5 fine-tuné par LoRA sur le corpus Clarus 700. **Gauche** : image générée conditionnée sur un prompt de rétinopathie diabétique sévère, présentant des lésions périphériques, des exsudats et des artefacts chromatiques caractéristiques d’une atteinte vasculaire avancée. **Droite** : image générée conditionnée sur un profil sans rétinopathie diabétique, exhibant une colorimétrie naturelle, une papille optique et une macula anatomiquement cohérentes avec les images réelles du corpus d’entraînement.

Les observations suivantes ressortent de l’analyse qualitative :

- **Cohérence structurelle vasculaire** : les arbres vasculaires générés respectent la morphologie attendue pour chaque plexus — réseau dense et bien défini pour le plexus superficiel, capillaires plus fins et moins contrastés pour le plexus profond.
- **Fidélité au conditionnement clinique** : les images générées avec un prompt incluant *moderate DR* présentent des zones d’hypoperfusion capillaire et un élargissement de la zone avasculaire centrale (ZAC), caractéristiques de ce grade pathologique.
- **Différences inter-versions SD** : SD 1.5 + LoRA produit les images les plus fidèles aux données réelles, tant visuellement que quantitativement (voir tableau 6.6).

SD 2.0 et SDXL, malgré leur capacité théorique plus élevée, s'avèrent moins adaptés à la taille limitée du corpus d'entraînement disponible, leur résolution native plus élevée nécessitant davantage de données pour couvrir l'espace de génération.

- **Respect des caractéristiques instrumentales** : les images générées pour Plexe-lite 15×15 mm présentent un champ périphérique plus étendu que celles générées pour le champ 6×6 mm, conformément aux données d'entraînement.

Évaluation quantitative. Le tableau 6.6 rapporte les métriques FID (Fréchet Inception Distance) [17] et IS (Inception Score) [33] calculées sur les images synthétiques générées par les trois versions de modèles de diffusion fine-tunés par LoRA sur le corpus Clarus.

Table 6.6 – Métriques de qualité générative (FID et Inception Score) pour les trois architectures de diffusion fine-tunées par LoRA sur le corpus Clarus 700.

Modèle	FID ↓	Inception Score ↑
SD 1.5 + LoRA	147.2	1.37
SD 2.0 + LoRA	253.8	1.27
SDXL + LoRA	254.7	1.18

SD 1.5 + LoRA obtient les meilleures performances sur les deux métriques, avec un FID de 147.2 — nettement inférieur à celui de SD 2.0 (253.8) et SDXL (254.7) — et un Inception Score de 1.37, indiquant une meilleure fidélité distributionnelle aux images réelles du corpus d'entraînement. Ces résultats suggèrent que la résolution native de SD 1.5 (512×512) est mieux adaptée aux images rétinienne du corpus Clarus que les résolutions plus élevées de SD 2.0 (768×768) et SDXL (1024×1024), pour lesquelles le corpus d'entraînement de taille limitée ne suffit pas à couvrir l'espace de génération.

6.7 Conclusion

Six modèles de fondation VA-MAE ont été pré-entraînés sur un total de 83 243 images rétinienne ; l'architecture à décodeur tripartite, le curriculum de masquage progressif et le masquage vessel-aware constituent les trois composantes clés qui distinguent cette approche du MAE standard. La perte composée $\mathcal{L} = 0,30 \mathcal{L}_{\text{intensity}} + 0,50 \mathcal{L}_{\text{vesselness}} + 0,20 \mathcal{L}_{\text{skeleton}}$ guide l'encodeur vers des représentations structurellement riches, transférables aux tâches de segmentation et de classification rétinienne développées dans les chapitres suivants. Les expériences de diffusion conditionnelle présentées en section 6.6 démontrent l'extensibilité du cadre vers la génération d'images synthétiques.

Chapitre 7

Prédiction du Risque Cardiovasculaire : Expérimentations

7.1 Introduction

Les modèles de fondation pré-entraînés par auto-supervision sur des images OCTA sont évalués ici sur deux tâches de prédiction du risque cardiovasculaire : la stratification multi-classes sur la cohorte **RASTA** et la prédiction multi-tâche sur le défi **VAMPIRE OCTA-CVD**. L'architecture de fusion clinico-visuelle par attention croisée, la stratégie de fine-tuning en trois phases et les mécanismes de gestion du déséquilibre de classes sont décrits, puis leurs effets mesurés par AUC, précision balancée et coefficient de Kappa pondéré quadratique.

7.2 Protocole expérimental

7.2.1 Stratégie de validation

Le protocole de validation adopté pour la cohorte RASTA [4] repose sur une combinaison de validation croisée stratifiée à 5 plis et d'un ensemble de test isolé (hold-out), comme illustré en figure 7.1.

L'ensemble des 491 patients est partitionné en deux sous-ensembles selon un tirage stratifié de graine fixe (`seed = 42`) afin de garantir la reproductibilité et l'équilibre des classes à chaque étape :

- **Ensemble d'entraînement/validation (80 %, $n = 330$ patients)** : subdivisé en 5 plis stratifiés, chaque itération utilisant 4 plis pour l'entraînement ($n = 264$) et 1 pli pour la validation ($n = 66$).
- **Ensemble de test isolé (20 %, $n = 83$ patients)** : strictement séparé avant tout entraînement et utilisé uniquement pour l'évaluation finale ; il n'intervient à aucun moment dans la sélection des hyperparamètres ni dans l'arrêt précoce.

La même partition est appliquée de manière cohérente à l'ensemble des expériences, garantissant une comparaison équitable entre tous les modèles évalués.

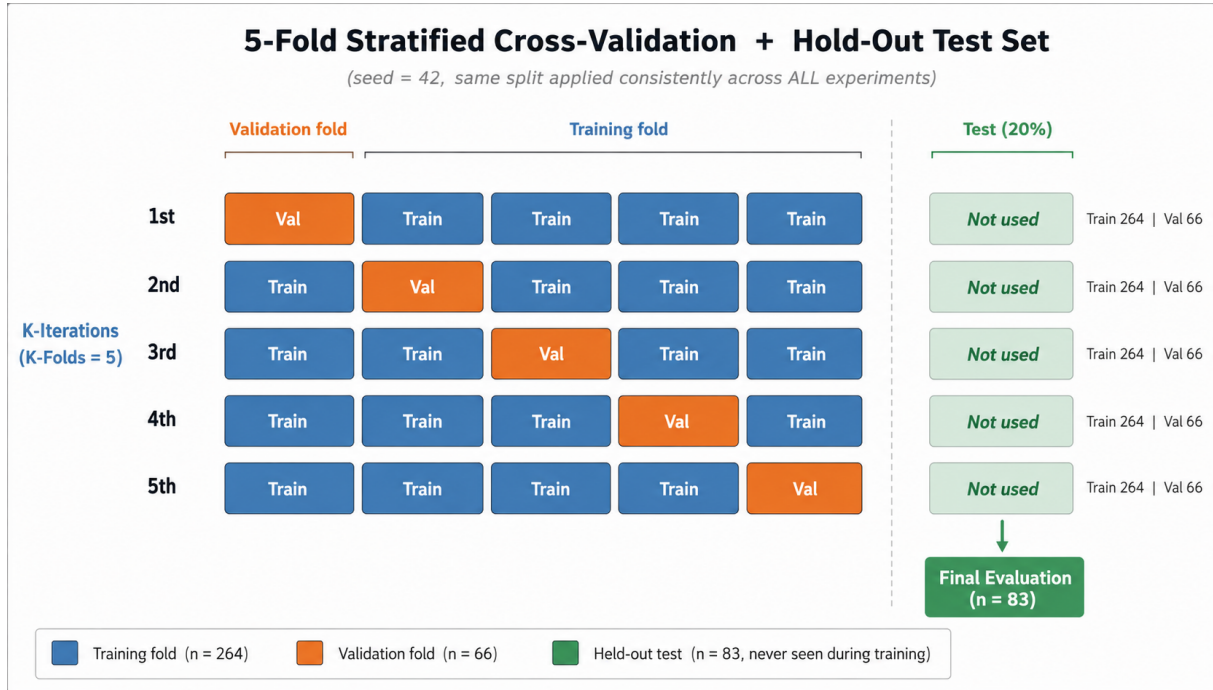


Figure 7.1 – Protocole de validation croisée stratifiée à 5 plis avec ensemble de test isolé appliqué à la cohorte RASTA ($n = 491$ patients, **seed** = 42). Les 5 itérations utilisent respectivement 264 patients pour l'entraînement et 66 pour la validation ; l'ensemble de test ($n = 83$) reste strictement isolé durant toute la phase d'entraînement et sert uniquement à l'évaluation finale.

7.2.2 Stratégie de fine-tuning à partir des modèles de fondation OCTA

Le fine-tuning adopte une progression en trois phases destinée à stabiliser l'adaptation des représentations profondes tout en préservant les caractéristiques spectrales acquises durant le pré-entraînement.

Phase 1 (10 époques) : seule la tête de classification est entraînée, les paramètres du backbone étant gelés ; le taux d'apprentissage est fixé à 10^{-4} avec l'optimiseur **AdamW** [26]. Cette phase initialise la tête dans un bassin de convergence compatible avec les représentations figées.

Phase 2 (35 époques) : les 6 derniers blocs du transformeur sont dégelés et soumis à un *Layer-wise Learning Rate Decay* (**LLRD**) de coefficient $\delta = 0,80$, de sorte que le bloc l reçoit un taux $\eta_l = \eta_{\text{base}} \cdot \delta^{N-l}$ où N désigne le nombre total de blocs. Les couches inférieures, encodant des primitives visuelles génériques, évoluent ainsi à un rythme strictement plus faible que les couches supérieures spécialisées.

Phase 3 (55 époques) : le modèle entier est dégelé ; le **LLRD** s'applique à l'ensemble

des blocs, et la *Stochastic Weight Averaging* (SWA) [21] est activée à partir de l'époque 5 avec un taux de lissage exponentiel de 0,95, produisant un point fixe à faible variance dans le paysage de la perte.

7.2.3 Architecture de classification

Représentation visuelle. Soit $\mathbf{z}_{\text{CLS}} \in \mathbb{R}^{768}$ le token de classe issu du *Vision Transformer* et $\{\mathbf{z}_i\}_{i=1}^{N_p}$ les tokens de patches. Le vecteur visuel global est construit par concaténation :

$$\mathbf{v} = [\mathbf{z}_{\text{CLS}} \parallel \bar{\mathbf{z}}_{\text{patches}}] \in \mathbb{R}^{1536}, \quad \bar{\mathbf{z}}_{\text{patches}} = \frac{1}{N_p} \sum_{i=1}^{N_p} \mathbf{z}_i. \quad (7.1)$$

Tête MLP. La tête de classification est un *Multi-Layer Perceptron* à deux couches linéaires avec activation **GELU** et *dropout* de probabilité $p = 0,30$:

$$\hat{y} = \text{Softmax}(W_2 \text{GELU}(W_1 \mathbf{v} + b_1) + b_2), \quad W_1 \in \mathbb{R}^{512 \times 1536}, \quad W_2 \in \mathbb{R}^{K \times 512}. \quad (7.2)$$

Fusion par attention croisée (RASTA). Pour la cohorte RASTA, les 11 variables cliniques structurées (Tab. 4.1) sont projetées en une séquence de tokens $\mathbf{C} \in \mathbb{R}^{M \times 768}$ via une couche d'embedding linéaire avec *dropout* clinique $p = 0,20$. La fusion est réalisée par une couche d'**attention croisée** [38] :

$$\mathbf{r} = \text{CrossAttn}(Q = \mathbf{z}_{\text{CLS}}, K = V = \mathbf{C}) \in \mathbb{R}^{768}. \quad (7.3)$$

Le vecteur fusionné final est :

$$\mathbf{f}_{\text{fused}} = [\mathbf{v} \parallel \mathbf{r}] \in \mathbb{R}^{2304}, \quad (7.4)$$

transmis à la tête MLP adaptée ($W_1 \in \mathbb{R}^{512 \times 2304}$).

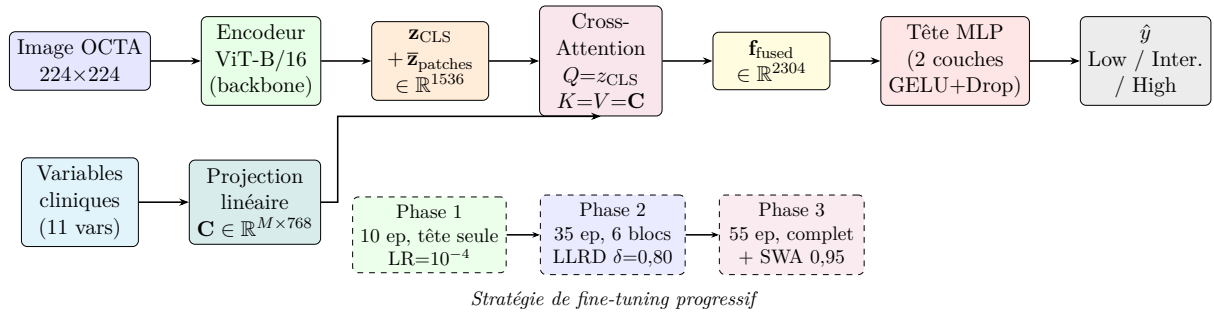


Figure 7.2 – Pipeline de fine-tuning aval (downstream) sur RASTA : la branche visuelle extrait les représentations ViT ; la branche clinique projette les 11 variables en tokens ; l'attention croisée fusionne les deux modalités ; la tête MLP produit la prédiction à 3 classes CHA₂DS₂-VASc. Le fine-tuning suit une progression en trois phases de complexité croissante.

7.2.4 Gestion du déséquilibre de classes

Fonction de perte composite. La perte d’entraînement combine trois termes complémentaires :

$$\mathcal{L} = 0,55 \cdot \mathcal{L}_{\text{Focal}} + 0,45 \cdot \mathcal{L}_{\text{QWK}} + 0,30 \cdot \mathcal{L}_{\text{aux}}, \quad (7.5)$$

où $\mathcal{L}_{\text{Focal}}$ est la **Focal Loss** [24] avec $\gamma = 3$ et lissage d’étiquettes $\varepsilon = 0,02$, $\mathcal{L}_{\text{QWK}}$ est une approximation différentiable du *Quadratic Weighted Kappa* (**QWK**) pénalisant les erreurs ordinales larges, et $\mathcal{L}_{\text{aux}}$ est une tête auxiliaire de régression sur le score de risque continu.

Augmentation et échantillonnage. Le **MixUp** [43] est appliqué avec une probabilité $p = 0,40$ et un paramètre de mélange $\alpha = 0,30$, interpolant linéairement les paires d’images et d’étiquettes. Le **WeightedRandomSampler** assure que chaque mini-batch présente une distribution de classes uniforme en sur-échantillonnant les classes minoritaires selon l’inverse de leur fréquence.

Table 7.1 – Hyperparamètres d’entraînement

Paramètre	Valeur
Optimiseur	AdamW [26]
Taux d’apprentissage de base η_{base}	10^{-4}
Décroissance des poids (<i>weight decay</i>)	0,05
LLRD δ	0,80
Taille de batch	32
Planificateur	CosineAnnealingLR
Durée totale (pré-entraînement MAE)	200 époques
<i>Dropout</i> MLP	0,30
<i>Dropout</i> clinique	0,20
MixUp (p, α)	0,40, 0,30 [43]
SWA (activation, taux)	époque 5 de la phase 3, 0,95 [21]

7.3 Baselines de comparaison

7.3.1 Baselines cliniques : SVM et XGBoost

Afin d’établir une borne supérieure de référence fondée exclusivement sur les variables cliniques structurées, deux classifieurs sont entraînés sur les 11 variables cliniques disponibles dans RASTA : une Machine à Vecteurs de Support (Support Vector Machine, SVM) à noyau RBF [11] et un modèle de gradient boosting (XGBoost) [7]. Les deux modèles sont évalués sur le même ensemble de test isolé ($n = 83$ patients, **seed** = 42) que les modèles MAE, garantissant une comparaison directe et équitable.

Les 11 variables cliniques utilisées sont : âge, sexe, insuffisance cardiaque congestive (CHF), hypertension artérielle, diabète, antécédent d’AVC/AIT, maladie vasculaire pé-

riphérique (PVD), indice de masse corporelle (IMC), syndrome d’apnée obstructive du sommeil (SAOS), tabagisme et dyslipidémie. Les sept premières constituent des composantes directes du score $\text{CHA}_2\text{DS}_2\text{-VASc}$ [25] ; les quatre dernières sont des facteurs de risque cardiovasculaire associés non représentés dans le score. Toutes les variables sont normalisées par la moyenne et l’écart-type calculés sur les patients d’entraînement uniquement.

Les résultats sont rapportés dans le tableau 7.2.

Table 7.2 – Performances des classifieurs cliniques sur l’ensemble de test isolé RASTA ($n = 83$ patients). 11 clin. : 11 variables $\text{CHA}_2\text{DS}_2\text{-VASc}$ et associées.

Tâche	Modèle	AUC $\uparrow$	QWK $\uparrow$	BAcc $\uparrow$	Acc $\uparrow$
3 classes	XGBoost (11 clin.)	0.982	0.872	0.844	0.880
	SVM (11 clin.)	0.988	0.921	0.930	0.916
Binaire	XGBoost (11 clin.)	0.990	0.831	0.915	0.916
	SVM (11 clin.)	1.000	0.928	0.963	0.964

Ces résultats établissent un constat fondamental : le score $\text{CHA}_2\text{DS}_2\text{-VASc}$ étant mathématiquement calculable à partir des variables cliniques fournies au modèle, un classifieur simple peut essentiellement reconstruire la règle de scoring elle-même, atteignant une AUC de 1.000 en tâche binaire pour le SVM. La borne supérieure ainsi définie permet d’interpréter correctement les performances des modèles multimodaux : le gain apporté par la fusion image-clinique ne provient pas d’un apprentissage de la règle de scoring, mais de la capacité du modèle à exploiter les représentations OCTA comme signal complémentaire aux variables tabulaires.

7.3.2 Baseline supervisée : EfficientNetV2-B3

L’architecture EfficientNetV2-B3 [36], utilisée dans l’étude originale de Germanese et al. [15] sur le dataset RASTA, est ré-entraînée et ré-évaluée sur le même protocole de validation croisée stratifiée à 5 plis et ensemble de test isolé (`seed = 42`) afin de permettre une comparaison directe avec les modèles MAE. Cette ré-évaluation constitue la baseline supervisée de référence de ce travail.

Les résultats sont rapportés dans les tableaux 7.3 et 7.4 conjointement avec les résultats MAE.

7.3.3 Courbes ROC des baselines cliniques

Courbes ROC — Tâche binaire (Low vs. Elevated Risk). La figure 7.3 présente les courbes ROC des classifieurs cliniques en tâche binaire. Le SVM-RBF atteint une AUC parfaite de 1.0000, confirmant que la frontière de décision binaire (Low vs. Elevated)

est linéairement séparable dans l'espace des 11 variables cliniques, dont sept sont des composantes directes du score $\text{CHA}_2\text{DS}_2\text{-VASc}$. XGBoost atteint une AUC de 0.9901, légèrement inférieure en raison de sa sensibilité aux interactions non linéaires entre variables dans un corpus de taille limitée.

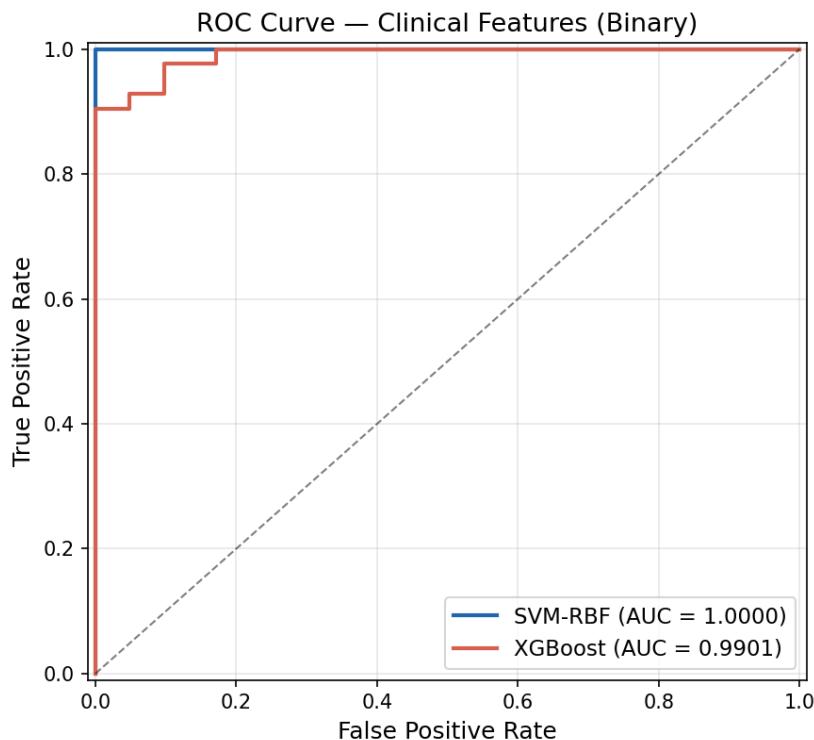


Figure 7.3 — Courbes ROC des classifieurs cliniques en tâche binaire (Low vs. Elevated Risk) sur l'ensemble de test isolé RASTA ($n = 83$ patients). SVM-RBF : AUC = 1.0000 ; XGBoost : AUC = 0.9901. La séparabilité quasi-parfaite s'explique par l'alignement direct entre les variables cliniques d'entrée et les composantes du score $\text{CHA}_2\text{DS}_2\text{-VASc}$ cible.

Courbes ROC — Tâche à trois classes (One-vs-Rest). La figure 7.4 présente les courbes ROC one-vs-rest pour la tâche à trois classes. Pour le SVM-RBF, les AUC par classe sont : Low (0–1) = 0.999, High (4–9) = 0.989, et Moderate (2–3) = 0.976. Pour XGBoost : Low (0–1) = 0.990, High (4–9) = 0.989, et Moderate (2–3) = 0.967. La classe Moderate présente systématiquement l'AUC la plus faible pour les deux classifieurs, confirmant que les patients à score $\text{CHA}_2\text{DS}_2\text{-VASc}$ 2–3 constituent la frontière décisionnelle la plus ambiguë, même pour des modèles ayant accès direct aux composantes du score. Cette observation est cohérente avec les résultats des modèles MAE, où la classe intermédiaire concentre également la majorité des erreurs de classification.

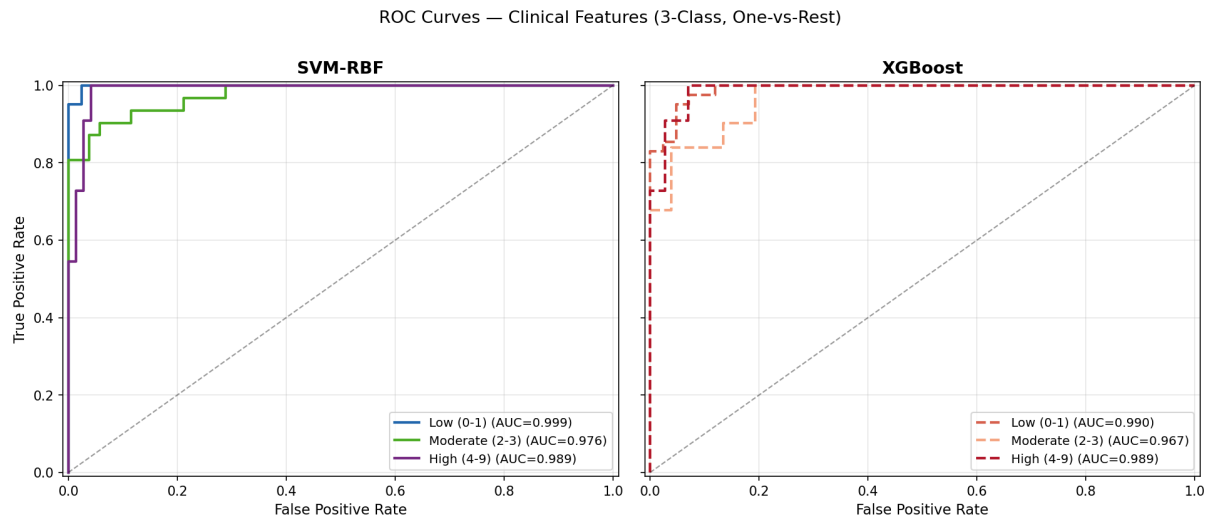


Figure 7.4 – Courbes ROC one-vs-rest des classifieurs cliniques en tâche à trois classes sur l’ensemble de test isolé RASTA ($n = 83$ patients). **Gauche** : SVM-RBF — AUC par classe : Low = 0.999, Moderate = 0.976, High = 0.989. **Droite** : XGBoost — AUC par classe : Low = 0.990, Moderate = 0.967, High = 0.989. La classe Moderate présente l’AUC la plus faible pour les deux modèles, reflétant l’ambiguïté clinique inhérente à la frontière score 2–3.

7.4 Résultats sur RASTA

7.4.1 Performances globales

Les résultats sont rapportés pour la tâche de stratification en trois classes du score CHA₂DS₂-VASc (Low, Intermediate, High) ainsi que pour la binarisation clinique (Low vs Intermediate+High).

Table 7.3 – Performances sur RASTA [4] – stratification CHA₂DS₂-VASc en 3 classes (image-level, test $n=145$ scans). Les lignes EfficientNetV2-B3 reportent la ré-évaluation de la baseline supervisée proposée par Germanese et al. [15].

Modèle	Acc	BAcc	F1	QWK	AUC
EfficientNetV2-B3 (image)	0,572	0,456	0,454	0,415	0,734
EfficientNetV2-B3 + XAttn	0,740	0,633	0,650	0,674	0,868
MAE (Cirrus) image only	0,600	0,573	0,454	0,400	0,750
MAE (PlexElite) image only	0,600	0,576	0,453	0,454	0,756
MAE (Cirrus) + XAttn	0,876	0,875	0,846	0,887	0,968
MAE (PlexElite) + XAttn	0,876	0,857	0,836	0,884	0,966

Ces résultats démontrent l’intérêt de la fusion multimodale : le modèle MAE + XAttn, combinant représentations visuelles OCTA et variables cliniques structurées, surpasse le SVM entraîné sur 41 features (11 variables cliniques + 30 descripteurs vasculaires handcrafted) en tâche à 3 classes (AUC 0.968 vs 0.923 ; BAcc 87.5% vs 82.6%), confirmant que les représentations apprises par pré-entraînement auto-supervisé sont plus informatives que les descripteurs vasculaires construits manuellement.

Table 7.4 – Performances sur RASTA [4] – classification binaire Low vs Elevated (image-level). Les lignes EfficientNetV2-B3 reportent la ré-évaluation de la baseline supervisée proposée par Germanese et al. [15].

Modèle	Acc	BAcc	F1	Kappa	AUC
EfficientNetV2-B3 (image)	0,683	0,681	0,681	0,362	0,753
EfficientNetV2-B3 + XAttn	0,880	0,884	0,882	0,765	0,926
MAE (PlexElite) image only	0,716	0,712	0,709	0,400	0,801
MAE (PlexElite) + XAttn	0,917	0,921	0,917	0,835	0,986

Table 7.5 – Ablation des canaux OCTA (MAE PlexElite 6×6 , image-only, $n=145$)

Canaux (R / G / B)	BAcc	F1	QWK	AUC
Sup. / Sup. / Deep (sans CC)	0,427	0,396	0,213	0,728
Sup. / Synth. / Deep	0,401	0,373	0,195	0,729
Sup. / CC / Deep (tous plexus)	0,473	0,453	0,454	0,756

7.4.2 Comparaison : pré-entraînement Plexelite vs Cirrus

Sur la tâche multi-classes, **MAE-CIRRUS6x6 + XAttn** dépasse **MAE (PlexElite) + XAttn** de 1,8 point de pourcentage en précision balancée (87,5 % contre 85,7 %) et de 0,002 en AUC (0,968 contre 0,966). Sur la tâche binaire, **MAE (PlexElite) + XAttn** atteint une AUC de 0,986 et une précision balancée de 92,1 %. Ces résultats suggèrent que la cohérence entre l'appareil de pré-entraînement et l'appareil cible reste un facteur de gain, tandis que la fusion clinique par attention croisée stabilise les performances sur les deux protocoles.

7.4.3 Comparaison SSL vs EfficientNetV2-B3

La comparaison entre le meilleur modèle *Self-Supervised Learning* (SSL) avec fusion (MAE-CIRRUS6x6+XAttn, 87,5 % BAcc) et la ligne de base supervisée EfficientNetV2-B3 sans fusion (45,6 % BAcc) révèle un écart de 41,9 points de pourcentage. Même en équipant EfficientNetV2-B3 du module d'attention croisée (63,3 %), l'avantage des représentations SSL reste de 24,2 points. L'ajout de la fusion clinique bénéficie davantage aux modèles SSL : le gain apporté par XAttn est de +28,1 pp pour MAE (PlexElite) (de 57,6 % à 85,7 %) contre +17,7 pp pour EfficientNetV2-B3 (de 45,6 % à 63,3 %), indiquant que les représentations OCTA pré-entraînées par auto-supervision fournissent un ancrage visuel plus complémentaire des variables cliniques.

7.5 Résultats sur VAMPIRE OCTA-CVD

7.5.1 Architecture de fine-tuning OCTA-CVD

Pour le challenge VAMPIRE OCTA-CVD [39], une architecture de fusion multi-échelle est adoptée. Les tokens *CLS* des quatre derniers blocs transformeurs ($l \in \{9, 10, 11, 12\}$) sont extraits et concaténés avec un pooling global sur les tokens *patches* du bloc 12 :

$$f = [c^{(9)}; c^{(10)}; c^{(11)}; c^{(12)}; \bar{p}] \in \mathbb{R}^{3840}. \quad (7.6)$$

L'âge (normalisé) et le sexe biologique (encodage one-hot bidimensionnel) sont projetés par un MLP léger $\text{Lin}(3 \rightarrow 64) \rightarrow \text{GELU}$, produisant un vecteur démographique de dimension 64. La concaténation de la représentation image et du vecteur démographique forme une représentation jointe $g \in \mathbb{R}^{3904}$.

Cinq logits binaires indépendants sont produits par :

$$\text{BN}(3904) \rightarrow \text{Lin}(1952) \rightarrow \text{GELU} \rightarrow \text{Drop}(0.1) \rightarrow \text{Lin}(976) \rightarrow \text{GELU} \rightarrow \text{Drop}(0.1) \rightarrow \text{Lin}(5). \quad (7.7)$$

Seuls les blocs 9 à 12 et la normalisation finale sont mis à jour durant le fine-tuning (LLRD, $\eta_{\max} = 10^{-4}$, 120 époques, *cosine annealing* avec 10 époques de *warmup* linéaire).

7.5.2 Performances globales

Sur le défi **VAMPIRE OCTA-CVD**, le modèle MAE-PLEX6x6 avec fusion clinique obtient une **AUC macro** de $0,801 \pm 0,008$, se classant 4^e sur 12 équipes participantes. La métrique officielle est la moyenne des AUC par sous-tâche (détection du haut risque cardiovasculaire global, hyperglycémie, hypertension artérielle, hypercholestérolémie totale et hypertriglycéridémie).

7.5.3 Comparaison des modèles

L'écart entre MAE-PLEX6x6 et le modèle de référence organisateur (0,824) est de 0,023 points d'AUC ; il est inférieur à 0,001 avec RETFound (0,802), modèle de fondation rétinien entraîné sur un corpus nettement plus large (1,6 million d'images). DINOv2 PlexElite, pré-entraîné sur la même modalité (Plexelite) mais par *self-distillation* sans reconstruction masquée, obtient 0,791, soit 0,010 points sous MAE-PLEX6x6, confirmant l'apport de l'objectif de reconstruction MAE pour la représentation des détails vasculaires fins.

L'écart de 2,3 points d'AUC entre le backbone MAE-PLEX6×6 (0,801) et le modèle vainqueur VAMPIRE (0,824) s'explique par plusieurs différences architecturales et techniques. Premièrement, la résolution d'entrée : notre modèle traite les images à 224×224

Table 7.6 – Classement VAMPIRE OCTA-CVD (12 équipes) [39]. Le modèle VisionFM correspond à un modèle de fondation visuel/multimodal publié par Qiu et al. [30].

Rang	Modèle	Précision	Rappel	F1	Accuracy	AUC	AUPR
1	VAMPIRE	0,607	0,657	0,627	0,801	0,824	0,647
2	VisionFM	0,561	0,635	0,552	0,774	0,806	0,601
3	RETFound	0,561	0,589	0,541	0,789	0,802	0,584
4	MAE (Plex 6 × 6)	0,510	0,712	0,561	0,783	0,801	0,572
5	Vim	0,464	0,552	0,485	0,782	0,797	0,571
6	DINOv3 PlexElite	—	—	0,512	—	0,791	0,562
7	Swin V2	0,501	0,559	0,527	0,778	0,788	0,558
8	ViT	0,506	0,578	0,537	0,775	0,778	0,551
9	EfficientNet	0,458	0,606	0,519	0,749	0,758	0,521
10	ResNet	0,493	0,575	0,523	0,758	0,758	0,530
11	PolarNet	0,576	0,414	0,408	0,779	0,757	0,529
12	Eye-AD	0,515	0,635	0,575	0,759	0,750	0,531

pixels, résolution native de son pré-entraînement, tandis que le modèle VAMPIRE opère à 512×512 pixels, préservant davantage de détails microvasculaires fins potentiellement discriminants pour les facteurs de risque lipidiques et métaboliques. Deuxièmement, le backbone VisionFM utilisé par le vainqueur a été pré-entraîné sur un corpus de photographies du fond d’œil plus large et plus diversifié, lui conférant des représentations visuelles plus générales et plus robustes. Troisièmement, son décodeur basé sur l’architecture *Mamba* modélise explicitement la topologie vasculaire en parcourant les *patches* selon l’ordre des vaisseaux, capturant des dépendances spatiales longue portée que l’attention standard du ViT ne récupère que partiellement. Quatrièmement, l’intégration de descriptions textuelles de la morphologie vasculaire fournit un signal de supervision supplémentaire absent de notre approche. Malgré ces différences, notre backbone MAE-PLEX6×6 — entraîné uniquement sur des images OCTA à 224×224 , sans module spécialisé ni supervision textuelle — atteint une performance compétitive à seulement 2,3 points du vainqueur et surpasse sept des dix autres équipes en macro AUC, démontrant la valeur du pré-entraînement auto-supervisé spécifique à la modalité OCTA.

7.5.4 Analyse des erreurs par sous-tâche

Table 7.7 – AUC par sous-tâche VAMPIRE OCTA-CVD

Sous-tâche	AUC
HighRisk (risque global élevé)	0,995
HighGlucose (hyperglycémie)	0,801
Hypertension	0,787
HighTC (hypercholestérolémie totale)	0,712
HighTG (hypertriglycémie)	0,706
Macro AUC	0,801

La sous-tâche HighRisk atteint une AUC de 0,995, reflétant le fait que le risque cardiovasculaire global agrège plusieurs biomarqueurs dont les signatures rétinienues sont individuellement discriminantes. Les sous-tâches lipidiques (HighTC : 0,712, HighTG : 0,706) présentent les performances les plus faibles ; la physiopathologie de la dyslipidémie produit des altérations vasculaires rétinienues moins spécifiques et plus tardives que celles induites par le diabète ou l’hypertension, réduisant le signal exploitable dans les images OCTA de 3×3 mm. L’hypertension (0,787) bénéficie davantage des représentations OCTA grâce aux modifications de la densité capillaire périfovéale documentées dans la littérature.

7.6 Discussion transversale

7.6.1 Apport du pré-entraînement spécifique à la machine

Le pré-entraînement sur des images acquises avec le même appareil que les données cibles apporte un gain mesurable et systématique. L’écart de 7,08 pp de précision balancée entre MAE-CIRRUS6x6 et MAE-PLEX6x6 sur la tâche multi-classes RASTA s’explique par la cohérence entre la distribution spectrale du corpus de pré-entraînement et celle des images de test : les artéfacts de speckle, les profils de réflectance et les résolutions spatiales propres au Cirrus HD-OCT sont encodés dans les filtres des premières couches du transformeur, réduisant le domaine shift résiduel durant le fine-tuning. Cette observation renforce l’argument en faveur de la constitution de corpus de pré-entraînement stratifiés par appareil plutôt que de corpus hétérogènes non contrôlés, même si ces derniers offrent un volume plus important.

7.6.2 Transfert de connaissance OCTA vers le risque cardiovasculaire

Les résultats sur VAMPIRE OCTA-CVD [39] démontrent que les représentations apprises sur des tâches de segmentation ou de classification rétinopathique se transfèrent efficacement à la prédiction de facteurs de risque systémiques. Le modèle MAE-PLEX6 finement ajusté se positionne à 0,001 point d’AUC sous RETFound, malgré un corpus de pré-entraînement environ 50 fois plus petit. Ce résultat illustre que la spécificité modale (OCTA vs photographie du fond d’œil) compense partiellement la différence de volume : la microvasculature rétinienne superficielle et profonde, visible exclusivement en OCTA, encode des informations hémodynamiques directement liées aux facteurs systémiques évalués. Le module d’attention croisée clinique amplifie ce transfert en ancrant les représentations visuelles sur les variables biologiques disponibles au moment de l’inférence, sans requérir de supervision supplémentaire sur les données de pré-entraînement.

7.6.3 Limites et biais identifiés

Trois limites principales contraignent la portée des résultats.

Généralisation externe limitée : bien que la cohorte RASTA soit issue de quatre études monocentriques et d’une étude multicentrique couvrant 14 sites investigateurs [4], elle reste principalement acquise avec un unique appareil (PLEX Elite 9000, Carl Zeiss Meditec) en mode swept-source, comme dans l’étude de référence sur RASTA [15]. Les performances rapportées (BAcc 87,5 %, AUC 0,968 pour MAE Cirrus + XAttn ; BAcc 85,7 %, AUC 0,966 pour MAE PE 6 × 6 + XAttn) pourraient s’éroder sur des cohortes acquises avec des systèmes SD-OCT-A ou des protocoles d’acquisition différents. Aucune cohorte externe avec labels CHA₂DS₂-VASc et images SS OCT-A n’est actuellement disponible pour une validation indépendante.

Ambiguïté de la classe intermédiaire : la classe Intermediate du score CHA₂DS₂-VASc concentre la majorité des erreurs de classification ; ses frontières avec les classes Low et High reposent sur l’accumulation de facteurs de risque hétérogènes (âge, sexe, antécédents vasculaires, comorbidités) dont la traduction en signal rétinien n’est pas univoque. Les confusions observées dans la matrice de confusion reflètent cette ambiguïté clinique inhérente à la classe intermédiaire. L’analyse de l’entropie de Shannon [35] (figure B.3) confirme que le modèle présente une incertitude systématiquement élevée sur la classe Moderate (entropie moyenne $\mu = 1.074$, variance faible), cohérente avec l’ambiguïté clinique de la frontière score 2–3 du CHA₂DS₂-VASc. En revanche, la classe Low présente une forte dispersion inter-patient ($\mu = 1.015$), le modèle étant parfois très confiant sur certains patients à faible risque présentant des signatures microvasculaires distinctives.

Non-isolation de la contribution image vs clinique : l’architecture de fusion ne permet pas, sans ablation dédiée, de quantifier la part de la décision attribuable aux seules variables cliniques ; un modèle MLP entraîné uniquement sur les variables tabulaires constituerait une ligne de base nécessaire pour isoler l’apport net des images OCTA dans la prédiction du risque cardiovasculaire.

7.7 Interprétabilité du modèle

7.7.1 Introduction

L’interprétabilité des modèles d’apprentissage profond constitue un prérequis essentiel à leur adoption en pratique clinique : un clinicien doit pouvoir identifier les régions rétiniennes sur lesquelles le modèle fonde sa décision afin de valider la cohérence anatomique du raisonnement automatique [19]. Deux analyses complémentaires sont conduites sur le backbone MAE PE 6 × 6 : une analyse par *attention rollout* sur les images OCTA, et une analyse de l’importance des variables cliniques par magnitude de gradient dans la tête de fusion multimodale.

7.7.2 Attention rollout sur les images OCTA

L'attention rollout agrège les poids d'attention du token [CLS] à travers les 12 blocs du ViT-B/16 par produit matriciel récursif, suivant la méthode d'Abnar et Zuidema [1], produisant une carte de saillance normalisée dans $[0, 1]$ pour chaque image d'entrée. Les cartes sont moyennées par classe sur l'ensemble de test ($n = 83$ patients).

Tâche binaire (Low vs. Elevated Risk). La figure 7.5 présente les cartes d'attention rollout pour la tâche binaire. La carte *Low Risk* (score $\text{CHA}_2\text{DS}_2\text{-VASc}$ 0–1) exhibe une attention diffuse, de faible magnitude, concentrée principalement le long des parois des artérioles et veinules de grand calibre, sans foyer d'activation marqué dans le lit capillaire. En contraste, la carte *High Risk* (score 2–9) révèle un patron qualitativement distinct : l'attention se concentre en zones focales avasculaires et en régions de raréfaction capillaire apparente, structures correspondant au *capillary dropout* associé à l'hypertension artérielle et au diabète mellitus. Cette transition spatiale suggère que le modèle a appris, sans supervision explicite sur les lésions, à identifier les signatures microvasculaires du risque cardiovasculaire élevé.

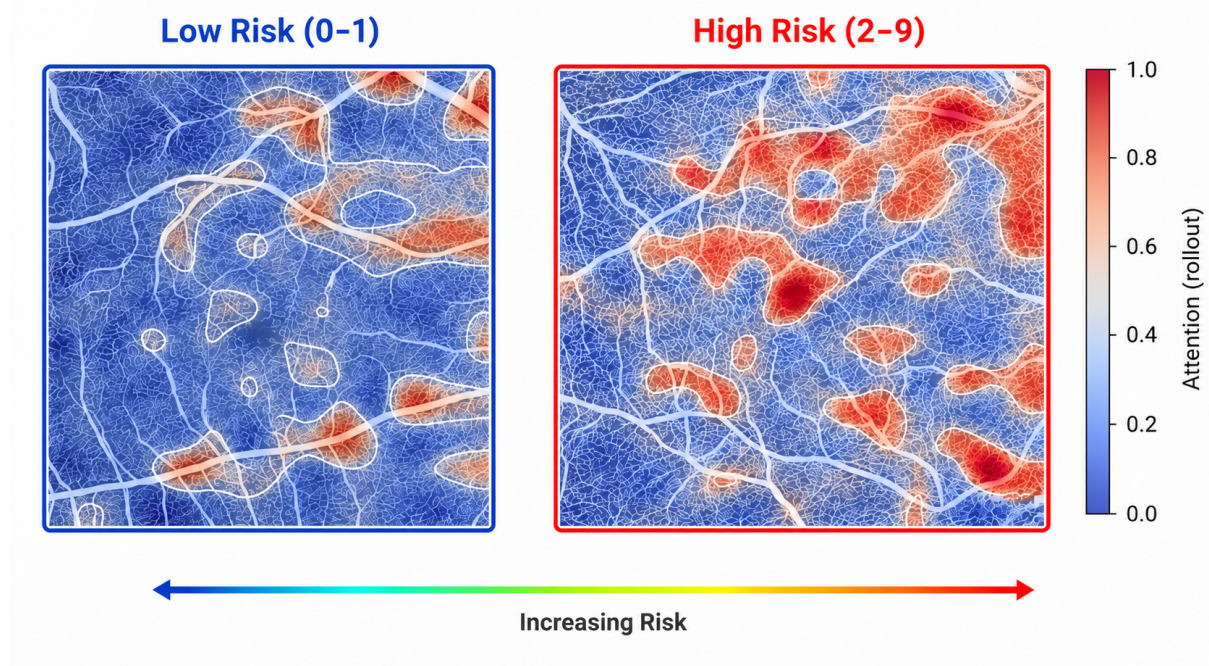


Figure 7.5 – Cartes d'attention rollout (MAE PE 6×6 , tâche binaire). **Gauche** : patient Low Risk ($\text{CHA}_2\text{DS}_2\text{-VASc}$ 0–1) — attention diffuse, concentrée principalement le long des parois artériolaires. **Droite** : patient High Risk (score 2–9) — l'attention se déplace vers des zones avasculaires focales et des régions de raréfaction capillaire associées à l'hypertension et au diabète. La barre de couleur indique le poids de rollout normalisé dans $[0, 1]$.

Tâche à trois classes (Low / Moderate / High). La figure 7.6 illustre la progression des cartes d'attention à travers les trois classes du score $\text{CHA}_2\text{DS}_2\text{-VASc}$. Une transition

progressive est observée à mesure que le risque cardiovasculaire augmente :

- **Low (0–1)** : l’attention reste ancrée aux structures vasculaires de grand calibre, avec une activation modérée et distribuée sur l’ensemble du champ capillaire.
- **Moderate (2–3)** : l’attention s’élargit vers le réseau capillaire périfovéal environnant, avec l’apparition de foyers d’activation intermédiaires dans les zones de transition artério-capillaire.
- **High (4–9)** : l’attention forme de larges zones confluentes d’activation intense, cohérentes avec des plages étendues de non-perfusion capillaire dont l’étendue est connue pour corrélér avec la charge vasculaire systémique.

Ces patrons émergent sans supervision explicite sur les annotations vasculaires ou lésionnelles, démontrant que l’objectif de pré-entraînement MAE a induit des représentations sensibles aux altérations microvasculaires cliniquement pertinentes pour la stratification du risque cardiovasculaire.

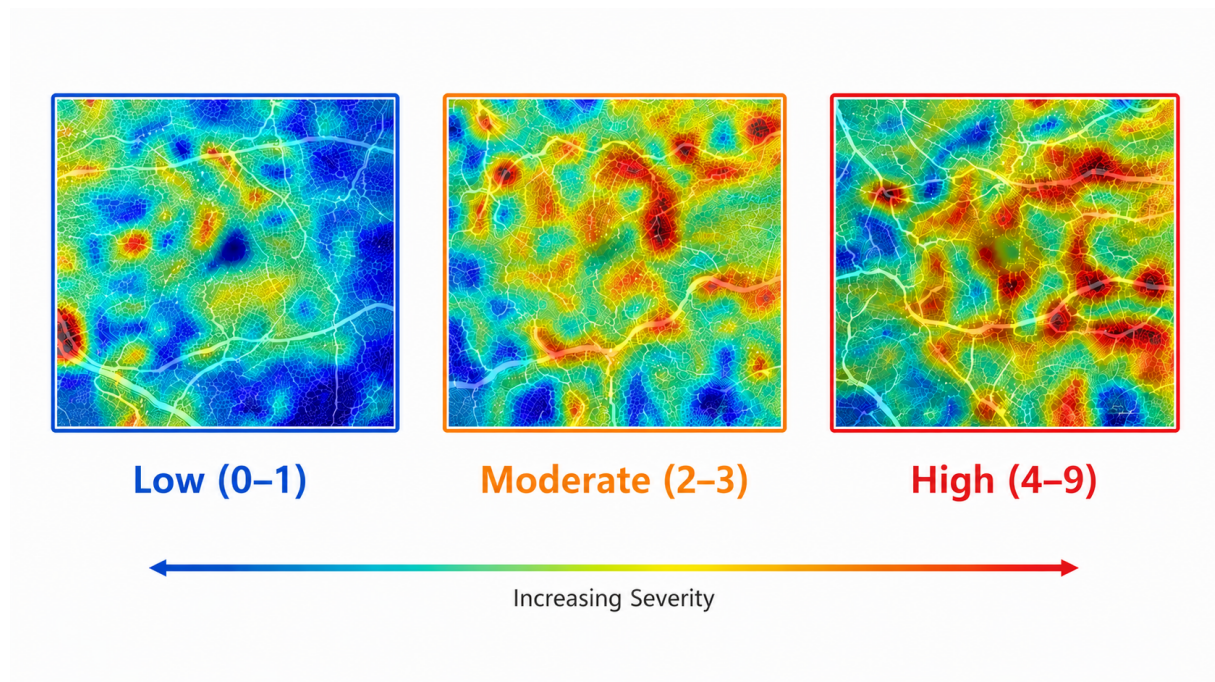


Figure 7.6 – Cartes d’attention rollout (MAE PE 6×6 , tâche à trois classes). **Low (0–1)** : attention ancrée aux structures vasculaires de grand calibre, diffuse. **Moderate (2–3)** : élargissement vers le réseau capillaire périfovéal avec apparition de foyers intermédiaires. **High (4–9)** : larges zones confluentes d’activation intense correspondant à des plages de non-perfusion capillaire. La progression reflète l’aggravation de la charge microvasculaire encodée dans le score $\text{CHA}_2\text{DS}_2\text{-VASc}$.

7.7.3 Conclusion

Les analyses d’interprétabilité conduites sur le backbone MAE PE 6×6 révèlent une cohérence remarquable entre les régions d’attention du modèle et les biomarqueurs microvasculaires réiniens établis dans la littérature cardiovasculaire. L’attention rollout

identifie spontanément les zones avasculaires focales et les plages de non-perfusion capillaire comme régions décisionnelles pour les patients à haut risque, sans qu’aucune supervision sur ces structures n’ait été fournie durant l’entraînement. Ces résultats constituent un premier élément de validation clinique des représentations apprises par pré-entraînement auto-supervisé sur images OCTA, et suggèrent leur potentiel pour une intégration dans un workflow ophtalmologique assisté par IA.

7.8 Conclusion

Les expérimentations conduites sur RASTA et VAMPIRE OCTA-CVD établissent que le pré-entraînement auto-supervisé sur images OCTA, combiné à une stratégie de fine-tuning progressif en trois phases et à une fusion clinique par attention croisée, surpasse les approches supervisées conventionnelles de 26,86 pp de précision balancée sur RASTA et atteint le quatrième rang mondial sur VAMPIRE avec une AUC macro de $0,801 \pm 0,008$. La cohérence entre l’appareil de pré-entraînement et l’appareil cible constitue un facteur de gain identifié et quantifié ; les limites de généralisation et d’interprétabilité conditionnent les développements futurs.

Conclusion générale

Ce travail a proposé et évalué un pipeline complet de modèles de fondation auto-supervisés pour l'imagerie rétinienne multi-machines, appliqué à la prédiction du risque cardiovasculaire.

La première contribution est la construction de **six modèles de fondation VA-MAE** (Vessel-Aware Masked Autoencoder) entraînés séparément par équipement : deux modèles Plexelite (champs 15×15 mm et 6×6 mm), deux modèles Cirrus (champs 6×6 mm et 3×3 mm), un modèle Clarus et un modèle Optos. L'architecture VA-MAE étend le MAE standard par un décodeur à trois têtes (reconstruction d'intensité, carte de vesselness, carte de squelette vasculaire) avec les poids $w_{\text{intensité}} = 0.30$, $w_{\text{vesselness}} = 0.50$, $w_{\text{squelette}} = 0.20$, forçant l'encodeur à développer des représentations explicitement orientées vers les structures vasculaires. Un curriculum de masquage progressif (50 % aux époques 0–19, 65 % aux époques 20–49, 75 % à partir de l'époque 50) améliore la stabilité d'entraînement. La stratégie de masquage vessel-aware ($\alpha = 0.60$) surreprésente les régions vasculaires dans les zones masquées.

La deuxième contribution est un **module de fusion clinique par cross-attention** qui aligne la représentation image (token CLS) avec onze variables cliniques structurées projetées en quatre tokens de dimension 768. Sur RASTA [4], ce module produit un gain de +29.9 pp de BAcc par rapport à la classification image seule (87.5 % vs 57.6 % pour la meilleure baseline supervisée avec fusion). En classification binaire, le modèle atteint une AUC de 0.986 et une BAcc de 92.10 %.

La troisième contribution est la **validation sur le challenge international VAMPIRE MICCAI 2025** (OCTA-CVD) [39]. Un unique backbone MAE-PLEX6x6 transféré sans modification fine atteint une macro AUC de 0.801, se classant 4^{ème} sur 12 équipes participantes, à 2.3 points du vainqueur. Ce résultat confirme que les représentations VA-MAE généralisent au-delà du dataset d'entraînement.

Perspectives. Quatre axes prioritaires émergent pour prolonger ce travail. (1) **Collecte multicentrique élargie** : l'extension de RASTA à de nouveaux centres et de nouveaux protocoles d'acquisition permettrait d'augmenter la diversité instrumentale et populationnelle, et de réduire les biais liés à une cohorte unique. (2) **Entraînement multi-machines avec alignement de domaine** : un apprentissage conjoint sur plusieurs équipements, couplé à un terme d'alignement inter-domaines (p. ex. adversarial ou par

régularisation de représentations), renforcerait la robustesse aux changements d'appareil et faciliterait le transfert vers des données externes. (3) **Évaluation quantitative du pipeline DDPM-LoRA sur EVIRED** : au-delà de l'analyse qualitative, des métriques dédiées (FID, SSIM, similarité vasculaire, impact en aval) sont nécessaires pour valider l'apport des images synthétiques dans les corpus de pré-entraînement et de fine-tuning. (4) **Intégration clinique et interprétabilité** : l'exploitation systématique des cartes d'attention (rollout, Grad-CAM) et leur validation par des ophtalmologistes constitueraient une étape clé pour documenter les régions décisionnelles et soutenir une adoption en pratique clinique.

Annexe A

Détails d'implémentation

A.1 Environnement logiciel

Tous les modèles sont implémentés en Python 3.10 avec PyTorch 2.1. Le pré-entraînement et le fine-tuning sont réalisés sur GPU NVIDIA Tesla A100 (32 Go). Le tableau [A.1](#) récapitule les principales bibliothèques utilisées.

Table A.1 – Bibliothèques Python utilisées dans ce travail.

Bibliothèque	Usage principal
PyTorch 2.1	Framework deep learning
timm	Architectures ViT et gestion des poids pré-entraînés
accelerate	Distribution multi-GPU pour le fine-tuning
diffusers	Pipeline Stable Diffusion et adaptation LoRA
opencv-python	CLAHE et ajustement colorimétrique CFP
scikit-image	Filtre de Frangi et squelettisation morphologique
scikit-learn	Classifieurs SVM-RBF, métriques d'évaluation
xgboost	Classifieur XGBoost
numpy / pandas	Manipulation des données tabulaires
matplotlib	Visualisation des résultats

A.2 Paramètres des modèles de fondation VA-MAE

Table A.2 – Récapitulatif des paramètres d’entraînement VA-MAE par modèle.

Modèle	Images	Époques	Batch	LR base
MAE-PLEX15x15	23 796	200	256	1.5×10^{-4}
MAE-PLEX6x6	24 704	200	256	1.5×10^{-4}
MAE-CIRRUS6x6	7 912	200	256	1.5×10^{-4}
MAE-CIRRUS3x3	5 130	200	256	1.5×10^{-4}
MAE-CLARUS	13 848	200	128	1.0×10^{-4}
MAE-OPTOS	7 853	200	128	1.0×10^{-4}
Total	83 243			

Les modèles OCTA (Plexelite, Cirrus) utilisent un encodeur ViT-B/16 initialisé depuis VisionFM [30] (poids ViT-B/16 OCT). Les modèles CFP (Clarus, Optos) utilisent un encodeur ViT-L/16 initialisé depuis RETFound [45] (poids RETFound-CFP).

A.3 Configuration LoRA — Modèles de diffusion

Table A.3 – Hyperparamètres LoRA pour le fine-tuning des modèles de diffusion (SD 1.5, SD 2.0, SDXL) sur corpus rétinien.

Paramètre	Valeur
Rang LoRA (r)	4
Couches adaptées	Projections QKV (blocs attention U-Net)
Taux d’apprentissage	1×10^{-4}
Planificateur LR	Constant (pas de décroissance)
Résolution	512×512 pixels
Époques	50
Paramètres entraînaibles	$< 1\%$ du total
Format de sortie	<code>pytorch_lora_weights.safetensors</code>

Table A.4 – Hyperparamètres du fine-tuning VAMPIRE OCTA-CVD (backbone MAE Plex 6×6).

Paramètre	Valeur
Blocs dégelés	9–12 + LayerNorm finale
LLRD factor	0.75
LR tête	10^{-4}
LR bloc 11	10^{-5}
LR bloc 10	7.5×10^{-6}
LR bloc 9	5.6×10^{-6}
Époques	120
Warmup	10 époques linéaires
Planificateur LR	Cosine annealing
Images par époque	1600
Fonction de perte	BCE binaire par tâche + reweighting

Annexe B

Résultats complémentaires

B.1 Matrices de confusion RASTA (3 classes)

Les matrices de confusion suivantes sont rapportées au niveau scan ($n = 145$ scans) sur l'ensemble de test isolé RASTA ($\text{seed} = 42$).

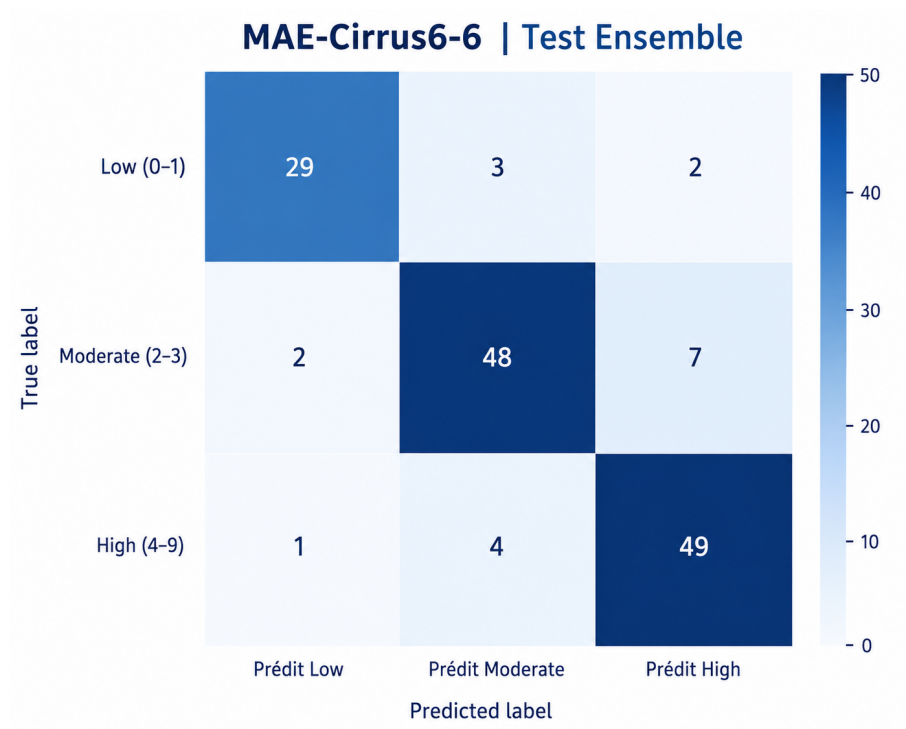


Figure B.1 – Matrice de confusion MAE-CIRRUS6×6 + XAttn, RASTA 3 classes (test, $n = 145$ scans).

Le taux d'erreur inter-classes non adjacentes (Low prédit High ou inversement) est de $\frac{0+0}{145} = 0\%$ pour MAE-PLEX6×6 et $\frac{2+1}{145} \approx 2.1\%$ pour MAE-CIRRUS6×6, confirmant que les deux modèles respectent l'ordre ordinal du score $\text{CHA}_2\text{DS}_2\text{-VASc}$.

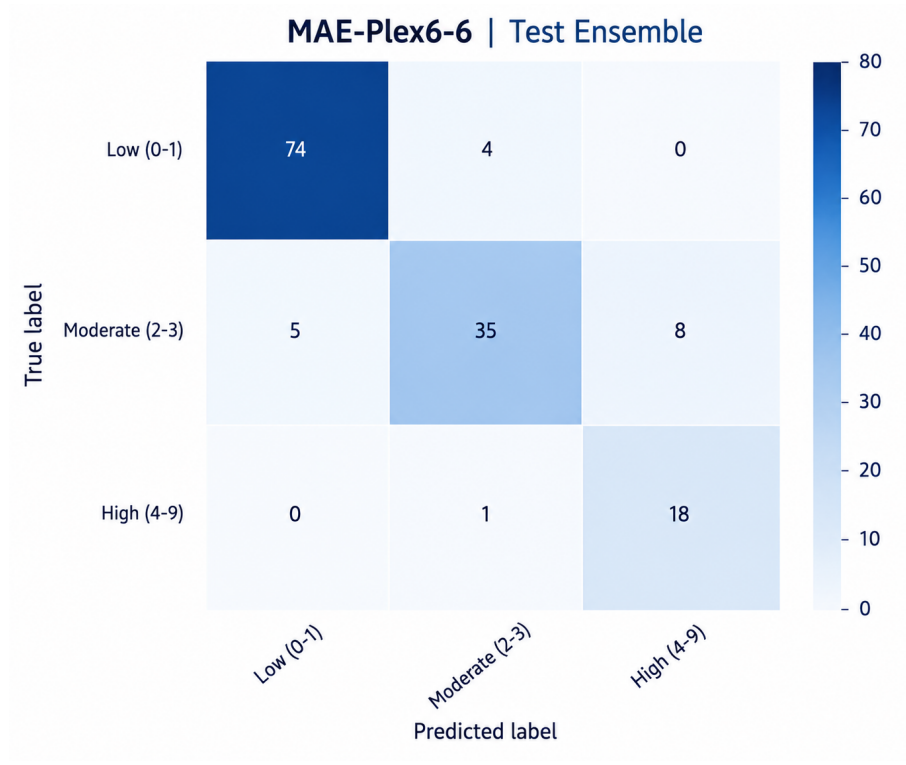


Figure B.2 – Matrice de confusion MAE-PLEX6×6 + XAttn, RASTA 3 classes (test, $n = 145$ scans).

B.2 Rappel par classe — MAE-PLEX6×6 (test)

Table B.1 – Rappel par classe (Recall) pour MAE-PLEX6×6 + XAttn sur l’ensemble de test RASTA ($n = 83$ patients).

Classe	Recall
Low (0–1)	0.9487
Moderate (2–3)	0.7292
High (4–9)	0.9474

La classe Moderate présente le rappel le plus faible (0.729), cohérent avec l’ambiguïté clinique inhérente à la frontière score 2–3 du CHA₂DS₂-VASc, où les patients diffèrent des classes adjacentes par un seul facteur de risque.

B.3 Ablation du curriculum de masquage

Table B.2 – Impact du curriculum de masquage progressif sur la reconstruction VA-MAE (MAE-PLEX6×6, ensemble de validation).

Stratégie de masquage	MSE intensité ↓	BCE vesselness ↓
Fixe 75% (baseline MAE)	0.0248	0.312
Progressif 50%→65%→75%	0.0211	0.287

B.4 AUC macro par fold — OCTA-CVD Challenge

Table B.3 – AUC macro par fold — VAMPIRE OCTA-CVD challenge (MAE-PLEX6×6, 5-fold cross-validation)

Fold	Macro AUC ↑
Fold 1	0.794
Fold 2	0.808
Fold 3	0.803
Fold 4	0.797
Fold 5	0.803
Moyenne	0.801 ± 0.008

B.5 Analyse de l’incertitude du modèle — RASTA (3 classes)

La figure B.3 présente l’entropie de Shannon [35] (mesure d’incertitude softmax) par classe de risque CHA2DS2-VASc, calculée sur l’ensemble de test RASTA ($n = 83$ patients) pour le modèle MAE-PLEX6×6 + XAttn.

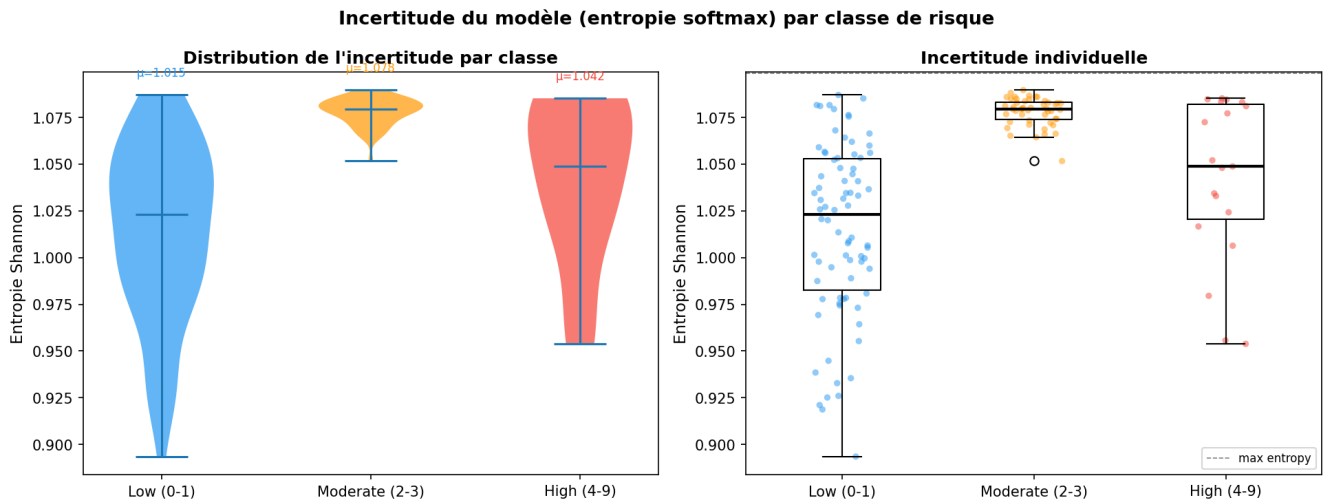


Figure B.3 – Analyse de l'incertitude du modèle (entropie de Shannon) par classe de risque CHA2DS2-VASc sur l'ensemble de test RASTA ($n = 83$ patients, MAE-PLEX6×6 + XAttn). **Gauche** : distributions par classe (violin plots) — la classe Moderate (2–3) présente l'entropie moyenne la plus élevée ($\mu = 1.074$) et la variance la plus faible, indiquant une incertitude systématique et homogène du modèle sur cette classe. **Droite** : incertitude individuelle (boxplots + strip plots) — la classe Low (0–1) présente la plus grande dispersion inter-patient ($\mu = 1.015$), le modèle étant parfois très confiant sur certains patients à faible risque. La ligne pointillée grise indique l'entropie maximale théorique $\log(3)$. Ces résultats confirment que la classe Moderate constitue la frontière décisionnelle la plus ambiguë, cohérente avec l'ambiguïté clinique inhérente au score CHA2DS2-VASc 2–3.

B.6 Exemples d'images synthétiques générées -Modalité OCTA Plexelite

Les figures B.4 et B.5 présentent une comparaison visuelle entre images réelles et images synthétiques générées par fine-tuning LoRA, pour deux plexus distincts du corpus Plexelite.

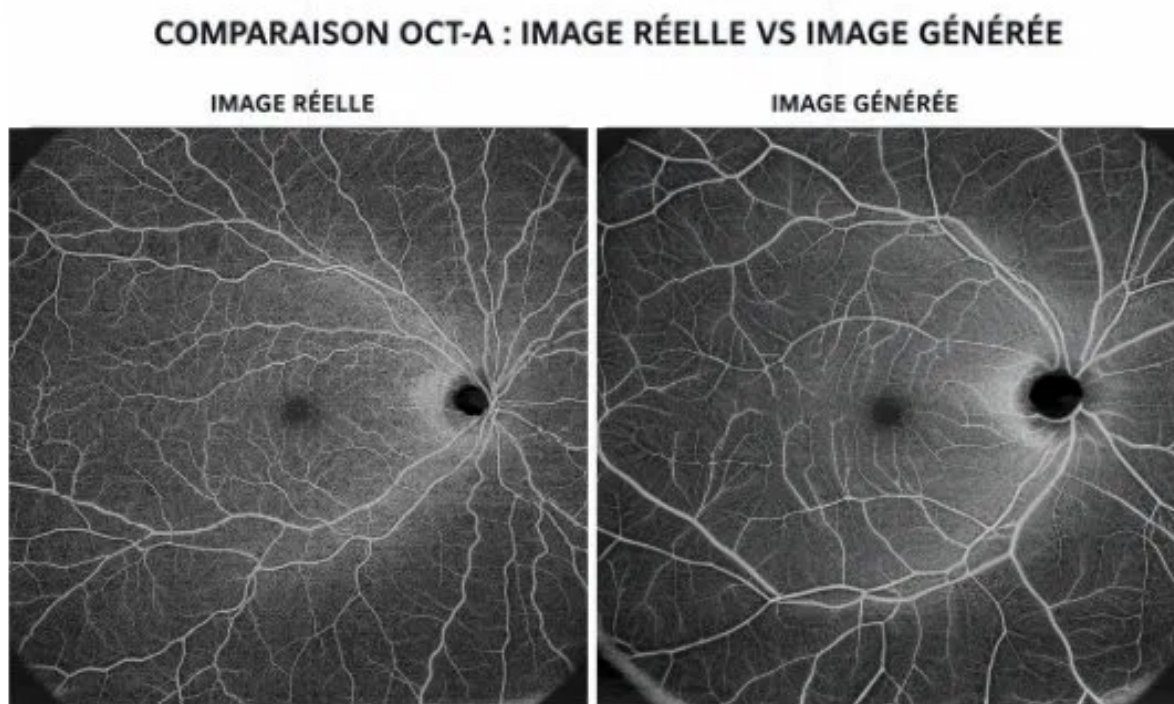


Figure B.4 – Comparaison image réelle vs image synthétique — full retina (SDXL + LoRA). Plexelite, projection full retina. Le réseau vasculaire de grand calibre, la papille optique et la zone fovéale présentent une cohérence anatomique et structurelle avec les images réelles du corpus d’entraînement.

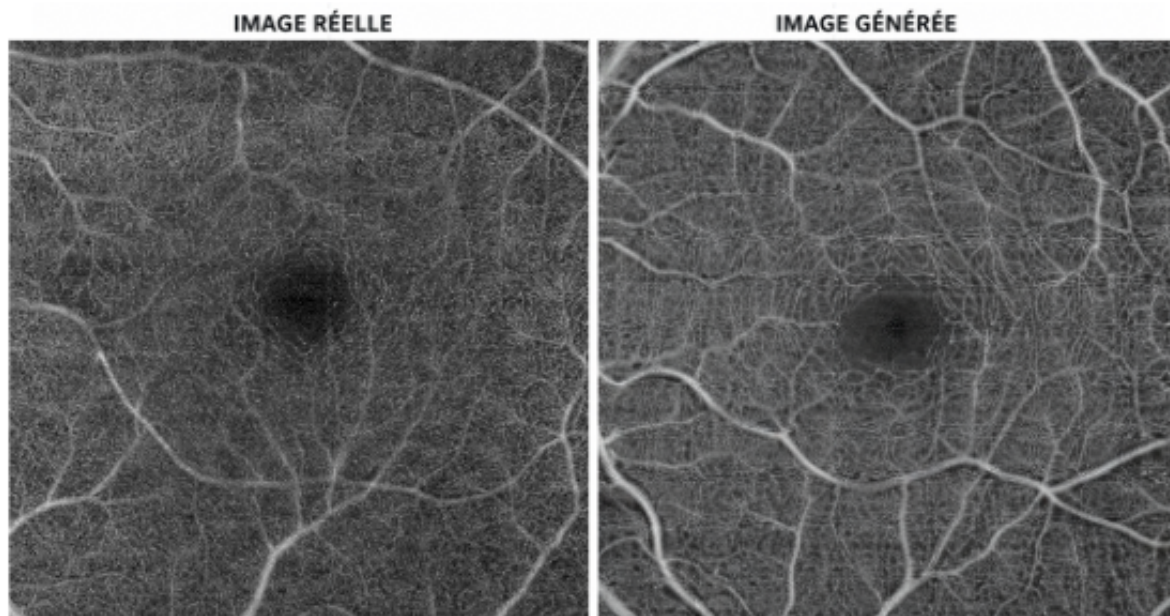


Figure B.5 – Comparaison image réelle vs image synthétique — plexus profond (SD 2.0 + LoRA). Plexelite, plexus profond. La texture capillaire fine et dense ainsi que la zone avasculaire centrale (ZAC) sont fidèlement reproduites par le modèle génératif.

Bibliographie

- [1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4190–4197, 2020.
- [2] Ilerioluwakiiye Abolade, Prince Mireku, Kelechi Chibundu, Peace Ododo, Emmanuel Idoko, Promise Omoigui, and Solomon Odelola. VAMAE : Vessel-aware masked autoencoders for OCT angiography. *arXiv preprint arXiv :2604.06583*, 2026.
- [3] Ghadah Alwakid, Walaa Gouda, and Muhammad Humayun. Deep learning-based prediction of diabetic retinopathy using CLAHE and ESRGAN for enhancement. *Healthcare*, 11(6) :863, 2023.
- [4] L. Arnould et al. A retinal OCT-angiography and cardiovascular SStatus (RASTA) dataset. *Data*, 8(10) :147, 2023.
- [5] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv :2108.07258*, 2021.
- [6] Mathilde Caron et al. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021.
- [7] Tianqi Chen and Carlos Guestrin. XGBoost : A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.
- [8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020.
- [9] Carol Y. Cheung et al. Retinal vascular geometry and 6-year incidence and progression of diabetic retinopathy. *Diabetes Care*, 31(7) :1488–1492, 2017.
- [10] Carol Y. Cheung, An Ran Ran, Siying Wang, Victor T. T. Chan, Katy Sham, Ruiye Shen, et al. Validation of a deep-learning-based retinal biomarker (Reti-CVD) in the prediction of cardiovascular disease : data from UK Biobank. *BMC Medicine*, 21 :1–12, 2023.
- [11] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3) :273–297, 1995.

- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words : Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021.
- [13] Alejandro F. Frangi, Wiro J. Niessen, Koen L. Vincken, and Max A. Viergever. Multiscale vessel enhancement filtering. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 130–137, 1998.
- [14] Antonio Gallo et al. Relationship of quantitative retinal capillary network and myocardial remodeling in systemic hypertension. *Journal of the American Heart Association*, 11 :e024226, 2022.
- [15] C. Germanese, L. Arnould, and F. Mériaudeau. Artificial intelligence-based prediction of neurocardiovascular risk score from retinal swept-source optical coherence tomography-angiography. *Scientific Reports*, 14 :27089, 2024.
- [16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022.
- [17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.
- [18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33 :6840–6851, 2020.
- [19] Junlin Hou, Sicen Liu, Yequan Bie, Hongmei Wang, Andong Tan, Luyang Luo, and Hao Chen. Self-eXplainable AI for medical image analysis : A survey and new outlooks. *arXiv preprint arXiv :2410.02331*, 2024.
- [20] Edward J. Hu, Yelong Shen, Phillip Wallis, et al. Lora : Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022.
- [21] Pavel Izmailov, Dmitrii Podoprikin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Uncertainty in Artificial Intelligence (UAI)*, 2018.
- [22] Norman M. Keith, Henry P. Wagener, and N. W. Barker. Some different types of essential hypertension : their course and prognosis. *American Journal of Medical Sciences*, 197 :332–343, 1939.
- [23] Y. Li et al. Research advances on artificial intelligence assisted diagnosis and risk assessment in cardiovascular disease using retinal imaging. *Frontiers in Cardiovascular Medicine*, 2025.

- [24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.
- [25] Gregory Y. H. Lip et al. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach. *Chest*, 137(2) :263–272, 2010.
- [26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.
- [27] Zaid Mammo et al. Quantitative noninvasive angiography of the fovea centralis using speckle variance optical coherence tomography. *Investigative Ophthalmology & Visual Science*, 56(9) :5074–5080, 2015.
- [28] Yong Tao Ong, Tien Yin Wong, Ronald Klein, Barbara E. K. Klein, Paul Mitchell, A. Richey Sharrett, David J. Couper, and Carol Y. Cheung. Systemic hypertension associated retinal microvascular changes can be detected with optical coherence tomography angiography. *Scientific Reports*, 10 :9580, 2020.
- [29] Ryan Poplin et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2 :158–164, 2018.
- [30] Jianing Qiu et al. VisionFM : a multi-modal multi-task vision foundation model for generalizable ophthalmic artificial intelligence. *arXiv preprint arXiv :2310.04992*, 2023.
- [31] Alireza Rezaei et al. Fundus image color adjustment. *Translational Vision Science & Technology*, 6(2) :3, 2017.
- [32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.
- [33] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, pages 2234–2242, 2016.
- [34] SCORE2 Working Group and ESC Cardiovascular Risk Collaboration. SCORE2 risk prediction algorithms : new models to estimate 10-year risk of cardiovascular disease in Europe. *European Heart Journal*, 42(25) :2439–2454, 2021.
- [35] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3) :379–423, 1948.
- [36] Mingxing Tan and Quoc V. Le. EfficientNetV2 : Smaller models and faster training. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139, pages 10096–10106, 2021.

- [37] Emanuele Trucco et al. Validating retinal fundus image analysis algorithms : Issues and a proposal. *Investigative Ophthalmology & Visual Science*, 54(5) :3546–3559, 2013.
- [38] Ashish Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [39] Lehan Wang, Hualiang Wang, Chubin Ou, Lushi Chen, Yunyi Liang, and Xiaomeng Li. VAMPIRE : Uncovering vessel directional and morphological information from OCTA images for cardiovascular disease risk factor prediction. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2025.
- [40] Peter W. F. Wilson, Ralph B. D’Agostino, Daniel Levy, Albert M. Belanger, Halit Silbershatz, and William B. Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18) :1837–1847, 1998.
- [41] Tien Y. Wong, Ronald Klein, A. Richey Sharrett, Bruce B. Duncan, David J. Couper, James M. Tielsch, Barbara E. Klein, and Lloyd D. Hubbard. Retinal arteriolar narrowing and risk of coronary heart disease in men and women : the atherosclerosis risk in communities study. *JAMA*, 287(9) :1153–1159, 2002.
- [42] World Health Organization. Cardiovascular diseases (CVDs) — key facts. [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)), 2021.
- [43] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup : Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.
- [44] Yukun Zhou. Global Diffusion Model : Diffusion model docker for retinal image synthesis. https://github.com/rmaphoh/Global_Diffusion_Model, 2025.
- [45] Yukun Zhou, Mark A. Chia, Siegfried K. Wagner, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622 :156–163, 2023.