الجمهورية الجزائرية الديمقراطية الشعبية

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

وزارة التعليم العالي والبحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جامعة أبو بكر بلقايد – تلمسان -

Université Aboubekr Belkaïd – Tlemcen – Faculté de TECHNOLOGIE

MÉMOIRE

Présenté pour l'obtention du **diplôme** de **MASTER**

En : Génie Biomédical

Spécialité : Informatique biomédicale et hospitalière

Par : CHERRAK Mardjana

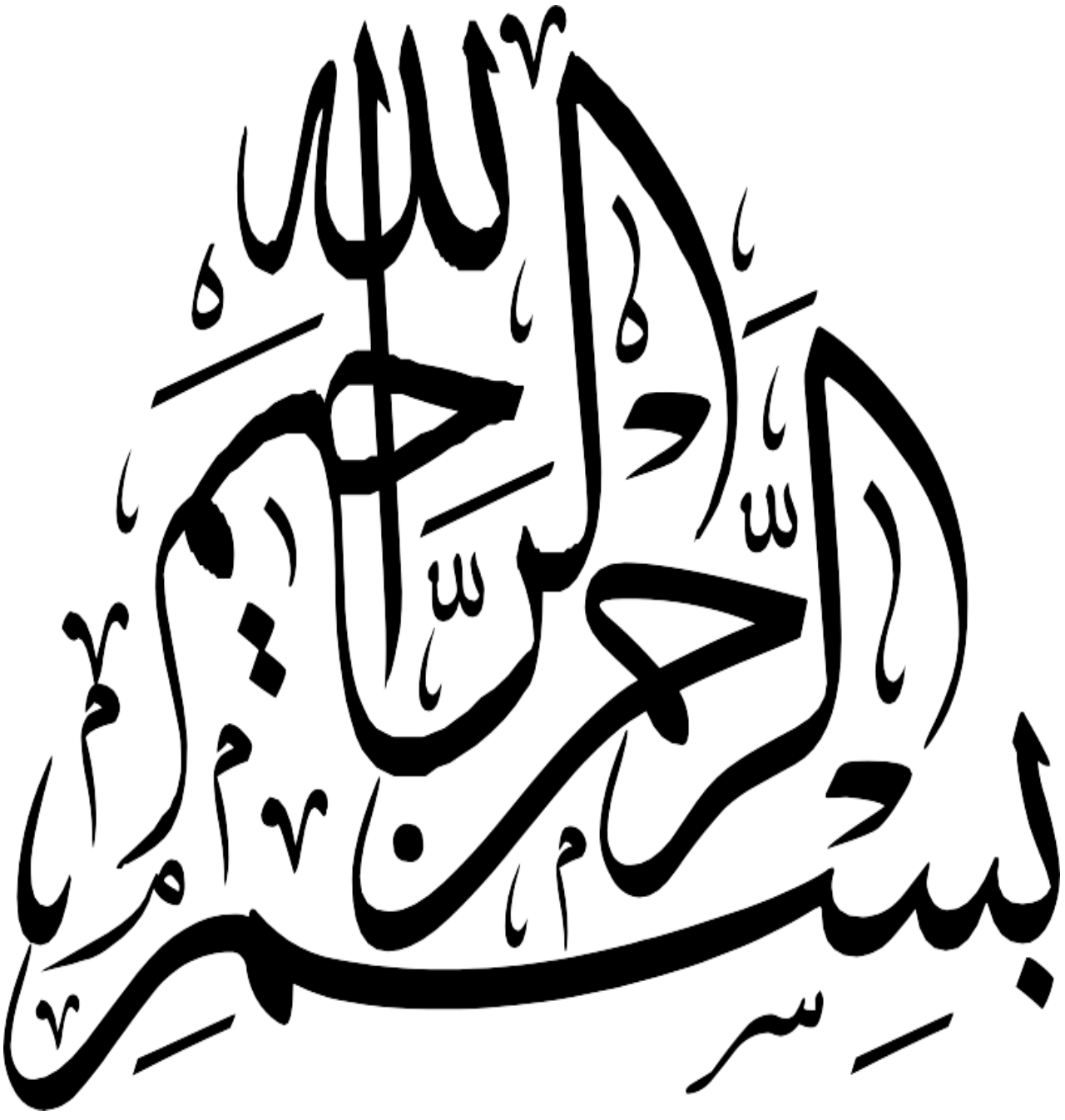
Sujet

Approches d'apprentissage profond pour la détection précoce des maladies : Jeux de données déséquilibrés

Soutenu publiquement, le 27/06/2026, devant le jury composé de :

M/GAOUAR Adil	MCB	Université de Tlemcen	Président
Mme/BENSMAIL Ilham	MCB	Université de Tlemcen	Examinatrice
Mme/BELAROUCI Sara	MCB	Université de Tlemcen	Encadrante
Mme/YOUBI Fatiha	Docteur	Université de Tlemcen	Co-Encadrante
M/ CHIKH Mohammed Amine	Prof	Université de Tlemcen	Invité

Année universitaire: 2025-2026


بسم الله الرحمن الرحيم

Dédicace

بسم الله الرحمن الرحيم

﴿وَمَا تَوْفِيقِي إِلَّا بِاللَّهِ ۚ عَلَيْهِ تَوَكَّلْتُ وَإِلَيْهِ أُنِيبُ﴾

الحمد لله الذي بنعمته تتم الصالحات، والصلاة والسلام على سيدنا محمد صلى الله عليه وسلم.

ها أنا اليوم أقف عند نهاية هذا الطريق، وقد ظننته طويلًا وثقيلًا، لكنه مرّ... ومرّ معه كل التعب الذي صار الآن ذكرى جميلة، لم تكن الرحلة سهلة، لكن لطف الله كان أكبر من كل الصعاب، فالحمد لله الذي يسر البدايات وبلغنا النهايات بفضله وكرمه.

إلى من أحمل اسمه بكل فخر وامتنان... إلى **أبي العزيز**،

ذلك الرجل الذي لم يكن مجرد أب، بل كان سندًا صامتًا، وقوة لا تُقال بل تُحس،كان يفتح لي أبواب الأمل حين تُغلق كل الأبواب، ويعلّمني أن الطريق لا يُهزم مهما طال.

وإلى **أمي**... إلى نبض قلبي الحقيقي،

إلى الدعاء الذي لم ينقطع يومًا، وإلى القلب الذي كان يسبقني في الخوف ويحميني دون أن أرى. كانت أمي دائمًا الوطن الذي أعود إليه مهما ابتعدت.

وإلى أخي وتوأم روحي **المقداد يوسف**،

الذي لم يكن مجرد أخ، بل كان ظلّي حين يثقلني الواقع، وضحكتي حين تضيق بي الدنيا.

كان يزرع فيّ الحياة من جديد كلما ظننت أنني فقدتها، ويقف بجانبي حين كنت أقاوم فكرة الاستسلام.

وإلى أخي الصغير **أحمد سعيد**،

بهجة البيت ونقطة الضوء التي تخفف ثقل الأيام، وابتسامته التي كانت كفيلة بإعادة شيء من التوازن لكل شيء.

وإلى **عائلتي الكبيرة**... من أجداد وأعمام وعمات وخالات،

الذين شكلوا حولي دائرة من الدعاء والاطمئنان، وجعلوا من الانتماء قوة أستند إليها في كل خطوة.

وإلى **صديقاتي** اللواتي كنّ أكثر من مجرد رفقة،

كنّ حضنًا معنويًا في لحظات الضعف، ودليلًا على أن بعض العلاقات تُشبه النجاة أكثر مما تُشبه الصداقة.

أثبتن لي أن الصديق الحقيقي لا يظهر في القمم فقط، بل يُعرف في لحظات الانهيار.

وإلى زميلي **خابو عبد الحق**،

الذي كان أكثر من زميل دراسة، كان بوصلة حقيقية أعادتني إلى الطريق كلما ضعت وسط التفاصيل والتعب، وشارك معي هذا المسار بكل ما فيه من صعوبة وإنجاز.

وإلى مشرفتي **بلعروسي سارة**،

التي لم تكن فقط إشرافًا أكاديميًا، بل كانت إيمانًا يُزرع في داخلي حين كان الشك يحاول أن يملأ الفراغ.

شكراً لها لأنها آمنت بي حين كنت أحتاج فقط إلى من يؤمن.

وأخيرًا، إلى كل من مرّ في هذه الرحلة، ولو بكلمة، أو موقف، أو دعاء صادق،

لكم أثر لا يُنسى، وشكر لا يكفيه التعبير، لكنه يبقى محفوظًا في القلب ما حييت.

شراك مرجانة

Remerciements

Avant tout, je remercie Dieu, le Tout-Puissant, de m'avoir accordé la force, la patience et la persévérance nécessaires pour mener à bien ce modeste travail.

Je tiens à exprimer ma profonde gratitude à mon encadrante, ***Mme BELAROUCI Sara****, pour son accompagnement précieux tout au long de ce projet. Je la remercie sincèrement pour sa disponibilité, ses conseils avisés, son soutien constant et la confiance qu'elle m'a accordée. Ses orientations et son expertise ont grandement contribué à la réalisation de ce travail.*

J'adresse également mes vifs remerciements à ma co-encadrante, ***Mme YOUBI Fatiha****, pour son aide, ses recommandations pertinentes et son suivi tout au long de cette étude.*

Je remercie chaleureusement les honorables membres du jury qui ont accepté d'évaluer ce travail. Mes sincères remerciements vont à ***M. GAOUAR Adil****, pour avoir accepté de présider ce jury, ainsi qu'à* ***Mme BENSMAIL Ilham****, pour avoir accepté d'examiner ce mémoire. Je les remercie pour l'intérêt porté à ce travail et pour leurs remarques constructives qui contribueront à son amélioration.*

J'adresse également mes remerciements à ***M. CHIKH Mohammed Amine*** *pour sa présence et l'honneur qu'il me fait en participant à cette soutenance.*

Je remercie enfin l'ensemble des enseignants de l'Université de Tlemcen pour la qualité de leur enseignement et les connaissances qu'ils m'ont transmises tout au long de mon parcours universitaire.

Enfin, j'exprime ma reconnaissance à toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de ce projet de fin d'études.

Merci à tous

Résumé

Le diagnostic précoce est un enjeu de santé publique, mais l'intelligence artificielle appliquée à l'imagerie se heurte au déséquilibre des classes, qui biaise les modèles vers les cas majoritaires. Ce mémoire propose un pipeline hybride corrigeant ce biais et assurant la transparence, combinant trois leviers (suréchantillonnage ciblé, augmentation des données et pondération adaptative des classes) dans un apprentissage en deux phases, avec EfficientNetB2 comme architecture de référence, sélectionnée parmi sept réseaux de neurones convolutifs.

Les expérimentations, menées sur le jeu de données Alzheimer MRI pour la classification multiclasse des stades de démence, ont permis d'atteindre une accuracy de 96,72 %, un F1-score de 97,54 % et un rappel de 97,63 %, résultats comparables ou supérieurs à des architectures d'ensemble (Deep Ensemble) nettement plus lourdes et complexes. La capacité de généralisation de l'approche a été confirmée sur trois jeux de données indépendants (tumeurs cérébrales binaires, tumeurs cérébrales multiclasses, ostéoarthrite du genou), démontrant ainsi la transversalité pathologique du pipeline proposé.

Sur le plan de l'interprétabilité, un protocole de validation rigoureux de l'IA explicable (XAI) a été mis en place. Fondé sur la méthode Grad-CAM, il combine des tests de cohérence mathématique, une analyse d'occlusion causale et un masquage anatomique. Ce cadre a permis de démontrer que les décisions du réseau reposent sur des biomarqueurs anatomocliniques pertinents, dont la cohérence a été validée qualitativement par une radiologue spécialiste.

Ce travail, en alliant précision diagnostique, robustesse et transparence décisionnelle, propose un cadre méthodologique opérationnel pour le développement d'outils d'intelligence artificielle fiables et interprétables en routine clinique. Il ouvre par ailleurs des perspectives prometteuses pour l'extension de cette approche à d'autres pathologies et modalités d'imagerie, ainsi que pour son intégration dans des dispositifs médicaux certifiés.

Mots-clés : Réseaux de neurones convolutifs, déséquilibre des classes, EfficientNetB2, maladie d'Alzheimer, imagerie médicale, Grad-CAM, explicabilité (XAI), validation clinique qualitative.

Abstract

Early diagnosis is a public health priority, yet artificial intelligence applied to medical imaging is hindered by class imbalance, biasing models toward majority classes. This thesis presents a hybrid pipeline that mitigates this bias while ensuring transparency, combining targeted oversampling, data augmentation, and adaptive class weighting within a two-phase learning framework. EfficientNetB2 is selected as the reference architecture from a benchmark of seven convolutional neural networks.

Experiments on the Alzheimer MRI dataset for multi-class dementia classification achieved 96.72% accuracy, 97.54% F1-score, and 97.63% recall, outperforming significantly heavier ensemble architectures (Deep Ensemble). Generalizability was confirmed on three independent datasets (binary and multi-class brain tumors, knee osteoarthritis), demonstrating the approach's pathological transversality.

For interpretability, a rigorous XAI validation protocol based on Grad-CAM was established, combining mathematical consistency tests, causal occlusion analysis, and anatomical masking. This framework confirmed that the network's decisions rely on relevant anatomo-clinical biomarkers, qualitatively validated by a specialist radiologist.

By combining diagnostic accuracy, robustness, and decision transparency, this work provides an operational framework for reliable and interpretable AI tools in clinical routine. It also opens promising perspectives for extending the approach to other pathologies and imaging modalities, and for integration into certified medical devices.

Keywords: Convolutional Neural Networks, class imbalance, EfficientNetB2, Alzheimer's disease, medical imaging, Grad-CAM, explainability (XAI), qualitative clinical validation

ملخص

يُعدّ التشخيص المبكر أولوية في مجال الصحة العامة، ورغم ذلك، فإن تطبيقات الذكاء الاصطناعي في التصوير الطبي تواجه عائقاً يتمثل في **عدم توازن البيانات** (*Class Imbalance*)، مما يسبب انحيازاً في النماذج لصالح الفئات الأكثر تكراراً (الأغلبية). تقدم هذه الأطروحة خط معالجة هجين (*Hybrid Pipeline*) يحدّ من هذا الانحياز مع ضمان الشفافية، حيث يدمج بين تقنيات الإفراط في أخذ العينات المستهدفة (Targeted Oversampling)، وتعزيز البيانات (*Data Augmentation*)، والوزن التكيفي للفئات (Adaptive Class Weighting) ضمن إطار تعلم ثنائي المراحل. وقد تم اختيار نموذج ***EfficientNetB2*** كبنية مرجعية بعد إجراء مقارنة معيارية شملت سبع شبكات عصبية تلافيفية (CNNs).

أظهرت التجارب التي أُجريت على قاعدة بيانات الصور بالرنين المغناطيسي لمرض آلزهايمر (*Alzheimer MRI dataset*) لتصنيف الخرف متعدد الفئات دقة بلغت 96.72 %، ومقياس F1 بنسبة 97.54 %، وحساسية (*Recall*) بنسبة 97.63 %، متفوقةً بشكل ملحوظ على بنى النماذج التجميعية (*Deep Ensemble*) الأكثر تعقيداً وضخامة. كما تم تأكيد القدرة على التعميم (*Generalizability*) عبر ثلاث قواعد بيانات مستقلة (أورام الدماغ الثنائية ومتعددة الفئات، والفصال العظمي للركبة)، مما يثبت كفاءة هذا النهج عبر مختلف الأمراض (*Pathological Transversality*).

ومن أجل تحقيق قابلية التفسير، تم وضع بروتوكول تحقق صارم للذكاء الاصطناعي القابل للتفسير يعتمد على تقنية *Grad-CAM*، حيث يجمع بين اختبارات الاتساق الرياضي، وتحليل الحجب السببي (*Causal Occlusion Analysis*)، والقناع التشريحي (*Anatomical Masking*). وقد أكد هذا الإطار أن قرارات الشبكة العصبية تعتمد على مؤشرات حيوية تشريحية وسريرية ذات صلة، وتم التحقق منها نوعياً بواسطة طبيب أشعة متخصص.

من خلال الجمع بين دقة التشخيص، والمتانة، وشفافية القرار، يقدم هذا العمل إطاراً تشغيلياً لأدوات ذكاء اصطناعي موثوقة وقابلة للتفسير في الممارسات السريرية اليومية. كما يفتح آفاقاً واعدة لتوسيع نطاق هذا النهج ليشمل أمراضاً وأساليب تصوير طبي أخرى، ودمجه في الأجهزة الطبية المعتمدة.

الكلمات المفتاحية: الشبكات العصبية التلافيفية، عدم توازن البيانات، EfficientNetB2، مرض آلزهايمر، التصوير الطبي، Grad-CAM، الذكاء الاصطناعي القابل للتفسير (XAI)، التحقق السريري النوعي.

Table des matières

Dédicace ..... **I**
Remerciements ..... **II**
Résumé ..... **III**
Abstract ..... **IV**
ملخص ..... **V**
Table des matières VI
Liste des figures ..... **XI**
Liste des tableaux ..... **XV**
Liste d'abréviations XVI
Introduction générale 1
Chapitre 1 : Jeux de données médicaux déséquilibrés 5
1. Introduction 6
2. Définition et caractéristiques des jeux de données déséquilibrés 6
2.1. Déséquilibre inter-classes 6
2.2. Rareté des événements cliniques 8
2.3. Distribution non uniforme des données 9
3. Déséquilibre dans le diagnostic médical 12
3.1. Cas du diagnostic précoce 12
3.2. Imagerie médicale et données cliniques 12
3.3. Conséquences des erreurs sur les classes minoritaires 14
4. Problèmes induits par le déséquilibre de classes 14
4.1. Faible représentation des classes minoritaires 14
4.2. Chevauchement entre classes 15
4.3. Données bruitées et ambiguës 16
4.4. Complexité des frontières de décision 17
5. Impact du déséquilibre sur les modèles d'apprentissage profond 17
5.1. Biais des modèles vers les classes majoritaires 19
5.2. Dégradation des performances sur les classes rares 19
5.3. Limites de l'exactitude comme métrique unique 19
6. Limites des approches existantes 21
6.1. Dépendance au type de données 21
6.2. Difficulté de généralisation 21
6.3. Risque de surapprentissage 22
7. Conclusion 23

Chapitre 2 : Stratégies de gestion du déséquilibre et Apprentissage profond 24
1. Introduction 25
2. Stratégies de gestion du déséquilibre 25
2.1. Méthodes basées sur les données 26
2.1.1. Techniques de suréchantillonnage 26
2.1.2. Techniques de sous-échantillonnage 27
2.1.3. Augmentation de données 28
2.1.4. GANs 29
2.1.5. Limites de l'augmntation excessive des données 29
2.2. Méthodes algorithmiques 30
2.2.1. Apprentissage sensible au coût 30
2.2.2. Focal Loss 31
2.2.3. Classification hiérarchique 31
2.2.4. Apprentissage en deux phases 31
2.3. Stratégies hybrides 31
2.3.1. Combinaison des approches au niveau des données et des algorithmes 32
2.3.2. Importance et limites des approches hybrides 32
3. Fondements de l'apprentissage profond 32
3.1. Intelligence artificielle 33
3.2. Apprentissage automatique 33
3.3. Apprentissage profond 34
3.4. Réseaux de neurones convolutifs 35
3.4.1. Structure générale des CNNs 35
3.4.2. Principales couches d'un CNN 36
3.4.3. Fonctions de perte 41
3.4.4. Algorithmes d'optimisation 42
3.5. Hyperparamètres et stratégie d'entraînement 42
3.6. Apprentissage par transfert 42
3.7. Architectures CNN de classification les plus utilisées 43
3.7.1. EfficientNetB2 44
3.7.2. ResNet50 45
3.7.3. DenseNet121 46
3.7.4. InceptionV3 46
3.7.5. VGG19 47
3.7.6. Xception 47
3.7.7. MobileNetV2 48
4. Métriques d'évaluation des performances 49

4.1. Matrice de confusion 49
4.2. Métriques de classification 49
4.3. Courbe ROC et AUC 50
4.4. Courbe PRC 50
4.5. Courbe DCA 51
5. L'intelligence artificielle interprétable (Explainable AI, XAI) en médecine 51
5.1. Principe de Grad-CAM 51
5.2. Interprétation visuelle des décisions cliniques 53
5.3. Limites intrinsèques des cartes de chaleur 53
5.4. Validation quantitative et causale des méthodes XAI 54
5.4.1. Tests de cohérence élémentaires (*Sanity Checks*) 54
5.4.2. Test d'occlusion et validation causale 55
5.4.3. Analyse de superposition géométrique : IoU et coefficient de Dice 55
6. Analyse de la littérature et positionnement 55
6.1. Travaux existants et études récentes 56
6.2. Positionnement de la méthode proposée 59
7. Conclusion 59
Chapitre 3 : Méthodologie expérimentale, Résultats et Discussion 60
1. Introduction 61
2. Environnement de développement et outils technologiques 62
3. Description des jeux de données 63
3.1. Jeu de données Alzheimer MRI Preprocessed Dataset 63
3.2. Jeu de données Brain MRI Images 63
3.3. Jeu de données Brain Tumor MRI Dataset 64
3.4. Jeu de données Osteoarthritis Prediction 65
4. Approche Méthodologique Proposée 65
4.1. Pipeline de prétraitement et équilibrage des classes 66
4.1.1 Chargement et organisation du jeu de données 66
4.1.2. Division stratifiée et étanche de la cohorte expérimentale 67
4.1.3. Algorithme de Suréchantillonnage Sélectif et Contrôlé 68
4.1.4. Pipeline d'Augmentation Visuelle et Prétraitement Asynchrone 69
4.1.5. Optimisation Matérielle et Parallélisation du Flux 70
4.1.6. Pondération Résiduelle Compensatoire 71
4.2. Stratégie d'apprentissage par transfert et optimisation convolutive 71
4.2.1. Architecture de la tête de classification 72
4.2.2. Stratégie d'entraînement en deux phases 72
4.3. Phase d'Explicabilité Post-Hoc et de Validation Quantitative 73

4.3.1. Principe et implémentation de Grad-CAM 73
4.3.2. Cadre de validation quantitative 74
4.3.3. Filtrage anatomique — Brain-Masked Grad-CAM 75
5. Résultats expérimentaux et discussion 76
5.1. Expérimentation 1 : Sélection de l'Architecture de Référence 76
5.2. Expérimentation 2 : Évaluation des stratégies de gestion du déséquilibre des classes 80
5.2.1. Étude expérimentale et analyse multi-niveaux de l'approche hybride proposée 84
5.2.2. Analyse quantitative des courbes de convergence et validation clinique du modèle optimal 85
5.2.3. Analyse visuelle des résultats de prédiction du modèle proposé 88
5.2.4. Comparaison de nos résultats avec l'état de l'art 89
5.3. Expérimentation 3 : Analyse de l'interprétabilité et validation de la robustesse des explications par Grad-CAM 90
5.3.1. Validation de la robustesse de Grad-CAM : Tests de cohérence élémentaires 92
5.4.2. Validation anatomique et causale par Brain-Masking 95
5.4. Expérimentation 4 : Évaluation de la généralisabilité et de la transversalité pathologique 98
5.4.1. Analyse globale multi-métriques des scénarios de neuroimagerie (Brain MRI et Brain Tumor) 98
5.4.2. Validation de la transversalité anatomoclinique (Osteoarthritis Prediction Dataset) 99
5.4.3. Analyse qualitative par Grad-CAM et validation par expertise clinique 101
6. Conclusion 103
Conclusion générale 105
Annexe Oasis Med —Plateforme d'Aide à la Décision Clinique 109
1. Présentation Générale de la Plateforme 110
2. Interfaces de la plateforme 110
2.1. Tableau de bord principal 110
2.2. Interface de gestion des patients 111
2.3. Gestion des examens en attente (Pending Cases) 111
2.4. Interface de création d'un nouveau cas clinique 112
2.5. Chargement de l'image médicale 112
2.6. Résultats de l'analyse IA et visualisation Grad-CAM 113
2.7. Validation XAI et explication clinique 113
2.8. Rapport médical 114
2.9. Validation finale et signature du rapport 115
2.10. Système de notifications 116
2.11. Rapports finalisés 116

2.12. Paramètres (Settings).....117
Références bibliographiques.....119

Liste des figures

Figure 1.1:Comparaison entre une répartition équilibrée et une structure déséquilibrée des classes au sein d'un jeu de données médicales. 7
Figure 1.2: Exemples de régions tumorales faiblement représentées dans des images IRM cérébrales. 10
Figure 1.3:llustration de la variabilité inter-scanners dans les acquisitions IRM cérébrales. ..13
Figure 1.4:Représentation schématique de la faible proportion des classes minoritaires dans une base de données déséquilibrée. 15
Figure 1.5:Représentation schématique du phénomène de chevauchement entre classes au sein de l'espace de caractéristiques. 16
Figure 1.6:Frontières de décision linéaires et non linéaires selon la séparabilité des classes dans un espace de caractéristiques bidimensionnel. 17
Figure 1.7: Évolution des métriques d'évaluation en fonction du seuil de décision dans un contexte de fort déséquilibre de classes. 20
Figure 1.8:Illustration du phénomène de surapprentissage dans les modèles entraînés sur des données déséquilibrées. 22
Figure 2.1:. Principe du suréchantillonnage aléatoire traditionnel. 26
Figure 2.2:Principe du sous-échantillonnage aléatoire traditionnel. 28
Figure 2.3:Architecture générale d'un GAN : compétition entre le générateur et le discriminateur pour produire des images synthétiques réalistes. 29
Figure 2.4:Relation hiérarchique de l'intelligence artificielle : IA → Machine Learning → Deep Learning. 32
Figure 2.5:La différence entre l'apprentissage automatique traditionnel et l'apprentissage profond. 34
Figure 2.6:. Illustration d'un exemple d'architecture CNN pour la classification d'images. 35
Figure 2.7:Exemple d'opération de convolution 36
Figure 2.8:Principe de fonctionnement de la fonction ReLU : annulation des valeurs négatives et maintien des valeurs positives. 37
Figure 2.9:Les courbes des fonctions d'activation. 38
Figure 2.10:Les deux principaux types de pooling : Max Pooling et Average Pooling. 38
Figure 2.11:Architecture d'une couche entièrement connectée 39

Figure 2.12:Structure d'un réseau de neurones avant (A) et après (B) l'application de la régularisation par Dropout. 40
Figure 2.13:Principe du Transfer Learning. 42
Figure 2.14:Principe du compound scaling d'EfficientNet 44
Figure 2.15: Architecture détaillée duResNet50 44
Figure 2.16: Architecture DenseNet121 45
Figure 2.17:Architecture générale du module InceptionV3 46
Figure 2.18:Architecture séquentielle et profonde du réseau VGG19. 46
Figure 2.19:Architecteur du réseau Xception 47
Figure 2.20:Architecture et structure des blocs résiduels inversés de MobileNetV2. 47
Figure 2.21:Pipeline complet de la méthode Grad CAM : de l'image d'entrée aux gradients de la dernière couche convolutionnelle, jusqu'à la heatmap superposée à l'image originale. 51
Figure 2.22:Exemple illustrant une heatmap médicalement pertinente et une heatmap influencée par des facteurs de confusion 53
Figure 3.1:Exemples d'images issues de la base de données Alzheimer MRI Preprocessed. . 62
Figure 3.2:Exemples d'images issues de la base de données Brain MRI Images (Tumor and No Tumor) 63
Figure 3.3:Exemples d'images issues de la base de données Brain Tumor MRI Dataset. 63
Figure 3.4:Exemples d'images issues de la base de données Osteoarthritis Prediction 64
Figure 3.5:Organigramme général du protocole méthodologique développé pour la classification multiclasses des stades d'Alzheimer à partir des image IRM. 65
Figure 3.6:Distribution des classes avant rééquilibrage 66
Figure 3.7:Distribution des classes après division stratifiée 67
Figure 3.8:Distribution des classes (train) avant et après un Suréchantillonnage Sélectif et Contrôlé 68
Figure 3.9:Exemples d'images après application Augmentation. 68
Figure 3.10:Matrices de confusion des différents modèles testés 78
Figure 3.11:Matrices de confusion d'EfficientNetB2 78
Figure 3.12:Confrontation des courbes d'apprentissage (accuracy et perte) entre les architectures VGG19 et EfficientNetB2 au cours des phases d'entraînement et de validation. 79
Figure 3.13:Comparaison des profils de distribution des classes nosologiques avant et après application des méthodes de suréchantillonnage. 81
Figure 3.14:Évolution comparative de l'accuracy et de la perte pour les stratégies de suréchantillonnage au cours des époques. 82

Figure 3.15:Matrices de confusion obtenues sur le jeu de test : (a) configuration initiale avant traitement du déséquilibre et (b) après optimisation par notre approche hybride unifiée. 84
Figure 3.16:Évolution des courbes d'apprentissage (accuracy et perte) au cours des deux phases du protocole Two-Phase Learning pour le modèle hybride optimal. 85
Figure 3.17:Courbes ROC et Précision-Rappel du modèle optimal pour les quatre classes de démence. 86
Figure 3.18:Analyse des courbes de décision (DCA) évaluant le bénéfice net clinique du modèle pour chaque stade diagnostique. 87
Figure 3.19:Exemples d'inférences correctes générées par notre modèle hybride sur des coupes IRM réelles du jeu de test, pour les quatre stades de démence. 88
Figure 3.20:Visualisation de l'interprétabilité par cartes d'activation Grad CAM pour les quatre stades cliniques de la maladie d'Alzheimer. 90
Figure 3.21:Évaluation de la robustesse par le Test 1. 92
Figure 3.22:Évaluation de la robustesse par le Test 2. 92
Figure 3.23:Évaluation de la robustesse par le Test 3. 93
Figure 3.24:Évaluation de la robustesse par le Test 4. 93
Figure 3.25:Évaluation de la validité causale par le Test 5 (Test d'occlusion globale).....94
Figure 3.26:Comparaison qualitative du pipeline Brain-Masked Grad-CAM. 95
Figure 3.27:. Métriques comparatives de généralisation et de transversalité pathologique sur trois bases de données indépendantes.....100
Figure 3.28:Visualisation de l'interprétabilité par Grad-CAM pour la détection de tumeurs cérébrales multiclasses. 101
Figure 3.29:Visualisation de l'interprétabilité par Grad-CAM pour la détection de l'ostéoarthrite du genou. 102
Figure A.1 Tableau de bord principal avec statistiques cliniques et indicateurs de performance IA……110
Figure A.2 Interface de liste des patients avec informations démographiques et accès rapide aux dossiers……..……111
Figure A.3 Interface des cas en attente avec filtres par statut et niveaux d'urgence……..111
Figure A.4 Interface de création de cas — étape 1 : dossier patient et sélection du type d'examen IA……112
Figure A.5 Interface de téléchargement d'image médicale avec prévisualisation et métadonnées DICOM…….…….112
Figure A.6 Station de travail IA — résultats de prédiction et visualisation Grad-CAM avec régions anatomiques clés……113

Figure A.7 Interface de validation XAI avec résultats des tests de robustesse et explication clinique par IA……114
Figure A.8 Vue détaillée du système d'explication clinique basé sur des règles (panneau droit) ……114
Figure A.9 Interface de rapport médical avec résumé IA, diagnostic, commentaires et recommandations cliniques…….115
Figure A.10 Interface de validation finale avec actions de décision clinique et signature numérique du médecin……115
Figure A.11 Centre de notifications avec alertes cliniques classées par priorité et statut de lecture……..116
Figure A.12 Interface des rapports finalisés avec statistiques globales et liste des comptes rendus……..116
Figure A.13 Interface des Informations du compte……..117
Figure A.14 : Interface de sécurité et accès……..117
Figure A.15 Interface des Notifications……...118
Figure A.16 Interface des Préférences d'affichage……..118

Liste des tableaux

Tableau 1.1:Les principaux types de déséquilibre de données rencontrés en contexte médical. 11
Tableau 1.2:Impacts des différents types de déséquilibre sur les modèles d'apprentissage profond. 18
Tableau 2.1:Synthèse comparative des principales architectures CNN utilisées en imagerie médicale.....43
Tableau 2.2:Structure d'une matrice de confusion binaire.....48
Tableau 3.1:Configuration des hyperparamètres pour la phase comparative.....75
Tableau 3.2:Performances comparatives des différentes architectures CNN.....76
Tableau 3.3: Synthèse comparative des performances des stratégies d'équilibrage avec l'architecture EfficientNetB2. 80
Tableau 3.4 :Comparaison de notre méthode avec l'état de l'art: 88
Tableau 3.5:Résultats des tests de cohérence élémentaires pour la validation de Grad CAM. 91
Tableau 3.6:Indicateurs de performance du test Brain-Masked Grad-CAM par classe diagnostique.....96
Tableau 3.7: Métriques comparatives de généralisation et de transversalité pathologique sur trois bases de données indépendantes.....97

Liste d'abréviations

ADNI:	Alzheimer's Disease Neuroimaging Initiative
AI Act:	Artificial Intelligence Act (réglementation européenne sur l'intelligence artificielle)
API:	Application Programming Interface (interface de programmation d'application)
AUC:	Area Under the Curve (aire sous la courbe)
AUTOTUNE:	paramètre d'optimisation automatique du parallélisme (tf.data.AUTOTUNE)
BMC:	BioMed Central (éditeur scientifique)
CAM:	Class Activation Mapping (cartographie d'activation de classe)
CNN:	Convolutional Neural Network (réseau de neurones convolutif)
CPU:	Central Processing Unit (unité centrale de traitement)
CW:	Class Weights (pondération adaptative des classes)
DA:	Data Augmentation (augmentation des données)
DCA:	Decision Curve Analysis (analyse par courbe de décision)
DGAN-WP-TL:	cadre combinant un GAN à divergence (Divergence GAN), une perte de Wasserstein avec pénalité de gradient (WP) et le Transfer Learning (TL)
ECG:	Électrocardiogramme
EfficientNetB2:	architecture de réseau de neurones convolutif de la famille EfficientNet, second variant
ENN:	Edited Nearest Neighbours (approche des voisins édités)
FC:	Fully Connected (couche entièrement connectée)
FDA:	Food and Drug Administration (agence américaine des produits alimentaires et médicamenteux)
FFST:	Fast Finite Shearlet Transform (transformée de Shearlet finie rapide)
G-mean:	Geometric Mean (moyenne géométrique de la sensibilité et de la spécificité)
GAN:	Generative Adversarial Network (réseau antagoniste génératif)
GLCM:	Gray-Level Co-occurrence Matrix (matrice de co-occurrence des niveaux de gris)
GPU:	Graphics Processing Unit (unité de traitement graphique)
Grad-CAM:	Gradient-weighted Class Activation Mapping
HOG:	Histogram of Oriented Gradients (histogramme de gradients orientés)
IA:	Intelligence Artificielle

IEEE:	Institute of Electrical and Electronics Engineers
IR:	Imbalance Ratio (ratio de déséquilibre)
IRM:	Imagerie par Résonance Magnétique
k-NN:	k-Nearest Neighbors (k plus proches voisins)
MAD:	Mean Absolute Deviation (écart absolu moyen)
MAMIR:	Multi-stage Adaptive Multi-class Imbalanced Resampling
MCC:	Matthews Correlation Coefficient (coefficient de corrélation de Matthews)
MRI:	Magnetic Resonance Imaging (imagerie par résonance magnétique)
OASIS:	Open Access Series of Imaging Studies
PR-AUC:	Precision-Recall - Area Under the Curve (aire sous la courbe précision-rappel)
PRC:	Precision-Recall Curve (courbe précision-rappel)
ROC:	Receiver Operating Characteristic
ROC-AUC:	Receiver Operating Characteristic - Area Under the Curve
RVB:	Rouge-Vert-Bleu (espace colorimétrique)
SGD:	Stochastic Gradient Descent (descente de gradient stochastique)
SMOTE:	Synthetic Minority Over-sampling Technique
SVM:	Support Vector Machine (machine à vecteurs de support)
TDM:	Tomodensitométrie (scanner)
TOS:	Target Oversampling (suréchantillonnage ciblé)
Two-PL:	Two-Phase Learning (apprentissage en deux phases)
VGG16 / VGG19:	Visual Geometry Group (architectures CNN à 16 ou 19 couches)
VRAM:	Video Random Access Memory (mémoire vidéo)
WCE:	Weighted Cross-Entropy (entropie croisée pondérée)
WGAN-GP:	Wasserstein GAN with Gradient Penalty (GAN de Wasserstein avec pénalité de gradient)
XAI:	Explainable Artificial Intelligence (intelligence artificielle explicable)



Introduction générale



Contexte

Chaque année, des millions de patients voient leur pronostic compromis par un diagnostic trop tardif. Le diagnostic précoce des maladies graves constitue donc un enjeu majeur de santé publique. L'intelligence artificielle, et en particulier l'apprentissage profond fondé sur les réseaux de neurones convolutifs (CNN), joue un rôle de premier plan en imagerie médicale en automatisant l'analyse et la classification des images. Elle constitue un outil d'aide au diagnostic incontournable pour améliorer la détection à un stade initial de pathologies telles que la maladie d'Alzheimer, les tumeurs cérébrales ou l'ostéoarthrite du genou. Cette détection précoce représente un enjeu médical et socio-économique considérable, car elle permet d'élargir les options thérapeutiques, d'optimiser la prise en charge des patients et de réduire significativement les coûts de santé associés à une prise en charge tardive.

Toutefois, les stades initiaux des maladies, cliniquement plus rares, sont systématiquement sous-représentés au sein des bases de données disponibles. Cette disparité numérique induit un risque élevé de faux négatifs (classant à tort un patient atteint comme sain), ce qui s'avère particulièrement critique en pratique médicale, où une détection tardive peut compromettre l'efficacité thérapeutique. Le développement de modèles capables de discriminer efficacement ces cas minoritaires constitue donc un défi essentiel pour le diagnostic assisté par ordinateur.

Problématique

Malgré les performances remarquables des modèles d'apprentissage profond, leur application clinique se heurte de manière récurrente à la distribution déséquilibrée des classes. Dans la majorité des bases de données médicales, les cas pathologiques sont en effet beaucoup moins nombreux que les cas témoins ou les sujets sains.

Ce déséquilibre sévère entraîne un biais d'apprentissage qui pousse les algorithmes à privilégier la classe majoritaire au détriment des catégories minoritaires, pourtant les plus déterminantes d'un point de vue diagnostique. En conséquence, un modèle peut afficher une exactitude globale (*accuracy*) trompeusement élevée tout en échouant systématiquement à détecter la pathologie recherchée, ce qui réduit fortement sa pertinence et sa fiabilité en contexte clinique réel.

Limites des approches existantes

Bien que de nombreuses méthodes aient été proposées pour traiter le déséquilibre des classes, plusieurs verrous subsistent dans la littérature actuelle. La plupart des études se limitent à l'exactitude globale, une métrique reconnue comme biaisée en contexte asymétrique, et n'exploitent que des stratégies de rééquilibrage isolées (rééchantillonnage seul ou pondération seule), dont l'efficacité demeure limitée lorsqu'elles sont appliquées indépendamment. Par

ailleurs, les mécanismes d'explicabilité visuelle y sont souvent insuffisamment validés sur le plan quantitatif, et les expérimentations se cantonnent généralement à une seule base de données ou à une unique pathologie, ce qui interroge la généralisabilité des résultats obtenus, et ne répondent pas aux exigences croissantes de transparence imposées par les réglementations récentes. Ces lacunes convergentes justifient le développement d'une approche hybride unifiée, combinant plusieurs techniques de rééquilibrage complémentaires, une évaluation multi-métrique standardisée et une analyse approfondie de l'interprétabilité du modèle.

Objectifs du travail

Ce travail s'inscrit dans une démarche de conception d'un pipeline d'aide au diagnostic robuste, capable de neutraliser le biais de distribution tout en garantissant la transparence des décisions du modèle. Les objectifs spécifiques de ce mémoire sont les suivants :

- Analyser les différentes formes de déséquilibre des classes et leur impact sur les fonctions de coût traditionnelles ;
- Identifier, parmi plusieurs architectures de pointe pré-entraînées, celle offrant le meilleur compromis entre performance et efficience computationnelle ;
- Développer et évaluer une stratégie hybride de gestion du déséquilibre, puis la confronter rigoureusement aux approches conventionnelles de la littérature ;
- Valider la capacité de généralisation et la transversalité de cette stratégie sur plusieurs bases de données médicales indépendantes ;
- Intégrer et valider quantitativement un mécanisme d'explicabilité fondé sur l'algorithme Grad-CAM, afin de certifier la pertinence clinique des prédictions du modèle.

Contributions principales

Les apports majeurs de ce travail se résument ainsi :

- Une étude comparative systématique de sept architectures CNN pré-entraînées, établissant la supériorité d'EfficientNetB2 pour le traitement de données médicales asymétriques.
- La conception d'une stratégie hybride originale combinant l'apprentissage en deux phases, le suréchantillonnage ciblé, l'augmentation des données et la pondération adaptative des classes, surpassant les méthodes de rééquilibrage appliquées individuellement.
- La validation de la robustesse de cette stratégie sur quatre bases de données médicales distinctes couvrant différentes tâches de classification et modalités d'imagerie (IRM et rayons X).

- La formulation d'un protocole complet d'explicabilité, s'appuyant sur un filtrage anatomique (Brain-Masked Grad-CAM) et sur des tests de validation quantitative et causale garantissant la pertinence clinique des cartes de saillance générées.

Organisation du mémoire

Pour répondre à ces problématiques, ce mémoire est structuré en trois chapitres principaux, organisés selon une progression allant du cadre théorique vers la validation expérimentale.

- Le **Chapitre 1** définit le phénomène de déséquilibre inter-classes, explicite les origines, notamment la rareté clinique des événements pathologiques, ainsi que ses conséquences structurelles sur les données, telles que le chevauchement de classes, l'ambiguïté des échantillons et la complexification des frontières de décision. Ce chapitre démontre l'impact de ce biais sur l'apprentissage des modèles d'apprentissage profond et met en évidence les limites de l'accuracy en tant que métrique d'évaluation, justifiant ainsi le recours à des approches méthodologiques alternatives.
- Le **Chapitre 2** passe en revue l'état de l'art des techniques de rééquilibrage, qu'elles agissent au niveau des données, des algorithmes ou selon des stratégies hybrides combinant les deux approches. Il rappelle les fondements théoriques des réseaux de neurones convolutifs et du transfert d'apprentissage, puis présente une étude comparative des architectures retenues (EfficientNetB2, ResNet50, DenseNet121 et VGG19). Ce chapitre introduit par ailleurs un ensemble de métriques robustes adaptées au contexte du déséquilibre, avant de formaliser le principe de l'IA explicable (XAI) à travers l'algorithme Grad-CAM et ses protocoles de validation associés.
- Le **Chapitre 3**, cœur applicatif de ce travail, décrit la configuration technologique adoptée, les bases de données exploitées, ainsi que notre pipeline méthodologique, structuré autour d'un prétraitement adaptatif, d'un entraînement séquentiel en deux phases et d'un mécanisme d'explicabilité Brain-Masked Grad-CAM. Ce chapitre expose et discute nos quatre contributions principales : la sélection empirique de l'architecture de référence, l'évaluation comparative de notre approche hybride, la validation causale de l'XAI, et la démonstration de la transversalité clinique du modèle proposé.

Ce travail est par ailleurs complété par une annexe présentant Oasis Med — Plateforme d'Aide à la Décision Clinique, un outil web qui opérationnalise le pipeline d'intelligence artificielle développé dans le cadre de ce mémoire.

Ce travail se clôt par une conclusion générale qui synthétise nos apports, discute les limites actuelles du système proposé et esquisse les perspectives de déploiement en milieu hospitalier.



Chapitre 1 :

Jeux de données médicaux déséquilibrés



1. Introduction

L'intelligence artificielle (IA), notamment via les réseaux de neurones convolutifs, joue un rôle croissant en imagerie médicale, qu'il s'agisse de la radiographie, du scanner, de l'échographie ou de la résonance magnétique (IRM). Face à la masse de données à analyser dans des délais restreints, ces modèles automatisent efficacement les tâches de classification et de détection, rivalisant parfois avec l'expertise humaine. Leur contribution majeure réside dans le diagnostic précoce, une étape cruciale pour optimiser le pronostic des patients et limiter les coûts de santé. Par exemple, pour la maladie d'Alzheimer, les premiers signes structurels sont visibles à l'IRM plusieurs années avant les manifestations cliniques, tandis que pour les tumeurs, une détection à un stade initial conditionne souvent la possibilité d'un traitement curatif.

Cependant, l'application des méthodes d'apprentissage profond en imagerie médicale se heurte à un défi majeur : le déséquilibre des classes. Les cas pathologiques sont naturellement bien moins fréquents que les cas sains ou bénins, notamment lorsqu'il s'agit de formes rares de maladies ou de stades avancés. Cette asymétrie, systématique dans les données cliniques, altère l'apprentissage automatique : un modèle entraîné sur de tels jeux de données risque de négliger les classes minoritaires, qui possèdent pourtant la plus forte importance médicale.

Ce chapitre explore ce problème dans sa globalité. Après avoir défini le déséquilibre des classes et analysé la vulnérabilité spécifique de l'imagerie médicale, nous évaluerons ses conséquences sur les modèles, en démontrant les biais des métriques d'évaluation classiques. Enfin, l'examen des limites des approches existantes ouvrira la voie aux méthodes spécialisées présentées au chapitre suivant.

2. Définition et caractéristiques des jeux de données déséquilibrés

Avant d’étudier les conséquences du déséquilibre des classes, il est nécessaire de mieux comprendre sa nature ainsi que les différentes formes qu'il peut prendre dans un contexte médical. Ce phénomène ne se manifeste pas toujours de manière identique, pouvant être global ou local, inter-classes ou intra-image, et son niveau de sévérité varie fortement d’un jeu de données à l'autre.

2.1. Déséquilibre inter-classes

La forme la plus directe du déséquilibre est celle que l'on observe entre les différentes catégories d'un problème de classification. On parle de déséquilibre inter-classes lorsque la distribution des échantillons n'est pas uniforme entre les classes à reconnaître. Ainsi, certaines catégories contiennent un très grand nombre d’exemples et sont qualifiées de classes majoritaires, tandis que d'autres en contiennent très peu, constituant les classes minoritaires [1].

Dans un tel contexte, le ratio de déséquilibre, défini comme le rapport entre l'effectif de la classe majoritaire et celui de la classe minoritaire, représente un indicateur majeur de la difficulté du problème. Ce ratio, souvent désigné sous le terme anglophone d'Imbalance Ratio (IR), peut être exprimé par la relation suivante [2]:

$$IR = \frac{N_{maj}}{N_{min}} \qquad (1.1)$$

Où N_{maj} et N_{min} désignent respectivement le nombre d'échantillons de la classe majoritaire et celui de la classe minoritaire. Un ratio de $10:1$ indique que la classe dominante est dix fois plus représentée que la classe rare. Un ratio de $100:1$ ou davantage définit un déséquilibre sévère, une situation courante en imagerie médicale. Par exemple, si un jeu de données contient 10 000 images normales et 100 images pathologiques, le ratio de déséquilibre est alors égal à $100:1$. La Figure1.1 illustre la différence entre une distribution équilibrée et une distribution déséquilibrée dans un contexte médical [2].

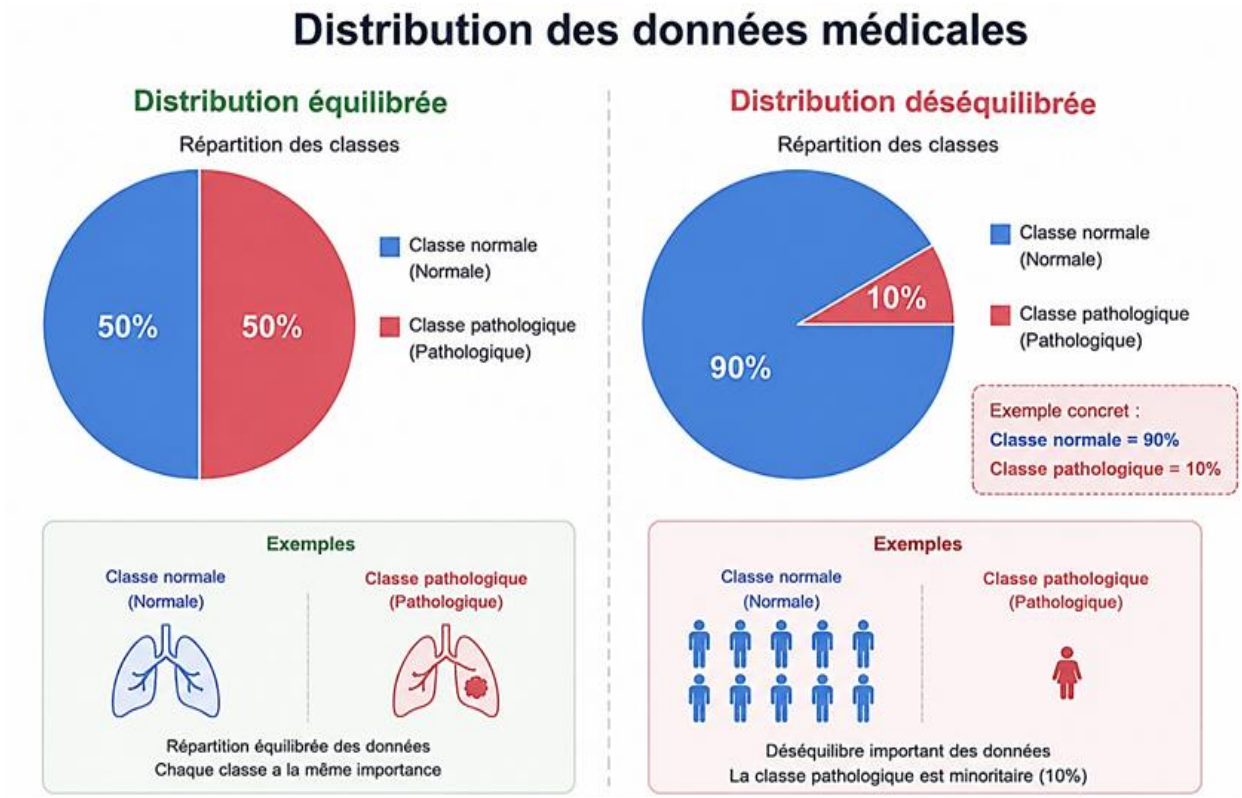


Figure 1.1 : Comparaison entre une répartition équilibrée et une structure déséquilibrée des classes au sein d'un jeu de données médicales.

Ce phénomène s'illustre de manière concrète à travers plusieurs applications cliniques. Dans un jeu de données dédié à l'évaluation des stades de la maladie d'Alzheimer, les patients catégorisés comme « non-déments » peuvent représenter jusqu'à 50 % des observations, contre seulement 1 à 5% pour le stade « modérément dément », qui s'avère pourtant le plus critique sur le plan clinique.

De même, lors de la classification de tumeurs cérébrales, les coupes d'IRM exemptes d'anomalie s'avèrent nettement plus nombreuses que celles présentant une lésion tumorale. Cette asymétrie ne constitue pas un artefact lié au recueil des données : elle reflète fidèlement la prévalence réelle de ces pathologies au sein de la population.

Dans ce contexte, l'apprentissage du modèle est fortement biaisé. Un réseau de neurones entraîné de manière classique va ajuster ses poids pour maximiser la précision globale, qui dépend principalement des classes majoritaires. Comme la classe minoritaire contient très peu d'exemples, elle ne contribue que faiblement à la mise à jour des paramètres du modèle, ce qui limite l'apprentissage de ces cas spécifiques [3].

2.2. Rareté des événements cliniques

La sous-représentation des classes minoritaires dans les jeux de données médicaux n’est pas uniquement liée à des choix méthodologiques : elle est avant tout la conséquence directe de la rareté naturelle de certains événements cliniques. Les maladies graves, les stades avancés de pathologies chroniques et les anomalies rares sont, par définition, moins fréquents dans la population générale que les formes bénignes ou les sujets sains [3].

À cette rareté naturelle s'ajoutent des contraintes pratiques qui limitent encore davantage la constitution de bases de données équilibrées. L'annotation des images médicales requiert l'expertise de spécialistes (radiologues, neurologues, anatomopathologistes) dont le temps est limité et le coût élevé. Les données médicales sont également soumises à des réglementations éthiques strictes : le consentement des patients, la protection des données personnelles et les restrictions imposées par les comités d'éthique limitent l'accès à de grands volumes de données sensibles [4].

Dans le cas précis de la maladie d'Alzheimer, le stade le plus avancé, qui correspond à la démence modérée, est non seulement rare dans la population, mais aussi difficile à diagnostiquer avec certitude à partir d'images IRM seules. Les quelques dizaines de cas disponibles dans les bases de données publiques restent très limités par rapport aux milliers d’exemples correspondant aux stades précoces. Une situation similaire s'observe pour les petites tumeurs cérébrales, dont les images peuvent être difficiles à distinguer de simples variations anatomiques normales, et dont la présence n'est parfois confirmée qu'après une intervention chirurgicale ou une biopsie [4, 5].

Paradoxalement, ce sont ces cas rares qui sont les plus importants sur le plan clinique. Un modèle incapable de les identifier correctement a une utilité diagnostique limitée, quelle que soit sa performance sur les cas courants. C'est pourquoi la gestion de la rareté des classes minoritaires constitue un enjeu central dans la conception des systèmes d'aide au diagnostic [3].

2.3. Distribution non uniforme des données

Le déséquilibre ne se limite pas à une simple différence de volume entre les classes ; il peut également revêtir des formes plus complexes. En effet, une distribution non uniforme peut exister au sein même des données, ce qui signifie que certaines régions de l'espace de représentation sont bien couvertes par les échantillons d'entraînement, tandis que d'autres demeurent faiblement représentées [3].

Cette variation peut être observée à différents niveaux. Au niveau global, certains sous-groupes d'une même classe peuvent être sur-représentés : par exemple, une base de données de tumeurs cérébrales peut contenir davantage d'images de tumeurs de grande taille, facilement repérables, que de tumeurs de petite taille, pourtant plus difficiles à détecter. Au niveau local, des zones spécifiques de l'image peuvent contenir l'information pathologique sans que cette information soit équilibrée avec le contexte sain environnant [4].

Un aspect théorique important dans ce contexte est le déséquilibre au niveau des pixels. Dans de nombreuses images d'IRM cérébrales ou d'échographies, la région pathologique (qu'il s'agisse d'une tumeur, d'une zone d'atrophie ou d'une lésion) occupe une fraction très réduite de la surface totale de l'image, entourée d'une grande proportion de tissu sain ou de fond non informatif. Cette asymétrie spatiale constitue un défi majeur pour les algorithmes d'apprentissage, car l'information minoritaire est noyée dans une majorité de pixels non pertinents. Ce phénomène devient particulièrement critique dans les tâches de segmentation, où le modèle doit délimiter précisément les frontières de la lésion au sein d'une image presque entièrement composée de tissu normal. La Figure 1.2 illustre des exemples de déséquilibre spatial dans des images IRM cérébrales, où les régions pathologiques occupent une portion réduite de l'image par rapport au tissu sain environnant [6].

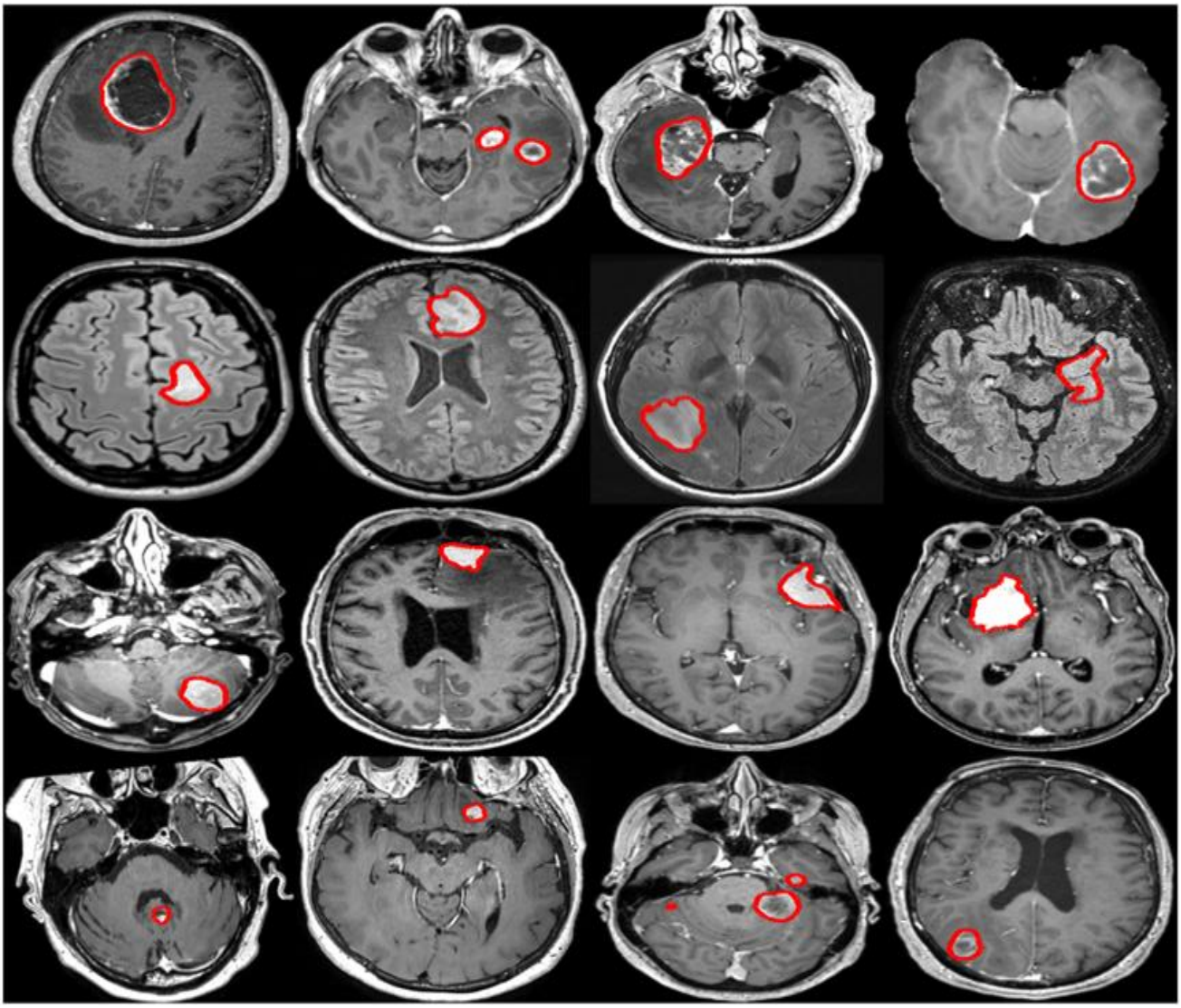


Figure 1.2: Exemples de régions tumorales faiblement représentées dans des images IRM cérébrales [4].

Ce déséquilibre spatial contribue également à compliquer l'interprétation des modèles : un réseau peut optimiser ses performances en se focalisant sur le tissu sain, qui est omniprésent, plutôt qu'en apprenant les caractéristiques fines des régions pathologiques. C'est l'un des défis que tentent de relever les méthodes d'explicabilité, dont l'objectif est précisément de vérifier que le modèle se concentre sur les régions pertinentes de l'image [7].

Les jeux de données médicaux peuvent ainsi présenter plusieurs formes de déséquilibre, chacune ayant des conséquences différentes sur les modèles d'apprentissage profond. Le tableau suivant résume les principaux types de déséquilibre rencontrés dans les applications médicales, ainsi que leurs impacts et les approches de correction associées.

Type de déséquilibre	Définition scientifique	Exemple médical réel
Déséquilibre proportionnel *(Proportion-based imbalance)*	Forte disparité du nombre d'échantillons entre les classes, où la classe minoritaire est sous-représentée par rapport à la classe majoritaire [8].	Diagnostic d'une maladie rare face à une large population saine (ex. détection du mélanome).
Déséquilibre extrême *(Extreme imbalance)*	Situation critique où la classe minoritaire représente moins de 1 % de la totalité du jeu de données (déséquilibre sévère) [9].	Enregistrements ECG de longue durée où les anomalies critiques, telles que les salves d'extrasystoles ventriculaires, sont extrêmement rares face aux battements normaux.
Distribution à longue traîne *(Long-tailed distribution)*	Distribution multi-classes où une minorité de catégories domine le jeu de données, tandis qu'une multitude d'autres classes ne possèdent que très peu d'exemples [10].	Classification automatique sur des radiographies thoraciques, où les cas d'épanchement pleural ou de cardiomégalie sont très fréquents, alors que la fibrose ou les nodules restent très rares.
Déséquilibre intrinsèque *(Intrinsic imbalance)*	Asymétrie naturelle directement liée à la rareté réelle de la pathologie au sein de la population, indépendamment du protocole de collecte [2].	Base de données sur les maladies génétiques rares ou certains sous-types spécifiques de cancers.
Déséquilibre extrinsèque *(Extrinsic imbalance)*	Asymétrie artificielle introduite par des facteurs externes ou des biais lors du processus de collecte, de numérisation ou de sélection des données, indépendamment de la prévalence réelle de la maladie [9].	Données multicentriques provenant de plusieurs hôpitaux ayant des critères d'admission ou des équipements différents.
Déséquilibre conditionnel *(Conditional imbalance)*	Déséquilibre de la variable cible qui dépend étroitement des variables explicatives, la probabilité d'occurrence étant quasi nulle pour la majorité des profils [11].	Pathologie spécifique se manifestant quasi exclusivement chez un groupe d'âge ou un genre particulier (ex. cancer de la prostate).
Déséquilibre local *(Local imbalance)*	Disparité numérique qui se manifeste spécifiquement au sein de sous-groupes ou de clusters de données l'échelle, sans être forcément visible à globale [12].	Analyse de sous-types tumoraux spécifiques ou de variations anatomiques locales dans un jeu de données d'imagerie.

Tableau 1.1 : Les principaux types de déséquilibre de données rencontrés en contexte médical.

3. Déséquilibre dans le diagnostic médical

Les difficultés générées par le déséquilibre de classes deviennent particulièrement critiques lorsqu'elles sont replacées dans le contexte du diagnostic médical. Les enjeux ne sont plus seulement statistiques ou méthodologiques : ils touchent directement à la sécurité du patient et à la fiabilité des outils d'aide à la décision clinique [4].

3.1. Cas du diagnostic précoce

La détection précoce des maladies est l'un des domaines où l'IA médicale présente le plus fort potentiel. Plusieurs études montrent, en oncologie comme en neurologie, qu'une identification de la pathologie à un stade initial améliore considérablement les options thérapeutiques ainsi que le pronostic à long terme. C'est précisément dans ce contexte que le problème du déséquilibre de classes se révèle le plus préoccupant [5].

Les stades précoces d'une maladie correspondent presque toujours aux classes minoritaires dans les bases de données d'imagerie médicale. Dans un jeu de données sur la maladie d'Alzheimer, les patients au stade de déficit cognitif léger ou de démence initiale sont nettement moins représentés que les sujets sains. De même, dans une base de données oncologique, les petites lésions au stade initial sont moins fréquentes que les tumeurs à un stade avancé, dont les signes radiologiques sont plus marqués et plus aisés à annoter. Un modèle qui échoue à reconnaître ces classes minoritaires produit des faux négatifs : il manque certains cas pathologiques en les classifiant à tort comme normaux [1, 13].

Le faux négatif est la forme d'erreur la plus redoutée en médecine diagnostique. Contrairement à une fausse alerte, appelée faux positif, qui entraîne généralement de simples examens complémentaires, un diagnostic manqué peut conduire à un retard de prise en charge potentiellement irréversible. Dans ce contexte, l'objectif d'un modèle de classification médicale n'est pas simplement d'atteindre une performance globale élevée : il doit être particulièrement sensible aux classes pathologiques minoritaires, même au prix d'une légère augmentation du taux de fausses alertes [5].

Ce constat établit un lien direct entre la gestion du déséquilibre de classes et la pertinence clinique d'un système d'IA médicale. Un pipeline qui ne traite pas ce problème explicitement risque d'être statistiquement performant tout en restant médicalement peu utile, voire dangereux [3].

3.2. Imagerie médicale et données cliniques

L'imagerie médicale regroupe un ensemble de modalités variées, chacune produisant des données aux caractéristiques propres. L’IRM permet d'obtenir des images à haute résolution des tissus mous, particulièrement utilisées pour l’étude du cerveau. La radiographie et le scanner (TDM) fournissent

des informations sur la densité des tissus et des structures osseuses, tandis que l'échographie offre une imagerie en temps réel, adaptée à l'étude des organes abdominaux ou de la glande mammaire. Cette hétérogénéité des modalités entraîne une grande variabilité dans les bases de données disponibles [4].

À la diversité de ces modalités s'ajoute la variabilité intrinsèque des données cliniques. Les images acquises sur des scanners de marques différentes, selon des protocoles d'acquisition distincts ou dans des centres hospitaliers aux pratiques variées, présentent des variations visuelles importantes, même pour une pathologie identique. Des variations de contraste, de résolution, d'orientation de coupe ou la présence d'artefacts peuvent fortement influencer la représentation d'une lésion. Un modèle d'apprentissage profond entraîné sur une base de données trop homogène peut ainsi échouer à généraliser sur des images acquises dans un contexte différent. Figure 1.3 présente des exemples de variabilité inter-appareils dans les images IRM cérébrales, mettant en évidence les différences visuelles liées aux scanners et aux paramètres d'acquisition [3, 4].

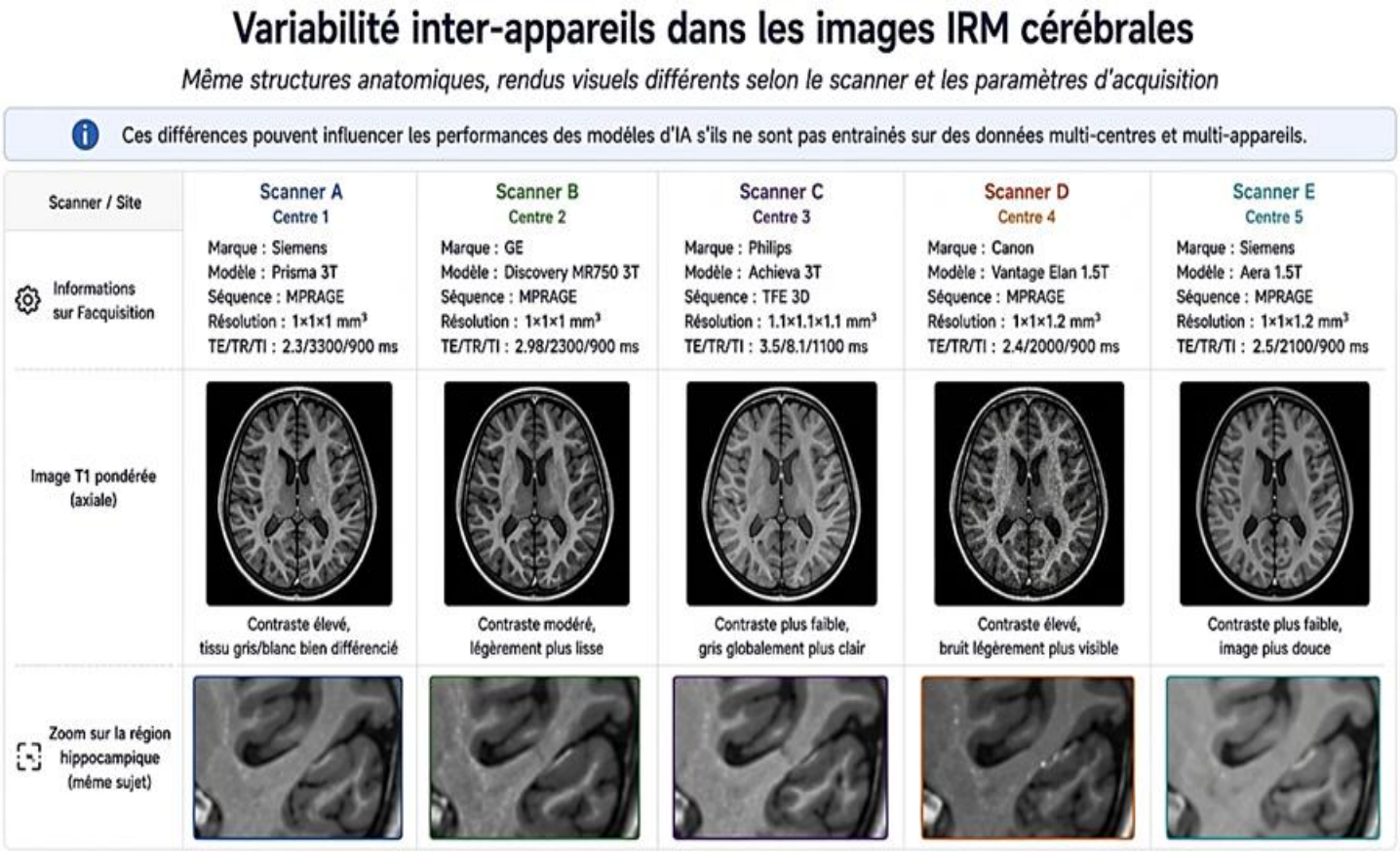


Figure 1.3 : llustration de la variabilité inter-scanners dans les acquisitions IRM cérébrales.

Par ailleurs, l'annotation des images médicales est un processus subjectif, soumis à une forte variabilité inter-observateur. Deux radiologues examinant la même IRM peuvent diverger sur l'existence ou les contours précis d'une lésion discrète. Cette subjectivité introduit un bruit de

labellisation qui pénalise de manière disproportionnée la classe minoritaire, dont les exemples sont déjà rares. L'ensemble de ces facteurs rend l'imagerie médicale particulièrement complexe pour l'apprentissage automatique, ce qui explique l'insuffisance des approches génériques [5].

3.3. Conséquences des erreurs sur les classes minoritaires

Pour bien comprendre les enjeux du déséquilibre, il est important d'analyser concrètement ce que signifie, sur le plan clinique, une mauvaise classification d'un exemple appartenant à la classe minoritaire. La disparité entre les conséquences des différents types d'erreurs constitue l'une des spécificités les plus critiques du diagnostic médical [5].

En médecine, ne pas détecter une pathologie est presque toujours plus grave que de générer une fausse alerte. Un patient atteint d'un stade précoce d'Alzheimer, s'il est classifié comme sain par un système d'IA, ne bénéficiera pas d'un suivi médical adapté qui aurait pourtant pu ralentir l'évolution de la maladie. De même, une petite tumeur cérébrale identifiée à tort comme un tissu sain manquera la fenêtre thérapeutique optimale pour une intervention chirurgicale, voire une radiothérapie ciblée. Ces scénarios illustrent pourquoi le rappel, défini comme la capacité du modèle à identifier correctement tous les cas positifs réels, constitue l'une des métriques les plus cruciales en imagerie médicale [5, 14].

La sensibilité clinique d'un modèle, représentant sa propension à ne pas manquer les cas pathologiques, dépend directement de sa capacité à appréhender les caractéristiques des classes minoritaires. Un faible rappel sur ces catégories, conséquence directe d'un déséquilibre non corrigé, signifie que le système génère systématiquement des faux négatifs sur les cas cliniquement prioritaires. Une telle situation s'avère inacceptable dans un contexte réel d'aide à la décision médicale [1, 13].

4. Problèmes induits par le déséquilibre de classes

Le déséquilibre de classes ne se limite pas à une simple inégalité de distribution : il provoque des difficultés majeures qui altèrent la qualité de l'apprentissage à différents niveaux. Cette section examine les principaux verrous pratiques associés à ce phénomène, en partant des enjeux liés à la représentation des données jusqu'aux effets sur la géométrie de l'espace de décision.

4.1. Faible représentation des classes minoritaires

Le premier problème est le plus direct : les classes minoritaires disposent de trop peu d'exemples pour que le modèle puisse en extraire des représentations robustes. Un réseau de neurones apprend en généralisant à partir de nombreux exemples variés. Avec seulement quelques dizaines d'images représentant un stade avancé d'Alzheimer ou une tumeur de petite taille, le modèle ne dispose pas d'une diversité suffisante pour capturer les différentes formes que peut prendre cette pathologie. La

Figure 1.4 illustre la faible représentation des classes minoritaires par rapport aux classes majoritaires dans un contexte d'apprentissage déséquilibré [1, 13].

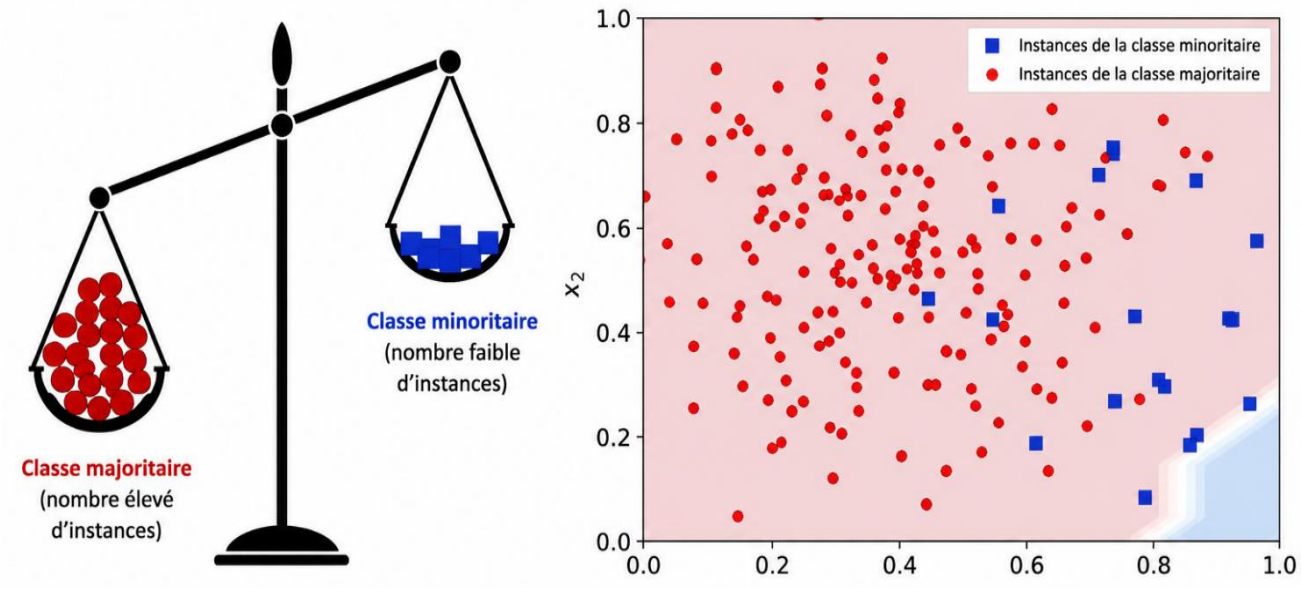


Figure 1.4: Représentation schématique de la faible proportion des classes minoritaires dans une base de données déséquilibrée [13].

Cette situation conduit à des représentations instables des classes minoritaires. Les caractéristiques apprises par le réseau pour identifier ces classes sont fragiles. En effet, elles reposent sur un nombre trop restreint de cas. Elles risquent alors de correspondre à des spécificités du jeu d'entraînement plutôt qu'à des propriétés réellement généralisables. Par conséquent, le modèle aura tendance à mal classer les exemples de classes minoritaires qu'il n'a jamais rencontrés auparavant. Ce problème survient même si ces nouveaux cas sont similaires aux exemples utilisés pendant l'entraînement [3].

4.2. Chevauchement entre classes

Un second problème majeur est celui du chevauchement entre classes. Dans de nombreuses applications médicales, les différentes catégories à distinguer présentent des caractéristiques visuelles similaires, sans frontière nette entre elles. Par exemple, un tissu légèrement pathologique peut ressembler à un tissu sain sur une IRM, surtout aux stades précoces de la maladie où les modifications morphologiques sont difficiles à observer. La Figure 1.5 illustre ce phénomène de chevauchement au sein d'un espace de représentation [15, 16].

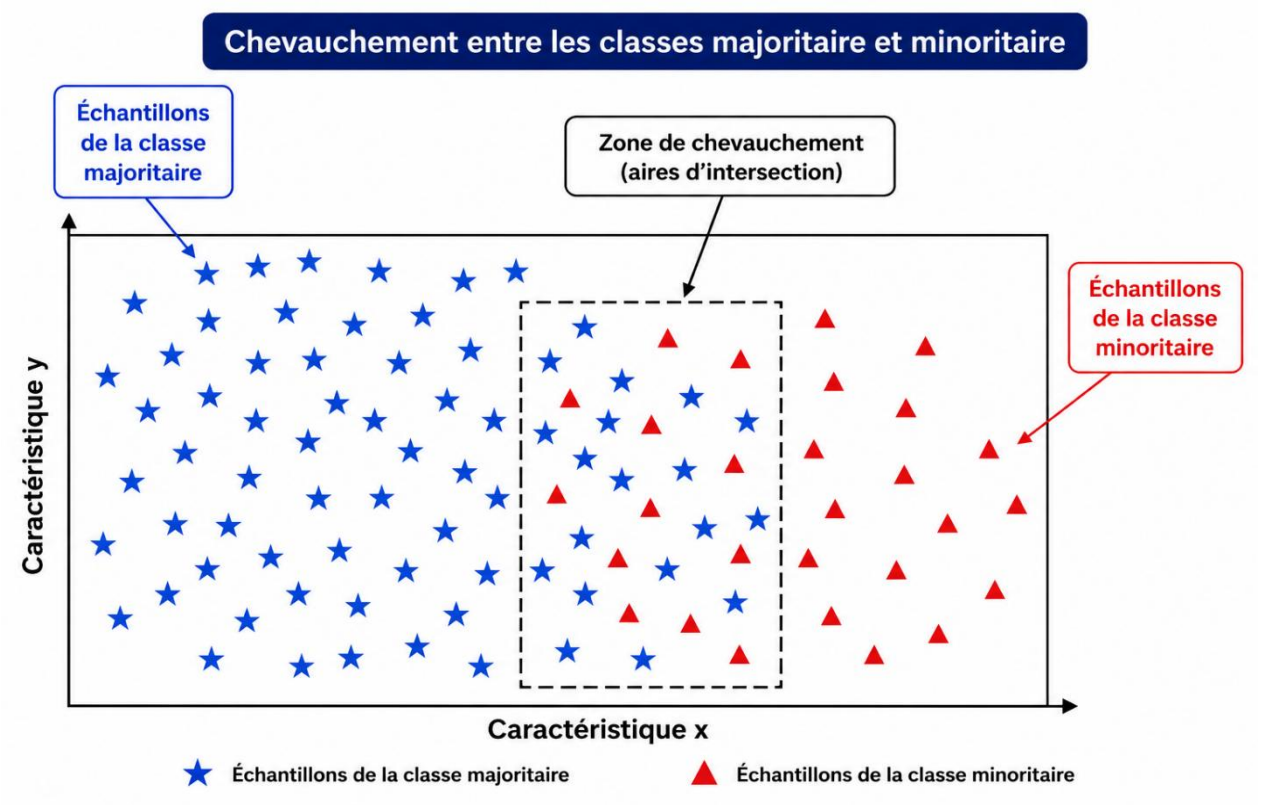


Figure 1.5: Représentation schématique du phénomène de chevauchement entre classes au sein de l'espace de caractéristiques.

Ce chevauchement est encore aggravé par le déséquilibre de classes. Lorsque la classe minoritaire est peu représentée, les algorithmes d'apprentissage ont tendance à déplacer la frontière de décision en faveur de la classe majoritaire. Il en résulte que le modèle attribue à cette classe majoritaire des exemples appartenant en réalité à la classe minoritaire, en raison de l'influence limitée de cette dernière sur les paramètres du réseau lors de l'entraînement [1, 3].

4.3. Données bruitées et ambiguës

La qualité des images médicales est loin d'être uniforme. Les artefacts d'acquisition, le faible contraste de certaines lésions, les variations liées aux équipements ou aux conditions d'examen, ainsi que l'incertitude dans l'annotation introduisent un niveau important de bruit dans les données. Ce bruit s'avère particulièrement préjudiciable pour les classes minoritaires [4].

En effet, pour une classe majoritaire disposant de milliers d'exemples, la présence de quelques dizaines d'images bruitées ou mal annotées a un impact limité sur la convergence du modèle. En revanche, pour une classe minoritaire ne comptant que quelques dizaines d'exemples au total, une poignée d'images incorrectes représente une fraction significative des données disponibles. Cela introduit une confusion importante dans les représentations apprises par le réseau. Ce problème est d'ailleurs amplifié par la variabilité inter-observateurs, une situation fréquente lors de l'annotation de lésions ambiguës [17].

4.4. Complexité des frontières de décision

La combinaison de la faible représentation, du chevauchement et du bruit complexifie fortement la frontière de décision que le modèle doit apprendre. Dans les espaces de caractéristiques des données médicales, les exemples des classes minoritaires se trouvent souvent isolés dans des régions denses ou mal délimitées, entourés d'exemples appartenant à la classe majoritaire. Afin d'illustrer cette difficulté, la Figure 1.6 met en contraste un cas idéal de frontière linéaire avec deux situations de non-linéarité, ces dernières étant nettement plus représentatives des distributions complexes observées dans les bases de données médicales déséquilibrées [15].

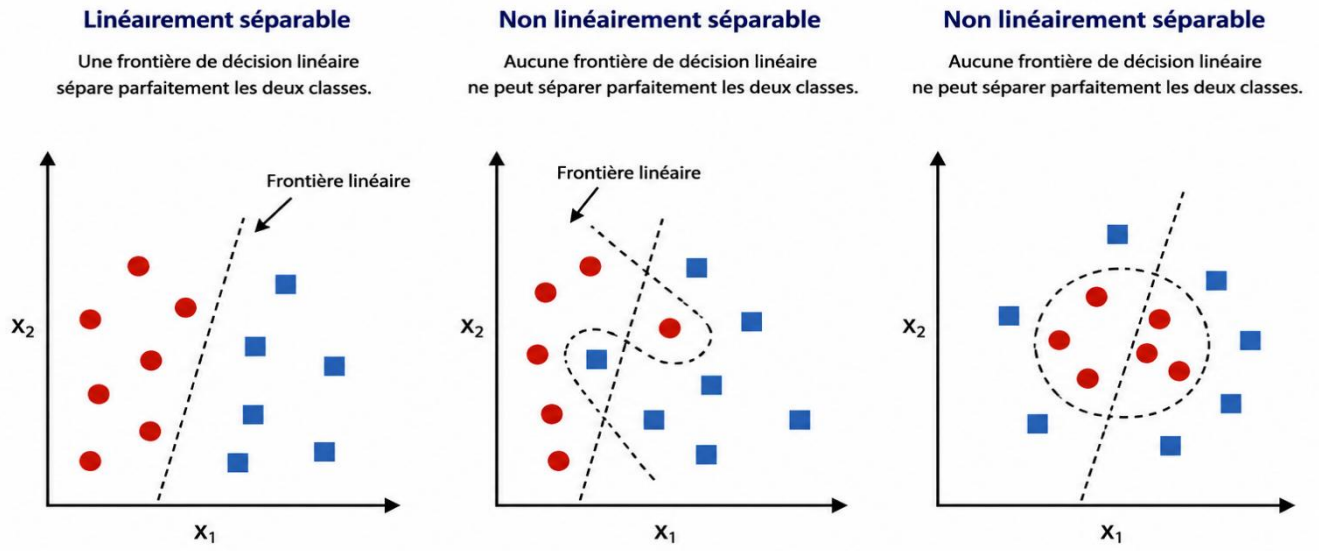


Figure 1.6: Frontières de décision linéaires et non linéaires selon la séparabilité des classes dans un espace de caractéristiques bidimensionnel.

Dans ce cas, les modèles d'apprentissage profond produisent des frontières de décision instables dans les zones peu documentées de l'espace des caractéristiques. La moindre variation du jeu d'entraînement fait alors osciller ces frontières, nuisant directement à la robustesse des prédictions. Ce phénomène engendre une forte incertitude : le modèle attribue de faibles scores de confiance aux exemples proches de la frontière. Cette situation est particulièrement critique en contexte clinique, où la fiabilité du diagnostic reste cruciale [1].

5. Impact du déséquilibre sur les modèles d'apprentissage profond

Les réseaux de neurones profonds ne sont pas naturellement conçus pour gérer le déséquilibre de classes. Leurs mécanismes d'optimisation standard visent à minimiser une fonction de perte globale, indépendamment de la contribution de chaque classe. Cette section analyse les effets directs de ce déséquilibre sur le comportement des modèles, depuis l'apparition des biais décisionnels jusqu'à la dégradation des métriques d'évaluation [10].

Ces perturbations influencent l'apprentissage profond de manière variable selon la distribution des données, la rareté des classes et la complexité des représentations à extraire. Certaines configurations introduisent un biais important en faveur des classes majoritaires, tandis que d'autres compromettent la capacité de généralisation ou réduisent fortement les performances sur les classes rares. Le tableau 1.2 synthétise les principaux impacts des différents types de déséquilibre sur les modèles d'apprentissage profond appliqués à l'imagerie médicale [10].

Type de déséquilibre	Impact sur les modèles d'apprentissage profond
Déséquilibre proportionnel	Le réseau développe un biais systématique envers la classe majoritaire, ce qui entraîne un rappel très faible pour la classe minoritaire.
Déséquilibre extrême	L'extraction de caractéristiques pour la classe rare devient quasi impossible ; le modèle converge souvent vers une prédiction uniforme de la classe majoritaire.
Distribution à longue traîne	Le modèle capture efficacement les concepts des classes fréquentes, mais néglige les classes rares situées dans la "queue" de la distribution.
Déséquilibre intrinsèque	Les performances restent structurellement limitées, même avec des architectures avancées, en raison d'un manque absolu de cas réels.
Déséquilibre extrinsèque	Le modèle assimile des biais liés aux conditions d'acquisition ou à la source des données plutôt que des caractéristiques médicales pertinentes.
Déséquilibre conditionnel	Le réseau éprouve des difficultés à capturer les corrélations et les structures conditionnelles rares au sein de la distribution.
Déséquilibre local	Les performances globales restent élevées, mais le modèle commet des erreurs critiques dans certaines régions spécifiques de l'espace des caractéristiques.

Tableau 1.2: Impacts des différents types de déséquilibre sur les modèles d'apprentissage profond.

L'analyse globale synthétisée dans ce tableau révèle que le déséquilibre n'affecte pas seulement la performance finale, mais perturbe en réalité l'ensemble du processus d'apprentissage. Afin de comprendre les mécanismes fondamentaux de ces défaillances, les sous-sections suivantes examinent de manière détaillée comment cette asymétrie des données induit d'abord un biais algorithmique lors du calcul des gradients, comment elle dégrade la reconnaissance des pathologies rares, et pourquoi elle rend les outils d'évaluation traditionnels obsolètes.

5.1. Biais des modèles vers les classes majoritaires

Quand une classe domine numériquement, cela affecte d'abord la rétropropagation du gradient, ce qui modifie la trajectoire d'optimisation du réseau.

Le mécanisme d'apprentissage d'un réseau de neurones repose sur la minimisation d'une fonction de perte qui mesure l'écart entre les prédictions et les étiquettes réelles. Lorsque les classes sont déséquilibrées, cette perte est dominée par les exemples de la classe majoritaire, en raison de leur prépondérance numérique. En conséquence, les gradients qui guident la mise à jour des paramètres du réseau reflètent principalement les caractéristiques des classes fréquentes [1].

Progressivement, le modèle apprend à reconnaître les classes majoritaires avec une grande précision, tout en développant une tendance systématique à classer les exemples ambigus dans ces mêmes classes dominantes. N'exerçant qu'une influence marginale sur le calcul des gradients, les classes minoritaires voient leurs caractéristiques apprises de manière insuffisante. Ce phénomène de biais décisionnel constitue l'une des conséquences les plus documentées du déséquilibre au sein de l'apprentissage profond [1, 13].

5.2. Dégradation des performances sur les classes rares

Ce biais mathématique en faveur de la classe majoritaire se traduit concrètement par une baisse drastique de l'efficacité du modèle face aux pathologies peu fréquentes. Cette dégradation des performances se manifeste principalement par un faible rappel, le modèle ne capturant qu'une fraction réduite des cas positifs réels. En contexte médical, cela engendre un taux élevé de faux négatifs, c'est-à-dire des patients malades non détectés [1].

Ce phénomène s'avère d'autant plus insidieux qu'il peut passer totalement inaperçu si l'on se contente d'évaluer le modèle à l'aide de métriques globales. Par exemple, un modèle obtenant une exactitude globale (accuracy) de 94 % peut simultanément omettre la moitié des cas de la classe minoritaire la plus critique. Cette performance agrégée masque ainsi une défaillance spécifique et critique sur les cas les plus importants sur le plan clinique [14].

5.3. Limites de l'exactitude comme métrique unique

La possibilité qu'une défaillance majeure soit masquée par une performance globale flatteuse met en lumière l'obsolescence des outils d'évaluation classiques et impose une redéfinition des critères de validation.

Bien que l'exactitude globale soit la métrique la plus intuitive, elle s'avère particulièrement trompeuse dans un contexte de données déséquilibrées. Pour obtenir une évaluation fidèle et

rigoureuse des performances d'un modèle en présence d'une telle asymétrie, il devient indispensable de recourir à des indicateurs plus informatifs [16].

Le rappel, ou sensibilité, mesure la proportion de vrais positifs correctement identifiés, reflétant ainsi la capacité du modèle à détecter les cas pathologiques. À l'inverse, la précision évalue la proportion de prédictions positives effectivement correctes, mesurant de ce fait la fiabilité des alertes générées par le réseau [16].

Le F1-score combine ces deux mesures en une unique métrique équilibrée (moyenne harmonique), offrant ainsi une évaluation bien plus représentative que l'exactitude. En complément, les courbes ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) et PR-AUC (Precision-Recall - Area Under the Curve) offrent une vision globale des compromis possibles entre sensibilité et spécificité à différents seuils de décision. Dans un contexte médical où le coût clinique d'un faux négatif dépasse largement celui d'un faux positif, le rappel doit systématiquement être considéré comme la métrique prioritaire [18].

Le paradoxe de l'exactitude en contexte déséquilibré

Pour illustrer cette limite, supposons un ensemble de données où 95 % des exemples correspondent à des patients sains et seulement 5 % à des patients atteints d'une pathologie. Un modèle naïf qui prédirait systématiquement la classe "sain" pour l'intégralité des entrées obtiendrait une exactitude globale de 95 %. Ce score, remarquable en apparence, masque le fait que le modèle ne détecte aucun patient malade. Sa valeur clinique est par conséquent nulle, et son déploiement s'avérerait extrêmement dangereux [14].

La Figure 1.7 met en évidence ce comportement critique lors de la variation du seuil de décision. On observe que l'exactitude globale se maintient à un niveau faussement optimal, tandis que le rappel et le F1- score chutent drastiquement, révélant l'incapacité du modèle à capturer efficacement la classe minoritaire.

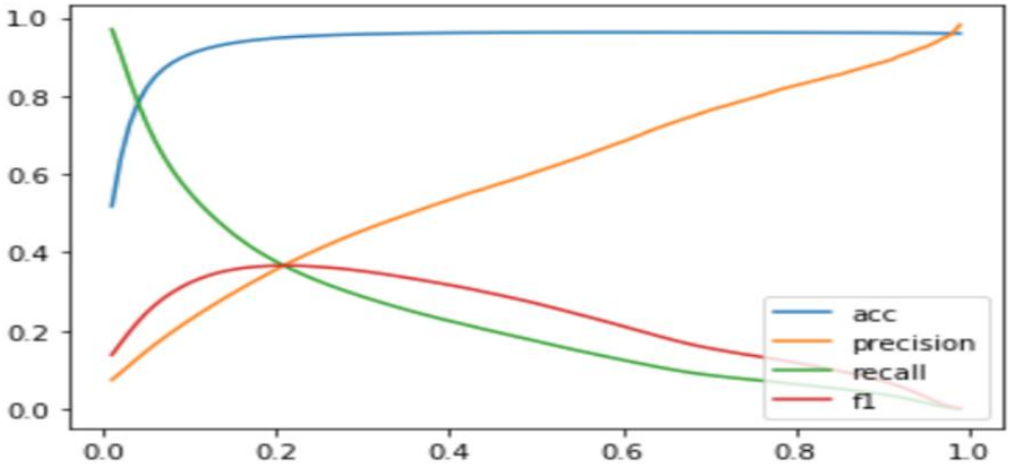


Figure 1.7: Évolution des métriques d'évaluation en fonction du seuil de décision dans un contexte de fort déséquilibre de classes[17].

6. Limites des approches existantes

Face au problème du déséquilibre de classes, diverses stratégies ont été proposées dans la littérature. Ces méthodes, qui seront détaillées dans le chapitre suivant, présentent néanmoins des limites structurelles qu'il est indispensable d'analyser objectivement. La compréhension de ces insuffisances s'avère essentielle pour concevoir des protocoles méthodologiques rigoureux et scientifiquement transparents.

6.1. Dépendance au type de données

L'une des premières limitations des méthodes de gestion du déséquilibre est leur dépendance stricte au contexte d'application. Une technique efficace sur un type particulier d'imagerie peut se révéler inefficace, voire contre-productive, sur une autre modalité. Les caractéristiques visuelles des images échographiques diffèrent fondamentalement de celles d'IRM cérébrales : le niveau de bruit, la résolution, le contraste et la texture des tissus ne sont pas comparables [4].

Cette dépendance s'étend également à la pathologie considérée. Une stratégie conçue pour traiter le déséquilibre dans un ensemble de données dédié à la classification des stades de la maladie d'Alzheimer peut ne pas convenir à la détection de tumeurs, car la nature des variations inter-classes est différente. Il n'existe pas de solution universelle au problème du déséquilibre en imagerie médicale, ce qui souligne l'importance d'une évaluation méthodique des approches pour chaque base de données [3].

6.2. Difficulté de généralisation

Un modèle entraîné sur un jeu de données spécifique peut échouer à produire des prédictions fiables lorsqu'il est appliqué à des données provenant d'un centre hospitalier différent, d'un scanner d'une marque distincte ou d'une population aux caractéristiques démographiques différentes. Ce

phénomène, connu sous le nom de décalage de domaine, est particulièrement prononcé en imagerie médicale en raison de la grande variabilité des protocoles d'acquisition et des équipements [4].

Les méthodes de traitement du déséquilibre n'échappent pas à cette règle. Un suréchantillonnage optimisé pour les caractéristiques visuelles d'un jeu de données peut introduire des biais non généralisables à d'autres contextes. Dès lors, un défi majeur consiste à concevoir des chaînes de traitement universelles et réutilisables, performantes sur différentes modalités et bases de données. C'est précisément l'un des objectifs visés dans les chapitres suivants de ce mémoire [3].

6.3. Risque de surapprentissage

Les techniques de rééquilibrage des classes, en particulier celles basées sur le suréchantillonnage, comportent un risque bien documenté de surapprentissage. Lorsqu'une classe minoritaire est rééquilibrée par la duplication successive des mêmes exemples, le réseau risque de mémoriser ces images spécifiques plutôt que d'en extraire des caractéristiques généralisables. Les performances mesurées sur le jeu de validation peuvent alors largement surestimer les capacités du modèle en environnement clinique [19].

Les techniques classiques d’augmentation de données, telles que les rotations, les symétries ou les variations de contraste, n’améliorent pas systématiquement les performances en contexte médical. C'est particulièrement vrai pour les images déjà prétraitées et normalisées. Dans cette configuration, une augmentation excessive risque d'introduire des variations artificielles s’éloignant des distributions médicales réelles [3].

Trouver le juste équilibre entre le rééquilibrage des classes et la préservation de la distribution naturelle des données constitue donc un défi méthodologique non trivial. Sans cette précaution, le modèle apprend excessivement les exemples présents dans les données d'entraînement sans parvenir à généraliser correctement sur de nouvelles données. La Figure 1.8 illustre ce phénomène de surapprentissage induit par un rééquilibrage artificiel des classes [3].

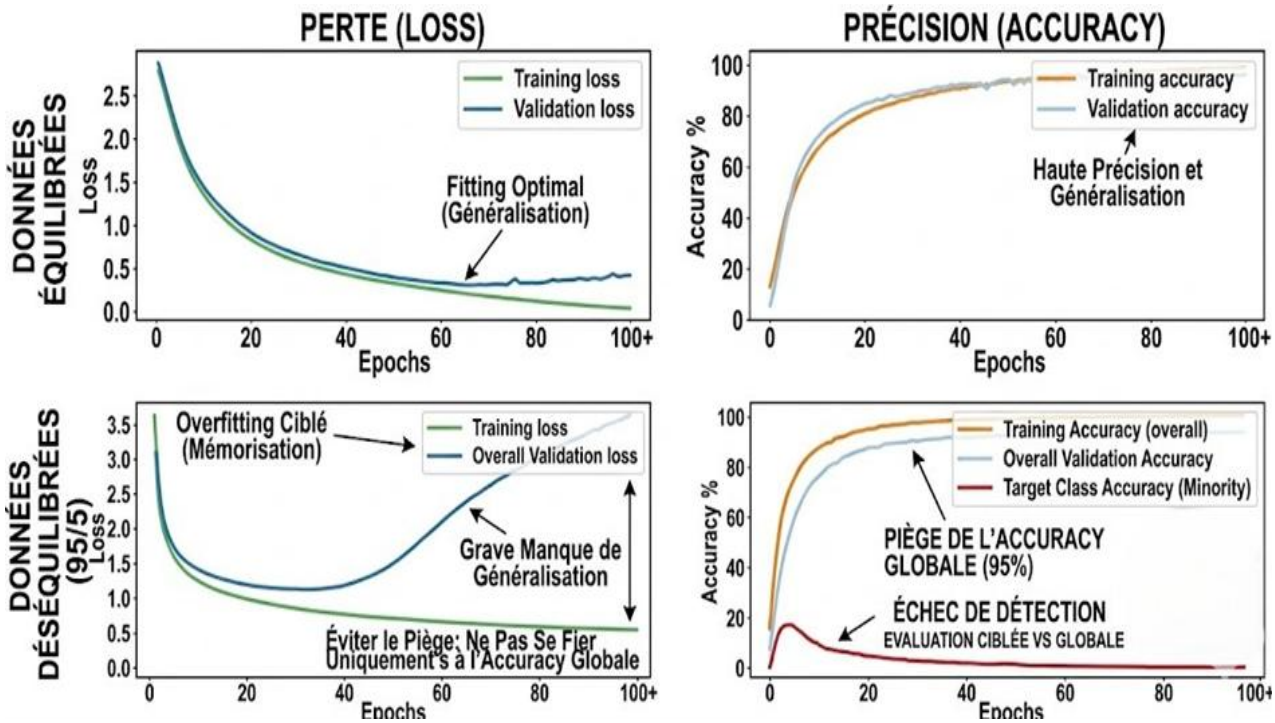


Figure 1.8:Illustration du phénomène de surapprentissage dans les modèles entraînés sur des données déséquilibrées.

7. Conclusion

Ce premier chapitre a permis de démontrer que le déséquilibre de classes en imagerie médicale constitue un phénomène endémique. Directement lié aux réalités épidémiologiques et aux contraintes logistiques de collecte, ce problème structurel ne peut être résolu par une simple accumulation de données et impose le recours à des solutions algorithmiques spécialisées.

L'analyse des mécanismes d'optimisation a également prouvé que les modèles standards développent un biais systématique en faveur des classes majoritaires. Sans correction, ils maximisent une exactitude globale trompeuse qui masque l'omission de cas pathologiques critiques, rendant indispensable l'usage de métriques ciblées comme le rappel et le F1-score. Enfin, l'examen des approches de rééquilibrage actuelles a révélé des limites majeures, notamment une forte dépendance au contexte, des faiblesses de généralisation et un risque élevé de surapprentissage.

Ces constats justifient pleinement le développement d'architectures plus robustes et ouvrent directement la voie au chapitre suivant. Ce dernier sera ainsi dédié aux techniques de rééquilibrage de données, suivies des fondements de l'apprentissage profond, des réseaux de neurones convolutifs ainsi que de l'apprentissage par transfert, pour se clore sur l'explicabilité par cartes d'activation de classe (Grad-CAM).



Chapitre 2 :

Stratégies de gestion du déséquilibre et Apprentissage profond



1. Introduction

L'analyse automatique des images médicales représente aujourd'hui l'un des domaines d'application les plus prometteurs de l'intelligence artificielle. Les réseaux de neurones profonds ont démontré une capacité remarquable à extraire des informations diagnostiques à partir d'images complexes, telles que les IRM cérébrales, les radiographies thoraciques ou les échographies, parfois avec une précision comparable à celle de spécialistes humains. Ces avancées ouvrent de nouvelles perspectives pour l'aide au diagnostic précoce de maladies comme la maladie d'Alzheimer ou les tumeurs cérébrales, pour lesquelles une détection rapide influence directement le pronostic du patient.

Cependant, l'application pratique de ces méthodes en milieu médical se heurte à plusieurs défis structurels. Le premier défi concerne le déséquilibre des classes : les ensembles de données cliniques reflètent la prévalence naturelle des pathologies, ce qui rend les cas rares systématiquement sous-représentés. Sans correction explicite, les modèles d'apprentissage profond développent un biais en faveur des classes majoritaires, compromettant leur utilité pour les cas cliniquement les plus importants. Le second défi est celui de la confiance : un médecin ne peut raisonnablement intégrer dans sa pratique clinique un outil dont il ne comprend pas les mécanismes de décision, aussi précis soit-il.

L’objectif de ce chapitre est de montrer comment articuler ces deux dimensions critiques, à savoir la rareté des cas pathologiques et la transparence des modèles, afin de concevoir des outils d’aide au diagnostic à la fois robustes et acceptables en pratique clinique. Pour ce faire, le chapitre s’organise en cinq parties progressives. La première présente les stratégies spécialisées de gestion du déséquilibre de classes, depuis les méthodes agissant directement sur les données jusqu’aux approches algorithmiques et hybrides. La deuxième rappelle les fondements théoriques de l’apprentissage profond ainsi que les principales architectures de réseaux de neurones convolutifs utilisées en imagerie médicale, éléments indispensables à la compréhension des sections suivantes. La troisième expose les métriques d’évaluation spécifiquement adaptées aux ensembles de données asymétriques. La quatrième est dédiée aux principaux outils d’explicabilité de l’intelligence artificielle. Enfin, la cinquième analyse de manière critique les travaux existants de la littérature afin de positionner la démarche méthodologique retenue dans ce mémoire.

2. Stratégies de gestion du déséquilibre

Face au problème du déséquilibre de classes, la littérature propose trois grandes familles de solutions : les méthodes basées sur les données, les méthodes algorithmiques et les stratégies hybrides. Chacune intervient à un niveau spécifique du pipeline d’apprentissage afin de limiter l’effet du déséquilibre sur les performances du modèle. Dans le domaine médical, le choix de la stratégie doit prendre en compte non seulement les performances du modèle, mais également

la plausibilité clinique des données générées ainsi que les caractéristiques de l'ensemble de données disponible.

Cette section présente successivement les trois grandes familles de solutions issues de la littérature. Nous commençons par les méthodes basées sur les données, puis aborderons les méthodes algorithmiques et enfin les stratégies hybrides.

2.1. Méthodes basées sur les données

Les méthodes de niveau données (data-level methods) agissent directement sur la composition de l'ensemble d'entraînement en modifiant la distribution des exemples par classe avant le début de l'apprentissage. Leur objectif est de rapprocher les effectifs des différentes classes afin de réduire le biais induit par le déséquilibre.

2.1.1. Techniques de suréchantillonnage

Les techniques de suréchantillonnage (Oversampling) ont pour objectif d'équilibrer la distribution des classes en augmentant artificiellement la présence de la classe minoritaire au sein du jeu de données d'entraînement. En pratique, cette augmentation peut s'effectuer soit par la duplication directe des données existantes, soit par la génération de nouvelles instances synthétiques. Dans le cadre de l'imagerie médicale, ces approches permettent d'éviter que le modèle ne néglige les pathologies rares. Les sections suivantes présentent l'évolution de ces méthodes, en partant des approches de duplication élémentaires jusqu'aux algorithmes d'interpolation plus complexes et adaptatifs.

2.1.1.1. Suréchantillonnage Aléatoire

Le suréchantillonnage aléatoire (Random Oversampling) consiste à dupliquer des exemples de la classe minoritaire jusqu'à atteindre un effectif cible. Simple à mettre en œuvre, cette technique préserve l'intégrité des données originales, un aspect important en imagerie médicale, comme l'illustre la Figure 2.1. Son principal inconvénient est le risque de surapprentissage (overfitting), car la répétition des mêmes exemples peut conduire le modèle à mémoriser les données plutôt qu'à en apprendre des caractéristiques généralisables, notamment lorsque la classe minoritaire contient peu d'exemples distincts [1, 20].

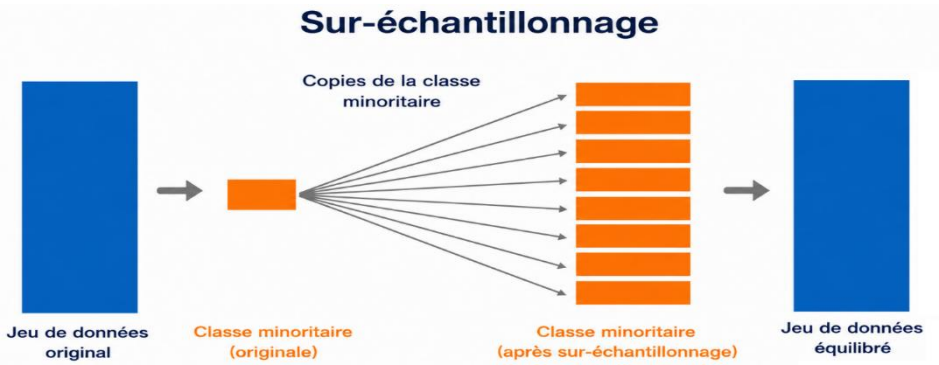


Figure 2.1 : Principe du suréchantillonnage aléatoire traditionnel [20].

2.1.1.2. Suréchantillonnage Ciblé

Le suréchantillonnage ciblé (Target Oversampling, TOS) constitue une extension du suréchantillonnage aléatoire. Au lieu d'appliquer un même ratio à toutes les classes minoritaires, cette approche définit un effectif cible spécifique pour chaque classe en fonction du degré de déséquilibre observé. Les classes les plus rares peuvent ainsi être suréchantillonnées de manière contrôlée, limitant le risque de surcompensation et de création d'un biais inverse. Cette flexibilité en fait une solution adaptée aux ensembles de données médicaux présentant plusieurs niveaux de déséquilibre [1].

2.1.1.3. SMOTE

Proposé par Chawla et al. [16], SMOTE (Synthetic Minority Over-sampling Technique) génère de nouveaux exemples synthétiques par interpolation linéaire entre un exemple minoritaire et l'un de ses k plus proches voisins dans l'espace des caractéristiques. Pour chaque exemple x_i, un nouvel exemple $x_{nouveau}$ est créé selon :

$$x_{nouveau} = x_i + \lambda(\tilde{x} - x_i) \qquad (2.1)$$

Où $\tilde{x}$ est un voisin sélectionné aléatoirement et λ un coefficient aléatoire compris entre 0 et 1. Cette approche enrichit la représentation de la classe minoritaire sans recourir à la simple duplication. Toutefois, dans le contexte de l'imagerie médicale, l'interpolation peut produire des structures anatomiques peu réalistes, limitant son application directe sur les images brutes [14].

2.1.1.4. ADASYN

ADASYN (Adaptive Synthetic Sampling) étend SMOTE en générant davantage d'exemples synthétiques dans les régions où la classification est la plus difficile. Le nombre d'échantillons créés dépend de la proportion de voisins appartenant à la classe majoritaire autour de chaque exemple minoritaire. Cette stratégie favorise un meilleur apprentissage des frontières de décision tout en évitant de suréchantillonner les régions déjà bien représentées [21].

2.1.2. Techniques de sous-échantillonnage

À l'opposé du suréchantillonnage, le sous-échantillonnage (Undersampling) consiste à réduire l'effectif des classes majoritaires. Le sous-échantillonnage aléatoire (Random Undersampling), illustré par la Figure 2.2, supprime aléatoirement des exemples de la classe dominante jusqu'à obtenir un ratio plus équilibré entre les classes [22].

Cette approche présente toutefois un inconvénient majeur dans le domaine médical : elle peut entraîner une perte d'information significative dans un contexte où les données annotées sont rares et coûteuses à acquérir. Des méthodes plus sélectives, telles que les liens de Tomek (Tomek Links) ou la méthode des plus proches voisins édités (Edited Nearest Neighbours, ENN), cherchent à limiter cette

perte en supprimant principalement les exemples situés près de la frontière de décision. Néanmoins, leur application directe aux images médicales reste délicate en raison de la forte dimensionnalité des données [23, 24].

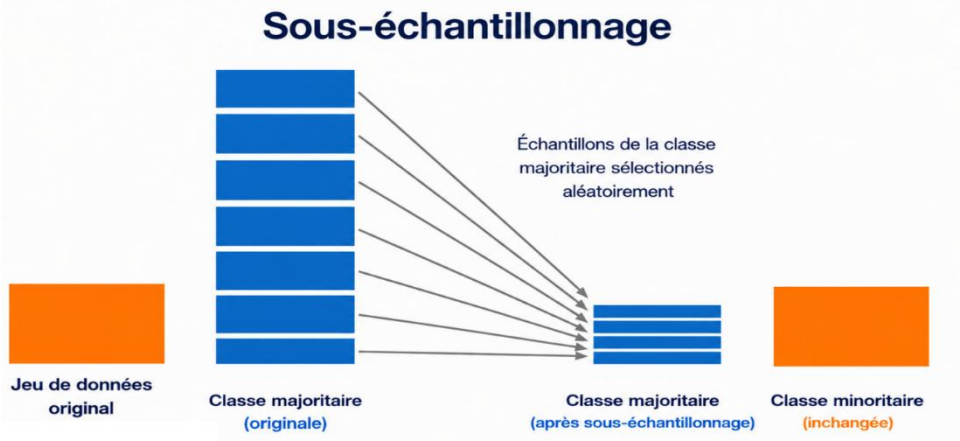


Figure 2.2: Principe du sous-échantillonnage aléatoire traditionnel [24].

2.1.3. Augmentation de données

L'augmentation de données (Data Augmentation, DA) enrichit artificiellement l'ensemble d'entraînement par l'application de transformations aux images existantes. Dans le contexte du déséquilibre de classes, ces transformations sont généralement appliquées de manière préférentielle aux classes minoritaires afin d'accroître leur diversité sans recourir à une simple duplication.

Les transformations les plus couramment utilisées en imagerie médicale sont les suivantes [25]:

- **Rotation** : permet d'améliorer la robustesse du modèle face aux variations d'orientation des acquisitions.
- **Zoom et recadrage aléatoire** : simulent les variations de champ de vue et d'échelle.
- **Symétrie horizontale et verticale (***Flipping***)** : utilisée lorsque ces transformations restent anatomiquement plausibles.
- **Bruit gaussien** : modélise le bruit d'acquisition inhérent aux modalités d'imagerie et suit généralement une loi normale.
- **Variation de contraste et de luminosité** : simule les différences entre les scanners ou les protocoles d'acquisition.
- **Déformations élastiques (***Elastic Deformations***)** : transformations non rigides reproduisant les variations morphologiques naturelles entre les patients, particulièrement adaptées à l'imagerie anatomique.

Il convient toutefois de souligner que l'augmentation de données ne garantit pas toujours une amélioration des performances. Lorsque l'ensemble de données a déjà été fortement prétraité et

normalisé, certaines transformations peuvent introduire des artefacts artificiels qui s'éloignent de la distribution réelle des images cliniques et dégradent la capacité de généralisation du modèle [25].

2.1.4. GANs

Les réseaux génératifs adversariaux (Generative Adversarial Networks, GANs), proposés par Goodfellow et al, permettent de générer des images synthétiques à l'aide d'un générateur et d'un discriminateur entraînés de manière compétitive. En imagerie médicale, ils sont de plus en plus utilisés pour augmenter la diversité des classes minoritaires [26].

Le principe fondamental de cette approche repose sur un jeu à somme nulle entre les deux sous-réseaux, comme l'illustre la Figure 2.3. Le générateur tente de créer des images fausses mais réalistes à partir d'un bruit aléatoire, tandis que le discriminateur apprend à distinguer les images réelles de l'ensemble de données des échantillons synthétiques.

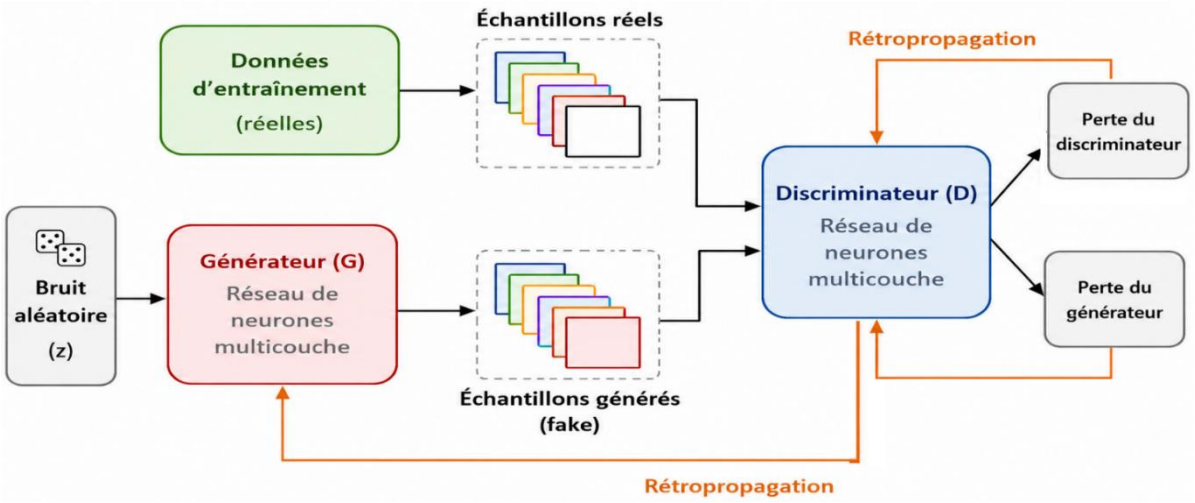


Figure 2.3: Architecture générale d'un GAN : compétition entre le générateur et le discriminateur pour produire des images synthétiques réalistes [26].

Cependant, malgré leur potentiel, l'entraînement des GANs reste particulièrement instable et la validation clinique des structures anatomiques générées constitue encore une limite importante pour leur déploiement pratique [26, 27].

2.1.5. Limites de l'augmntation excessive des données

Après avoir présenté les techniques d'augmentation classiques et génératives, il convient d'en examiner les limites, en particulier lorsqu'elles sont appliquées de manière excessive.

Malgré leurs avantages, les techniques d'augmentation présentent des risques importants lorsqu'elles sont poussées à l'extrême. Une manipulation trop répétitive ou disproportionnée peut favoriser le surapprentissage, en exposant le modèle à des exemples trop similaires. Concernant les méthodes génératives, notamment les GANs, elles peuvent produire des images visuellement très réalistes mais totalement dépourvues de plausibilité clinique, par exemple en créant des structures

anatomiques inexistantes. Enfin, si les données originales contiennent déjà des artefacts ou des erreurs d'annotation, l'augmentation risque d'amplifier ces imperfections et d'introduire un biais supplémentaire dans le modèle. Ces risques sont particulièrement critiques dans le domaine médical, où la stricte validité clinique des données est essentielle [1, 27].

2.2. Méthodes algorithmiques

Les méthodes algorithmiques interviennent directement au niveau du processus d'apprentissage, sans modifier l'ensemble de données. Elles adaptent la manière dont le modèle traite les exemples de chaque classe pendant l'entraînement en ajustant le processus d'optimisation afin de réduire les effets du déséquilibre.

2.2.1. Apprentissage sensible au coût

Deux techniques principales relèvent de l'apprentissage sensible au coût (Cost-Sensitive Learning): l'attribution de poids aux classes et l'entropie croisée pondérée. Nous les détaillons ci-dessous.

2.2.1.1. Poids des classes (Class Weights)

L'apprentissage sensible au coût associe à chaque classe un poids inversement proportionnel à sa fréquence dans l'ensemble de données. Les classes rares reçoivent ainsi un poids plus élevé, amplifiant leur contribution aux gradients d'optimisation. Le poids w_k d'une classe *k* est calculé selon :

$$W_k = \frac{N_{total}}{K \times N_k} \qquad (2.2)$$

Où N_{total} est le nombre total d'exemples, K le nombre de classes, et N_k l'effectif de la classe k. Cette formule garantit que la contribution pondérée de chaque classe à la mise à jour des paramètres est équilibrée [28].

2.2.1.2. Entropie croisée pondérée (Weighted Cross Entropy)

La fonction de perte standard en classification multiclasse est l'entropie croisée (*Cross Entropy*) définie par :

$$CE_{(p,y)} = -\sum_k y_k \cdot \log(p_k) \qquad (2.3)$$

L'entropie croisée pondérée intègre les poids de classe pour amplifier l'impact des erreurs sur les classes minoritaires :

$$WCE_{(p,y)} = -\sum_k w_k \cdot y_k \cdot \log(p_k) \qquad (2.4)$$

En multipliant chaque terme par le poids w_k, les erreurs sur les classes rares contribuent davantage à la mise à jour des poids du réseau, corrigeant le biais naturel vers les classes fréquentes [29].

2.2.2. Focal Loss

La *Focal Loss*, proposée par Lin et al. [30], constitue une extension de l'entropie croisée conçue pour traiter le déséquilibre de classes en se concentrant sur les exemples difficiles à classifier. Elle introduit un terme modulateur qui réduit la contribution des exemples bien classifiés et augmente celle des exemples incertains :

$$FL_{(p_t)} = -\alpha_t \cdot (1 - p_t)^{\gamma} \cdot \log(p_t) \qquad (2.5)$$

Où p_t est la probabilité associée à la vraie classe, γ est le paramètre de focalisation (généralement (γ = 2)) et α_t est un facteur de pondération par classe. Cette approche est particulièrement adaptée aux applications médicales, où les classes minoritaires correspondent souvent à des cas pathologiques subtils et difficiles à détecter [14, 30].

2.2.3. Classification hiérarchique

La classification hiérarchique constitue une stratégie efficace pour traiter les problèmes de déséquilibre sévère. Elle décompose une tâche complexe en plusieurs niveaux de décision successifs. Dans une première étape, un modèle distingue les cas normaux des cas pathologiques. Dans une seconde étape, un modèle spécialisé affine la classification des cas pathologiques en différentes sous-classes. Cette approche réduit la complexité de chaque sous-problème et améliore la sensibilité aux classes minoritaires, ce qui la rend particulièrement adaptée aux applications d'imagerie médicale [31].

2.2.4. Apprentissage en deux phases

L'apprentissage en deux phases (Two-Phase Learning, Two-PL) est une stratégie d'entraînement progressif. Dans une première phase, le modèle est entraîné sur l'ensemble des données afin d'apprendre une représentation générale et robuste des données. Dans une seconde phase, une étape d'affinage (fine-tuning) est réalisée sur un ensemble de données rééquilibré ou avec des poids de classes renforcés, afin d'adapter les frontières de décision aux classes minoritaires [32].

Cette approche combine les connaissances acquises lors de la première phase avec une adaptation ciblée aux classes rares, offrant une meilleure stabilité que l'entraînement direct sur des données déséquilibrées. Elle est conceptuellement proche de l'apprentissage par transfert (Transfer Learning), où une phase de pré-entraînement est suivie d'une phase d'affinage spécialisée [32].

2.3. Stratégies hybrides

Après avoir présenté séparément les méthodes basées sur les données et les méthodes algorithmiques, nous examinons maintenant les approches qui les combinent : les stratégies hybrides.

2.3.1. Combinaison des approches au niveau des données et des algorithmes

En pratique, aucune méthode isolée ne permet de résoudre parfaitement tous les problèmes de déséquilibre de classes. Les stratégies hybrides combinent donc des approches complémentaires agissant à différents niveaux du pipeline d'apprentissage afin de tirer parti de leurs avantages respectifs.

La combinaison la plus courante associe un suréchantillonnage contrôlé des classes minoritaires à une pondération des classes ou à une Focal Loss. Le rééquilibrage des données réduit le biais structurel de l'ensemble de données, tandis que l'adaptation de la fonction de perte maintient une sensibilité accrue aux classes rares durant l'optimisation. D'autres combinaisons, telles que l'utilisation conjointe de SMOTE avec la Focal Loss, ont également montré leur efficacité pour améliorer l'apprentissage des classes minoritaires [28].

2.3.2. Importance et limites des approches hybrides

Les stratégies hybrides offrent une grande flexibilité, car chaque composante peut être ajustée selon le degré de déséquilibre et les caractéristiques de l'ensemble de données. Elles sont particulièrement adaptées aux applications médicales, où les classes minoritaires correspondent souvent aux cas cliniquement les plus critiques.

Cependant, une combinaison excessive de techniques peut entraîner une surcompensation du déséquilibre et rendre le modèle trop sensible aux classes minoritaires au détriment des classes majoritaires. Pour cette raison, l'évaluation de ces approches doit s'appuyer sur des métriques adaptées, notamment le F1-score, le rappel par classe et la matrice de confusion, plutôt que sur la seule exactitude globale (accuracy) [28].

3. Fondements de l'apprentissage profond

Après avoir présenté les stratégies de gestion du déséquilibre des classes, il est nécessaire d'aborder les fondements de l'apprentissage profond sur lesquels elles reposent. L'intelligence artificielle (AI), l'apprentissage automatique (Machine Learning) et l'apprentissage profond (Deep Learning) constituent une hiérarchie imbriquée, où chaque niveau est une spécialisation du précédent. Cette organisation, illustrée par la Figure 2.4, permet de situer précisément les méthodes utilisées dans ce mémoire [33].

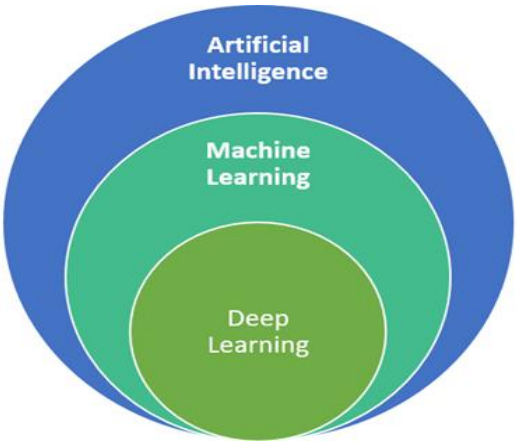


Figure 2.4: Relation hiérarchique de l'intelligence artificielle : IA → Machine Learning → Deep Learning [33].

Cette section introduit donc les concepts d'intelligence artificielle, d'apprentissage automatique et d'apprentissage profond, ainsi que leur hiérarchie.

3.1. Intelligence artificielle

L'intelligence artificielle désigne l'ensemble des méthodes et techniques informatiques visant à reproduire ou automatiser certaines capacités associées à l'intelligence humaine, telles que l'apprentissage, la reconnaissance de formes et la prise de décision.

En médecine, l'IA a progressivement trouvé des applications dans l'aide au diagnostic, la planification thérapeutique, l'analyse d'images médicales et la prédiction de l'évolution des pathologies. Son principal intérêt réside dans sa capacité à traiter rapidement de grandes quantités de données complexes et à fournir une assistance aux professionnels de santé [34].

Dans le domaine de l'imagerie médicale, les systèmes d'IA ne remplacent pas le clinicien : ils constituent des outils d'aide à la décision capables de détecter des anomalies, de mettre en évidence des zones d'intérêt et de proposer un diagnostic différentiel. La décision médicale finale demeure sous la responsabilité exclusive du professionnel de santé [4].

3.2. Apprentissage automatique

L'apprentissage automatique est un domaine de l'intelligence artificielle permettant aux systèmes d'apprendre à partir de données sans programmation explicite. Il comprend trois paradigmes principaux [33] :

- **L'apprentissage supervisé** consiste à entraîner un modèle à partir de données annotées, c'est-à-dire des exemples pour lesquels la réponse correcte est déjà connue. Le modèle apprend à associer des entrées à des sorties en minimisant ses erreurs.

- **L'apprentissage non supervisé** s'applique à des données sans étiquettes. Le modèle doit découvrir seul les structures, regroupements ou patterns cachés dans les données, sans indication préalable sur ce qu'il doit trouver.
- **L'apprentissage par renforcement** repose sur l'interaction avec un environnement : un agent effectue des actions, reçoit des récompenses ou des pénalités selon leurs effets, et ajuste progressivement sa stratégie pour maximiser ses gains à long terme.

Dans la classification d'images médicales, l'approche principalement utilisée est l'apprentissage supervisé, où des experts fournissent des annotations permettant d'associer des caractéristiques visuelles à des catégories cliniques. Les méthodes classiques telles que les SVM, les arbres de décision, le k-NN et les forêts aléatoires reposent sur une extraction manuelle des caractéristiques. Toutefois, leur efficacité est limitée face à la complexité et à la haute dimensionnalité des images brutes, ce qui a conduit au développement des méthodes d'apprentissage profond [33].

3.3. Apprentissage profond

L'apprentissage profond est une sous-branche de l'apprentissage automatique basée sur l'utilisation de réseaux de neurones artificiels composés de plusieurs couches de traitement. Le terme « profond » fait référence à la profondeur de ces architectures, où chaque couche apprend des caractéristiques de plus en plus abstraites et complexes à partir des données d'entrée [35].

Contrairement aux approches classiques qui nécessitent une étape manuelle d'extraction des caractéristiques (*feature extraction*) avant la classification, les modèles d'apprentissage profond réalisent ces deux étapes de manière simultanée et automatique. La Figure 2.5 met en évidence cette différence fondamentale : le réseau apprend par lui-même à extraire les représentations les plus discriminantes directement à partir des données brutes.

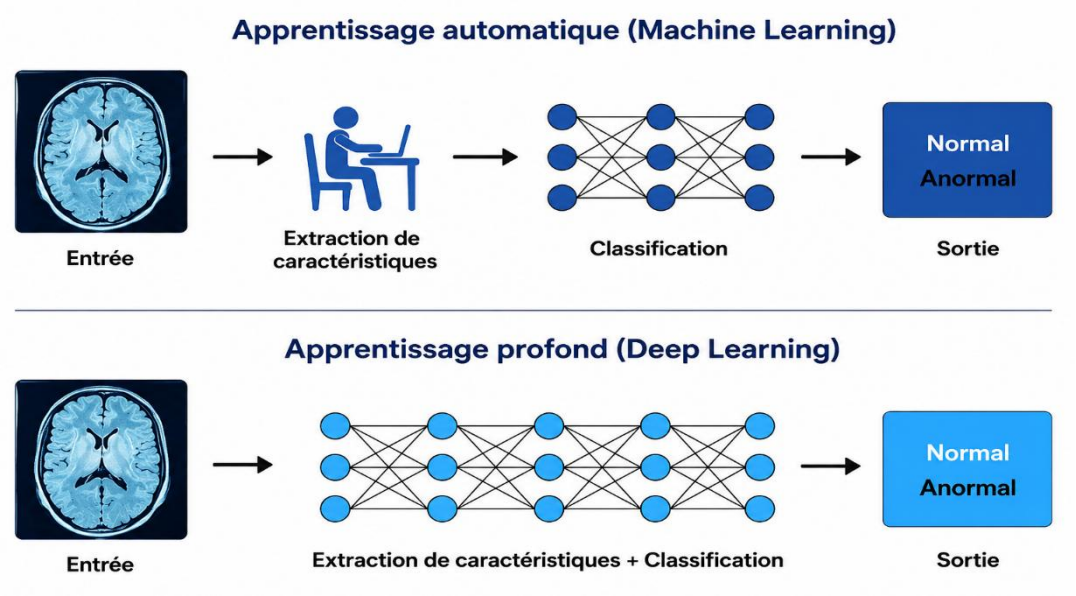


Figure 2.5: La différence entre l'apprentissage automatique traditionnel et l'apprentissage profond [35].

Cette capacité d'auto-apprentissage est particulièrement adaptée à l'analyse d'images médicales (telles que les IRM ou les échographies), dont la richesse structurelle, les détails anatomiques et les textures fines sont extrêmement complexes à décrire et à modéliser manuellement [35].

Toutefois, l'application de réseaux de neurones profonds standards sur des images brutes pose des défis majeurs, notamment en raison de l'explosion du nombre de paramètres à entraîner et de la perte des informations spatiales. Pour surmonter ces limites, des architectures spécialisées ont été développées afin de traiter efficacement les données bidimensionnelles et tridimensionnelles. Parmi elles, les réseaux de neurones convolutifs s'imposent comme la référence absolue en imagerie médicale, ce qui justifie leur présentation détaillée dans la section suivante.

3.4. Réseaux de neurones convolutifs

Les réseaux de neurones convolutifs (Convolutional Neural Networks, CNN) constituent l'architecture de référence pour l'analyse et la classification d'images. Inspirés du fonctionnement du cortex visuel biologique, ils sont capables d'apprendre automatiquement des représentations hiérarchiques à partir des données, sans nécessiter une étape manuelle d'extraction des caractéristiques.

3.4.1. Structure générale des CNNs

Un CNN est généralement composé de deux parties principales. La première, appelée extracteur de caractéristiques, utilise des couches de convolution et de mise en commun (pooling) pour détecter progressivement des motifs de complexité croissante tout en réduisant les dimensions spatiales de l'image. La seconde, appelée classifieur, exploite des couches entièrement connectées (Fully Connected) afin de transformer ces représentations en probabilités de classification. La Figure 2.6 illustre cette organisation globale.

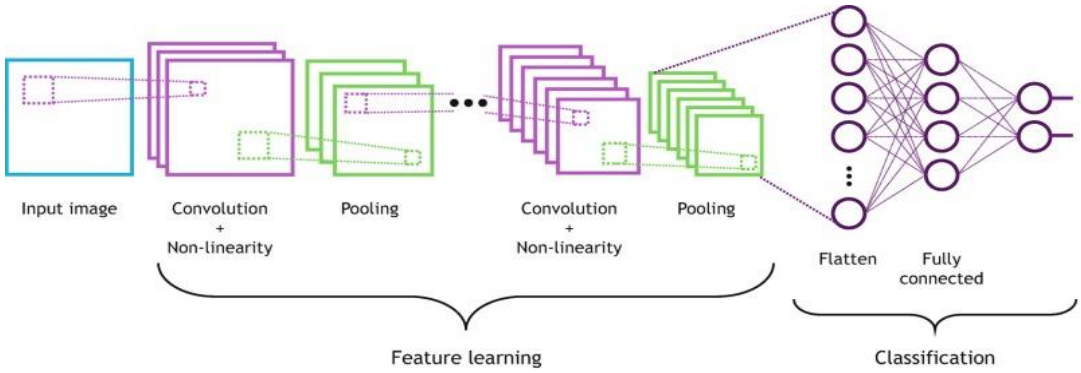


Figure 2.6: Illustration d'un exemple d'architecture CNN pour la classification d'images [36].

Dans le domaine de l'imagerie médicale, cette architecture permet d'extraire automatiquement des informations pertinentes à partir des images IRM, scanners ou échographies, améliorant ainsi les performances des systèmes d'aide au diagnostic [36].

3.4.2. Principales couches d'un CNN

La puissance et la flexibilité des CNN résultent de la combinaison de plusieurs types de couches, chacune jouant un rôle spécifique dans le traitement de l'information. Cette section présente chacune de ces couches en détail, en insistant sur leur rôle spécifique en imagerie médicale.

3.4.2.1. Couche de convolution

La couche de convolution est le composant fondamental d'un réseau CNN. Elle applique des filtres apprenables à l'image d'entrée afin de produire des cartes de caractéristiques. L'opération de convolution bidimensionnelle discrète est définie par :

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i-m, j-n) \cdot K(m,n) \qquad (2.6)$$

Où I désigne l'image d'entrée, K le filtre de convolution et $S(i,j)$ la valeur de la carte de caractéristiques à la position (i,j).

La Figure 2.7 illustre un exemple concret de cette opération.

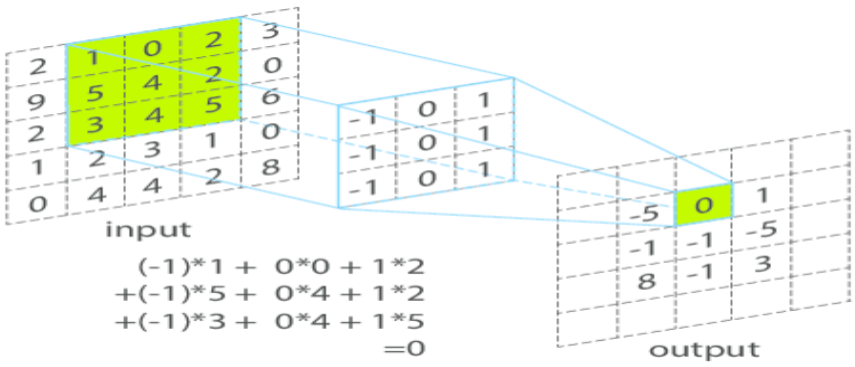


Figure 2.7: Exemple d'opération de convolution [36].

Les filtres des premières couches détectent des formes simples (contours, lignes, gradients de couleur). En revanche, les couches les plus profondes apprennent à détecter des structures

géométriques et anatomiques complexes, telles que les formes d'organes, les textures de tissus ou les motifs de lésions pathologiques [36, 37].

3.4.2.2. Fonctions d'activation

Après chaque opération de convolution, une fonction d'activation non linéaire est appliquée pour permettre au réseau de modéliser des relations complexes. La fonction ReLU est aujourd'hui la plus largement utilisée dans les couches cachées :

Après chaque opération de convolution, une fonction d'activation non linéaire est appliquée pour permettre au réseau de modéliser des relations complexes. Cette section présente les deux fonctions les plus utilisées en imagerie médicale : ReLU (Rectified Linear Unit) pour les couches cachées et *Softmax* pour la classification finale.

1) La fonction ReLU

La fonction ReLU est actuellement la plus largement utilisée dans les couches cachées. Elle est définie par :

$$ReLU(x) = max(0, x) \qquad (2.7)$$

Où x désigne la valeur d'entrée du neurone (c'est-à-dire le résultat de l'opération de convolution avant activation).

Comme l'illustre la Figure 2.8, cette fonction est simple et efficace : elle renvoie x si x est positif, et 0 sinon. Elle évite le problème de saturation des gradients (phénomène classique avec les fonctions sigmoïde ou tangente hyperbolique) et accélère considérablement la convergence de l'entraînement [36].

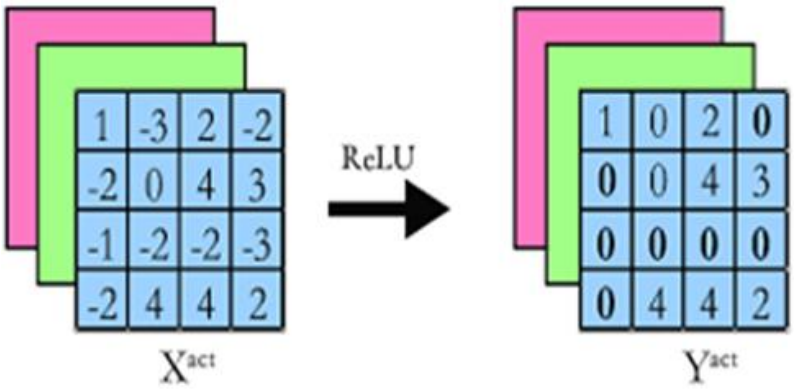


Figure 2.8 : Principe de fonctionnement de la fonction ReLU : annulation des valeurs négatives et maintien des valeurs positives [36].

2) La fonction Softmax

Pour les couches de sortie, notamment dans les tâches de classification multiclasse, on utilise généralement la fonction Softmax. Elle transforme les scores bruts fournis par le réseau en une distribution de probabilités :

$$Softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \qquad (2.8)$$

Dans cette équation, z_i représente le score associé à la classe i en sortie du réseau, K est le nombre total de classes, et l'exponentielle e^{z_i} garantit des valeurs positives. Le dénominateur $\sum_{j=1}^{k} e^{z_j}$ normalise ces valeurs pour que leur somme soit égale à 1. Ainsi, $Softmax(z_i)$ est la probabilité que l'image d'entrée appartienne à la classe i [36].

Cette fonction garantit que la somme des probabilités sur toutes les classes est égale à 1, facilitant ainsi l'interprétation des résultats.

3) **Autres fonctions d'activation**

Bien que ReLU et Softmax soient les standards en imagerie médicale, il existe d'autres fonctions d'activation, telles que la sigmoïde, la tangente hyperbolique (tanh) ou la Leaky ReLU. Chacune présente des propriétés spécifiques adaptées à des besoins d'apprentissage particuliers [36].

La Figure 2.9 illustre les courbes de ces différentes fonctions. On observe que ReLU et Leaky ReLU sont linéaires pour les valeurs positives, tandis que sigmoïde et tanh présentent une saturation progressive, ce qui explique leur moindre efficacité dans les couches profondes [36].

La même figure montre également le comportement de la fonction Softmax en fonction du nombre de classes. Le graphique correspondant indique comment la probabilité de sortie y_0 évolue lorsque l'entrée x_0 varie, les autres entrées étant fixées à 0. On constate que plus le nombre de classes augmente (de 2 à 10), plus la courbe s'aplatit, car la probabilité se répartit sur davantage de classes. Cette illustration complète la compréhension du rôle de Softmax dans les tâches de classification multiclasse.

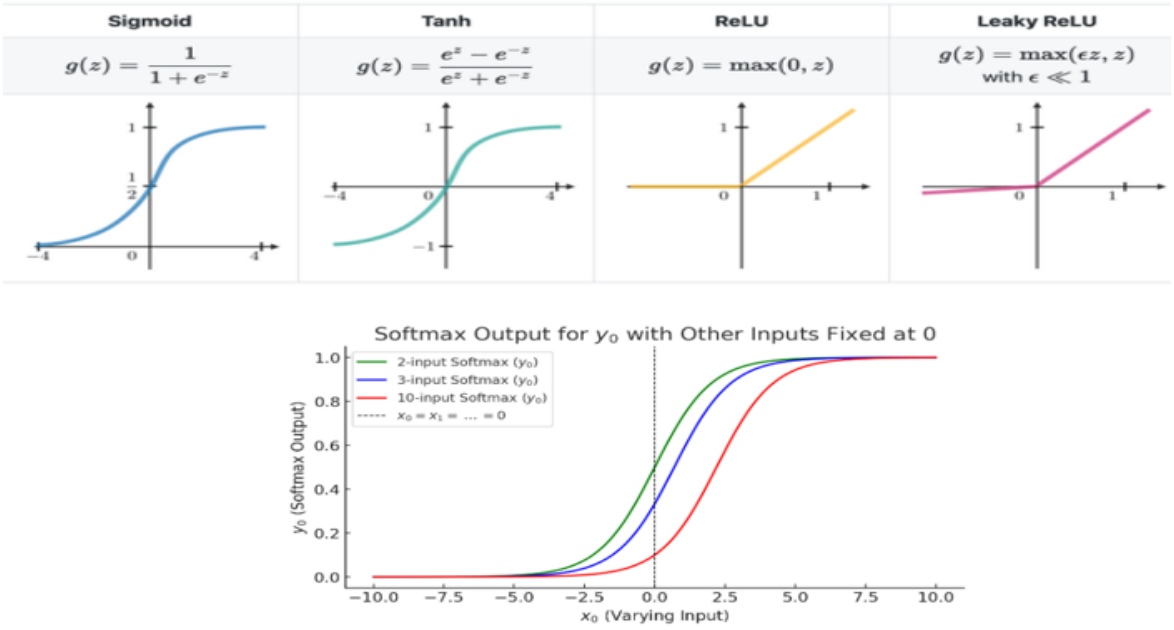


Figure 2.9: Les courbes des fonctions d'activation [36].

3.4.2.3. Couche de pooling

Les couches de pooling interviennent généralement après les fonctions d'activation pour réduire les dimensions spatiales (hauteur et largeur) des cartes de caractéristiques. Cette réduction permet de diminuer le nombre de paramètres à entraîner, de limiter le coût computationnel et de rendre le réseau plus tolérant aux légères translations ou déformations des structures dans l'image [36]. Cette propriété s'avère particulièrement utile en imagerie médicale, où une anomalie ou une lésion doit pouvoir être détectée indépendamment de sa position exacte sur la coupe anatomique.

Comme le montre la Figure 2.10, les deux techniques de sous-échantillonnage spatial les plus couramment utilisées sont les suivantes [36]:

- **Le Max Pooling** : il extrait uniquement la valeur maximale de chaque région de la carte, préservant ainsi les caractéristiques visuelles et les contrastes les plus marqués (bords, intensités fortes).
- **L'Average Pooling** : il calcule la valeur moyenne de la région, lissant ainsi l'information texturale globale.

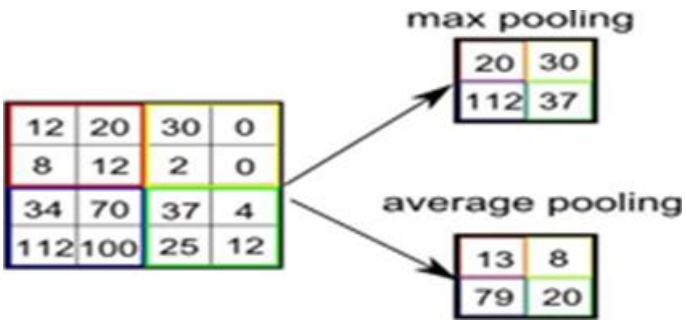


Figure 2.10: Les deux principaux types de pooling : Max Pooling et Average Pooling [36].

3.4.2.4. Couches entièrement connectées

Les couches entièrement connectées (Fully Connected, FC), également appelées couches denses, sont systématiquement positionnées à la fin de l'architecture du CNN pour assurer l'étape finale de classification. Une fois les opérations successives de convolution et de pooling terminées, les cartes de caractéristiques bidimensionnelles obtenues doivent être converties en un format compatible avec les réseaux de neurones classiques. Pour cela, on applique une opération d'aplatissement (flattening) qui réorganise l'ensemble des matrices de caractéristiques en un unique vecteur unidimensionnel. Ce vecteur est ensuite transmis aux couches entièrement connectées, où chaque neurone est connecté à l'intégralité des neurones de la couche précédente, comme l'illustre la Figure 2.11 [36].

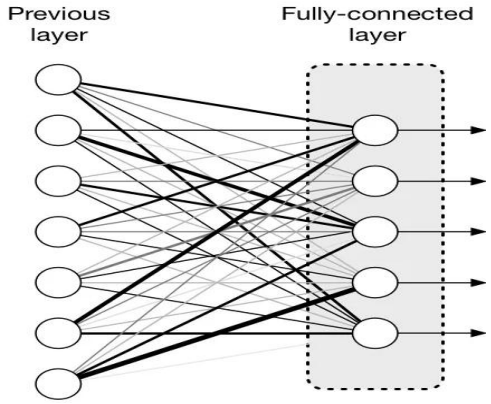


Figure 2.11 : Architecture d'une couche entièrement connectée [36].

Dans le cadre de l'imagerie médicale, ces couches jouent le rôle de classifieur global : elles combinent de manière non linéaire les caractéristiques de haut niveau extraites par la première partie du réseau (l'extracteur) pour formuler la décision finale. Pour les problèmes de classification multiclasse, la dernière couche entièrement connectée est associée à la fonction d'activation Softmax afin de générer le vecteur de probabilités diagnostiques [36].

3.4.2.5. Couche de Dropout

Le Dropout est une technique de régularisation incontournable conçue pour lutter contre le surapprentissage. Durant la phase d'entraînement, cette méthode consiste à désactiver « éteindre » de manière aléatoire et temporaire un certain pourcentage de neurones (défini par un taux p, généralement fixé entre 0.2 et 0.5) ainsi que leurs connexions associées à chaque itération [36].

La Figure 2.12 illustre parfaitement ce mécanisme en comparant l'architecture standard d'un réseau entièrement connecté (A) avec sa configuration après l'application du Dropout (B), où plusieurs neurones sont désactivés [36].

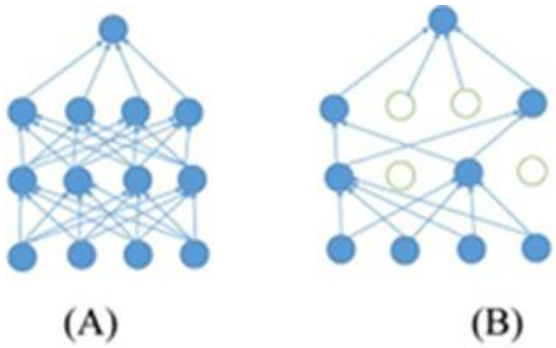


Figure 2.12 : Structure d'un réseau de neurones avant (A) et après (B) l'application de la régularisation par Dropout [36].

Cette stratégie repose sur un principe forcé de redondance : en modifiant continuellement la topologie du réseau, on empêche les neurones de développer des co-adaptations complexes et fragiles (où certains neurones compenseraient passivement les erreurs des autres). Le réseau est ainsi contraint d'apprendre des représentations internes beaucoup plus robustes, autonomes et distribuées, ce qui optimise sa capacité de généralisation [36].

En imagerie médicale, où les jeux de données souffrent fréquemment d'un nombre restreint d'échantillons ou de biais cliniques, l'intégration de couches de Dropout est particulièrement critique pour stabiliser la convergence et garantir la fiabilité du diagnostic sur de nouvelles données patients [36].

3.4.2.6. Couche de normalisation par lot

La normalisation par lot (Batch Normalization, BN) est une technique utilisée pour stabiliser et accélérer l'entraînement des réseaux de neurones profonds en normalisant les activations au sein de chaque mini-lot. Elle réduit le phénomène de dérive interne des covariables au cours de l'apprentissage, facilite la convergence et permet d'utiliser des taux d'apprentissage plus élevés. De plus, elle exerce un léger effet de régularisation et peut être combinée avec le Dropout pour améliorer la capacité de généralisation du modèle [38].

Après avoir présenté les différentes couches d'un CNN, nous abordons maintenant deux éléments essentiels à l'entraînement : les fonctions de perte et les algorithmes d'optimisation.

3.4.3. Fonctions de perte

La fonction de perte quantifie l'écart entre les prédictions du modèle et les vraies étiquettes de classification. Elle fournit le signal d'erreur qui guide directement la mise à jour des paramètres du réseau par rétropropagation du gradient [36].

Dans le contexte du déséquilibre des données, le choix de cette fonction est capital : une perte standard optimise la performance globale (l'exactitude) sans tenir compte de l'importance clinique des classes minoritaires. Les fonctions présentées précédemment dans la section 2.2 (entropie croisée,

entropie croisée pondérée et Focal Loss) constituent les options principales de ce travail. Leurs équations respectives, notées (2.3), (2.4) et (2.5), y ont été détaillées de manière rigoureuse.

3.4.4. Algorithmes d'optimisation

Les algorithmes d'optimisation ont pour objectif de minimiser la fonction de perte en mettant à jour de manière itérative les poids et les biais du modèle selon la descente de gradient. Parmi eux, la descente de gradient stochastique (Stochastic Gradient Descent, SGD) calcule les mises à jour à partir de mini-lots, ce qui accélère l'apprentissage et améliore l'efficacité computationnelle [36].

L'optimiseur Adam (Adaptive Moment Estimation) est actuellement l'un des algorithmes les plus largement utilisés en apprentissage profond. Il combine intelligemment le mécanisme de momentum (prise en compte des gradients passés) et l'adaptation automatique du taux d'apprentissage pour chaque paramètre. Cette double propriété offre une convergence particulièrement rapide et une excellente robustesse face aux gradients rares ou bruités, un avantage majeur dans les problèmes de classification médicale déséquilibrée [39].

Avant d'aborder les architectures avancées, il convient de préciser les réglages clés (hyperparamètres) et une stratégie d'apprentissage efficace (apprentissage par transfert).

3.5. Hyperparamètres et stratégie d'entraînement

Les performances d'un modèle d'apprentissage profond dépendent fortement de plusieurs hyperparamètres définis avant l'entraînement. Leur réglage est crucial pour assurer la stabilité du modèle [36].

- **Taux d'apprentissage (Learning Rate)** : c'est l'hyperparamètre le plus déterminant. Une valeur trop élevée peut provoquer une divergence (le modèle ne converge jamais), tandis qu'une valeur trop faible ralentit considérablement la convergence. Des stratégies adaptatives, telles que ReduceLROnPlateau, permettent d'ajuster automatiquement ce paramètre au cours de l'entraînement lorsque les performances stagnent.
- **Taille du batch (Batch Size)** : elle influence directement la stabilité du gradient et la capacité de généralisation du modèle.
- **Nombre d'époques (Epochs)** : il détermine la durée maximale de l'apprentissage.

Pour limiter le surapprentissage et améliorer la robustesse du réseau, ces paramètres sont souvent couplés à des mécanismes de contrôle comme l'arrêt précoce (Early Stopping), le Dropout ou la normalisation par lot, qui permettent d'interrompre l'entraînement dès que le modèle commence à perdre sa capacité de généralisation [36].

3.6. Apprentissage par transfert

Après avoir détaillé les hyperparamètres, nous abordons maintenant une stratégie puissante pour les applications médicales : l'apprentissage par transfert.

L'apprentissage par transfert consiste à réutiliser un modèle pré-entraîné sur une tâche source (souvent une base de données massive comme ImageNet) afin de l'adapter à une nouvelle tâche cible. Cette approche est particulièrement adaptée à l'imagerie médicale, où les ensembles de données annotés sont souvent rares et coûteux à constituer [36]. Le fonctionnement général de cette approche est schématisé sur la Figure 2.13, qui montre comment les poids pré-entraînés sur le jeu de données générique ImageNet sont réutilisés pour initialiser un modèle avant de l'adapter à une tâche d'imagerie médicale.

Le principe repose sur le transfert de connaissances : les premières couches d'un CNN entraîné sur des millions d'images apprennent des caractéristiques visuelles génériques (contours, textures, formes géométriques). Ces connaissances sont universelles et peuvent être réutilisées pour d'autres domaines.

L'adaptation à la tâche médicale se fait via une étape appelée fine-tuning (ajustement fin) :

- **Phase de gel (Freezing)** : on « gèle » les couches profondes du réseau (les poids sont verrouillés) pour conserver les caractéristiques génériques extraites.
- **Phase d'ajustement** : on débloque progressivement les dernières couches pour qu'elles apprennent les spécificités des images médicales (lésions, organes, pathologies).

Cette stratégie améliore les performances, accélère considérablement la convergence du réseau et réduit les besoins en données d'entraînement, faisant de l'apprentissage par transfert un outil indispensable pour les applications cliniques [36].

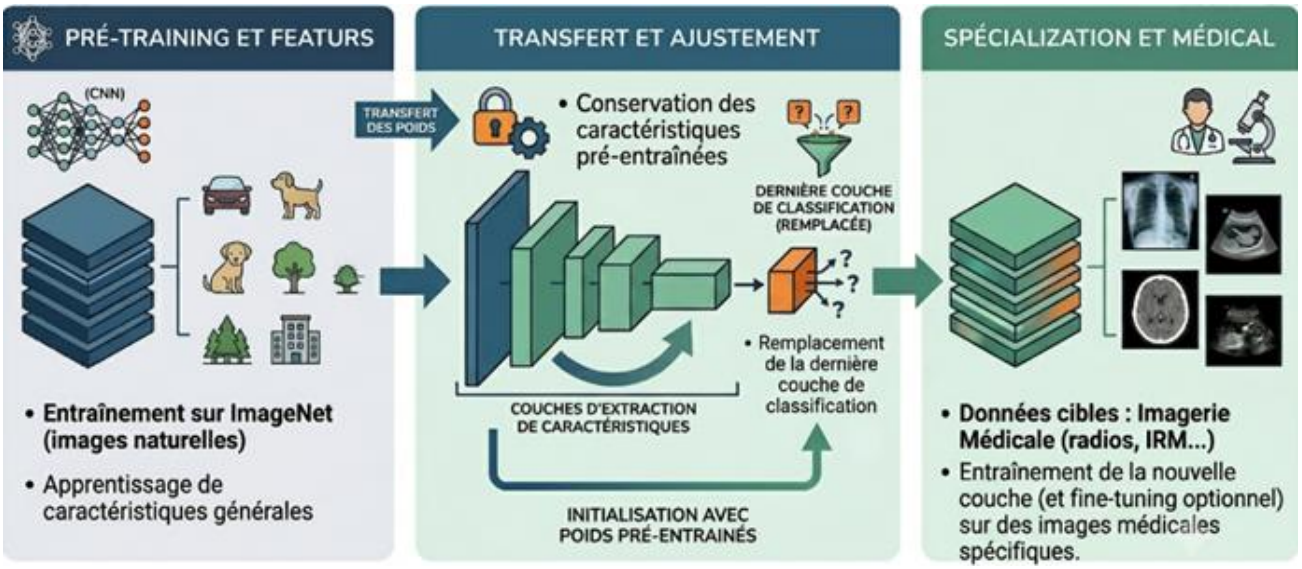


Figure 2.13: Principe du Transfer Learning.

3.7. Architectures CNN de classification les plus utilisées

Le domaine de la vision par ordinateur a vu émerger plusieurs architectures de réseaux de neurones convolutifs de référence. Celles-ci sont aujourd'hui largement adoptées en imagerie médicale grâce à leurs innovations architecturales, leur efficacité computationnelle et leur aptitude à l'ajustement fin sur

des ensembles de données limités. Ces modèles constituent le pilier des approches d'apprentissage par transfert et représentent les principales solutions utilisées pour la classification d'images médicales. Le Tableau 2.1 en présente une synthèse comparative avant leur description détaillée [36].

Architecture	Innovation principale	Avantage médical	Limite principale
EfficientNetB2	Mises à l'échelle structurées tridimensionnelles	Excellent rapport exactitude / taille, idéal pour la fine-tuning médical	Sensible au choix du taux d'apprentissage
ResNet50	Connexions résiduelles	Entraînement très stable, réduction de la dégradation du gradient sur les petits ensembles de données	Moins efficace qu'EfficientNet à taille de modèle équivalente
DenseNet121	Connexions denses entre toutes les couches	Réutilisation maximale des caractéristiques, limite le risque de surapprentissage	Consommation de mémoire GPU très élevée
InceptionV3	Convolutions multi-échelles en parallèle (Inception modules)	Capture les caractéristiques à plusieurs résolutions spatiales	Architecture complexe, temps d'entraînement élevé
VGG19	Convolutions 3×3 profondément empilées	Architecture simple, facile à appréhender et à adapter	Modèle très lourd, lent et gourmand en espace de stockage
Xception	Convolutions séparables en profondeur	Très efficace en calcul, bon compromis entre expressivité et coût	Moins robuste sur de très petits ensembles de données
MobileNetV2	Blocs résiduels inversés et convolutions séparables en profondeur	Modèle très léger, idéal pour les applications médicales embarquées	Capacité de représentation plus faible sur les tâches hautement complexes

Tableau 2.1 : Synthèse comparative des principales architectures CNN utilisées en imagerie médicale.

3.7.1. EfficientNetB2

L'architecture EfficientNet, introduite par Tan et Le, repose sur le concept novateur de mise à l'échelle composite (compound scaling). Contrairement aux approches traditionnelles qui se contentent d'augmenter arbitrairement la profondeur, la largeur ou la résolution d'un réseau, cette méthode optimise simultanément ces trois dimensions de manière équilibrée, selon un coefficient fixe déterminé par recherche d'architecture neuronale. EfficientNetB2, qui constitue le deuxième variant de cette famille de modèles, offre un excellent compromis entre l'exactitude globale et le nombre de paramètres adaptatifs. Ses performances surpassent régulièrement celles d'architectures beaucoup plus lourdes et complexes sur la base de données de référence ImageNet. Grâce à cette remarquable compacité et à sa capacité de généralisation, ce modèle s'avère particulièrement robuste et adapté pour l'étape

d'ajustement fin sur des ensembles de données médicaux de taille restreinte. Le principe fondamental de ce mécanisme est schématisé par la Figure 2.14 [40].

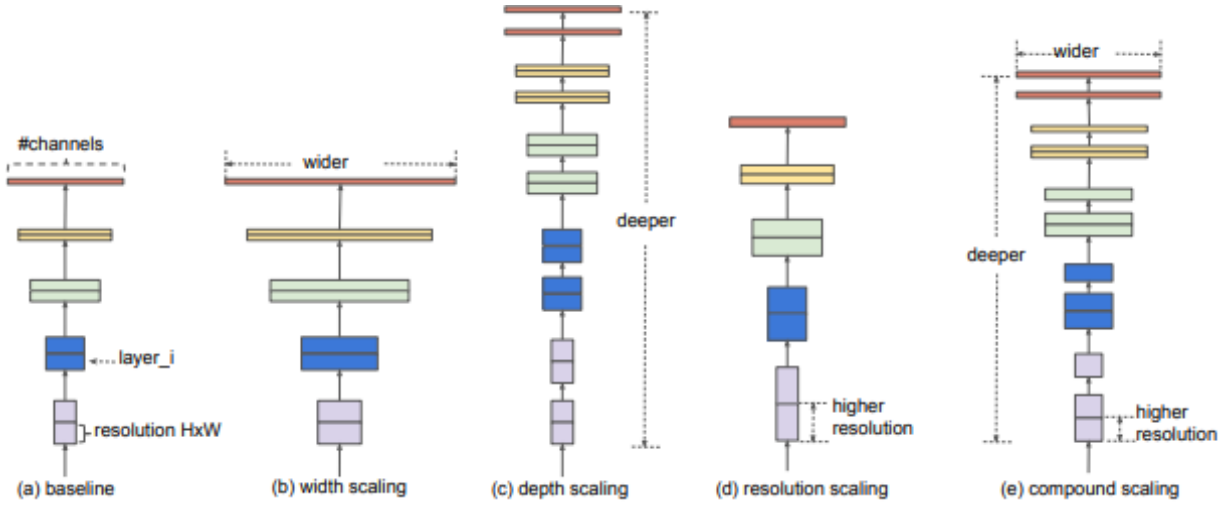


Figure 2.14: Principe du compound scaling d'EfficientNet [40].

3.7.2. ResNet50

L'architecture ResNet (Residual Networks), introduite par He et al., a permis de surmonter le problème majeur de la dégradation des performances dans les réseaux de neurones très profonds. Ce phénomène est résolu grâce à l'introduction de connexions résiduelles, également appelées connexions de saut (skip connections). Ces raccourcis permettent une meilleure propagation du signal d'erreur à travers le réseau en contournant certaines couches, limitant ainsi drastiquement le phénomène de disparition du gradient (vanishing gradient). Avec ses 50 couches de profondeur et environ 25,6 millions de paramètres adaptatifs, ResNet50 s'appuie sur des blocs spécifiques (comme les blocs Bottleneck) pour optimiser le coût de calcul. La Figure 2.15 illustre l'agencement détaillé de ces différents blocs au sein de l'architecture globale du réseau [41].

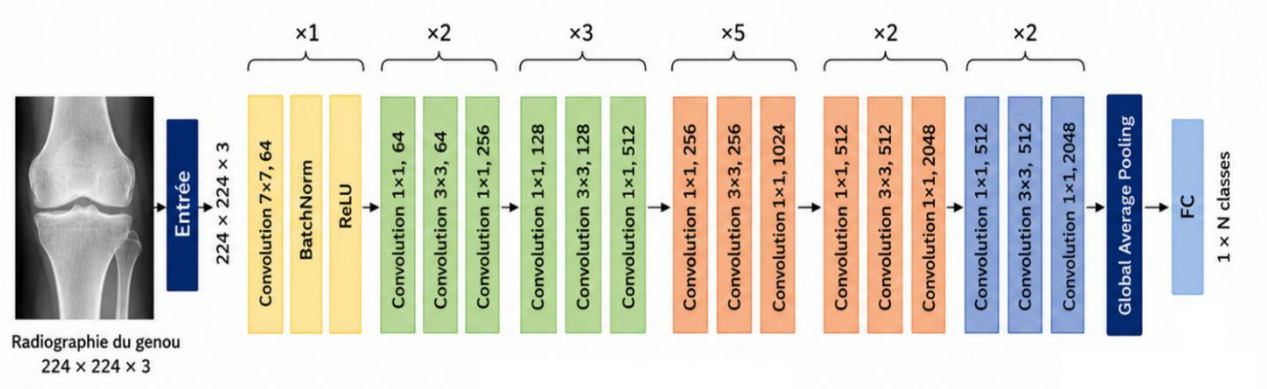


Figure 2.15: Architecture détaillée duResNet50 [36].

Grâce à cette robustesse structurelle et à sa remarquable stabilité lors de la convergence, ResNet50 s'impose comme l'une des architectures les plus exploitées en apprentissage par transfert médical. Elle garantit d'excellentes performances lors de l'étape d'ajustement fin sur des ensembles de données cliniques de taille réduite ou modérée [41].

3.7.3. DenseNet121

L'architecture DenseNet (Densely Connected Convolutional Networks), introduite par Huang et al., propose une approche novatrice où chaque couche est connectée à toutes les couches précédentes. Contrairement à ResNet qui réalise une sommation des valeurs, DenseNet opère par concaténation des cartes de caractéristiques. Ce mécanisme garantit une réutilisation maximale des caractéristiques à travers tout le réseau, renforce la propagation des fonctionnalités et optimise le flux du gradient. Cette connectivité dense continue est schématisée dans l'architecture globale présentée par la Figure 2.16 [36, 42].

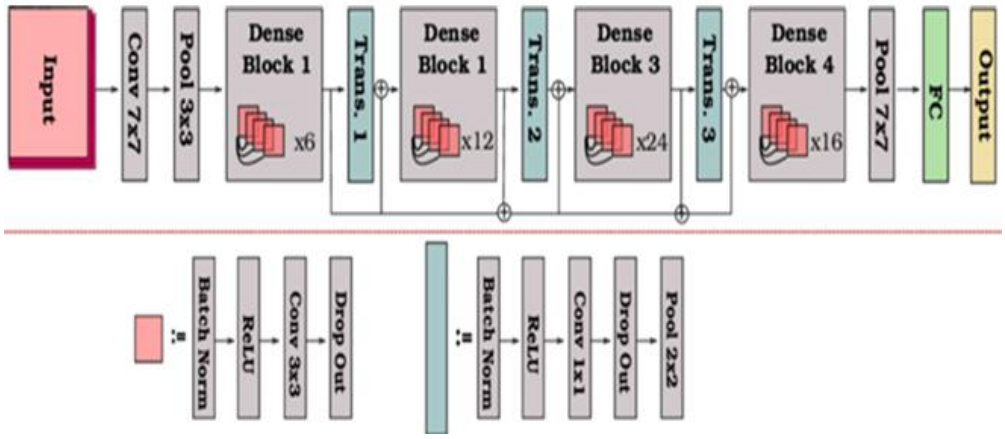


Figure 2.16: Architecture DenseNet121[42].

La version DenseNet121 (qui comprend 121 couches pour environ 8 millions de paramètres adaptatifs) allie efficacement une grande profondeur computationnelle et une remarquable économie de paramètres. Cette compacité structurelle limite fortement le risque de surapprentissage, rendant ce modèle particulièrement performant et adapté pour l'analyse d'ensembles de données médicaux de taille limitée ainsi que pour les tâches complexes de classification d'images cliniques [36, 42].

3.7.4. InceptionV3

L'architecture InceptionV3 introduit le concept de module Inception, qui consiste à appliquer en parallèle plusieurs convolutions de tailles différentes (1×1, 3×3, 5×5) sur un même niveau, avant de concaténer leurs sorties. Cette approche multi-échelles permet de capturer simultanément des caractéristiques visuelles à des résolutions variées. Cette propriété est particulièrement utile en imagerie médicale, où la taille et la morphologie des lésions pathologiques peuvent fortement varier d'un patient à l'autre. Le principe structurel de ce module est schématisé par la Figure 2.17 [36, 43].

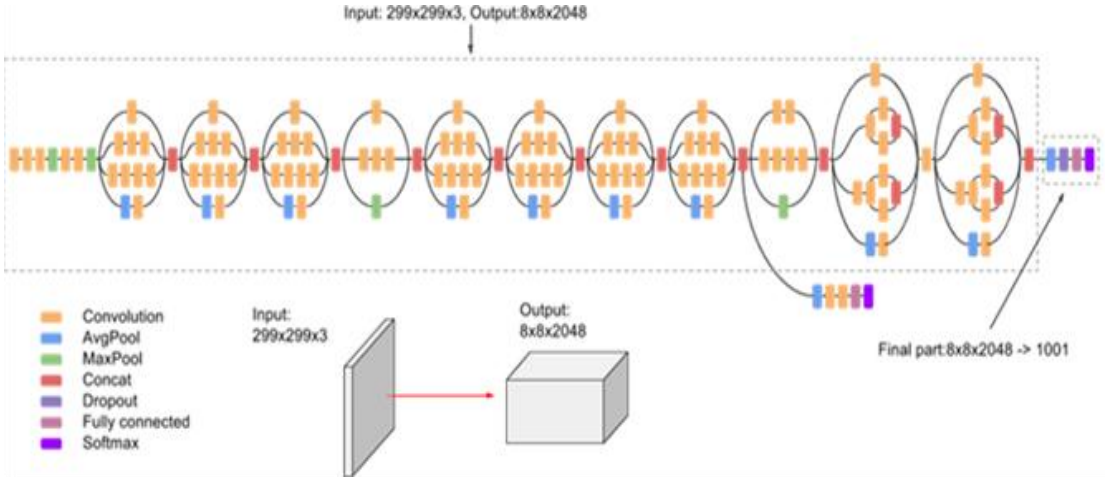


Figure 2.17: Architecture générale du module InceptionV3[43].

3.7.5. VGG19

L'architecture VGG19, proposée par Simonyan et Zisserman, se caractérise avant tout par sa grande simplicité structurelle. Elle repose sur un empilement régulier de convolutions de petite taille (3×3) organisées en blocs de plus en plus profonds. La Figure 2.18 présente cette configuration linéaire [37, 44].

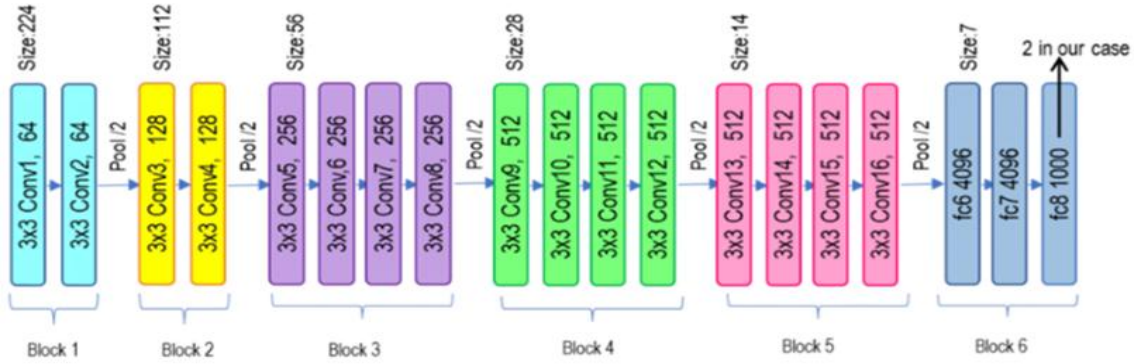


Figure 2.18: Architecture séquentielle et profonde du réseau VGG19 [44].

Grâce à cette régularité, VGG19 demeure une architecture de référence et une base classique pour l'apprentissage par transfert. Cependant, sa principale limitation réside dans son nombre massif de paramètres adaptatifs (environ 138 millions), ce qui engendre un coût mémoire et computationnel très élevé lors des phases d'entraînement [37, 44].

3.7.6. Xception

L'architecture Xception (Extreme Inception), introduite par Chollet, pousse plus loin le concept d'Inception en remplaçant les convolutions classiques par des convolutions séparables en profondeur (depthwise separable convolutions). Ce mécanisme dissocie complètement le filtrage spatial (appliqué canal par canal) de la combinaison des canaux via une convolution 1×1. L'organisation de ce traitement est illustrée par la Figure 2.19 [45, 46].

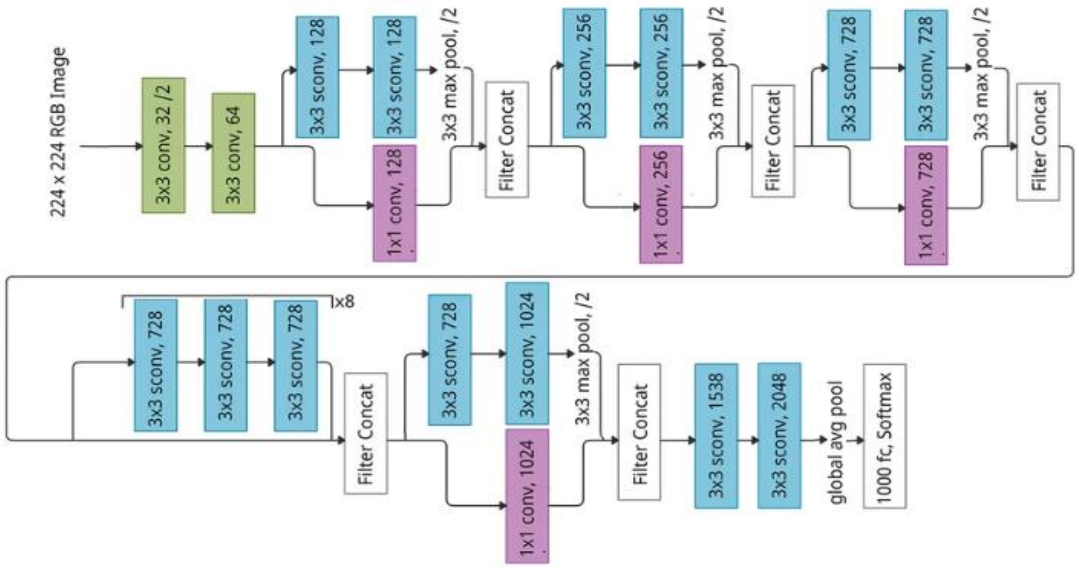


Figure 2.19: Architecteur du réseau Xception [46].

Cette dissociation réduit significativement le nombre d'opérations mathématiques et le coût computationnel tout en conservant une excellente capacité de représentation, offrant ainsi un outil hautement efficace pour la classification d'images médicales [45].

3.7.7. MobileNetV2

L'architecture MobileNetV2 est spécifiquement optimisée pour les environnements matériels disposant de ressources de calcul limitées. À l'instar de Xception, elle s'appuie sur des convolutions séparables en profondeur, mais y intègre des blocs résiduels inversés (inverted residual blocks) combinés à des couches de projection linéaire. Le flux de données à travers ces blocs est présenté dans la Figure 2.20 [36].

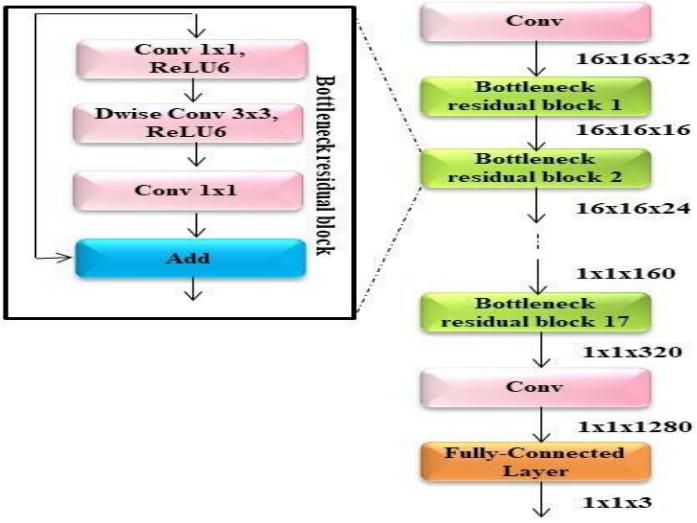


Figure 2.20: Architecture et structure des blocs résiduels inversés de MobileNetV2 [36].

Avec un volume restreint d'environ 3,4 millions de paramètres, ce modèle offre un compromis optimal entre exactitude globale, rapidité d'inférence et efficacité mémoire. Cette légèreté le rend particulièrement compétitif pour l'apprentissage par transfert appliqué aux systèmes d'imagerie médicale embarqués ou aux outils de diagnostic nomades [36].

Il convient de noter que ces architectures, bien que performantes, ont été initialement optimisées pour ImageNet. Leur adaptation à l'imagerie médicale peut nécessiter des modifications mineures, comme l'adaptation de la première couche convolutionnelle aux images monochromes.

4. Métriques d'évaluation des performances

L'évaluation des performances d'un modèle de classification médicale est une étape essentielle, particulièrement en présence de classes déséquilibrées. Dans ce contexte, certaines métriques conventionnelles, comme l'exactitude, peuvent être trompeuses en favorisant artificiellement la classe majoritaire au détriment de la classe pathologique minoritaire. Cette section présente les principales métriques d'évaluation, notamment celles adaptées aux problèmes de déséquilibre des données, ainsi que leur intérêt dans l'interprétation des performances cliniques [36].

4.1. Matrice de confusion

La matrice de confusion est le point de départ de toute évaluation de performances. Pour un problème de classification binaire, elle organise les prédictions du modèle en quatre quantités fondamentales [36] :

- **Vrais Positifs (VP)** : le modèle identifie correctement un cas pathologique.
- **Vrais Négatifs (VN)** : un cas sain (négatif) est correctement reconnu comme tel.
- **Faux Positifs (FP)** : un cas sain est classé à tort comme pathologique (fausse alerte).
- **Faux Négatifs (FN)** : un cas pathologique n'est pas détecté par le modèle (manque critique).

Ces quatre valeurs sont organisées dans la matrice de confusion présentée au Tableau 2.2.

	Prédiction positive	**Prédiction négative**
Réel positif	**VP**	**FN**
Réel négatif	**FP**	**VN**

Tableau 2.2: Structure d'une matrice de confusion binaire.

4.2. Métriques de classification

À partir des quatre valeurs de la matrice de confusion, plusieurs métriques mathématiques sont définies pour évaluer le comportement du classifieur [36].

- **Exactitude (Accuracy)** : mesure la proportion globale d'exemples correctement prédits, toutes classes confondues :

$$Accuracy = \frac{VP+VN}{VP+VN+FP+FN} \quad (2.9)$$

- **Rappel ou Sensibilité :** mesure la capacité du modèle à identifier l'intégralité des cas positifs réels. En imagerie médicale, c'est une métrique vitale pour s'assurer qu'aucune anomalie n'échappe au diagnostic :

$$\text{Rappel} = \frac{VP}{VP+FN} \quad (2.10)$$

- **Précision :** mesure la proportion de prédictions positives qui sont effectivement de vrais cas pathologiques :

$$Précision = \frac{VP}{VP+FP} \quad (2.11)$$

- **F1-score :** représente la moyenne harmonique de la Précision et du Rappel, offrant une métrique unique et robuste pour évaluer un modèle face à un fort déséquilibre de classes :

$$F1 - score = \frac{2\times(Precision\times Recall)}{(Precision+Recall)} \quad (2.12)$$

- **Spécificité :** mesure la capacité du réseau à reconnaître les cas sains et à ne pas générer de fausses alertes cliniques :

$$Specificité = \frac{VN}{VN+FP} \quad (2.13)$$

- **G-Mean (Moyenne Géométrique) :** particulièrement adapté aux architectures confrontées au déséquilibre des données, il évalue le compromis simultané entre la Sensibilité et la Spécificité, sans surévaluer la classe majoritaire :

$$G - Mean = \sqrt{\text{Rappel} \times Specificité} \quad (2.14)$$

En médecine, le Rappel demeure la métrique prioritaire pour limiter les non-détections de pathologies, tandis que le F1-score fournit une mesure finale équilibrée des performances sur l'ensemble des classes, y compris les minoritaires [36].

4.3. Courbe ROC et AUC

Courbe ROC (Receiver Operating Characteristic) et son AUC (Area Under the Curve) évaluent la capacité discriminante globale d'un modèle à différents seuils de décision. L'AUC varie entre 0 et 1, où une valeur de 1,0 représente un classifieur parfait [36].

4.4. Courbe PRC

La courbe précision-rappel (Precision-Recall Curve, PRC) est une représentation graphique utilisée pour évaluer les performances des modèles de classification binaire. Elle illustre le compromis entre la précision (en ordonnée) et le rappel (en abscisse) en fonction des différents seuils de décision, dont les valeurs sont généralement comprises entre 0 et 1 [47].

4.5. Courbe DCA

L'analyse par courbe de décision (Decision Curve Analysis, DCA) est une méthode d'évaluation des modèles de prédiction clinique reposant sur une approche diagnostique par étapes. Elle permet de comparer différentes stratégies de décision en tenant compte des seuils de probabilité et du bénéfice net associé à chacune d'elles [48].

Au-delà de la performance mesurée, la confiance clinique nécessite de comprendre les décisions du modèle, ce qui nous amène à l'explicabilité.

5. L'intelligence artificielle interprétable (Explainable AI, XAI) en médecine

Les architectures d'apprentissage profond présentées dans les sections précédentes offrent des performances remarquables pour la classification d'images médicales. Cependant, leur fonctionnement interne complexe demeure opaque, ce qui conduit souvent la communauté scientifique et médicale à les qualifier de « boîtes noires ». En contexte clinique, cette absence de transparence représente une limite majeure. Un professionnel de santé ne peut intégrer un outil d'aide au diagnostic dans sa routine thérapeutique sans comprendre, valider et assumer la responsabilité des bases d'une prédiction [49].

L'explicabilité s'impose ainsi comme un impératif à la fois clinique, éthique et réglementaire. Elle permet de :

- Renforcer la confiance des cliniciens en justifiant mathématiquement et visuellement les décisions du modèle ;
- Identifier d'éventuels biais d'apprentissage (comme l'utilisation de caractéristiques contextuelles non pertinentes médicalement) ;
- Répondre aux exigences strictes de transparence imposées par les nouvelles réglementations, telles que l'AI Act européen ou les directives de la FDA (Food and Drug Administration) pour les dispositifs médicaux basés sur l'intelligence artificielle [50, 51].

Dans ce cadre, les méthodes de XAI post-hoc, qui analysent le modèle après son entraînement, offrent des solutions robustes. Parmi elles, les approches de cartographie d'activation de classe (Class Activation Mapping, CAM), et plus particulièrement l'algorithme Grad-CAM, sont les plus largement adoptées pour inspecter les régions d'une image ayant guidé la décision du réseau [7, 52].

5.1. Principe de Grad-CAM

L'algorithme Grad-CAM (Gradient-weighted Class Activation Mapping), proposé par Selvaraju et al en 2017, est la méthode d'explicabilité de référence pour les CNN en imagerie médicale. Son principe fondamental consiste à exploiter les gradients de la classe cible calculés par rapport aux activations de la toute dernière couche convolutionnelle, afin d'identifier précisément les régions spatiales de l'image d'entrée les plus influentes dans la prédiction [7].

Pour une classe cible c, Grad-CAM détermine d'abord le poids d'importance neuronale α_k^c associé à la k-ième carte de caractéristiques A^k de la dernière couche convolutionnelle. Ce poids est calculé à partir de la moyenne globale des gradients selon l'équation suivante :

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{i,j}^k} \qquad (2.15)$$

Où y^c représente le score brut de la classe c avant l'application de la fonction Softmax, $A_{i,j}^k$ désigne la valeur d'activation de la carte k à la position spatiale (i,j) , et Z correspond au nombre total de pixels de la carte de caractéristiques ($Z = hauteur \times largeur$) [7].

La carte de chaleur Grad-CAM finale ($L_{Grad-CAM}^c$) est ensuite générée par une combinaison linéaire pondérée de l'ensemble des cartes de caractéristiques, sur laquelle est appliquée la fonction d'activation non linéaire ReLU :

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A^k) \qquad (2.16)$$

L'utilisation de la fonction ReLU est ici essentielle : elle permet de filtrer et de conserver uniquement les caractéristiques qui ont une contribution positive à la classe étudiée (celles qui augmentent le score y^c), en éliminant les pixels qui l'affaiblissent [7].

La carte d'activation bidimensionnelle obtenue est enfin redimensionnée par interpolation pour correspondre aux dimensions de l'image initiale, puis superposée à celle-ci sous la forme d'une carte de chaleur (heatmap) colorée. Le flux complet de ce traitement, de l'acquisition des gradients à la visualisation finale, est détaillé sur la Figure 2.21 [7].

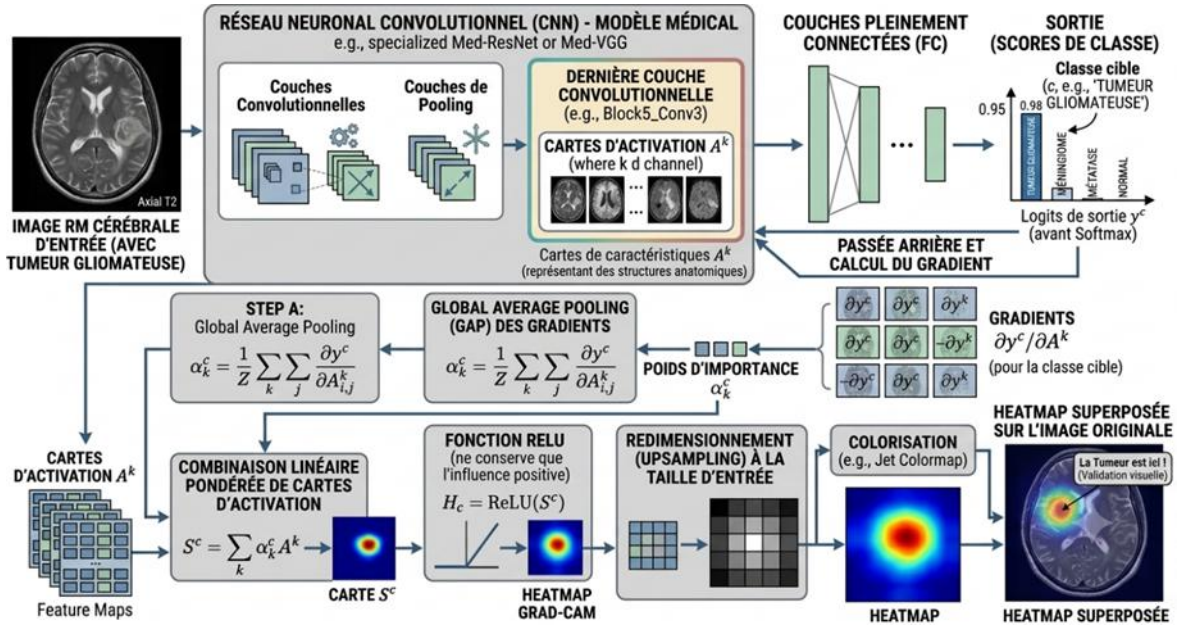


Figure 2.21: Pipeline complet de la méthode Grad CAM : de l'image d'entrée aux gradients de la dernière couche convolutionnelle, jusqu'à la heatmap superposée à l'image originale.

Le choix analytique de cibler la dernière couche convolutionnelle est particulièrement judicieux. C'est à ce niveau précis du réseau que se trouve le meilleur compromis entre la richesse sémantique (la compréhension des formes pathologiques complexes) et la préservation des informations de localisation spatiale, rendant l'interprétation diagnostique parfaitement fiable pour le clinicien [7].

5.2. Interprétation visuelle des décisions cliniques

Après avoir présenté le fonctionnement de Grad-CAM, nous montrons maintenant comment les cartes de chaleur produites sont interprétées en pratique clinique.

En imagerie médicale, les heatmaps générées par Grad-CAM transposent une prédiction numérique abstraite en une information topographique et spatiale directement intelligible par le corps médical. Elles permettent de vérifier instantanément si les régions de forte activation neuronale se superposent aux structures anatomiques ou aux lésions effectivement associées à la pathologie étudiée. Cette confrontation constitue une validation qualitative essentielle de la cohérence diagnostique du modèle. La Figure 2.22 illustre cette double réalité : à gauche, une heatmap médicalement pertinente où l'activation se concentre sur la lésion ; à droite, une heatmap perturbée par des facteurs de confusion (artefacts, texte, etc.), qui éclaire des régions non pertinentes [53].

Cette capacité d'inspection s'avère particulièrement cruciale pour des pathologies complexes telles que les tumeurs cérébrales, les maladies neurodégénératives (comme la maladie d'Alzheimer) ou les lésions musculosquelettiques, où les biomarqueurs visuels sont localisés. Une convergence visuelle sur les régions d'intérêt anatomiques suggère que le réseau extrait et exploite des caractéristiques sémantiques médicalement significatives. À l'inverse, une focalisation sur des zones périphériques ou saines peut révéler de profonds biais d'apprentissage ou la capture de facteurs de confusion contextuels [54].

5.3. Limites intrinsèques des cartes de chaleur

Malgré leur apport indéniable à la transparence des modèles, les cartes de chaleur Grad-CAM présentent des limites théoriques et pratiques qu'il convient de souligner [7] :

- **Résolution spatiale grossière** : puisqu'elles sont calculées à partir des gradients de la dernière couche convolutionnelle (qui a subi de multiples opérations de pooling), les heatmaps souffrent d'une faible résolution spatiale. Elles fournissent une zone d'intérêt floue et ne permettent en aucun cas une délimitation ou une segmentation précise des contours de la lésion.
- **Sensibilité aux artefacts et corrélations fallacieuses** : les cartes d'activation peuvent être fortement influencées par des artefacts d'acquisition (bruit de fond, variations de contraste, présence de texte ou de marqueurs de calibration sur les clichés). Le modèle risque alors de fonder sa décision sur des caractéristiques totalement étrangères à la pathologie, comme illustré sur le panneau droit de la Figure 2.22.

- **Illusion de robustesse** : une activation visuellement cohérente à l'œil nu ne garantit pas scientifiquement que le réseau a assimilé les véritables mécanismes causaux du diagnostic médical. Il peut s'agir d'une coïncidence statistique propre à l'ensemble de données d'entraînement.

Par conséquent, les heatmaps ne doivent pas être perçues comme une preuve absolue de vérité clinique, mais plutôt comme un outil d'audit et d'aide à l'interprétation. Elles doivent impérativement être couplées à des métriques quantitatives rigoureuses avant toute considération de déploiement en routine médicale [7].

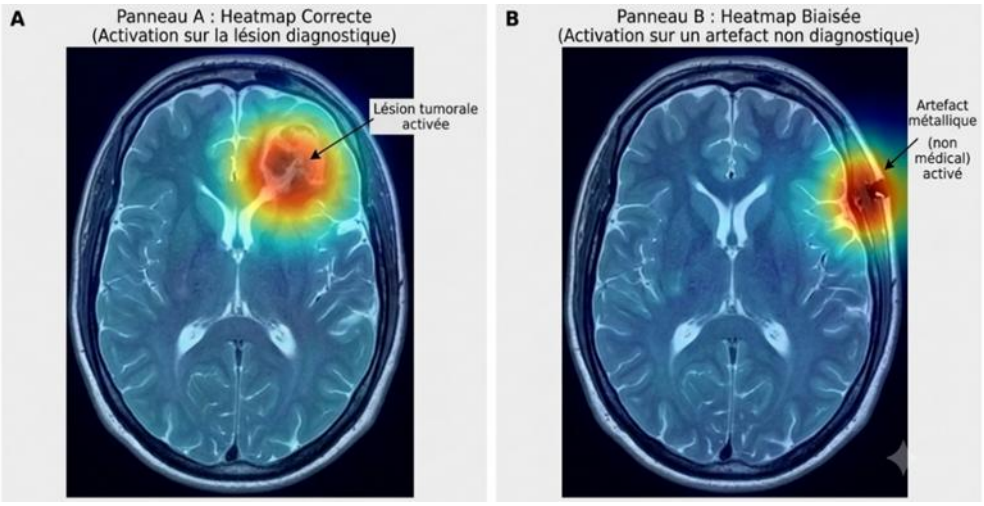


Figure 2.22: Exemple illustrant une heatmap médicalement pertinente et une heatmap influencée par des facteurs de confusion

5.4. Validation quantitative et causale des méthodes XAI

La simple inspection visuelle des cartes de chaleur est insuffisante pour garantir la crédibilité clinique d'un système d'IA médical. Adebayo et al ont en effet démontré que de nombreuses méthodes de cartes de saillance (saliency maps) produisent des visualisations de contours quasi identiques, même lorsque les poids du modèle sont totalement aléatoires. Cette observation fondamentale justifie le recours à des tests de validation rigoureux, objectifs et indépendants de l'œil humain [55].

5.4.1. Tests de cohérence élémentaires (*Sanity Checks*)

Les sanity checks constituent un ensemble de vérifications de cohérence fondamentales pour évaluer la fidélité d'une méthode d'explicabilité [55] :

- **Test de randomisation des paramètres du modèle** : il consiste à générer les heatmaps après avoir réinitialisé de manière aléatoire les poids du réseau. Si les cartes obtenues ne diffèrent pas significativement de celles du modèle entraîné, la méthode XAI ne reflète pas les connaissances acquises lors de l'apprentissage, mais traduit simplement une propriété géométrique de l'architecture ou des contrastes natifs de l'image.

- **Test de sensibilité aux classes** : il vérifie qu'une même image médicale engendre des heatmaps distinctes selon la classe d'intérêt ciblée (par exemple, une zone d'activation différente pour une classe « bénigne » versus « maligne »), prouvant que l'explication est bien conditionnée par la décision logique du réseau.
- **Test de cohérence avec l'arrière-plan** : il s'assure que les zones de forte activation évitent les régions non informatives ou vides de l'image, à l'instar du fond noir entourant une coupe d'imagerie par résonance magnétique.

5.4.2. Test d'occlusion et validation causale

Le test d'occlusion est une méthode d'évaluation causale directe. Si la région anatomique mise en évidence par la heatmap est réellement déterminante dans la décision du classifieur, son masquage doit induire une chute immédiate et significative du score de confiance de la prédiction [56].

En pratique, les pixels identifiés comme les plus critiques sont remplacés par une valeur neutre, puis la nouvelle probabilité de sortie est comparée à la valeur originale. Une chute d'assurance supérieure à un seuil critique, typiquement fixé à 15 %, confirme la pertinence causale de la région identifiée. Ce test permet de transformer une simple observation visuelle corrélative en une preuve d'impact causal [56].

5.4.3. Analyse de superposition géométrique : IoU et coefficient de Dice

Lorsque des masques de segmentation rigoureux annotés par des experts cliniciens sont disponibles, l'indice d'Intersection sur Union (Intersection over Union, IoU) et le coefficient de Dice permettent d'évaluer quantitativement la cohérence géométrique spatiale entre la carte d'explicabilité binarisée (A) et la zone pathologique réelle de référence (B) [57]:

$$\mathrm{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|} \qquad (2.17)$$

$$\mathrm{Dice}(A, B) = \frac{2|A \cap B|}{|A| + |B|} \qquad (2.18)$$

Dans la littérature, des valeurs d'IoU modérées sont généralement observées avec l'algorithme Grad-CAM. Cette tendance est normale et s'explique par le fait que cette méthode fournit une zone de localisation macroscopique et diffuse plutôt qu'une segmentation précise des contours lésionnels. Ainsi, l'IoU ne doit pas être interprété comme une mesure d'échec, mais comme un indicateur du caractère approximatif mais néanmoins cliniquement utile de la localisation [58].

6. Analyse de la littérature et positionnement

Après avoir présenté les concepts fondamentaux, les architectures et les méthodes d'évaluation, nous examinons maintenant les travaux récents appliqués à la maladie d'Alzheimer, puis nous positionnons notre approche par rapport à ces études.

6.1. Travaux existants et études récentes

Le domaine de l'aide au diagnostic de la maladie d'Alzheimer a connu des avancées considérables grâce à l'avènement de l'apprentissage profond, notamment à travers la classification d'images IRM en différents stades de démence. Cependant, la sensibilité de ces architectures connexionnistes face au déséquilibre des classes demeure un verrou méthodologique majeur, une problématique particulièrement saillante au sein de cohortes de référence telles que l'ADNI (Alzheimer's Disease Neuroimaging Initiative). Ce constat a naturellement orienté les recherches récentes vers l'exploration de deux axes synergiques : d'une part, les techniques d'augmentation de données synthétiques par le biais de réseaux antagonistes génératifs (GAN) et, d'autre part, le déploiement d'architectures avancées exploitant l'apprentissage par transfert.

S'inscrivant dans cette dynamique, Bouzidi et al. (2022) ont proposé une approche fondée sur l'apprentissage par transfert pour la classification multiclasse de la pathologie. En exploitant une base de données issue de la plateforme Kaggle, regroupant 6400 images IRM réparties en quatre classes nosologiques, les auteurs ont mis en confrontation six architectures pré-entraînées : ResNet50, VGG16, VGG19, Inception, Xception et MobileNet. Leurs investigations démontrent que le réseau VGG19 surpasse ses pairs en s'adjugeant une accuracy de 93,66%, confirmant ainsi l'efficacité du transfert de caractéristiques pour ce type d'imagerie [59]. Parallèlement, Ghaffari et al. (2022) ont développé un pipeline entièrement automatisé de détection et de classification. Leur méthodologie, adossée à la base de données ADNI, intègre un protocole rigoureux de prétraitement comprenant une phase de segmentation tissulaire via un réseau U-Net, une amélioration du contraste et une normalisation d'intensité des images. L'évaluation comparative de trois modèles pré-entraînés (ResNet101, Xception et InceptionV3) a mis en évidence la supériorité d'InceptionV3, ce dernier atteignant une accuracy de 93,75% et une AUC de 96,00 %, tout en validant une excellente capacité de généralisation lors de tests menés sur les bases de données externes OASIS et AIBL [60].

Dans une perspective complémentaire, Femmam et al. (2024) ont introduit une méthode reposant sur l'extraction multi-échelle de descripteurs morphologiques appliquée à la base OASIS. Afin de contrer l'impact du déséquilibre distributif des classes, les auteurs ont mis en œuvre des techniques de suréchantillonnage et de sous-échantillonnage, avant de comparer les performances des réseaux ResNet50, VGG16 et MobileNet. Le modèle ResNet50 a formalisé les meilleurs résultats, revendiquant une accuracy de 97,58% et un F1-score moyen de 98%, ce qui valide empiriquement l'intérêt de coupler des algorithmes de rééquilibrage de données aux architectures profondes [61]. De leur côté, Shahid et al. (2025) ont exploré l'impact de la fusion de multiples bases de données IRM, en l'occurrence l'ADNI et l'OASIS, pour optimiser la classification multiclasse. Confrontant plusieurs modèles d'apprentissage par transfert, dont ResNet152V2, VGG16, InceptionV3 et MobileNet, leurs conclusions indiquent que

l'architecture ResNet152V2 s'impose comme la plus performante. Ce modèle sécurise une accuracy d'environ 95,00 % sur l'ADNI, 93,00 % sur l'OASIS, et frôle les 96,00 % sur le jeu de données unifié, mettant ainsi en lumière le rôle crucial de la diversité et du volume des cohortes d'entraînement pour accroître la robustesse des modèles prédictifs [62]. Enfin, Meriem Saim (2025) a développé plusieurs approches complémentaires dédiées au diagnostic assisté par ordinateur. Ces travaux originaux combinent des méthodes d'extraction de caractéristiques géométriques et texturales classiques (telles que les descripteurs HOG, FFST et la matrice GLCM) à des techniques d'apprentissage profond basées sur le transfert de connaissances. Le déploiement d'une approche hybride, intégrant cette extraction automatique de caractéristiques et une phase d'optimisation bayésienne des hyperparamètres, a permis d'atteindre une accuracy remarquable de 98,45 %, mettant en évidence la forte synergie entre l'apprentissage automatique traditionnel et le Deep Learning [63].

Une attention particulière doit être portée aux investigations récentes qui partagent notre cadre expérimental en exploitant le jeu de données de référence Alzheimer Preprocessed Data, également structuré en quatre classes de démence. En premier lieu, l'approche par Deep Ensemble développée par Andrea Loddo et al. (2022) combine trois architectures pré-entraînées hétérogènes, à savoir AlexNet, ResNet-101 et InceptionResNetV2. Leur modèle surpasse les configurations isolées en atteignant une accuracy de 98,24%, ce qui démontre l'intérêt de la fusion d'architectures complémentaires pour stabiliser les frontières de décision. Bien que les auteurs soulignent la pertinence théorique des techniques de visualisation d'activations telles que Grad-CAM pour identifier l'atrophie de structures critiques comme l'hippocampe ou les ventricules latéraux, leur protocole n'intègre ni validation quantitative ni cadre formel d'explicabilité visuelle, ce qui en restreint l'applicabilité clinique directe [64]. En second lieu, Hina Tufail et al. (2024) ont introduit une architecture hybride associant le réseau InceptionV3 à un modèle convolutif sur mesure, couplé à un GAN dédié à l'augmentation synthétique des données. Si cette stratégie de régularisation permet d'obtenir une AUC honorable de 0,87, le rappel associé à la classe minoritaire Moderate Demented plafonne à une valeur critique de 33,00% en phase de test. Cette défaillance met en lumière une limite méthodologique fondamentale : la fidélité visuelle et texturale des échantillons générés par le GAN ne garantit en rien leur utilité sémantique pour le classifieur [65].

Pour pallier ce phénomène de distorsion, Marya Ryspayeva et Olga Salykova (2025) ont formalisé le cadre DGAN-WP-TL, matérialisé par un réseau WGAN-GP (perte de Wasserstein avec pénalité de gradient) adossé à une fonction de coût perceptuelle calibrée sur VGG19. Avant l'application de ce mécanisme d'augmentation, leur modèle de base ResNet-18 affichait une accuracy de 78,30% et un F1-score de 76,30%. L'injection contrôlée de 50 % d'images synthétiques a suscité une amélioration substantielle des performances, élevant l'accuracy à 85,70% et la spécificité à 93,90%. Néanmoins,

cette étude rapporte l'émergence d'un taux d'artefacts visuels s'élevant à 42,00%, se traduisant par de sévères déformations anatomo-pathologiques, tandis que l'absence de validation par l'intelligence artificielle explicable (XAI) fragilise la confiance médicale accordée à la plausibilité des représentations ainsi générées [66]. Par ailleurs, Mojgan Askarizadeh et al. (2025) ont exploité le potentiel de l'architecture EfficientNetV2S conjointe à une stratégie de rééquilibrage de distribution via un modèle WGAN-GP. Leur pipeline atteint des performances hautement compétitives, revendiquant une accuracy de 98,10% et un F1-score de 98,60% sur les quatre stades de la pathologie. En dépit de cette efficacité statistique, l'omission de méthodes de saillance ou d'interprétabilité post-hoc pour cartographier les invariants morphologiques exploités par le réseau constitue une lacune notable, faisant obstacle à toute forme de certification clinique [67]. Enfin, les travaux de Lilia Lazli et al. (2026) proposent l'algorithme MAMIR (Multi-stage Adaptive Multi-class Imbalanced Resampling), une technique de rééchantillonnage hybride novatrice combinant une estimation non paramétrique de densité et l'analyse des k plus proches voisins. Leur classifieur sécurise une accuracy de 95,00% corrélée à un coefficient de Matthews (MCC) de 0,927, un indicateur particulièrement rigoureux pour valider l'absence de biais adaptatif face à un déséquilibre de classes. Toutefois, à l'instar des paradigmes précédents, la mise en œuvre d'outils XAI quantitatifs demeure reléguée au rang de perspective de recherche future plutôt que d'être traitée comme un jalon empirique pleinement établi [68].

En somme, l'analyse approfondie de la littérature existante démontre que l'apprentissage par transfert constitue aujourd'hui le paradigme dominant pour la classification computationnelle de la maladie d'Alzheimer à partir d'images IRM. Les architectures de pointe issues des familles ResNet, VGG, Inception, Xception et EfficientNet s'imposent comme les extracteurs de caractéristiques les plus performants. Parallèlement, l'intégration de stratégies de rééchantillonnage, d'augmentation de données conventionnelle et, plus récemment, de techniques de génération synthétique par GAN permet d'élever significativement la sensibilité du diagnostic sur les classes minoritaires. Néanmoins, un triple verrou subsiste au sein de la plupart des études répertoriées : la fragilité de la généralisation des modèles face à des artefacts, la difficulté d'absorber des déséquilibres sévères sans biais adaptatif, et la quasi-absence d'intégration ou de validation quantitative des méthodes d'explicabilité visuelle.

C'est précisément dans cette perspective scientifique que s'inscrivent nos travaux. Notre contribution propose un pipeline unifié basé sur l'architecture EfficientNetB2, couplée à une stratégie hybride et rigoureuse de gestion du déséquilibre des classes et à une augmentation de données ciblée. La singularité de notre approche repose sur l'adjonction de protocoles d'explicabilité post-hoc formellement validés par des tests d'intégrité quantitative et causale, permettant d'optimiser simultanément la précision statistique de la classification et la fiabilité neuroanatomique du système d'aide au diagnostic.

6.2. Positionnement de la méthode proposée

L'analyse critique de l'état de l'art met en lumière une lacune systémique : le manque de validation formelle de *l'XAI*. Bien que les précisions dépassent parfois 98 %, aucun travail n'apporte la preuve rigoureuse (test d'occlusion ou validation experte) que le modèle ne souffre pas de biais de « double comptage » ou qu'il ne se focalise pas sur des artefacts plutôt que sur l'atrophie cérébrale réelle.

Notre méthode se distingue donc par l'intégration systématique de Grad-CAM couplée à des tests de validation quantitative (IoU, Dice, test d'occlusion) pour garantir la pertinence clinique des cartes d'activation. Nous proposons ainsi une approche où la performance diagnostique élevée s'accompagne d'une transparence vérifiable, répondant aux exigences éthiques et réglementaires des dispositifs médicaux assistés par IA.

7. Conclusion

Ce chapitre a exposé l'ensemble des bases théoriques et conceptuelles indispensables à la compréhension de la méthodologie expérimentale développée dans ce mémoire à travers cinq thématiques interconnectées. En premier lieu, l'étude des stratégies de gestion du déséquilibre de classes a permis de comparer les approches au niveau des données et les modifications algorithmiques, établissant que la Focal Loss et la pondération des classes constituent des leviers particulièrement robustes en imagerie clinique. En second lieu, les fondements de l'apprentissage profond ont été posés en détaillant l'architecture des CNN, leurs mécanismes d'optimisation et les principes de l'apprentissage par transfert, menant à l'identification d'EfficientNetB2 comme le modèle compact optimal pour les ensembles de données médicaux restreints. Ensuite, la définition des métriques d'évaluation en contexte asymétrique a réaffirmé la priorité clinique accordée au rappel et au F1-score face aux limites trompeuses de l'exactitude globale. De plus, l'introduction de l'intelligence artificielle interprétable à travers l'algorithme Grad-CAM a permis de formaliser un protocole de validation quantitative multi-niveaux, démontrant qu'une explication clinique ne peut se réduire à une simple inspection visuelle subjective. Enfin, l'analyse critique de la littérature récente a mis en lumière un manque systémique de validation causale des méthodes XAI dans l'état de l'art actuel. Ce socle théorique consolidé constitue ainsi la base sur laquelle s'appuie le chapitre suivant, dédié à la mise en œuvre pratique de notre pipeline expérimental pour le diagnostic de la maladie d'Alzheimer et la classification des tumeurs cérébrales.

Chapitre 3 :

Méthodologie expérimentale, Résultats et Discussion

1. Introduction

Ce chapitre constitue le pivot empirique et méthodologique de notre étude, marquant notre passage à la phase applicative. Après avoir établi les fondements théoriques de l'apprentissage profond, souligné les limites des modèles standards face au déséquilibre des données médicales, et démontré l'importance de l'intelligence artificielle explicable (XAI) lors du chapitre précédent, nous nous consacrons ici à la mise en œuvre pratique et à l'évaluation expérimentale de l'approche proposée dans ce travail.

L'objectif principal est de formaliser nos choix techniques et de valider empiriquement notre pipeline unifié, conçu pour diagnostiquer de manière robuste les stades pathologiques tout en surmontant le verrou des classes minoritaires. La finalité de cette démarche est d'évaluer, par une approche scientifique rigoureuse, la viabilité, la robustesse et la transparence de notre système d'aide au diagnostic par imagerie médicale.

Pour offrir une vision claire et hautement reproductible, la progression de ce chapitre s'articule autour de deux grands axes complémentaires : le cadre méthodologique (sections 2 à 4) et l'évaluation expérimentale (section 5). Le premier volet définit l'environnement de développement matériel et logiciel (Python, TensorFlow, accélérateurs GPU), ainsi que les caractéristiques hétérogènes et les défis morphologiques des quatre cohortes d'imagerie exploitées (maladie d'Alzheimer, tumeurs cérébrales et ostéoarthrite du genou) afin d'asseoir la transversalité de notre approche. Il détaille également, pas à pas, notre méthodologie algorithmique : pipeline de prétraitement adaptatif (suréchantillonnage sélectif et pondération adaptative), stratégie d'entraînement séquentiel (Two-Phase Learning), et déploiement de l'explicabilité visuelle par la méthode Grad-CAM. Le second volet confronte nos propositions aux réalités des données cliniques à travers quatre phases progressives, analysant successivement la sélection de l'architecture de référence (EfficientNetB2), l'efficacité de notre stratégie hybride de rééquilibrage face aux solutions de l'état de l'art, la robustesse mathématique et causale de nos explications (Brain-Masking et tests d'occlusion), et se concluant par une validation qualitative des cartes de saillance par une expertise clinique. En combinant ainsi optimisation quantitative et explicabilité post-hoc, ce chapitre s'attache à montrer que notre pipeline ne se limite pas à maximiser des métriques de classification brutes, mais apprend de véritables invariants anatomocliniques pertinents, condition sine qua non pour une future intégration en routine hospitalière.

2. Environnement de développement et outils technologiques

Le développement, la mise en œuvre et l'expérimentation des modèles proposés dans ce travail ont été réalisés en Python. Ce choix se justifie par son statut de standard de facto dans les domaines de l'intelligence artificielle, de l'analyse de données et de l'apprentissage profond, une popularité qu'il doit à sa syntaxe concise, **à** sa flexibilité et **à** la richesse de son écosystème logiciel [69].

Afin de structurer efficacement notre pipeline de traitement et de modélisation, nous avons exploité un ensemble de bibliothèques spécialisées et interconnectées :

- **NumPy** : utilisée pour la gestion des tableaux multidimensionnels (tenseurs) et l'exécution d'opérations mathématiques vectorielles de bas niveau, indispensables pour accélérer les calculs.
- **Pandas** : sollicitée pour la manipulation, le nettoyage et la structuration des données tabulaires sous forme de DataFrames, facilitant l'alignement et l'interrogation des variables cliniques ou d'imagerie.
- **Scikit-learn** : cette bibliothèque a été au cœur des étapes de prétraitement (normalisation, standardisation, encodage des variables) et de partitionnement des données (train/test split). Elle a également fourni les métriques d'évaluation nécessaires à la validation des performances (courbes ROC, matrices de confusion, F1-score).
- **TensorFlow et Keras** : TensorFlow a servi de moteur de calcul backend, tandis que Keras, son API de haut niveau, a permis de concevoir, compiler et entraîner les architectures de réseaux de neurones profonds de manière modulaire et efficiente.
- **Matplotlib et Seaborn** : ces bibliothèques ont permis de générer les représentations graphiques des résultats expérimentaux, telles que les courbes d'apprentissage (perte et précision au fil des époques), facilitant l'analyse visuelle du comportement des modèles et la détection d'éventuels phénomènes de surapprentissage.

L'ensemble des phases d'entraînement et d'évaluation a été effectué sur la plateforme cloud Kaggle Notebooks. Cet environnement de développement intégré présente plusieurs avantages pour la recherche en apprentissage profond [70] :

- Kaggle offre un accès gratuit à des accélérateurs matériels hautes performances, essentiels pour le calcul parallélisé requis par les réseaux de neurones complexes. Les expériences ont notamment exploité les GPU NVIDIA Tesla P100 (16 Go de VRAM) et NVIDIA T4 (deux unités), réduisant considérablement les temps d'entraînement.
- La plateforme garantit un environnement d'exécution standardisé et préconfiguré avec les dernières versions des frameworks d'apprentissage profond, éliminant ainsi les conflits de dépendances et assurant une reproductibilité stricte des expériences.

- Enfin, grâce à l'accès direct à de vastes répertoires de jeux de données publics, Kaggle facilite le versionnement du code et le partage collaboratif des pipelines de recherche, consolidant ainsi la rigueur méthodologique de notre démarche.

3. Description des jeux de données

Pour évaluer la robustesse et la capacité de généralisation des approches de deep learning développées dans ce travail, quatre bases de données médicales distinctes ont été exploitées. Ces jeux de données couvrent des modalités d'imagerie variées (IRM et radiographie) et présentent des défis méthodologiques hétérogènes, notamment en termes de classification multiclasse et de déséquilibre inter-classes.

3.1. Jeu de données Alzheimer MRI Preprocessed Dataset

Le premier cas d'étude est dédié au dépistage et à la détermination du stade de la neurodégénérescence. Le jeu de données public Alzheimer MRI Preprocessed Dataset est spécifiquement conçu pour la classification automatique de la maladie d'Alzheimer à partir d'images IRM cérébrales acquises en coupe axiale. Hébergé sur la plateforme Mendeley Data, ce corpus regroupe un total de 6400 images préalablement prétraitées et normalisées à une résolution de 128 × 128 pixels. Les échantillons sont distribués de manière très asymétrique entre quatre classes cliniques : Non Demented (3200 images), Very Mild Demented (2240 images), Mild Demented (896 images) et Moderate Demented (64 images) [71] (voir Figure 3.1 pour des exemples visuels).

Ce jeu de données se caractérise par un déséquilibre de distribution sévère, affichant un rapport d'asymétrie (Imbalance Ratio, IR) de 50 :1. Ce déséquilibre se traduit par une distribution à queue longue largement dominée par les stades sains ou précoces. Il est de nature intrinsèque, reflétant directement la rareté clinique et la difficulté de diagnostic des stades avancés de la pathologie.

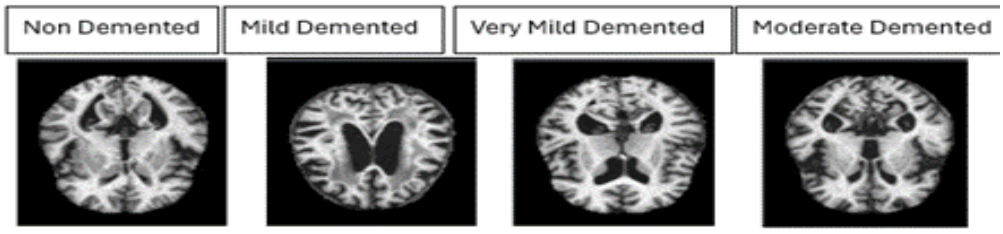


Figure 3.1: Exemples d'images issues de la base de données Alzheimer MRI Preprocessed [71].

3.2. Jeu de données Brain MRI Images

Tout en restant focalisé sur l'imagerie par résonance magnétique cérébrale, le second volet de notre étude passe de la neurodégénérescence à la neuro-oncologie avec le jeu de données Brain MRI Images (Tumor and No Tumor). Ce corpus binaire est constitué de 1421 images IRM réparties entre les classes No Tumor (1262 images) et Tumor (159 images) (voir Figure 3.2).

Cette répartition met en évidence un déséquilibre proportionnel notable, caractérisé par un IR d'environ 8 :1 entre la classe majoritaire et la classe minoritaire. À l'instar du cas précédent, ce déséquilibre est intrinsèque au domaine médical, puisque les examens de contrôle ne présentant aucune anomalie tumorale sont statistiquement plus fréquents dans les flux de routine clinique que les cas avérés de processus expansifs intracrâniens [72].

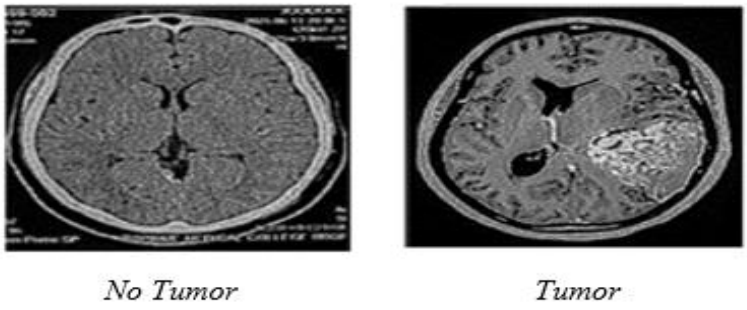


Figure 3.2: Exemples d'images issues de la base de données Brain MRI Images (Tumor and No Tumor) [72].

3.3. Jeu de données Brain Tumor MRI Dataset

Afin d'approfondir l'analyse neuro-oncologique et de confronter nos modèles à un scénario de diagnostic différentiel plus complexe, un troisième jeu de données a été intégré : Brain Tumor MRI Dataset. Contrairement au précédent, ce corpus propose une classification multiclasse détaillée et repose sur un volume de données nettement plus important, réunissant 11148 images IRM cérébrales pondérées en T1 avec injection de produit de contraste.

Les données sont segmentées en quatre catégories distinctes : Glioma (3621 images), Meningioma (2619 images), Pituitary (2574 images) et No Tumor (2334 images). Des exemples illustratifs sont présentés dans la Figure 3.3. Contrairement aux deux premiers jeux de données, la distribution des classes y est légèrement déséquilibrée (IR modéré d'environ 1,55 :1), ce qui permet néanmoins de comparer le comportement des modèles en l'absence de biais de distribution majeur [73]. Ce profil offre un point de comparaison crucial pour évaluer les architectures d'apprentissage dans un contexte proche de l'équilibre.

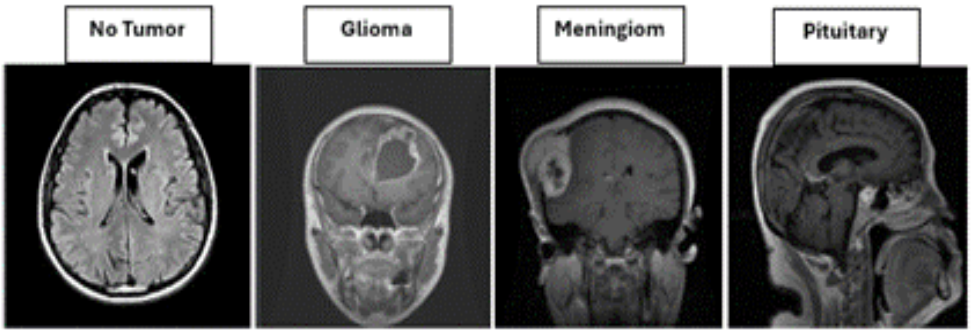


Figure 3.3: Exemples d'images issues de la base de données Brain Tumor MRI Dataset [73].

3.4. Jeu de données Osteoarthritis Prediction

Enfin, pour valider la polyvalence de nos approches au-delà de la sphère cérébrale et de la modalité IRM, le quatrième jeu de données, Osteoarthritis Prediction, introduit une problématique d'imagerie musculosquelettique par rayons X. Ce jeu de données est composé de 2350 images radiographiques du genou, réparties en deux classes pour la prédiction de l'arthrose : Osteoarthritis (1540 images) et Normal (810 images) (voir Figure 3.4).

L'analyse de cette distribution révèle un déséquilibre inversé par rapport aux cas précédents (la classe pathologique devient majoritaire), mais d'ampleur modérée, avec un IR d'environ 1,9 :1. Ce déséquilibre intrinsèque s'explique ici par un biais de recrutement fréquent dans les bases de données de recherche, où les clichés proviennent majoritairement de cohortes de patients déjà symptomatiques ou orientés vers des services orthopédiques spécialisés [74].

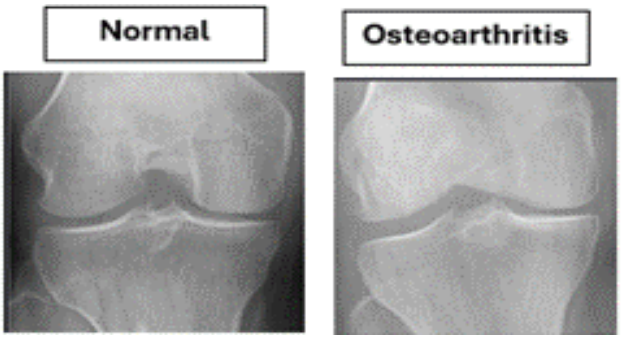


Figure 3.4: Exemples d'images issues de la base de données Osteoarthritis Prediction [74].

4. Approche Méthodologique Proposée

Cette section formalise le protocole expérimental développé pour la classification multiclasses des stades évolutifs de la maladie d'Alzheimer à partir d'images par résonance magnétique (IRM). L'architecture systémique proposée s'organise en trois phases séquentielles, elles-mêmes subdivisées en étapes interdépendantes. Cette architecture a été conçue pour répondre à trois objectifs majeurs : lever le verrou scientifique du déséquilibre clinique, optimiser l'apprentissage des représentations sémantiques de haut niveau et garantir l'interprétabilité des décisions conformément aux exigences médicales.

L'organigramme fonctionnel de la Figure 3.5 synthétise cette dynamique opérationnelle et explicite le séquençage des blocs macroscopiques constitutifs de notre approche, offrant ainsi une vue d'ensemble structurée avant d'aborder le détail de chaque module.

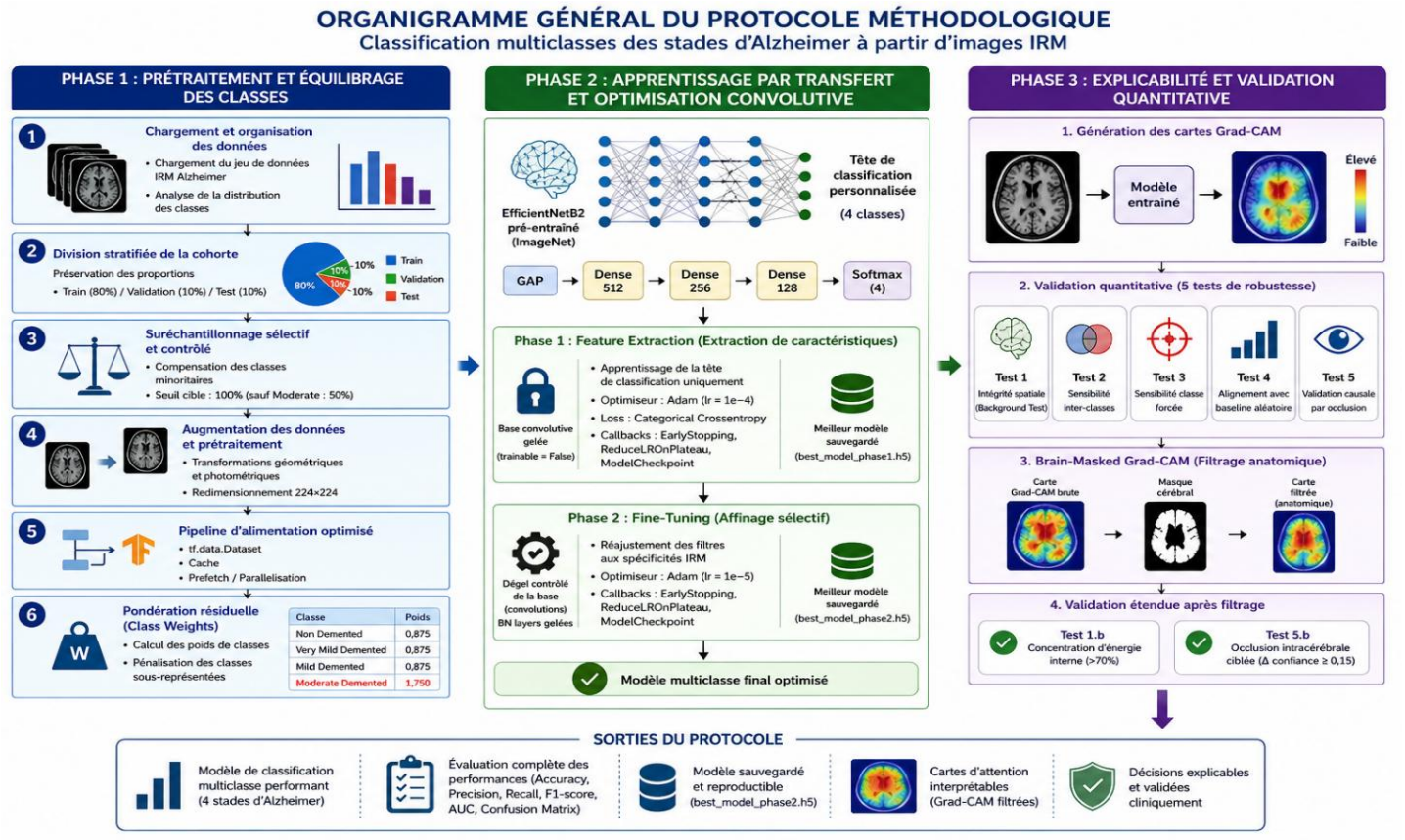


Figure 3.5: Organigramme général du protocole méthodologique développé pour la classification multiclasses des stades d'Alzheimer à partir des image IRM.

4.1. Pipeline de prétraitement et équilibrage des classes

La performance d'un modèle de deep learning en imagerie médicale repose sur deux piliers fondamentaux : la qualité du flux de données qui l'alimente et la maîtrise des biais statistiques inhérents aux cohortes cliniques. Pour concilier ces exigences, la présente section détaille l'intégralité du pipeline de prétraitement, depuis la division initiale des échantillons jusqu'à l'optimisation matérielle du chargement, en passant par des stratégies de rééquilibrage supervisé et de régularisation. Six modules composent cette chaîne de traitement, abordés séquentiellement ci-après.

4.1.1 Chargement et organisation du jeu de données

Le jeu de données utilisé dans cette étude est l'*Alzheimer MRI Preprocessed Dataset*. Les images y sont organisées par classe et chargées dans une structure de données (*dataframe*) qui facilite les opérations de prétraitement et de partitionnement ultérieures.

Une analyse préalable de la distribution des classes est réalisée afin de quantifier le déséquilibre existant. Comme l'illustre la Figure 3.6, la cohorte est marquée par une sous-représentation sévère de la classe *Moderate Demented*, dont le faible effectif est susceptible d'introduire un biais d'apprentissage en faveur des classes majoritaires (*Non Demented* et *Very Mild Demented*). Cette

disparité volumétrique justifie les stratégies de remédiation qui seront déployées dans les sections suivantes (suréchantillonnage et pondération résiduelle).

Enfin, pour se conformer à la topologie du classifieur multiclasse, les étiquettes catégorielles sont converties au format *one-hot encoding*. Cette représentation binaire permet d'exprimer la probabilité d'appartenance à chacune des quatre classes cliniques sous forme d'un vecteur de sortie à quatre dimensions, directement exploitable par la couche softmax du réseau EfficientNetB2.

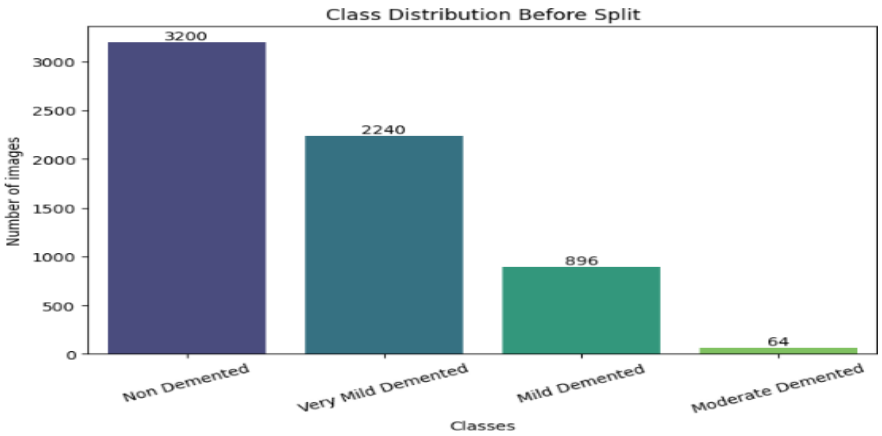


Figure 3.6: Distribution des classes avant rééquilibrage

4.1.2. Division stratifiée et étanche de la cohorte expérimentale

Afin d'assurer la validité interne du protocole d'évaluation et de prévenir tout phénomène de fuite d'information (*data leakage*), la population de données initiales (n = 6400 matrices d'imagerie) est fractionnée en trois sous-ensembles mutuellement exclusifs : entraînement, validation et test.

Dans le contexte d'une distribution de classes fortement asymétrique, une procédure de partitionnement aléatoire conventionnelle induit le risque d'isoler l'intégralité des échantillons d'une pathologie minoritaire au sein d'un unique sous-groupe, compromettant ainsi la convergence ou l'évaluation du modèle. Pour pallier cette dérive macroscopique, l'indice de la distribution probabiliste maximale de chaque instance est extrait par projection d'arguments maximaux ($np.argmax$) pour servir de descripteur de stratification.

La fonction de partitionnement applique une contrainte de préservation homothétique des proportions adaptatives ($stratify = labels_for_stratify$). Cette invariance garantit que la prévalence de chaque stade de la maladie d'Alzheimer (notamment la classe critique *Moderate Demented*) demeure strictement constante à travers les sous-cohortes, configurées selon un ratio de distribution de 80 % pour l'apprentissage, 10 % pour la validation continue et 10 % pour l'évaluation terminale (test). L'initialisation d'un état aléatoire déterministe ($random_state$) pérennise la topologie de ce découpage à des fins de reproductibilité stricte.

La Figure 3.7 détaille les effectifs obtenus pour chaque classe dans chacun des trois sous-ensembles, illustrant concrètement le respect de cette stratification proportionnelle.

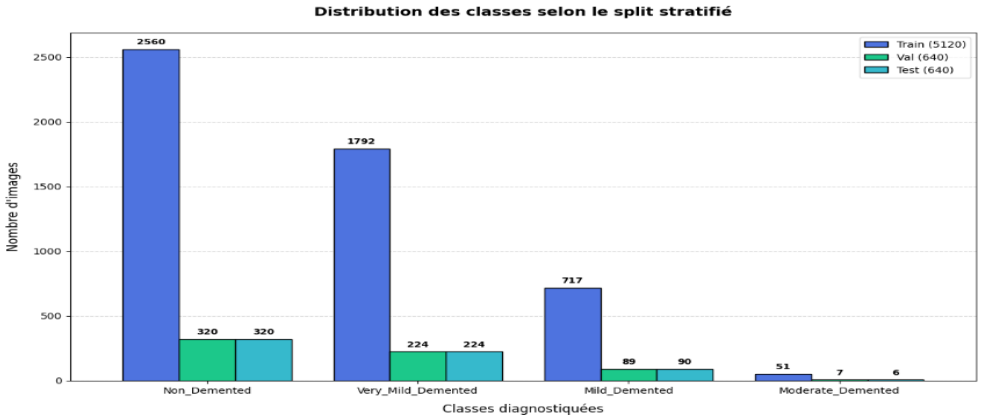


Figure 3.7: Distribution des classes après division stratifiée

4.1.3. Algorithme de Suréchantillonnage Sélectif et Contrôlé

Pour neutraliser les biais d'apprentissage inhérents aux disparités volumétriques inter-classes, le pipeline intègre une routine optimisée de suréchantillonnage minoritaire, formalisée par la fonction personnalisée $balanced_oversample$. Cette méthode résout les instabilités des techniques de duplication globales qui tendent à exacerber le surapprentissage par redondance uniforme.

L'algorithme segmente le jeu d'entraînement en structures de données isolées par classe. Pour chaque entité catégorielle, il évalue l'écart absolu entre l'effectif nominal empirique ($n_{original}$) et un seuil de saturation cible prédéfini (n_{target}). Lorsque le volume d'une classe s'avère déficitaire, une procédure de rééchantillonnage stochastique avec remplacement ($replace = True$) est instanciée de manière exclusive sur cette sous-population jusqu'à combler le déficit numérique exact :

$$\Delta n = n_{\text{target}} - n_{\text{original}} \quad (3.1)$$

Sur le plan de la stratégie clinique, l'effectif de la classe de contrôle sain (*Non Demented*) est érigé en référentiel de saturation pour les stades de démence très légère et légère. Néanmoins, pour préserver la stabilité du gradient face à la classe ultra-minoritaire (Moderate Demented), le plafond cible est délibérément bridé à 50 % de la valeur de référence :

$$n_{target}(Moderate) = [0.5 \times n(Non_Demented)] \quad (3.2)$$

Cette restriction prévient l'apparition d'un surapprentissage critique. Une réplication excessive de la signature d'un effectif initial restreint (n = 64) contraindrait le réseau à converger vers des artefacts de numérisation ou des bruits de fond électroniques singuliers, au détriment de l'assimilation des invariants morphologiques réels de l'atrophie corticale. Les sous-ensembles corrigés sont finalement agrégés et soumis à une permutation aléatoire globale afin de restaurer l'indépendance statistique intra-lot.

La Figure 3.8 illustre cette transformation en comparant les distributions avant et après suréchantillonnage, où l'on constate visuellement la compensation des classes minoritaires.

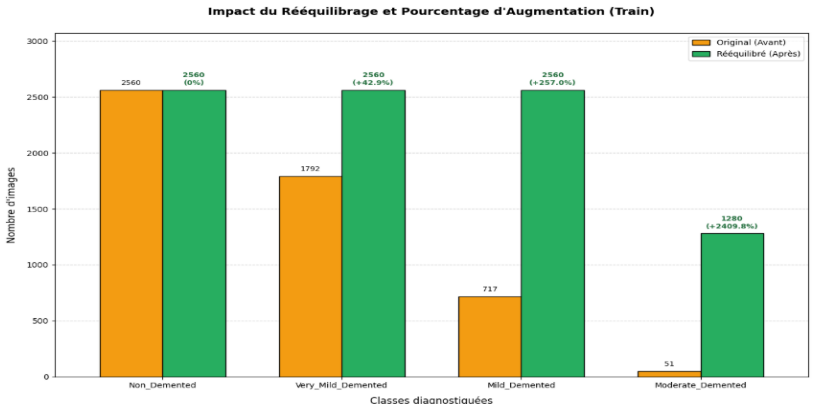


Figure 3.8: Distribution des classes (train) avant et après un Suréchantillonnage Sélectif et Contrôlé

4.1.4. Pipeline d'Augmentation Visuelle et Prétraitement Asynchrone

L'augmentation de données vise à accroître artificiellement la diversité des images d'entraînement afin d'améliorer la généralisation du modèle. Elle repose sur l'application de transformations aléatoires géométriques et photométriques, tout en préservant les caractéristiques anatomiques essentielles. Ces transformations sont générées dynamiquement à chaque époque, uniquement sur le flux d'apprentissage, et leurs paramètres sont volontairement limités afin de conserver le réalisme médical des images. Cette technique complète les méthodes de rééquilibrage des classes et renforce la robustesse du modèle face aux variations inter-instrumentales.

La Figure 3.9 illustre l'effet de ces transformations sur un échantillon source, montrant la diversité des variations introduites tout en maintenant l'intégrité des structures cérébrales.

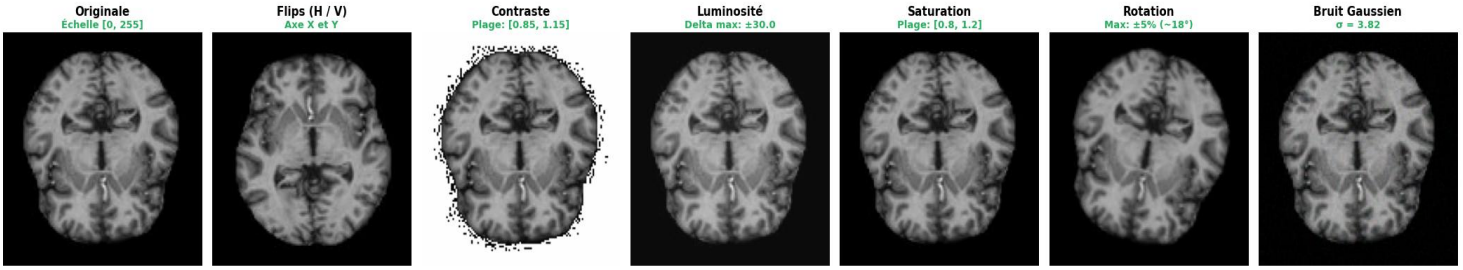


Figure 3.9: Exemples d'images après application Augmentation.

Le formalisme d'alimentation du réseau abandonne les architectures de génération de données séquentielles classiques au profit d'un graphe de flux optimisé sous l'API $tf.data.Dataset$. Ce pipeline orchestre de manière asynchrone trois primitives de transformation géométrique et stochastique :

- *Acquisition et Uniformisation Spatiale (load_image)* : Opère la lecture binaire des fichiers d'imagerie depuis l'espace de stockage et applique un décodage matriciel standardisé sur trois canaux (format pseudo-RVB indispensable à la topologie d'*EfficientNetB2*). Une interpolation bilinéaire est ensuite exécutée pour contraindre les dimensions spatiales à la résolution d'entrée requise de 224×224 pixels, garantissant ainsi une uniformité dimensionnelle stricte en amont du modèle.
- *Augmentation Stochastique Multimodale (augment_image)* : Implémentée de manière restrictive sur le flux d'apprentissage ($is_training = True$), cette phase applique un ensemble d'altérations stochastiques mimant les variations physiques et instrumentales de l'environnement clinique : des inversions spéculaires horizontales et verticales tirant parti de la symétrie bilatérale des structures encéphaliques ; des perturbations contrôlées du contraste, de la luminance et de la saturation afin de simuler la variabilité inter-constructeurs des scanners IRM ; des transformations affines de rotation et de zoom microscopique (via la sous-couche séquentielle $data_augmentation$) traduisant les inclinaisons d'assiette ou les légers défauts de centrage du patient dans le champ magnétique ; et l'injection d'un bruit blanc gaussien additif d'écart-type calibré (σ=0.015×255) contraignant les filtres convolutifs à s'abstraire du bruit de fond haute fréquence pour se focaliser sur la topographie des structures macroscopiques.
- *Normalisation Statistique Déportée (apply_preprocess)* : La fonction mathématique de mise à l'échelle spécifique à l'architecture $preprocess_input$ est rigoureusement positionnée en aval des altérations visuelles. Ce séquençage garantit que les données issues des opérations d'augmentation soient centrées et réduites selon la distribution originale d'ImageNet avant l'induction dans les couches de convolution profondes.

4.1.5. Optimisation Matérielle et Parallélisation du Flux

La fonction d'encapsulation $create_dataset$ configure la distribution des charges de calcul via le protocole d'optimisation dynamique de TensorFlow :

- *Parallélisation Multithread (num_parallel_calls = tf.data.AUTOTUNE)* : Les opérations de décodage, d'augmentation stochastique et de normalisation statistique sont distribuées de manière concurrente sur les différents cœurs du processeur (CPU), parallélisant la préparation des tenseurs.
- *Mécanisme d'Anticipation Asynchrone* (.prefetch(tf.data.AUTOTUNE)) : Cette technique supprime le goulot d'étranglement matériel traditionnel lié aux accès disques. Pendant que l'accélérateur matériel (GPU) traite le lot de tenseurs courant (B_t) pour la rétropropagation du

gradient, le processeur prépare et stocke en mémoire vive les lots ultérieurs (B_{t+1}, B_{t+2}). Cette architecture garantit un taux d'occupation optimal du GPU et une minimisation du temps de latence inter-époques.

4.1.6. Pondération Résiduelle Compensatoire

En dernier lieu, le protocole introduit une double sécurité algorithmique via le calcul d'un vecteur de pénalisation adaptative. Bien que le jeu de données d'apprentissage ait subi une correction volumétrique par suréchantillonnage, le bridage préventif appliqué à la classe Moderate Demented (fixé à 50 % de la classe de référence) maintient une asymétrie résiduelle mineure. Pour compenser ce déséquilibre persistant, le script évalue les fréquences de la nouvelle distribution à l'aide de la fonction $compute_class_weight$ de Scikit-learn, qui formalise un dictionnaire de poids d'ajustement. Les valeurs obtenues sont de 0,875 pour les classes Non Demented, Very Mild Demented et Mild Demented, et de 1,75 pour la classe Moderate Demented. Cette répartition attribue une pénalisation deux fois plus importante aux erreurs commises sur cette pathologie rare. Injectés au sein de la fonction de coût lors de l'appel de $.fit()$, ces coefficients rectificateurs modulent dynamiquement l'amplitude de la perte, forçant le gradient à corriger plus vigoureusement ses prédictions pour la catégorie critique. Ainsi, cette approche agit non plus sur la distribution des échantillons (comme l'oversampling), mais directement sur le signal d'erreur rétropropagé, parachevant l'équilibrage mathématique global du système d'apprentissage.

La combinaison du suréchantillonnage sélectif et de ces poids de classes constitue une stratégie redondante et robuste pour traiter le déséquilibre des effectifs cliniques. Cette double correction volumétrique et différentielle est encore renforcée, dans une phase ultérieure, par l'utilisation d'un apprentissage en deux phases (two-phase learning), qui sera détaillé dans la section suivante et qui permet d'affiner la régularisation du modèle en ajustant dynamiquement le taux d'apprentissage, améliorant ainsi la généralisation sur les classes sous-représentées et la stabilité globale de la convergence.

4.2. Stratégie d'apprentissage par transfert et optimisation convolutive

Le prétraitement des données et les stratégies de rééquilibrage étant désormais établis, la présente section expose l'architecture neuronale et le protocole d'optimisation mis en œuvre pour la classification des stades de la maladie d'Alzheimer. L'approche adoptée repose sur le paradigme du transfert d'apprentissage (Transfer Learning), en s'appuyant sur l'architecture EfficientNetB2 pré-entraînée sur le référentiel ImageNet. Afin d'adapter ce réseau aux spécificités de la classification multiclasse, les couches denses terminales d'origine ont été remplacées par une tête de classification personnalisée (Fully Connected Head).

4.2.1. Architecture de la tête de classification

La tête personnalisée est organisée selon un agencement séquentiel visant à condenser progressivement les descripteurs de haut niveau extraits par le bloc convolutif. Une opération de Global Average Pooling 2D aplatit d'abord les cartes de caractéristiques bidimensionnelles en un vecteur unidimensionnel, limitant ainsi l'explosion combinatoire du nombre de paramètres. S'ensuit un premier bloc dense de 512 neurones, accompagné d'une normalisation par lots, d'une activation ReLU et d'un Dropout de taux 0,4. Ce premier étage est relayé par un second bloc dense de 256 neurones, régulé par BatchNormalization, activation ReLU et Dropout de taux 0,3, qui stabilise la réduction progressive de la dimensionnalité. Un troisième bloc dense de 128 neurones, associé aux mêmes couches de régulation mais avec un Dropout réduit à 0,2, affine davantage les représentations. Enfin, la couche de sortie compose 4 neurones activés par une fonction Softmax, formalisant la distribution probabiliste des prédictions sur l'ensemble des classes diagnostiques.

4.2.2. Stratégie d'entraînement en deux phases

L'optimisation des poids synaptiques est conduite selon une stratégie séquentielle en deux phases distinctes et complémentaires.

1) Phase 1 : Extraction de caractéristiques (Feature Extraction)

Au cours de cette première phase, les filtres convolutifs du modèle de base EfficientNetB2 sont intégralement gelés ($trainable_base = False$). Le réseau joue le rôle d'extracteur de descripteurs figés, l'apprentissage étant restreint aux seuls paramètres de la tête de classification nouvellement greffée. Cette isolation préserve les représentations génériques acquises sur ImageNet contre les perturbations de gradient caractéristiques du début de l'entraînement. Le modèle est compilé avec la fonction de coût d'entropie croisée catégorielle (Categorical Crossentropy) et optimisé via l'algorithme Adam à un taux d'apprentissage initial de 10^{-4}.

La convergence est supervisée par trois rappels dynamiques (callbacks) :

- EarlyStopping : interrompt l'entraînement si la perte de validation ne s'améliore pas pendant 8 époques consécutives, avec restauration automatique des meilleurs poids.
- ReduceLROnPlateau : réduit le taux d'apprentissage par un facteur de 0,3 après 3 époques de stagnation, jusqu'à un seuil plancher de 10^{-7}.
- ModelCheckpoint : sauvegarde la configuration optimale du réseau dans le fichier best_model_phase1.h5 sur la base de la perte de validation minimale.

Le dictionnaire de pondération des classes ($class_weight = class_weights_dict$), calculé en amont, est transmis à la fonction $.fit()$ afin de contraindre l'optimiseur à accorder une importance proportionnelle aux classes cliniquement sous-représentées lors de la rétropropagation du gradient.

2) Phase 2 : Affinage sélectif (Fine-Tuning)

La seconde phase procède à une spécialisation en profondeur du modèle par dégel partiel et contrôlé de l'architecture. Pour prévenir l'effondrement des représentations associé à un dégel massif, le protocole applique une libération sélective des couches selon deux règles strictes. D'une part, toutes les sous-couches de type *BatchNormalization* restent explicitement gelées ($layer.trainable = False$), cette précaution étant critique en contexte d'affinage car elle empêche la déstabilisation des statistiques de moyenne et de variance globales face à des sous-lots d'entraînement de taille restreinte ($BATCH_SIZE = 16$). D'autre part, les filtres restants du modèle de base sont libérés ($layer.trainable = True$), permettant au réseau de réajuster ses champs récepteurs aux singularités morphologiques des structures cérébrales et de l'atrophie corticale propres aux images IRM. Le modèle est ensuite recompilé avec l'optimiseur Adam à un taux d'apprentissage très faible de 10^{-5}, garantissant des modifications chirurgicales des poids synaptiques sans altérer les filtres génériques stabilisés en phase 1.

Les critères de tolérance des rappels sont ajustés à la dynamique lente de l'affinage : la patience de l'EarlyStopping est portée à 10 époques, celle du ReduceLROnPlateau à 4 époques avec un seuil plancher abaissé à 10^{-8}, et la meilleure configuration est sauvegardée dans le fichier best_model_phase2.h5.

Cette stratégie en deux phases, combinant extraction figée puis affinage contrôlé des poids convolutifs, sera évaluée dans la section suivante au regard des métriques de performance et de la capacité de généralisation du modèle sur la cohorte de test.

4.3. Phase d'Explicabilité Post-Hoc et de Validation Quantitative

Au-delà de la performance prédictive brute, l'acceptabilité clinique d'un modèle d'apprentissage profond exige que ses décisions soient compréhensibles et justifiables. Cette section aborde ainsi la question cruciale de l'explicabilité des prédictions, en détaillant la mise en œuvre de l'algorithme Grad-CAM ainsi que le protocole rigoureux de validation quantitative qui l'accompagne.

4.3.1. Principe et implémentation de Grad-CAM

Le formalisme Grad-CAM constitue la brique centrale de notre approche d'interprétation visuelle. Afin de pallier l'opacité décisionnelle des architectures connexionnistes profondes, une composante d'explicabilité visuelle est intégrée au pipeline via l'algorithme Grad-CAM. Celui-ci localise les régions d'intérêt cérébrales ayant déterminé le score de confiance du réseau en isolant les cartes de caractéristiques de la dernière couche convolutive d'EfficientNetB2. Les gradients de la fonction de score, rétropropagés jusqu'à cette couche, sont pondérés par une opération de Global Average Pooling, puis combinés linéairement avec les activations spatiales. Un filtrage de type

ReLU élimine ensuite les contributions négatives, ne conservant que les structures morphologiques augmentant positivement le score de la classe cible.

4.3.2. Cadre de validation quantitative

Si les cartes de saillance offrent une représentation intuitive de l'attention du réseau, elles ne sauraient être prises pour argent comptant sans une évaluation objective de leur fiabilité. C'est pourquoi cinq protocoles de vérification quantitative ont été définis afin de sonder la robustesse et la sincérité des explications produites. La littérature récente signale en effet le risque de « faux positifs visuels » avec Grad-CAM, où les cartes thermiques se focalisent sur des artefacts sans lien avec les mécanismes internes du réseau. Pour certifier la fidélité des explications produites, ces cinq protocoles ont été implémentés.

1) **Test 1 — Intégrité spatiale (*Background Test*).**

Un masque binaire distingue le parenchyme cérébral du fond, défini par les pixels d'intensité inférieure à 15. Le ratio d'activation entre le cerveau et le fond est calculé ; la validation exige un ratio supérieur ou égal à 2,0, garantissant la concentration de la saillance sur la matière cérébrale.

2) **Test 2 — Sensibilité inter-classes (Class Sensitivity)**

Pour une même image IRM, la heatmap est calculée successivement pour chacune des quatre classes. Une matrice de corrélation de Spearman est établie entre les cartes aplaties. Une corrélation moyenne hors-diagonale faible (inférieure à 0,85) confirme que le réseau modifie bien sa configuration attentionnelle selon la classe explorée.

3) **Test 3 — Sensibilité de la classe forcée (Forced Class Test)**

Sur une image correctement classée, l'algorithme est contraint d'expliquer une classe erronée. L'écart absolu moyen entre la carte authentique et la carte altérée est calculé ; un score supérieur à 0,05 confirme la réactivité de Grad-CAM aux variations de l'index de classe.

4) **Test 4 — Alignement avec la ligne de base aléatoire (Random Baseline)**

Inspiré des travaux d'Adebayo et al [55], ce test compare les cartes thermiques du modèle entraîné à celles d'une réplique architecturale identique initialisée avec des poids aléatoires. Une divergence significative, caractérisée par un coefficient de corrélation inférieur à 0,50, prouve que les explications sont le produit de l'apprentissage et non de l'architecture seule.

5) **Test 5 — Validation causale par occlusion (Occlusion Test)**

Les pixels de forte saillance, correspondant aux 15 % les plus actifs, sont occultés par un masque opaque appliqué à l'image IRM d'entrée. Une chute de la confiance prédictive d'au moins 0,15 confirme que la région identifiée constitue un pivot informationnel réel pour le modèle.

4.3.3. Filtrage anatomique — Brain-Masked Grad-CAM

L'application des cinq protocoles précédents met en évidence une limite récurrente : les activations de Grad-CAM ne se limitent pas aux seules structures cérébrales et persistent sur des artefacts d'acquisition, des structures osseuses ou des bords de matrice, cliniquement non pertinentes. Afin de recentrer l'analyse sur les zones cliniquement pertinentes, un filtrage spatial anatomique a posteriori a été introduit, dont les principes et les protocoles de contrôle sont détaillés ci-après.

1) Extraction du parenchyme cérébral

Un seuillage adaptatif d'Otsu est appliqué sur le canal de luminance. L'intégrité topologique du masque est ensuite assurée par une fermeture morphologique (noyau elliptique d'un diamètre de 15 pixels) comblant les discontinuités intracérébrales, suivie d'une érosion fine (diamètre de 5 pixels) éliminant les structures crâniennes résiduelles. Une analyse en composantes connexes (selon une connectivité de 8 voisins) isole enfin le plus grand objet topologique, correspondant au cerveau.

2) Projection et renormalisation

La carte de saillance brute est multipliée pixel à pixel par le masque cérébral booléen, puis renormalisée afin de préserver sa dynamique visuelle, en divisant chaque pixel par l'intensité maximale de la carte filtrée (augmentée d'une faible constante pour éviter une division par zéro).

Sur cette base filtrée, deux protocoles étendus sont adossés au pipeline :

- **Test 1.b — Concentration d'énergie interne.** Le ratio d'énergie est mesuré comme la part de l'intensité totale des activations située à l'intérieur du parenchyme, avant et après filtrage. Un verdict Improved est prononcé si la carte filtrée concentre plus de 70 % de son énergie dans les structures cérébrales, avec un gain positif par rapport à la carte brute.
- **Test 5.b — Occlusion intracérébrale ciblée.** Ce protocole corrige le biais du Test 5 classique, où les pixels les plus actifs pouvaient se situer hors du cerveau. L'occlusion est ici restreinte au 20e percentile supérieur des activations parmi les seuls pixels validés par le masque cérébral. De plus, le remplissage par pixels noirs est remplacé par un Neutral Fill , une substitution par la valeur moyenne intra-canal de l'image originale, afin d'éviter l'introduction de fausses discontinuités de contraste. La validation causale est confirmée par une chute de confiance d'au moins 0,15.

Fort de ce cadre méthodologique complet, nous procédons désormais à l'évaluation expérimentale du modèle, dont les résultats sont présentés dans la section suivante, depuis les courbes d'apprentissage jusqu'aux performances finales sur le jeu de test.

5. Résultats expérimentaux et discussion

Ce chapitre expose et discute les performances empiriques de notre approche à travers une série de quatre expérimentations progressives, allant de la validation quantitative à l'évaluation qualitative clinique. Pour introduire cette analyse, la première phase expérimentale présentée ci-dessous se concentre sur l'évaluation comparative de plusieurs architectures de réseaux de neurones profonds afin de définir la base convolutive optimale de notre pipeline.

5.1. Expérimentation 1 : Sélection de l'Architecture de Référence

Cette première phase expérimentale vise à identifier l'architecture de réseau de neurones convolutifs la plus performante pour servir de modèle de référence (*backbone*). Une étude comparative standardisée a ainsi été menée sur sept architectures de pointe, pré-entraînées sur la base de données ImageNet et largement plébiscitées dans la littérature pour leurs performances en vision par ordinateur et en imagerie médicale : MobileNetV2, DenseNet121, InceptionV3, Xception, ResNet50, VGG19 et EfficientNetB2. Pour cette comparaison, la base de données disponible a été partitionnée en trois sous-ensembles distincts : 70 % pour l'entraînement, 15 % pour la validation et 15 % pour le test.

Afin de garantir l'équité de cette confrontation, le bloc de classification terminal (*top classifier*) a été unifié pour l'ensemble des architectures. Après avoir extrait les bases convolutionnelles de chaque réseau (en désactivant la tête d'origine via $include_top = False$), nous y avons greffé la structure de classification standardisée décrite en sous-section 4.2.1. Durant cette phase, les filtres convolutifs du modèle de base ont été intégralement gelés ($trainable = False$) : le réseau joue ainsi le rôle d'extracteur de descripteurs figés, l'apprentissage étant restreint aux seuls paramètres de la tête de classification adossée.

Enfin, pour assurer la parfaite reproductibilité de cette comparaison, tous les modèles ont été entraînés dans un environnement logiciel strictement identique, avec une configuration d'hyperparamètres d'optimisation uniforme, dont le détail est consigné dans le Tableau 3.1.

Hyperparamètre	**Valeur / Option**
Taille du lot (Batch size)	32
Nombre maximum d'époques	50
Taux d'apprentissage (Learning rate)	0.0001
Algorithme d'optimisation	Adam
Fonction de perte (Loss function)	Categorical_crossentropy

Tableau 3.1: Configuration des hyperparamètres pour la phase comparative.

Les performances globales de chaque architecture ont été rigoureusement évaluées sur le jeu de données de test indépendant à l'aide de trois métriques principales : l'accuracy, le rappel et le F1-score. Les résultats quantitatifs sont consignés dans le Tableau 3.2.

Modèle	**Accuracy (%)**	**Rappel (%)**	**F1-score (%)**
MobileNetV2	64,17	41,90	42,92
InceptionV3	66,67	43,13	43,64
DenseNet121	76,56	63,31	68,08
Xception	77,29	58,92	63,12
ResNet50	80,63	70,37	75,63
VGG19	83,65	72,56	**78,28**
EfficientNetB2	**83,92**	**73,30**	78,14

Tableau 3.2: Performances comparatives des différentes architectures CNN

L'analyse des résultats consignés dans le Tableau 3.2 révèle des disparités significatives entre les performances des architectures évaluées, permettant de segmenter les comportements des modèles en trois catégories distinctes. En premier lieu, les architectures légères et modulaires, représentées ici par MobileNetV2 et InceptionV3, affichent les performances les plus faibles du panel, avec une accuracy globale de 66,67 % et un rappel critique inférieur à 44,00 %. Une telle faiblesse démontre que ces structures, bien qu'optimisées pour la sobriété computationnelle et le déploiement embarqué, s'avèrent structurellement insuffisantes pour capturer la complexité morphologique ainsi que les variations géométriques subtiles des structures anatomiques cérébrales aux premiers stades de la pathologie.

En second lieu, les architectures de complexité intermédiaire, à savoir DenseNet121, Xception et ResNet50, témoignent d'une nette progression algorithmique en franchissant le seuil des 75,00 % d'accuracy. Au sein de ce groupe, le réseau ResNet50 se distingue particulièrement en atteignant un F1-score de 75,63 %. Ce résultat confirme empiriquement l'apport des connexions résiduelles (skip connections), qui permettent d'atténuer le phénomène d'évanouissement du gradient et favorisent ainsi une extraction de caractéristiques de haut niveau à la fois plus stable et plus profonde.

En dernier lieu, l'examen des configurations de pointe met en évidence un duel remarquable entre VGG19 et EfficientNetB2. Le modèle VGG19 délivre de très solides performances globales, affichant une accuracy de 83,65 % et un F1-score de 78,28 %, ce qui s'explique par sa profondeur et la régularité géométrique de ses filtres convolutifs de petite taille. Néanmoins, l'architecture EfficientNetB2 surpasse l'ensemble des réseaux testés en s'adjugeant la meilleure accuracy du benchmark à 83,92 % et, surtout, le meilleur rappel à 73,30 %. Dans un contexte d'aide au diagnostic

clinique, la maximisation du rappel constitue une priorité absolue, car elle permet de minimiser le taux de faux négatifs, écartant ainsi le risque majeur de non-détection d'un patient atteint de démence.

Afin de mener une inspection plus granulaire et d'évaluer de manière détaillée la distribution des erreurs de prédiction pour chaque classe nosologique, les matrices de confusion associées aux sept architectures évaluées sont mises en confrontation à la Figure 3.10. L'examen de ces matrices met en exergue les difficultés communes rencontrées par les extracteurs de caractéristiques figés, en particulier lors de la distinction entre les classes adjacentes ou intermédiaires telles que Very Mild Demented et Mild Demented. Pour les architectures les moins performantes, comme MobileNetV2 et InceptionV3, on observe un taux de confusion alarmant : un volume important de patients Very Mild Demented (respectivement 91 et 82 patients) est classé à tort comme sains (Non Demented). De plus, InceptionV3 et MobileNetV2 échouent massivement à identifier la classe Moderate Demented, n'enregistrant strictement aucune prédiction correcte (0 patient) pour cette catégorie critique, ce qui s'avère rédhibitoire dans un contexte de dépistage clinique.

Bien que les modèles intermédiaires tels que DenseNet121, Xception et ResNet50 améliorent significativement la détection des patients sains (franchissant le cap des 400 vrais positifs), ils maintiennent un biais de classification notable aux frontières des phases de transition de la maladie. La classe Mild Demented reste largement sous-diagnostiquée, ResNet50 et DenseNet121 classant par exemple une proportion substantielle de ces patients (54 et 50) dans la catégorie inférieure Very Mild Demented. Le modèle VGG19 se distingue par une nette amélioration de la sensibilité globale, totalisant 474 prédictions exactes pour la classe Non Demented et parvenant à sécuriser 76 diagnostics corrects pour le stade Mild Demented. Cependant, ce modèle affiche un coût computationnel disproportionné. Intégrant plus de 143 millions de paramètres exploitables, la lourdeur structurelle de VGG19 induit une latence d'inférence élevée et accroît le risque de surapprentissage sur des cohortes de données cliniques restreintes, limitant ainsi sa transférabilité pratique sur un outil d'aide au diagnostic en temps réel.

À l'inverse, l'architecture EfficientNetB2 s'impose comme la solution la plus équilibrée et la plus robuste du benchmark, comme l'illustre de manière isolée sa matrice de confusion (Figure 3.11). Ce modèle parvient à maximiser le nombre de vrais positifs pour la classe saine avec 450 classifications exactes, tout en réduisant drastiquement les faux négatifs cliniquement dangereux. En effet, pour la phase hautement critique et subtile qu'est le stade Very Mild Demented, EfficientNetB2 valide correctement 262 cas, surpassant VGG19 (250) et Xception (230). De surcroît, il obtient le meilleur score de sensibilité pour le stade Mild Demented avec 81 prédictions exactes. Cette efficacité algorithmique remarquable est obtenue avec une fraction du volume de paramètres de VGG19, une

prouesse directement imputable à son mécanisme d'ajustement composite, qui optimise conjointement la profondeur, la largeur et la résolution spatiale des filtres convolutifs.

En vertu de cette supériorité statistique sur le rappel, de sa robustesse intrinsèque face au surapprentissage et de sa remarquable efficacité computationnelle, EfficientNetB2 est officiellement retenu comme l'architecture de référence (backbone) pour la suite de nos travaux.

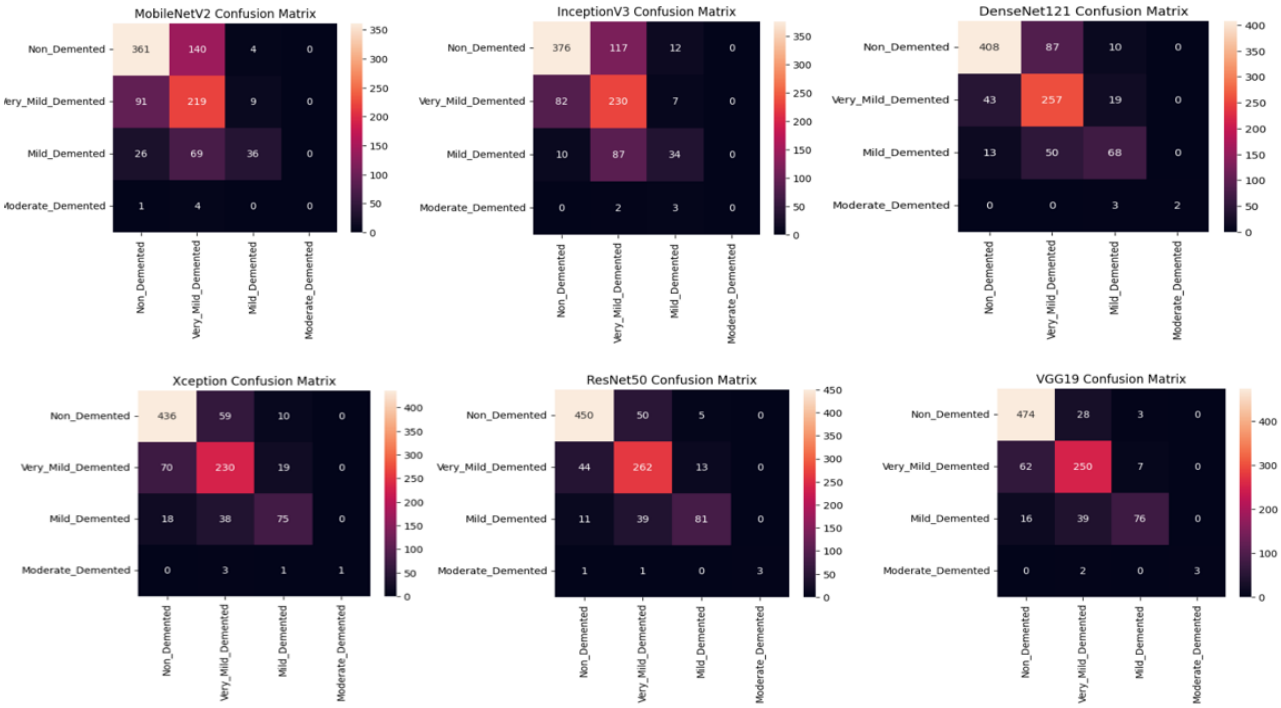


Figure 3.10: Matrices de confusion des différents modèles testés

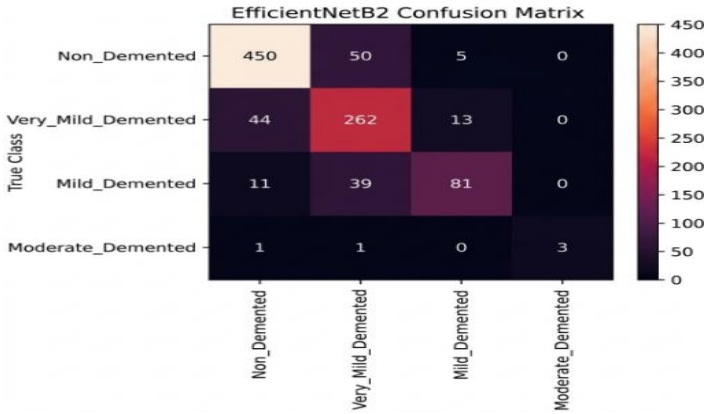


Figure 3.11: Matrices de confusion d'EfficientNetB2

Pour consolider ce choix, l'examen comparatif des profils d'évolution de l'accuracy et de la fonction de perte (loss) sur 50 époques, mis en confrontation à la Figure 3.12, révèle des trajectoires dynamiques radicalement opposées entre les deux architectures dominantes du benchmark. D'une part, le modèle VGG19 manifeste une convergence initiale rapide mais subit une divergence

structurelle critique dès la vingtième époque : tandis que ses métriques d'entraînement progressent linéairement jusqu'à frôler les 92 %, son accuracy de validation stagne et oscille de manière erratique entre 81% et 84%. Cet écart de généralisation croissant, également visible sur sa fonction de perte, trahit un phénomène marqué de surapprentissage induit par son sur paramétrage massif (143 millions de paramètres) qui pousse le réseau à mémoriser le bruit statistique des données d'entraînement. D'autre part, l'architecture EfficientNetB2 fait preuve d'une robustesse et d'une stabilité asymptotique remarquables, ses trajectoires d'apprentissage et de validation évoluant en parfaite synergie pour atteindre un minimum de perte optimal proche de 0,20. Grâce à son mécanisme d'ajustement composite qui équilibre la profondeur et la résolution spatiale des filtres, ce réseau limite drastiquement la redondance informationnelle et se focalise exclusivement sur des invariants morphologiques réels. En somme, EfficientNetB2 s'impose comme le cœur décisionnel optimal pour notre pipeline d'aide au diagnostic, puisqu'il égale la capacité prédictive de VGG19 tout en garantissant une convergence stable et une immunité manifeste face au surapprentissage.

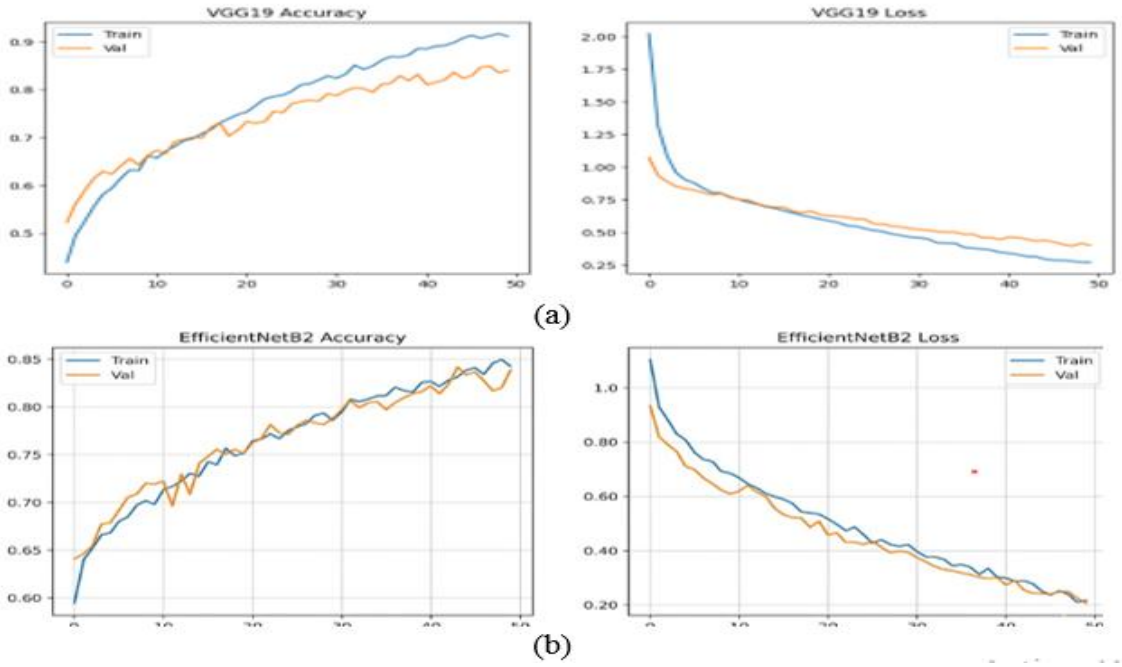


Figure 3.12: Confrontation des courbes d'apprentissage (accuracy et perte) entre les architectures VGG19 et EfficientNetB2 au cours des phases d'entraînement et de validation.

5.2. Expérimentation 2 : Évaluation des stratégies de gestion du déséquilibre des classes

Cette seconde phase expérimentale évalue l'efficacité de différentes approches de remédiation face au déséquilibre sévère des classes dans le cadre du système d'aide à la décision clinique. Pour assurer une comparabilité rigoureuse, toutes les stratégies ont été mises en œuvre sur une architecture unique, EfficientNetB2, dont les hyperparamètres ont été maintenus constants depuis la phase de sélection. Le protocole d'entraînement a été uniformisé par l'augmentation à 150 époques, couplée à un arrêt précoce (*Early Stopping)* pour prévenir le surapprentissage. Les

performances quantitatives (accuracy, rappel, F1-score, précision, spécificité et G-mean) sont regroupées dans le Tableau 3.3.

Stratégie évaluée	Accuracy (%)	Précision (%)	Rappel (%)	F1-score (%)	Spécificité (%)	G-mean (%)
Sans équilibrage	79,22	84,19	74,76	78,39	91.00	81.73
Augmentation de données	87,97	90,81	88,65	89,72	87,29	87,97
Suréchantillonnage aléatoire	87,91	89,84	89,92	89,88	85,90	87,89
Suréchantillonnage ciblé + pondération des classes	86,77	88,34	89,24	88,79	84,30	86,73
Focal loss	88,44	83,92	75,95	79,74	90,93	83,10
Classification hiérarchique	76,77	74,04	63,47	68,35	90,07	75,61
Classification hiérarchique + pondération des classes	87,92	89,32	89,28	89,30	86,56	87,91
Notre approche	**96,72**	**97,46**	**97,63**	**97,54**	**98,62**	**98,12**

Tableau 3.3: Synthèse comparative des performances des stratégies d'équilibrage avec l'architecture EfficientNetB2.

La première stratégie évaluée repose sur l'augmentation de données géométriques et texturales. Les transformations appliquées ne sont pas arbitraires : elles simulent des variations physiques réalistes observées lors des examens IRM, telles que les fluctuations de contraste inter-constructeurs, les micromouvements du patient ou le bruit thermique des bobines. Cette approche enrichit l'espace des caractéristiques tout en préservant l'intégrité des structures anatomiques, favorisant ainsi la généralisation du modèle. Comme l'indique le Tableau 3.3, cette stratégie améliore nettement les performances initiales du système : l'accuracy globale passe de 79,22 % (configuration sans équilibrage) à 87,97 %, tandis que le rappel bondit de 74,76 % à 88,65 %. Cette progression se confirme de manière homogène sur les autres métriques clés, le F1-score s'élevant à 89,72 % et le G-mean atteignant 87,97 %. Néanmoins, on observe une légère contre-performance sur la spécificité, qui fléchit de 91,00 % à 87,29 %, traduisant une propension marginale du modèle à générer de faux positifs sous l'effet des transformations géométriques. L'ensemble de données source ayant déjà subi un prétraitement standardisé, seules ces variantes réalistes ont été retenues, excluant les altérations artificielles risquant de créer de faux marqueurs d'atrophie.

S'inspirant de ces premiers gains, deux approches de réorganisation distributionnelle ont été comparées : le suréchantillonnage aléatoire et l'association suréchantillonnage ciblé avec pondération des classes, toutes deux conçues pour maximiser la sensibilité du réseau aux classes minoritaires. Si le suréchantillonnage aléatoire affiche des métriques élevées, avec une accuracy de 87,91 %, un rappel de 89,92 % et un F1-score de 89,88 %, sa spécificité recule à 85,90 % (contre 91,00 % pour la configuration native), ce qui s'explique par sa tendance à exacerber le risque de

surapprentissage en raison de la duplication d'échantillons identiques dans les couches profondes. La Figure 3.13 illustre la topologie de la distribution des classes avant et après l'application de ces méthodes, mettant en évidence la réduction de la variance inter-classe.

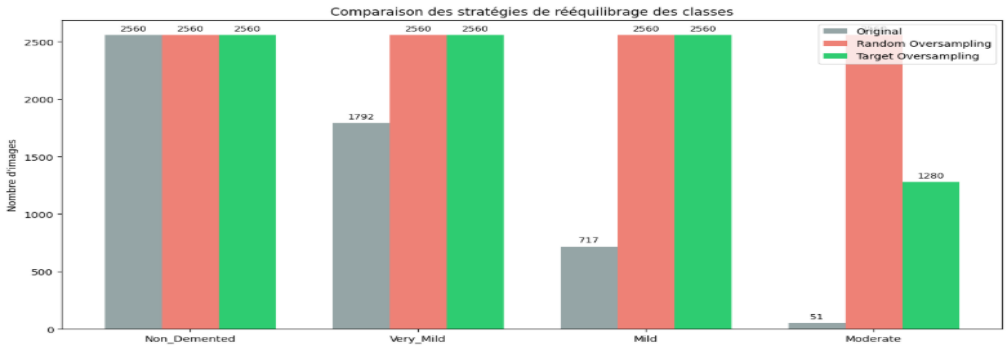


Figure 3.13: Comparaison des profils de distribution des classes nosologiques avant et après application des méthodes de suréchantillonnage.

L'analyse des trajectoires d'optimisation montre que la combinaison du suréchantillonnage ciblé et de la pondération des classes offre une meilleure stabilité et une robustesse accrue face au surapprentissage, malgré des performances brutes légèrement inférieures, marquées par une accuracy de 86,77%, un rappel de 89,24% et un F1-score de 88,79%. Sur le plan de la sélectivité, cette approche affiche une spécificité de 84,30%, ce qui correspond à son G-mean associé de 86,73%. Cette stratégie a néanmoins été privilégiée pour le modèle de base, car elle garantit le meilleur compromis entre pouvoir discriminant et régularisation des poids, un critère essentiel pour le diagnostic clinique. L'analyse des courbes de convergence (Figure 3.14) confirme cette différence de comportement : le suréchantillonnage aléatoire atteint rapidement des performances élevées, mais les oscillations marquées sur le jeu de validation trahissent un début de surapprentissage. À l'inverse, l'approche ciblée pondérée présente une progression plus régulière et des écarts réduits entre les phases d'entraînement et de validation, gage d'une meilleure généralisation.

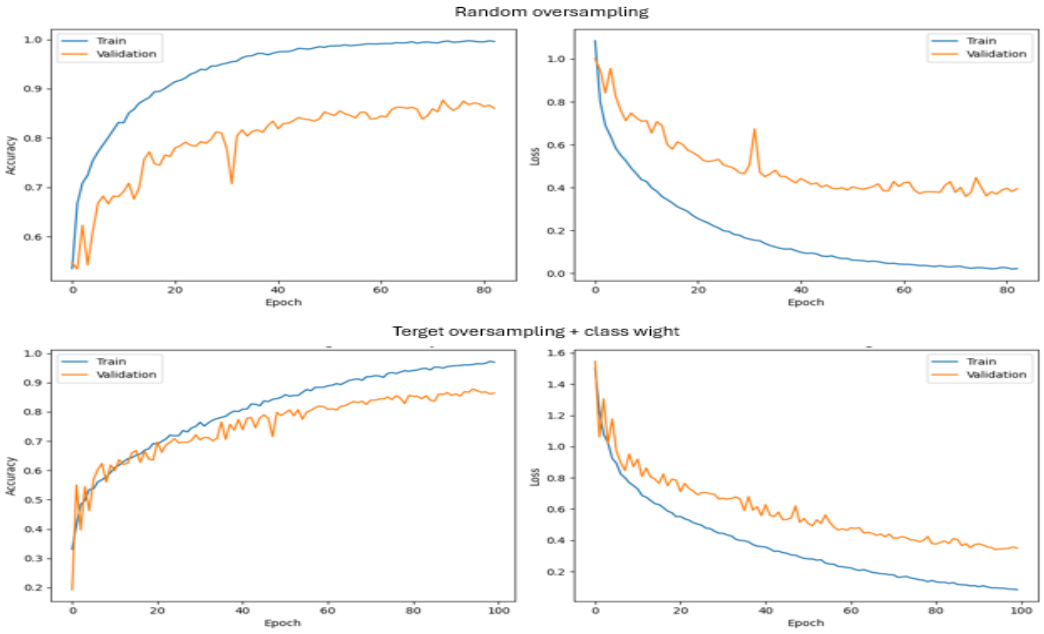


Figure 3.14: Évolution comparative de l'accuracy et de la perte pour les stratégies de suréchantillonnage au cours des époques.

Par ailleurs, la mise en œuvre de la perte focale (Focal Loss) apporte un gain notable en accuracy (88,44 %) par rapport à la configuration native, tout en préservant une spécificité remarquablement élevée de 90,93 %, très proche des 91,00 % initiaux. Cependant, son efficacité globale reste mitigée si on la compare aux méthodes de suréchantillonnage. En effet, son F1-score plafonne à 79,74 %, son rappel s'effondre à 75,95 % et son G-mean se limite à 83,10 %. Ce comportement asymétrique (une excellente spécificité mais un faible rappel) s'explique par le fait qu'une part de ces exemples dits « difficiles » correspond à des artefacts ou à des zones d'ambiguïté aux frontières des stades de démence ; la Focal Loss contraint alors le modèle à une prudence excessive, limitant les faux positifs au détriment de la détection des cas réels.

Quant à la classification hiérarchique, ses performances sont particulièrement décevantes lorsqu'elle est utilisée isolément, chutant à une accuracy de 76,77 %, un rappel de 63,47 % et un F1-score de 68,35 %. Bien qu'elle maintienne une spécificité solide à 90,07 %, son G-mean s'effondre à 75,61 %. Cette détérioration résulte de la propagation d'erreurs en cascade : une mauvaise orientation lors de la première décision binaire (sain contre pathologique) compromet irrémédiablement la caractérisation fine du stade de démence. Cependant, l'adjonction de pondérations au sein de cette structure (Hierarchical Classification + Class Weights) redresse spectaculairement la situation : l'accuracy remonte à 87,92 %, le rappel s'établit à 89,28 % et le F1-score atteint 89,30 %, tandis que la spécificité se stabilise à 86,56 % pour un G-mean équilibré de 87,91 %. Cela confirme qu'une décomposition hiérarchique est pertinente à la condition d'être couplée à un contrôle rigoureux de l'apprentissage des sous-branches.

Enfin, notre approche repose sur une stratégie hybride en quatre volets : le suréchantillonnage ciblé (TOS), l'augmentation des données (DA), la pondération adaptative des classes (CW) et l'apprentissage en deux phases (Two-PL). Elle surpasse de loin l'intégralité des configurations testées en maximisant simultanément toutes les dimensions d'évaluation : elle atteint une accuracy de 96,72 %, un rappel de 97,63 %, un F1-score de 97,54 %, et surtout une spécificité souveraine de 98,62 %, matérialisée par un G-mean d'excellence de 98,12 %. Ce succès empirique valide notre hypothèse centrale : seule une synergie entre la restructuration temporelle de l'apprentissage et le rééquilibrage géométrique permet d'éliminer le verrou du déséquilibre des classes sans aucun compromis, garantissant à la fois un rappel clinique maximal et un taux de faux positifs quasi nul.

5.2.1. Étude expérimentale et analyse multi-niveaux de l'approche hybride proposée

Notre approche hybride ayant obtenu les meilleures performances lors des évaluations comparatives, une analyse multi-niveau a été menée pour caractériser la dynamique fine du modèle au-delà des seules métriques quantitatives globales. Sur le plan de l'optimisation, le paradigme de l'apprentissage en deux phases repose sur un découplage temporel des gradients. Lors de la première phase, seule la couche de classification finale est entraînée, tandis que les couches d'extraction de caractéristiques restent gelées. Cette configuration initiale permet une convergence rapide et une projection optimale du vecteur de caractéristiques vers l'espace des classes nosologiques, sans altérer les représentations spatiales génériques préalablement acquises par le réseau.

Lors de la seconde phase, un ajustement fin et progressif est appliqué à l'ensemble du réseau, à l'exception des couches de normalisation par lots, maintenues gelées. Ce choix méthodologique est crucial en imagerie médicale : il fige les statistiques internes du modèle (moyenne et variance mobiles calculées lors du pré-entraînement), évitant ainsi les dérives de distribution et les instabilités stochastiques lors de l'adaptation aux spécificités texturales des coupes IRM. Cette double régularisation structurelle explique la stabilité de la convergence et la décroissance régulière de la perte observées au fil des époques.

Le rapport de classification confirme la robustesse de notre approche : le modèle consolidé atteint une accuracy de 96,72 % et un F1-score de 97,54 %, témoignant d'une excellente discrimination sur l'ensemble des classes pathologiques. Cette performance globale remarquable est directement visualisée à travers les matrices de confusion exposées à la Figure 3.15. Sur la matrice (b), qui matérialise le comportement macroscopique de notre approche hybride finalisée, on constate que la quasi-totalité des prédictions se concentre rigoureusement sur la diagonale principale, traduisant un taux d'erreur inter-classe extrêmement faible par rapport à la configuration initiale (a).

Le résultat le plus probant concerne la classe critique Moderate Demented, qui ne compte qu'un effectif restreint de 6 patients au sein du jeu de test. Classiquement, une telle rareté provoque un

effondrement du rappel ou une inflation des faux positifs ; notre modèle hybride y obtient pourtant une précision et un rappel parfaits de 100 %, ne concédant aucune fausse orientation. Cette performance atteste de l'efficacité conjointe du suréchantillonnage ciblé, de l'augmentation des données et des pondérations asymétriques des coûts, qui orientent fermement la fonction de perte vers les invariants morphologiques des classes rares, évitant ainsi leur éclipse par la classe majoritaire saine.

De même, les stades intermédiaires Very Mild Demented et Mild Demented, dont les frontières anatomiques sont particulièrement complexes à délimiter en neuroimagerie en raison du continuum d'atrophie voxel à voxel, conservent des scores de sensibilité remarquables, oscillant entre 96 % et 97 %. Ces résultats confirment que notre pipeline ne se contente pas de mémoriser des corrélations superficielles ou des artefacts de numérisation, mais extrait des descripteurs phénotypiques fins et cliniquement discriminants, en parfaite adéquation avec les signatures anatomocliniques de la maladie d'Alzheimer.

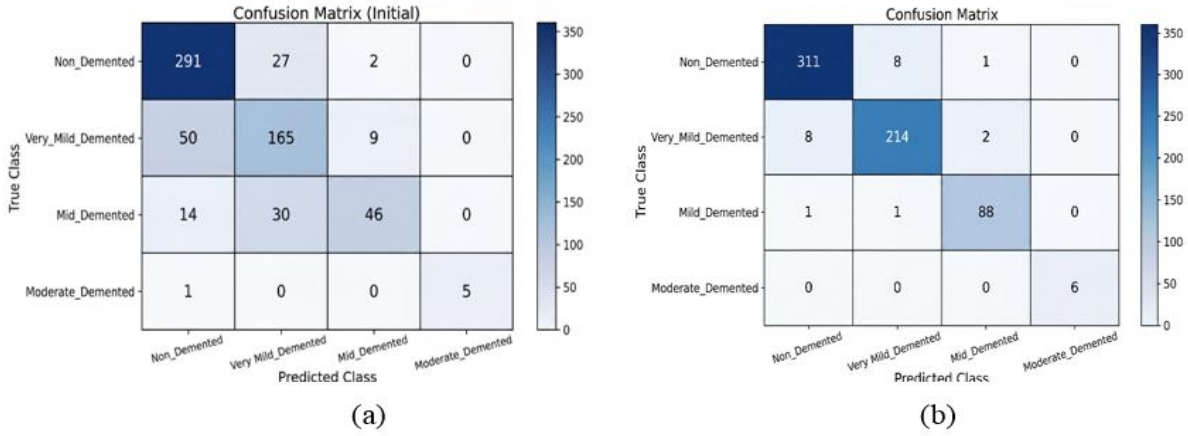


Figure 3.15: Matrices de confusion obtenues sur le jeu de test : (a) configuration initiale avant traitement du déséquilibre et (b) après optimisation par notre approche hybride unifiée.

5.2.2. Analyse quantitative des courbes de convergence et validation clinique du modèle optimal

L'examen des trajectoires d'optimisation de l'approche hybride unifiée révèle une dynamique d'apprentissage cohérente avec le protocole en deux phases adopté. Durant la Phase 1 (extraction de caractéristiques), un écart de généralisation modéré est observable : l'accuracy d'entraînement progresse régulièrement jusqu'à environ 0,63, tandis que la courbe de validation, plus instable, plafonne autour de 0,55. Cet écart, attendu à ce stade puisque seule la tête de classification est entraînée, ne traduit pas un surapprentissage mais les limites inhérentes à un apprentissage restreint aux paramètres terminaux du réseau.

La Phase 2 (fine-tuning) marque en revanche un changement net de comportement : les courbes d'accuracy et de perte, pour l'entraînement et la validation, se rapprochent étroitement et progressent de façon quasi parallèle jusqu'à converger respectivement autour de 0,96-0,98 (accuracy) et 0,10-0,15 (perte). Cette convergence resserrée, obtenue malgré le dégel partiel des couches convolutives, confirme l'efficacité du protocole de gel/dégel progressif : la transition entre les deux phases ne provoque aucune instabilité ni saut brutal de gradient. Cette dynamique est clairement visualisée sur la Figure 3.16, qui confirme la stabilité de la convergence avec un écart minimal entre les courbes d'entraînement et de validation dès le début de la seconde phase.

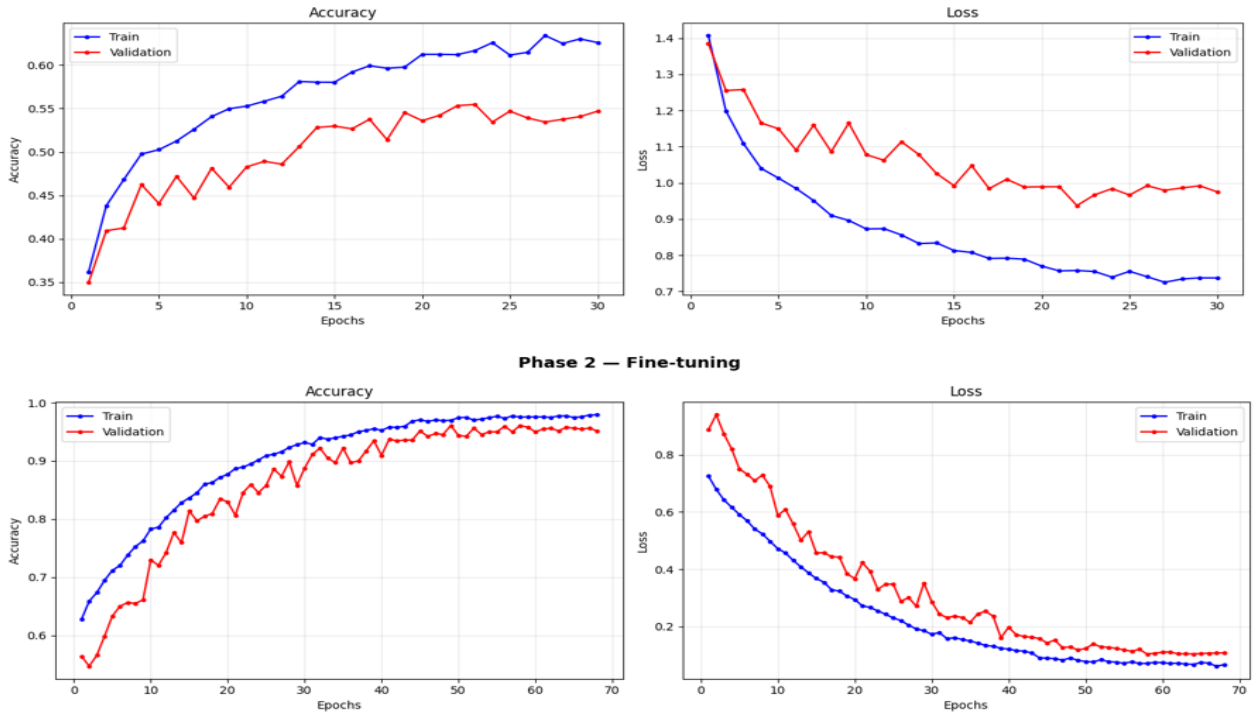


Figure 3.16: Évolution des courbes d'apprentissage (accuracy et perte) au cours des deux phases du protocole Two-Phase Learning pour le modèle hybride optimal.

Ce constat de stabilité est renforcé par l'analyse des courbes ROC et PRC présentée dans la Figure 3.17. La courbe ROC démontre une capacité de séparation quasi parfaite entre les quatre classes, avec des AUC variant de 0,9950 à 1,0000 et une moyenne micro-pondérée de 0,9985. Le modèle atteint une AUC de 0,9968 pour la classe Non Demented, 0,9950 pour Very Mild Demented, 0,9955 pour Mild Demented et 1,0000 pour la classe minoritaire Moderate Demented. Cette performance confirme la puissance discriminative du système, indépendamment du seuil de décision retenu. Elle garantit, en pratique clinique, une identification fine des structures saines face aux différents stades d'atrophie, minimisant ainsi le risque de faux diagnostics. La courbe PRC corrobore la résilience du réseau face au déséquilibre distributionnel sévère, en se concentrant sur l'interaction directe entre faux positifs et faux négatifs, sans être influencée par la surreprésentation

de la classe majoritaire. Les indices de précision moyenne sont tout aussi élevés : 0,9984 pour les sujets sains, 0,9930 pour le stade très léger, 0,9971 pour le stade léger et 1,0000 pour le stade modéré, soit une moyenne globale de 0,9957. Ces scores indiquent que le modèle préserve simultanément une précision élevée et un rappel satisfaisant, y compris pour les sous-groupes cliniques les plus rares, suggérant que la pondération des classes a efficacement compensé le déficit d'échantillons sans dégrader la spécificité globale du système.

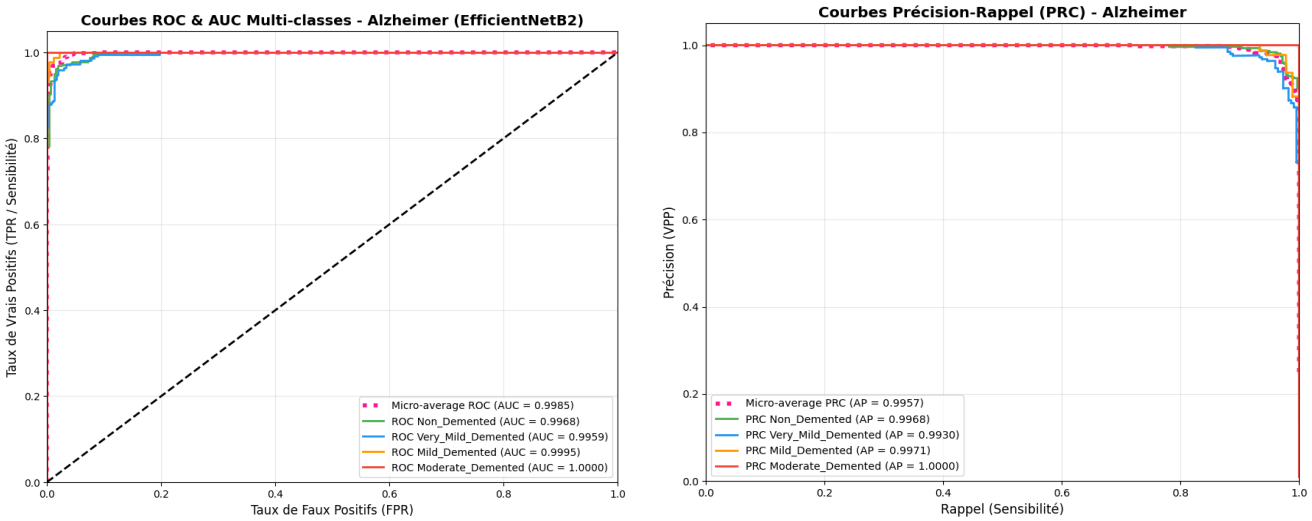


Figure 3.17: Courbes ROC et Précision-Rappel du modèle optimal pour les quatre classes de démence.

Pour concrétiser ces performances en termes d'impact clinique réel, la Figure 3.18 présente l'analyse par courbe de décision (DCA). Cette méthodologie estime le bénéfice net standardisé en confrontant le système à deux stratégies de référence : l'intervention systématique (tous les patients considérés comme atteints) et l'abstention totale (tous considérés comme sains). Les profils DCA obtenus pour chaque stade montrent que la courbe du modèle hybride se maintient au-dessus des deux lignes de référence sur une large plage de seuils de probabilité, y compris pour la classe minoritaire Moderate Demented où le bénéfice net reste positif alors que les deux stratégies extrêmes deviennent rapidement non avantageuses. Cela suggère que l'intégration de cet outil dans un flux de travail clinique pourrait maximiser le bénéfice net tout en réduisant conjointement les faux positifs et les faux négatifs, pour l'ensemble des quatre stades diagnostiques considérés.

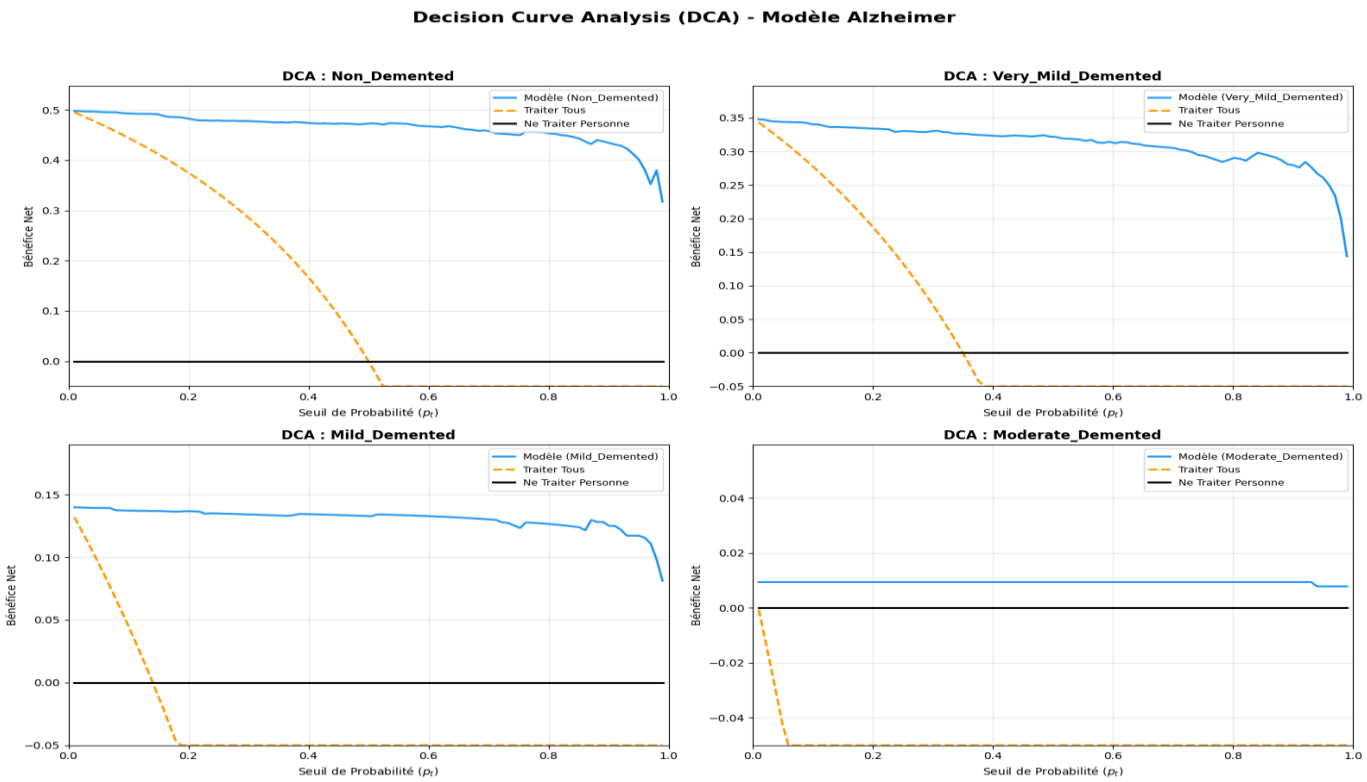


Figure 3.18: Analyse des courbes de décision (DCA) évaluant le bénéfice net clinique du modèle pour chaque stade diagnostique.

Les courbes d'apprentissage confirment la stabilité et l'absence de surapprentissage ; les courbes ROC/PRC valident la puissance discriminative et la robustesse au déséquilibre ; les courbes DCA établissent l'utilité clinique concrète du système. Ensemble, ces résultats positionnent notre approche comme un outil fiable et opérationnel pour le diagnostic assisté de la maladie d'Alzheimer.

5.2.3. Analyse visuelle des résultats de prédiction du modèle proposé

Afin de valider qualitativement l'efficacité décisionnelle du pipeline proposé, une inférence directe a été menée sur un ensemble d'échantillons représentatifs issus du jeu de test indépendant. Les résultats, illustrés par la Figure 3.19, montrent une concordance parfaite entre les diagnostics prédits par le réseau et les étiquettes réelles pour les quatre stades de démence : Non Demented, Very Mild Demented, Mild Demented et Moderate Demented. Cette inspection visuelle confirme que le modèle discrimine efficacement les variations morphologiques subtiles de l'atrophie cérébrale, tout en validant la reproductibilité des performances macroscopiques détaillées précédemment.

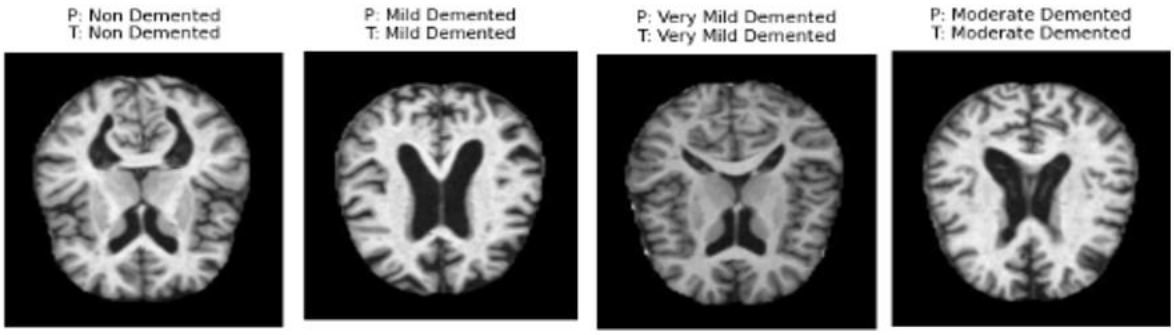


Figure 3.19: Exemples d'inférences correctes générées par notre modèle hybride sur des coupes IRM réelles du jeu de test, pour les quatre stades de démence.

5.2.4. Comparaison de nos résultats avec l'état de l'art

Afin de situer et d'évaluer rigoureusement la contribution de notre approche hybride unifiée, une confrontation systématique a été menée par rapport à plusieurs travaux récents de la littérature (2022-2026) dédiés à la classification multi-classes des stades de la maladie d'Alzheimer, dans des contextes de fort déséquilibre distributionnel. Le Tableau 3.4 synthétise cette étude comparative en présentant, pour chaque méthode, l'architecture de réseau de neurones profonds utilisée, les stratégies de compensation du déséquilibre adoptées ainsi que les métriques de performance clés.

Auteurs	**Architecture**	**Gestion déséquilibre**	**Accuracy (%)**	**F1-score (%)**	**Rappel (%)**
Askarizade (2025) [67]	EfficientNetV2S	WGAN-GP	98,1	98,6	98,4
Andrea Loddo et al. (2022) [64]	Deep-Ensemble	Augmentation	98,24	96,3	97,3
Lazli et al. (2026) [68]	Ensemble DL	MAMIR	95,0	90,1	90,7
Ryspayeva (2025) [66]	ResNet18	DGAN-WP-TL	85,7	85,4	85,7
Tufail et al. (2024) [65]	InceptionV3 + CNN	GAN	68,0	56	55
Notre méthode	**EfficientNetB2 + *Two-PL***	**TOS+ DA + CW**	**96,72**	**97,54**	**97,63**

Tableau 3.4Comparaison de notre méthode avec l'état de l'art.

L'analyse comparative des données de l'état de l'art révèle que notre pipeline se positionne parmi les solutions les plus compétitives et équilibrées de la littérature. Certes, les travaux d'Askarizade (2025) [67] et d'Andrea Loddo et al. (2022) [64] affichent des valeurs d'accuracy ou de F1-score légèrement supérieures (respectivement 98,1 % et 98,24 %). Toutefois, ces performances maximales reposent sur des architectures d'ensembles profonds ou des modèles génératifs complexes à gradients pénalisés (WGAN-GP). Ces approches, non seulement extrêmement lourdes sur le plan

computationnel, sont également sujettes aux instabilités d'entraînement chroniques (comme l'effondrement de mode) et complexes à déployer en routine clinique hospitalière.

À l'inverse, l'approche de Tufail et al. (2024) [65], exploitant un GAN conventionnel, montre des limites structurelles claires avec une accuracy qui chute à 68,0 %, mettant en exergue la difficulté critique à générer des images de synthèse médicales parfaitement fidèles sur le plan anatomique.

Dans ce panorama, notre méthode se distingue par une synergie algorithmique hautement efficiente : l'intégration d'un extracteur de caractéristiques léger (EfficientNetB2) optimisé par un protocole d'apprentissage en deux phases, combiné de manière transparente au suréchantillonnage ciblé, à l'augmentation de données et à la pondération adaptative des classes. Ce pipeline obtient un excellent compromis clinique, avec une accuracy de 96,72 %, un F1-score de 97,54 % et un rappel exceptionnel de 97,63 %. Ces résultats démontrent rigoureusement qu'il est possible d'égaler, voire de surpasser, la puissance discriminative de modèles nettement plus lourds sans faire le sacrifice de l'interprétabilité, de la sobriété computationnelle ou de la stabilité mathématique. Cette convergence de performances confirme la viabilité et la robustesse de notre pipeline pour la prise en charge clinique des données de neuroimagerie fortement asymétriques.

5.3. Expérimentation 3 : Analyse de l’interprétabilité et validation de la robustesse des explications par Grad-CAM

Bien que les indicateurs quantitatifs macroscopiques démontrent une excellente convergence, la validation d’un système d’aide au diagnostic en routine clinique ne peut se limiter à la seule performance numérique. Afin d’instaurer une relation de confiance avec le corps médical et de répondre aux exigences réglementaires de transparence, il est nécessaire d’élucider le processus décisionnel du réseau. Dans ce cadre, la méthode de cartographie d’activation par gradient pondéré (Grad-CAM) a été mise en œuvre. Cette technique d’explicabilité post-hoc permet d’atténuer l’effet « boîte noire » des architectures convolutives profondes en projetant les gradients de la dernière couche convolutionnelle sous forme de cartes thermiques bidimensionnelles.

La Figure 3.20 illustre l’application de Grad-CAM sur des coupes IRM caractéristiques des quatre classes nosologiques de la cohorte. L’échelle chromatique adoptée reflète l’intensité de la contribution : les zones hyperactivées en rouge/orange désignent les foyers d’intérêt ayant guidé la classification du réseau, tandis que les régions en bleu témoignent d’une neutralité décisionnelle.

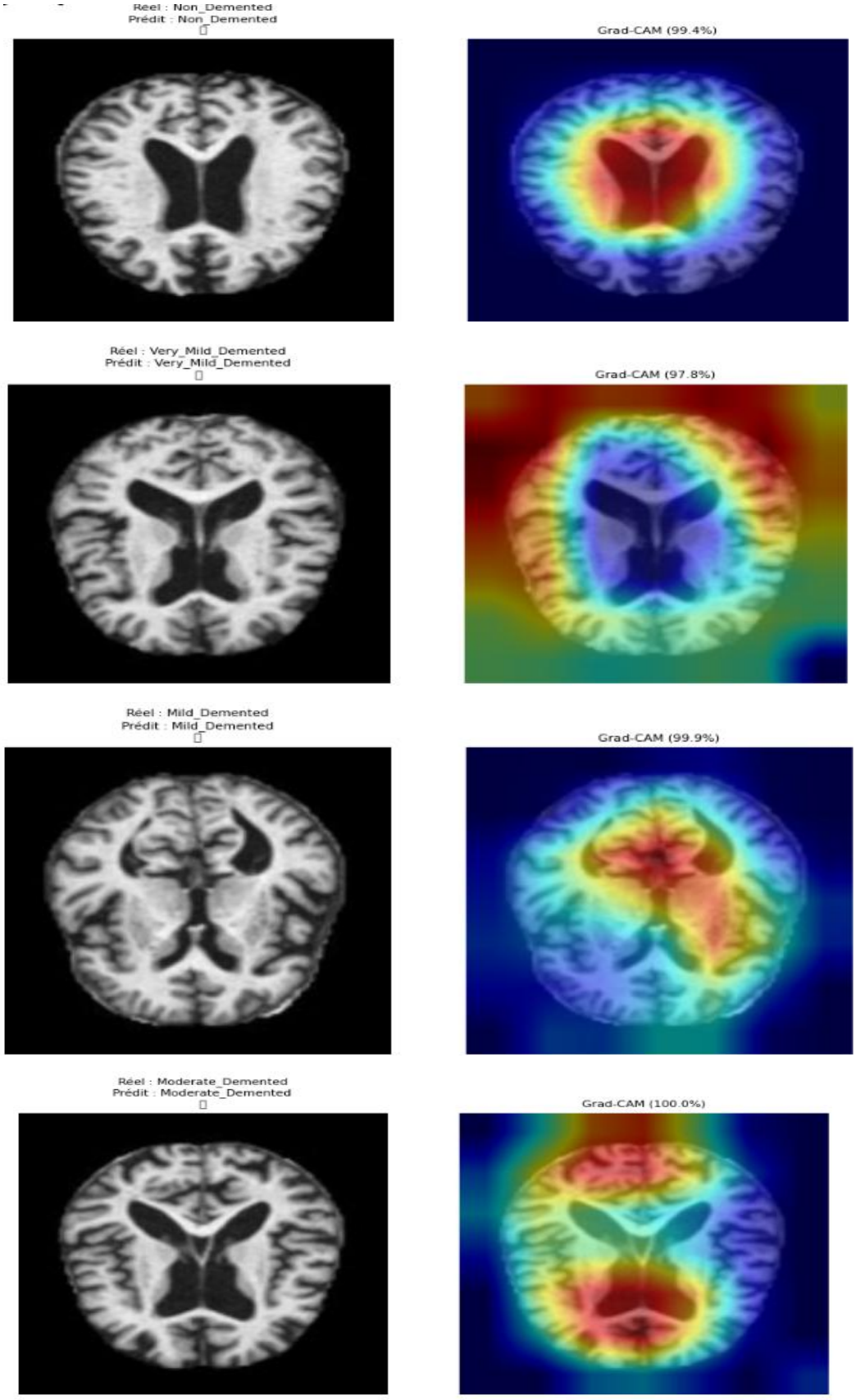


Figure 3.20: Visualisation de l'interprétabilité par cartes d'activation Grad CAM pour les quatre stades cliniques de la maladie d'Alzheimer.

En résumé, l'analyse d'interprétabilité par Grad-CAM confirme la cohérence clinique du modèle tout au long de la progression de la maladie d'Alzheimer. Au stade sain (non dément), l'attention du réseau se porte de manière symétrique sur les structures anatomiques de référence, reflétant l'intégrité des régions typiquement affectées par la pathologie. À mesure que la maladie évolue vers les stades légers et modérés, la focalisation se déplace progressivement vers les zones d'atrophie corticale et la dilatation ventriculaire, deux biomarqueurs radiologiques caractéristiques de la neurodégénérescence.

Cette dynamique décisionnelle suggère que le modèle ne se base pas sur des motifs aléatoires ou des biais de l'arrière-plan, mais qu'il extrait des biomarqueurs anatomocliniques pertinents, conformes aux connaissances médicales actuelles. *Cette pertinence diagnostique a été qualitativement confrontée à l'expertise du* ***Dr Sanhadji*** *Sara, médecin spécialiste en radiologie à l'établissement hospitalier* ***d'El Abiodh Sidi Cheikh****, qui a corroboré l'adéquation des zones d'activation avec les signes radiologiques évocateurs de la pathologie.*

5.3.1. Validation de la robustesse de Grad-CAM : Tests de cohérence élémentaires

Malgré l'obtention d'une validation clinique qualitative par un médecin expert, nous avons poursuivi l'évaluation en ajoutant quatre tests de cohérence (*sanity checks*) inspirés des travaux d'Adebayo et al. (2018) [55]. L'objectif est de démontrer scientifiquement la fiabilité des cartes Grad-CAM selon trois niveaux d'analyse : mathématique, causal et anatomique.

Ces quatre tests permettent de vérifier la robustesse mathématique de la méthode Grad-CAM et de confirmer que ses cartes d'explicabilité dépendent réellement des paramètres du modèle entraîné ainsi que de la classe prédite, et non d'un simple traitement de contours ou d'un artefact visuel. Le Tableau 3.5 présente le résumé des résultats obtenus, confirmant la validation de la méthode pour chacun de ces tests fondamentaux.

Test	Métrique utilisée	Valeur obtenue	Seuil de tolérance	Résultat
Test 1: Background	Distance de Frobenius entre la heatmap d'origine et celle générée sur fond noir	9,91	> 2,0	Succès
Test 2: Class Sensitivity	Corrélation de Spearman entre les heatmaps de deux classes différentes	0,045	< 0,85	Succès
Test 3: Forced Class	Distance de Jensen-Shannon entre les distributions d'activation des classes réelle et forcée	0,4166	> 0,05	Succès
Test 4: Random Baseline	Similarité cosinus entre la heatmap du modèle entraîné et celle d'un modèle aux poids aléatoires	0,1218	< 0,5	Succès

Tableau 3.5:Résultats des tests de cohérence élémentaires pour la validation de Grad CAM.

L'ensemble de ces résultats confirme que les cartes d'activation produites par Grad-CAM ne sont ni des artéfacts de traitement d'image ni des corrélations fortuites, mais bien des indicateurs fidèles de la stratégie décisionnelle du réseau. Le modèle satisfait ainsi aux exigences fondamentales de reproductibilité et de transparence requises pour une éventuelle intégration clinique.

Pour mieux comprendre ces résultats, examinons en détail le fonctionnement et l'objectif de chacun de ces quatre tests, en commençant par l'évaluation du fond de l'image.

Le **Test 1 (Background Test)** confirme la robustesse de la méthode Grad-CAM, avec un ratio moyen cerveau/fond de 9,91, ce qui est largement supérieur au seuil de tolérance fixé (>2). Ce résultat mathématique indique que les activations de Grad-CAM se concentrent de manière stricte sur les structures cérébrales d'intérêt en ignorant le fond de l'image. Cela valide la fiabilité de nos cartes de saillance, prouvant qu'elles ciblent le tissu biologique et non des artefacts extérieurs (Figure 3.21).

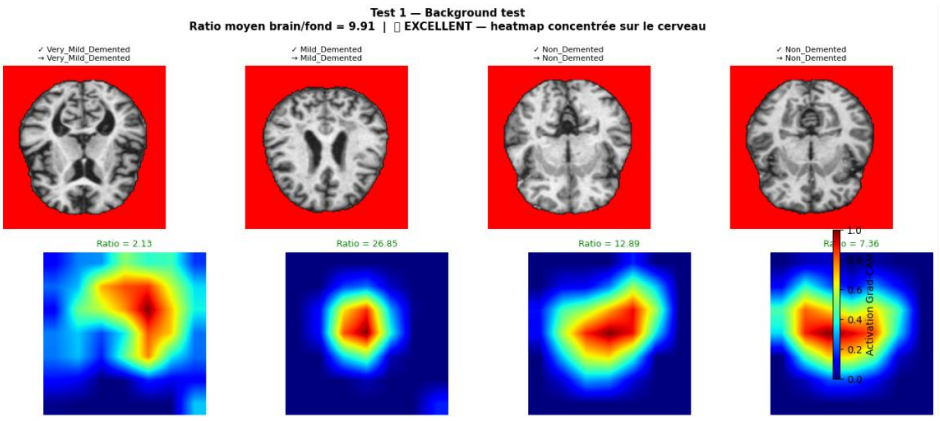


Figure 3.21:Évaluation de la robustesse par le Test 1.

Le **Test 2 (Class Sensitivity)** montre une très faible corrélation entre les différentes cartes Grad-CAM, avec une valeur de seulement 0,045, ce qui est bien inférieur au seuil critique de 0,85. Cette faible corrélation est positive : elle confirme que le modèle modifie de manière significative ses régions d'attention en fonction de la classe prédite. Le réseau active ainsi des zones cérébrales distinctes pour chaque catégorie (par exemple, des zones différentes pour un stade sain et un stade avancé), ce qui valide la spécificité et la cohérence de son comportement décisionnel (Figure 3.22).

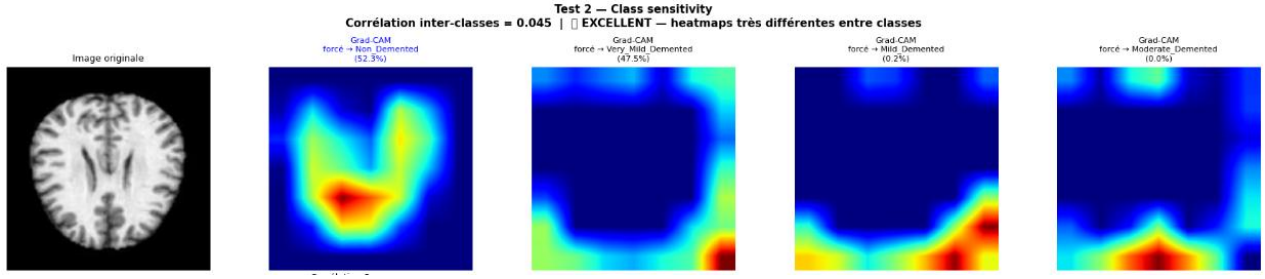


Figure 3.22:Évaluation de la robustesse par le Test 2.

Le **Test 3 (Forced Class Test)** montre un écart absolu moyen ou MAD (Mean Absolute Deviation) élevé de 0,4166, ce qui est largement supérieur au seuil minimal exigé (> 0,05). Ce résultat indique qu'un changement forcé de classe provoque une modification importante et visible des cartes Grad-CAM. Cela confirme mathématiquement que les explications générées sont fortement dépendantes de la classe sélectionnée et restent logiquement liées à la décision prise par le modèle (Figure 3.23).

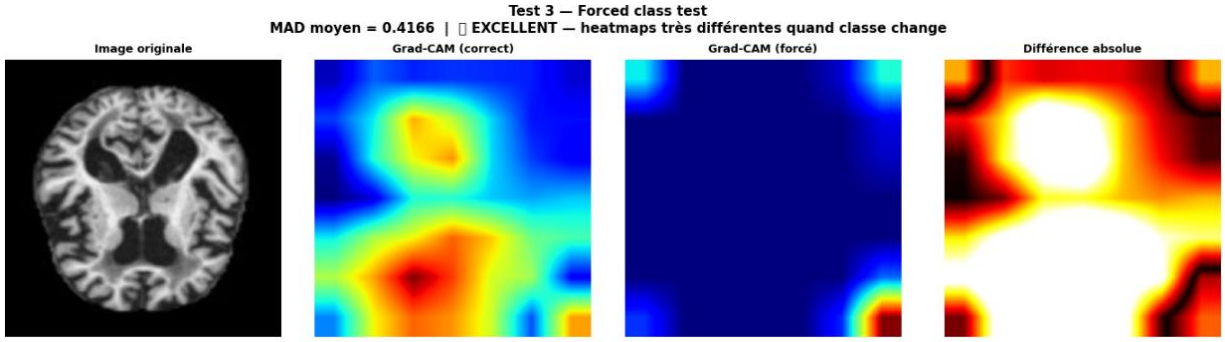


Figure 3.23:Évaluation de la robustesse par le Test 3.

Le **Test 4 (Random Baseline Test)** montre une faible corrélation de 0,1218, ce qui est bien inférieur au seuil limite fixé (< 0,5) entre notre modèle entraîné et un modèle témoin dont les poids ont été rendus aléatoires. Cette faible valeur est cruciale : elle confirme que les cartes Grad-CAM dépendent directement de l'apprentissage et des caractéristiques extraites par le réseau, et non de sa seule architecture géométrique. Cela valide définitivement la fiabilité et la fidélité des explications générées par rapport au savoir acquis par le modèle (Figure 3.24).

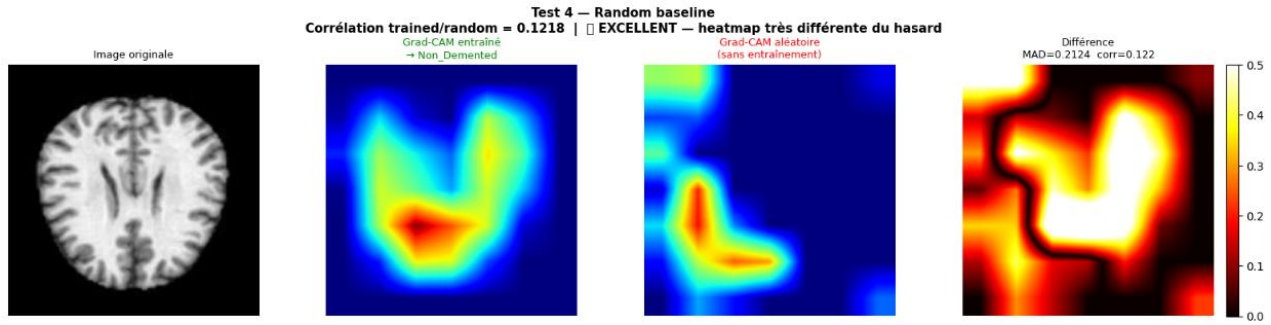


Figure 3.24:Évaluation de la robustesse par le Test 4.

Ainsi, les critères de validation mathématique sont tous validés avec la mention excellent. Cela prouve que la méthode Grad-CAM ne produit pas de visualisations génériques ou de simples cartes de contours basées sur les contrastes de l'image ; elle traduit fidèlement le raisonnement interne du réseau de neurones convolutifs.

Cependant, la seule validation mathématique de ces cartes ne suffit pas à garantir leur utilité pratique. Pour s'assurer que les zones suractivées dictent réellement le choix du réseau et qu'elles

correspondent à des réalités médicales, il est nécessaire de valider ce comportement sous un angle causal, puis anatomique.

Le **Test 5 (Test d'occlusion globale)** vient quant à lui confirmer la validité causale des cartes Grad-CAM. En pratique, la suppression (l'occlusion) des 20% de pixels identifiés comme les plus importants par Grad-CAM entraîne une forte baisse de la confiance du modèle, avec une chute moyenne de 69,16% et un taux de réussite du test de 100%. Ce résultat dépasse largement le seuil critique fixé à 15%. Cela démontre concrètement que les régions mises en évidence par les cartes thermiques sont essentielles à la décision du modèle et jouent un rôle causal direct dans la classification finale (Figure 3.25).

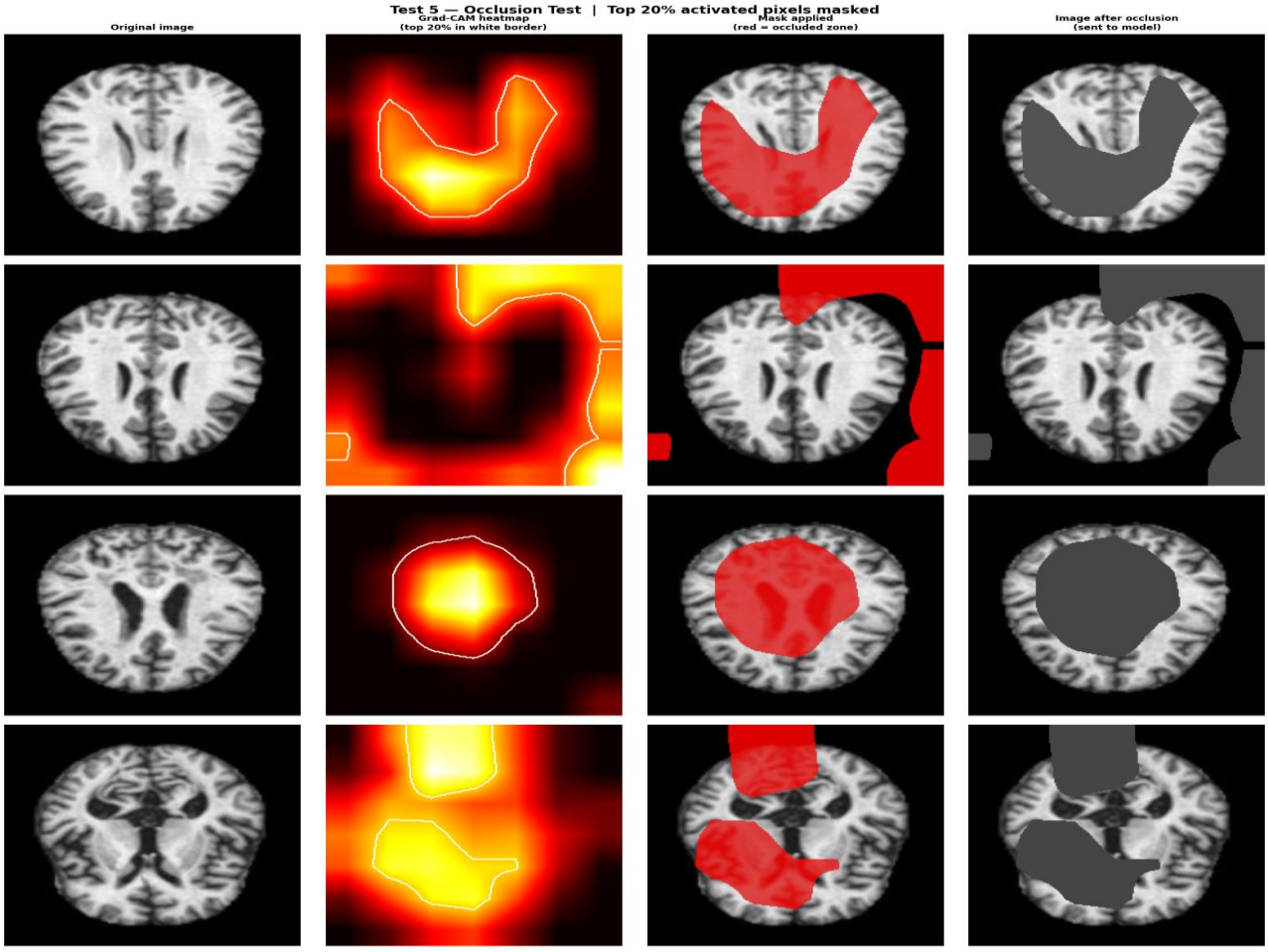


Figure 3.25:Évaluation de la validité causale par le Test 5 (Test d'occlusion globale)

5.4.2. Validation anatomique et causale par Brain-Masking

Le Brain-Masking constitue une étape essentielle pour améliorer la fiabilité clinique et anatomique des explications Grad-CAM. En éliminant les facteurs de confusion situés hors du cerveau (arrière-plan, os du crâne et artefacts d'imagerie), cette technique permet de concentrer les cartes de saillance uniquement sur les structures cérébrales pertinentes, conformément aux recommandations de Zech et al. (2018) [52].

La Figure 3.26 présente, de gauche à droite : l'image IRM originale, le masque cérébral généré par la méthode de seuillage d'Otsu et opérations morphologiques, la carte Grad-CAM brute (activations globales), la carte Grad-CAM filtrée (limitée au tissu cérébral), et enfin l'image occultée mettant en évidence les régions critiques supprimées pour le test de validité causale. Les quatre lignes correspondent aux différents stades de la cohorte.

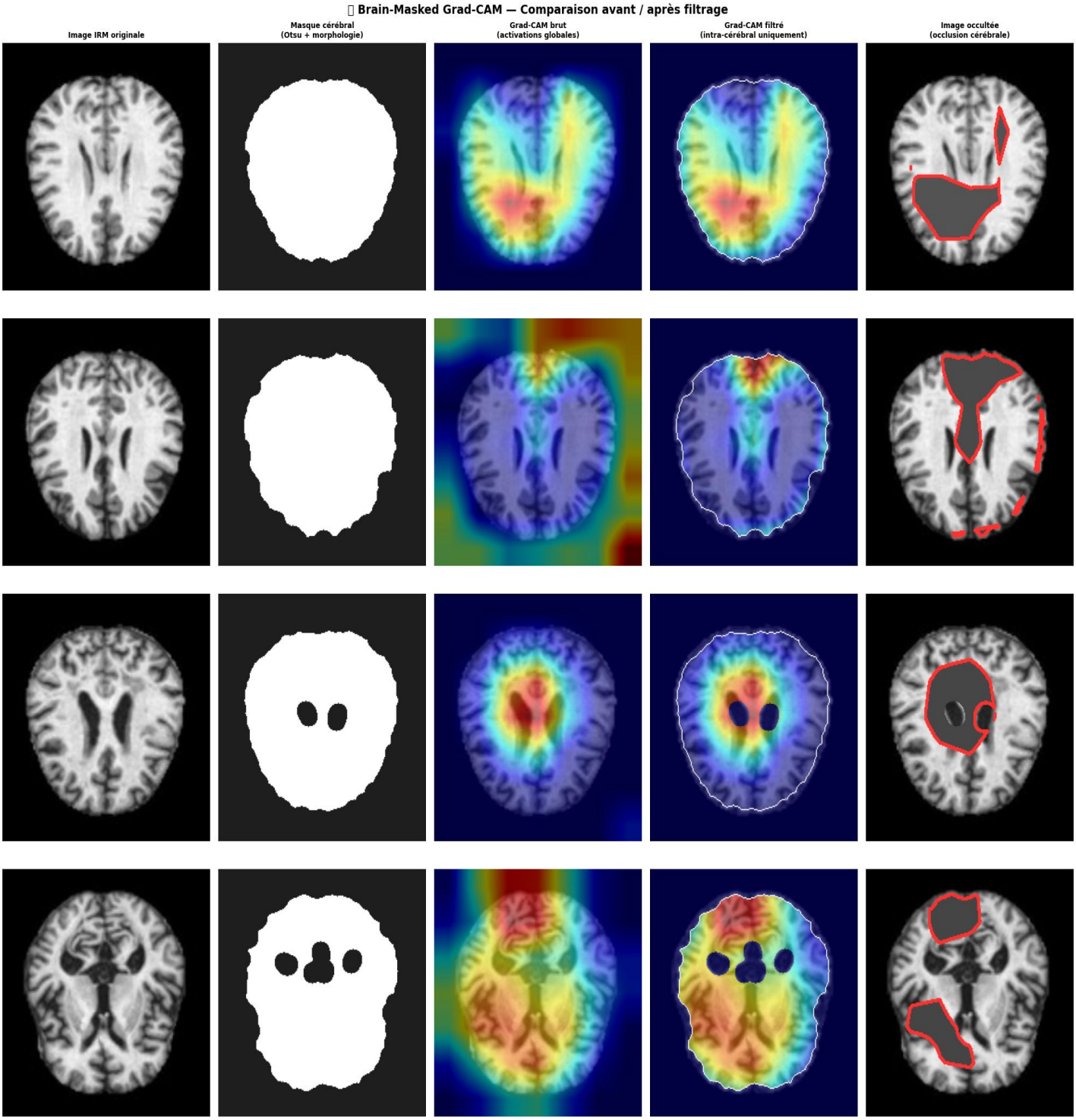


Figure 3.26:Comparaison qualitative du pipeline Brain-Masked Grad-CAM.

Le Tableau 3.6 présente une synthèse des principaux indicateurs causaux et énergétiques calculés à partir des 36 images IRM analysées, couvrant les quatre classes diagnostiques de notre base de données.

Classe	Nombre d'images	Énergie brute	Énergie filtrée	Gain net	Chute de confiance moyenne	Taux de réussite (PASSED)	Verdict Énergie
Non demented	10	79,3%	100%	+20,7%	70,3% (±39,9%)	80%	Amélioré
Very mild demented	10	32,9%	100%	+67,1%	66,7% (±36,3%)	80%	Amélioré
Mild demented	10	53,2%	100%	+46,8%	70,4% (±34,8%)	80%	Amélioré
Moderate demented	6	55,2%	100%	+44,8%	43,6% (±34,8%)	67%	Faible
Moyenne globale	36	55,1%	100%	+44,9%	62,8% (±36,4%)	77%	Excellent

Tableau 3.6:Indicateurs de performance du test Brain-Masked Grad-CAM par classe diagnostique

Les résultats du test Brain-Masked Grad-CAM sur ces 36 images montrent une bonne performance globale et confirment la forte capacité d'explicabilité de notre modèle. L'application du masque cérébral a permis d'éliminer complètement les activations parasites hors du cerveau, faisant passer l'énergie d'attention interne de 55,1% à 100%, soit un gain net moyen de +44,9%. Cela prouve que le modèle concentre ses choix sur des régions anatomiquement pertinentes.

Par ailleurs, l'analyse causale basée sur le test d'occlusion après filtrage confirme la robustesse des cartes d'attention générées. La chute moyenne de confiance globale de 62,8% après le masquage des 20% de pixels les plus contributifs démontre que les prédictions du réseau reposent effectivement sur des structures cérébrales déterminantes. Le taux de réussite global de 77% (soit 28 images sur 36) indique une bonne cohérence d'ensemble. Elle est particulièrement marquée dans les classes Non Demented, Very Mild Demented et Mild Demented, où le taux de succès atteint 80%. Ce score suggère une détection efficace des biomarqueurs précoces, tels que l'atrophie hippocampique ou la dilatation ventriculaire. En revanche, la classe Moderate Demented présente une performance légèrement en retrait (67%), probablement en raison de la redondance spatiale des lésions et de la diffusion des caractéristiques pathologiques à l'ensemble du cerveau, ce qui réduit l'impact d'une occlusion seulement partielle.

L'analyse des cas d'échec (22% des images, classées comme « Faible ») révèle néanmoins plusieurs limites méthodologiques. Certaines prédictions très confiantes (hyper-confiantes) restent stables malgré l'occlusion, ce qui montre un effet de saturation du modèle dans ses décisions.

D'autres cas s'expliquent par la redondance des signaux pathologiques dans les deux hémisphères : une occlusion partielle ne suffit pas toujours à perturber la prédiction globale. Enfin, les artefacts visuels introduits lors du masquage (le remplacement des pixels coupés) peuvent parfois générer de faux motifs que le réseau interprète à tort comme des informations pertinentes.

Dans l'ensemble, ces résultats s'intègrent dans un pipeline complet d'explicabilité comprenant des validations mathématiques et causales successives. Le test Brain-Masked Grad-CAM confirme ainsi la robustesse, la cohérence anatomique et la fiabilité des décisions du modèle. Le verdict final "Excellent" atteste que l'approche proposée constitue un cadre d'explicabilité solide, renforçant son potentiel d'utilisation en contexte clinique comme outil d'aide au diagnostic de la maladie d'Alzheimer.

5.4. Expérimentation 4 : Évaluation de la généralisabilité et de la transversalité pathologique

Afin de valider la robustesse, la flexibilité et la capacité de généralisation hors-domaine de notre protocole hybride, une série d'évaluations a été menée sur trois bases de données médicales indépendantes, distinctes de la cohorte d'apprentissage principale. Cette démarche vise à mesurer l'invariance des performances de notre méthode face aux variabilités inter-pathologies, inter-constructeurs et inter-protocoles d'acquisition. Les résultats quantitatifs, comparant les configurations natives (sans équilibrage) aux configurations optimisées (avec notre approche), sont synthétisés dans le Tableau 3.7.

Base de données	Configuration	Accuracy (%)	Rappel (%)	Précision (%)	F1-score (%)	Spécificité (%)	G-mean (%)
Brain MRI Images (Binaire)	Sans équilibrage	92,31	68,75	64,71	66,67	95,28	80,93
	Avec équilibrage	91,61	87,50	58,33	70,00	92,13	89,78
Brain Tumor MRI Dataset (Multi-classes)	Sans équilibrage	97,10	97,24	96,94	97,08	99,04	98,14
	Avec équilibrage	97,42	97,55	97,41	97,46	99,14	98,34
Osteoarthritis Prediction (Binaire)	Sans équilibrage	79,57	77,71	90,37	83,56	83,33	80,47
	Avec équilibrage	91,49	92,86	94,08	93,46	88,86	90,85

Tableau 3.7: Métriques comparatives de généralisation et de transversalité pathologique sur trois bases de données indépendantes

5.4.1. Analyse globale multi-métriques des scénarios de neuroimagerie (Brain MRI et Brain Tumor)

Dans le premier scénario, fondé sur la base de données Brain MRI Images pour la classification binaire des tumeurs cérébrales, l'intégration de notre approche hybride modifie sensiblement l'équilibre entre la sensibilité et la spécificité du modèle. Bien que l'accuracy globale diminue légèrement de 92,31 % à 91,61 % (−0,70 %), cette variation résulte d'un repositionnement plus

équilibré des frontières de décision. En effet, le modèle initial présentait un biais en faveur de la classe majoritaire (cas sains), ce qui conduisait à une surestimation de ses performances globales. L'approche proposée corrige ce déséquilibre en améliorant significativement la détection de la classe minoritaire pathologique : le rappel passe de 68,75 % à 87,50 %, soit un gain de 18,75 %. Cette amélioration se traduit également par une augmentation notable du G-mean, qui progresse de 80,93 % à 89,78 % (+8,85 %), témoignant d'un meilleur équilibre entre les performances obtenues sur les deux classes.

Cette augmentation de la sensibilité s'accompagne d'un ajustement modéré de la précision, qui passe de 64,71 % à 58,33 %, ainsi que d'une légère diminution de la spécificité, de 95,28% à 92,13%. Cette réorganisation entraîne une hausse contrôlée du nombre de faux positifs, mais permet simultanément une réduction substantielle des faux négatifs. En conséquence, le F1-score progresse de 66,67 % à 70,00 % (+3,33 %), confirmant la pertinence du compromis obtenu. Dans un contexte clinique tel que l'oncologie, où l'absence de détection d'une tumeur peut avoir des conséquences graves pour le patient, cette réduction des faux négatifs constitue un avantage majeur, justifiant largement l'augmentation limitée des faux positifs.

Dans le second scénario, consacré à la classification multi-classes du Brain Tumor MRI Dataset, le modèle de référence affichait déjà des performances élevées avec une accuracy de 97,10 %. L'application de notre stratégie hybride n'a engendré aucune dégradation des performances et a permis une amélioration cohérente de l'ensemble des indicateurs évalués. L'accuracy atteint 97,42 % (+0,32 %), tandis que le F1-score s'élève à 97,46 % (+0,38 %). Cette amélioration s'accompagne d'une progression du rappel (97,55 %), de la précision (97,41 %) et de la spécificité (99,14 %), conduisant à un G-mean de 98,34 %. Ces résultats mettent en évidence la capacité de l'approche proposée à affiner les frontières de décision tout en préservant la qualité des représentations apprises par le réseau. Ils démontrent également que la combinaison des techniques d'équilibrage et d'augmentation des données améliore la robustesse du modèle, même lorsque les performances de base du réseau sont déjà très élevées.

5.4.2. Validation de la transversalité anatomoclinique (Osteoarthritis Prediction Dataset)

La démonstration la plus probante de la transversalité anatomoclinique et de l'indépendance sémantique de notre approche hybride provient des résultats obtenus sur la base Osteoarthritis Prediction Dataset, consacrée à la gradation de l'ostéoarthrite du genou. Cette application représente un changement de domaine particulièrement exigeant, dans la mesure où les images radiographiques musculosquelettiques présentent des caractéristiques anatomiques, texturales et pathologiques très différentes de celles observées en neuroimagerie. Les structures analysées, les niveaux de contraste ainsi que les marqueurs diagnostiques (tels que le pincement de l'interligne articulaire, la sclérose

sous-chondrale ou la présence d'ostéophytes) diffèrent profondément des signatures visuelles associées aux tumeurs cérébrales.

Dans sa configuration initiale, le modèle était fortement affecté par le déséquilibre des classes, ce qui limitait sa capacité à identifier correctement les stades minoritaires de la maladie. Les performances atteignaient une accuracy de 79,57 % et un rappel de 77,71 %, traduisant une difficulté à reconnaître de manière fiable les cas pathologiques les moins représentés. Cette situation est caractéristique des bases de données médicales réelles, où certaines catégories cliniques demeurent sous-représentées.

L'intégration de notre pipeline hybride génère ici les gains les plus marqués de l'ensemble des expérimentations. L'accuracy progresse de 79,57% à 91,49 %, soit un gain de 11,92% de pourcentage, tandis que le rappel atteint 92,86 %, correspondant à une amélioration de 15,15%. Ces résultats témoignent d'une augmentation substantielle de la capacité du modèle à détecter correctement les différents niveaux de sévérité de l'arthrose.

Plus remarquable encore, cette hausse de la sensibilité ne s'effectue pas au détriment de la capacité discriminante du réseau. Contrairement à de nombreuses stratégies classiques de rééquilibrage, qui améliorent le rappel au prix d'une augmentation significative des faux positifs, notre approche permet une progression simultanée de la précision, qui passe de 90,37 % à 94,08 % (+3,71 %), et de la spécificité, qui s'élève de 83,33 % à 88,86 % (+5,53 %). Cette amélioration conjointe traduit une meilleure séparation des classes et une réduction des erreurs de classification.

L'amélioration globale du système est également confirmée par l'augmentation du F1-score, qui atteint 93,46 % contre 83,56 % dans la configuration initiale, soit un gain de 9,90%, ainsi que par la progression du G-mean à 90,85 % (+10,38%). Ces indicateurs démontrent que le modèle parvient à maintenir un équilibre élevé entre sensibilité et spécificité, tout en améliorant de manière homogène ses performances sur l'ensemble des classes.

Ces résultats suggèrent que la combinaison du suréchantillonnage ciblé, de l'augmentation des données et des pondérations adaptatives favorise l'apprentissage de représentations robustes et discriminantes, capables de capturer des caractéristiques pathologiques pertinentes indépendamment du contexte anatomique étudié. Les gains observés sur une pathologie musculosquelettique, après avoir été validés sur des tâches de neuroimagerie, démontrent que l'efficacité de l'approche ne dépend ni d'un organe particulier ni d'un type spécifique de biomarqueur visuel.

En conclusion, cette expérimentation confirme le caractère agnostique, généralisable et transversal de notre méthode hybride. Sa capacité à améliorer simultanément plusieurs indicateurs de performance dans des contextes cliniques très différents montre qu'elle constitue une solution

robuste au problème du déséquilibre des données médicales. Ces propriétés en font une approche particulièrement prometteuse pour le développement de systèmes d'aide au diagnostic médical fiables, interprétables et applicables à grande échelle.

5.4.3. Analyse qualitative par Grad-CAM et validation par expertise clinique

Pour prouver la robustesse de notre approche, il est essentiel de démontrer que le modèle ne se limite pas à une seule application ou à une unique base de données, mais qu'il est capable de généraliser son apprentissage à différentes structures anatomiques et pathologies. C'est dans ce cadre que nous avons appliqué la méthode Grad-CAM sur trois jeux de données distincts (tumeurs cérébrales binaires, tumeurs cérébrales multiclasses et ostéoarthrite du genou) afin d'analyser les régions sur lesquelles le modèle se base pour formuler ses décisions.

Les résultats qualitatifs, illustrés par les Figures 3.27, 3.28 et 3.29, montrent que le modèle se concentre principalement sur des zones d'intérêt parfaitement pertinentes sur le plan médical et clinique, confirmant ainsi la validité et la transversalité de son apprentissage face à des critères morphologiques très variés.

Cette pertinence diagnostique et la correspondance avec les réalités anatomiques du terrain ont été rigoureusement validées par l'expertise clinique de la Dr. Sanhadji Sara. Cette dernière a confirmé la cohérence des régions et des structures mises en évidence par les cartes thermiques. L'obtention de repères visuels aussi précis sur des types d'imageries et des pathologies si différentes confirme la forte capacité de généralisation de notre pipeline et son potentiel d'adaptation en contexte clinique multi-expertises.

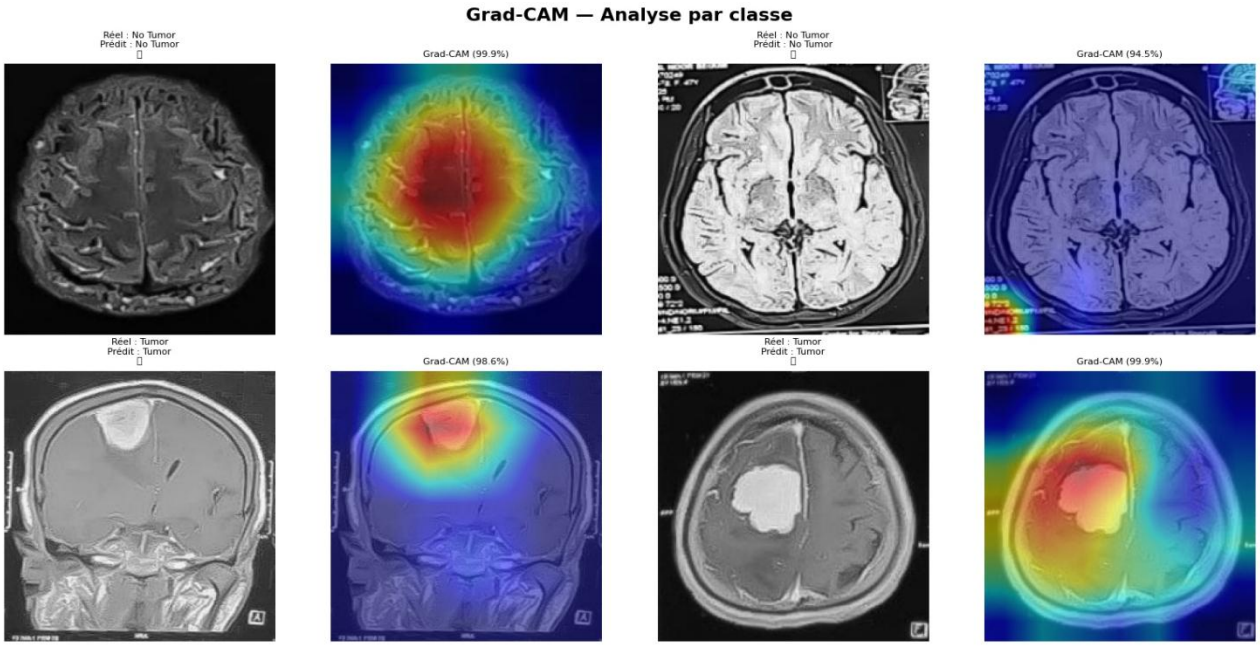


Figure 3.27:. Métriques comparatives de généralisation et de transversalité pathologique sur trois bases de données indépendantes

La figure présente l'image IRM d'origine suivie de sa carte thermique correspondante. La ligne supérieure montre des cas sains (No Tumor) où l'attention se distribue sur le parenchyme normal. La ligne inférieure met en évidence des cas pathologiques (Tumor), illustrant une focalisation précise et intense des activations (zones rouges) directement sur la masse tumorale.

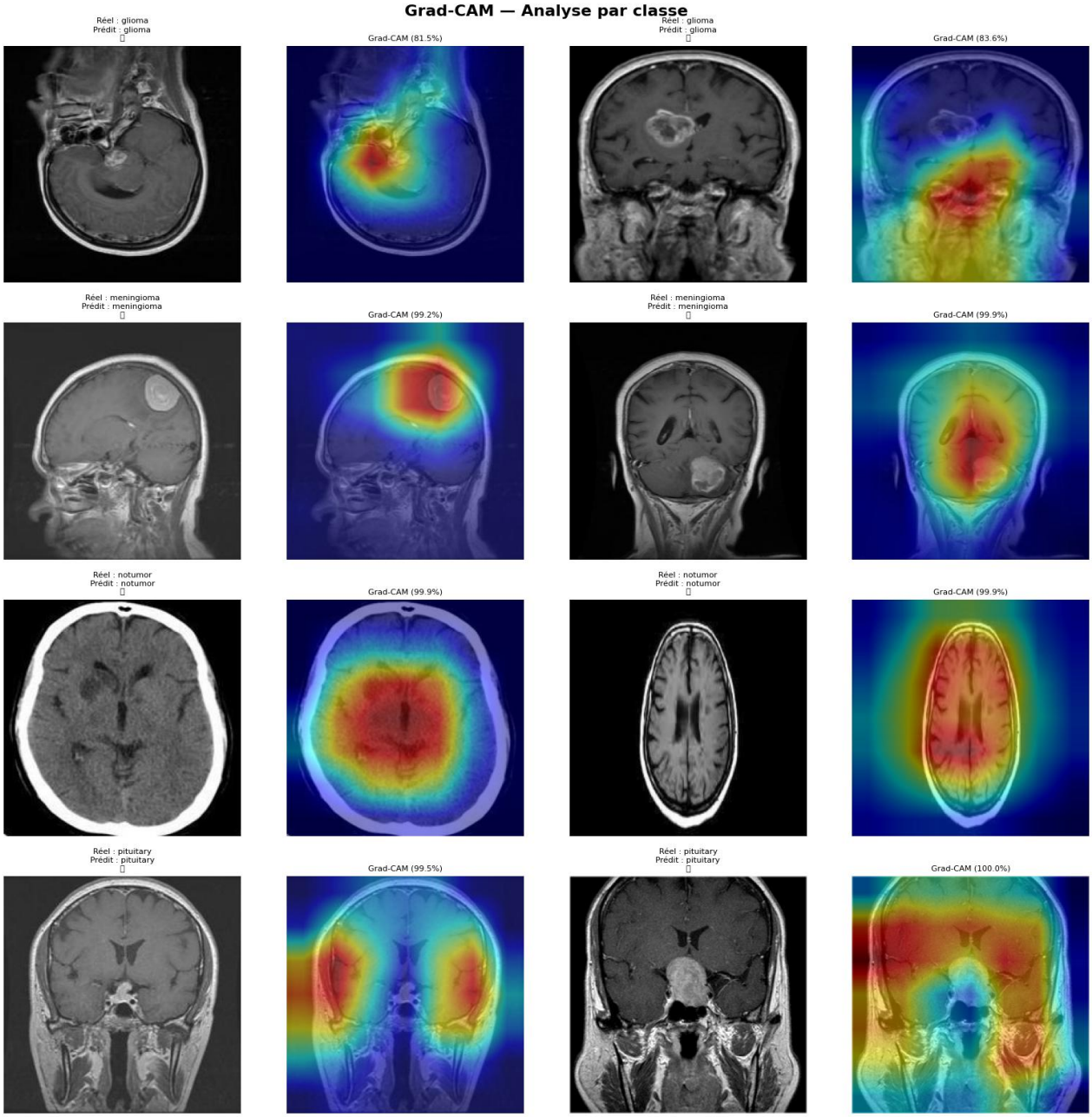


Figure 3.28:Visualisation de l'interprétabilité par Grad-CAM pour la détection de tumeurs cérébrales multi classes.

Les lignes illustrent la capacité du modèle à isoler des biomarqueurs topographiques spécifiques selon la catégorie diagnostique (notamment sur des régions profondes/ventriculaires, sur des localisations périphériques de type méningiome, ou sur la région de la selle turcique pour les

adénomes hypophysaires). La forte concentration des gradients confirme que l'apprentissage repose sur des caractéristiques anatomocliniques distinctes.

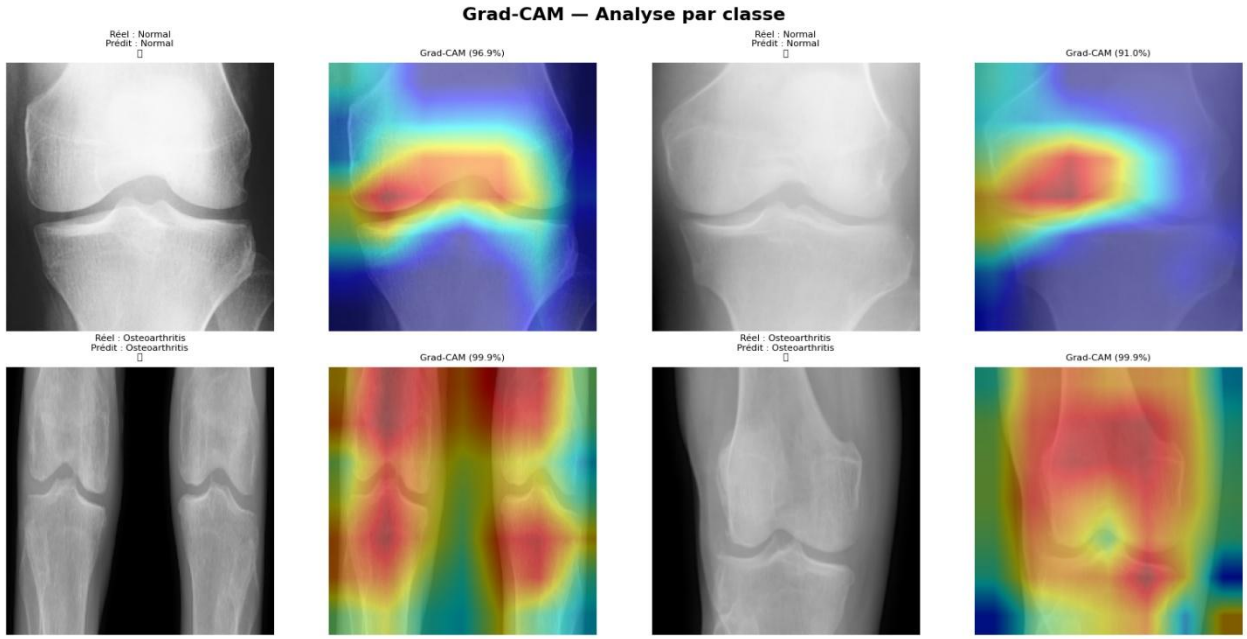


Figure 3.29:Visualisation de l'interprétabilité par Grad-CAM pour la détection de l'ostéoarthrite du genou.

Les lignes comparent les examens radiographiques de genoux sains (Normal, haut) et atteints d'arthrose (Osteoarthritis, bas). Pour les cas pathologiques, les zones de forte chaleur se concentrent spécifiquement au niveau de l'interligne articulaire, reflétant de manière cohérente les zones de pincement, de condensation sous-chondrale ou de formation d'ostéophytes caractéristiques de la maladie.

6. Conclusion

En conclusion, ce chapitre a permis de valider à la fois la pertinence théorique et l'efficacité pratique du pipeline d'aide au diagnostic proposé. En naviguant de la structuration des flux de données jusqu'à la confrontation clinique, les différentes sections ont mis en lumière la capacité de notre modèle à surmonter les défis complexes de l'imagerie médicale contemporaine.

Sur le plan purement quantitatif, la sélection rigoureuse d'EfficientNetB2 comme cœur décisionnel, optimisée par notre approche hybride unifiée (TOS + DA + CW + Two-PL), a permis d'effacer les biais induits par le déséquilibre sévère des classes nosologiques. Avec une accuracy de 96,72 % et un rappel hautement critique de 97,63 %, notre pipeline se positionne à la pointe de l'état de l'art, égalant ou surpassant des architectures d'ensembles ou des modèles génératifs nettement plus lourds et instables, tout en préservant une grande sobriété computationnelle.

Sur le plan de l'explicabilité, la mise en place d'un cadre de validation strict combinant tests mathématiques, masquage anatomique (Brain-Masking) et protocoles d'occlusion causale a permis

d'écarter le risque de "faux positifs visuels". Les résultats prouvent que le modèle oriente ses gradients de manière sincère et stable. Cette rigueur mathématique trouve son couronnement dans la validation qualitative menée par le Dr. Sanhadji Sara, qui a certifié la parfaite cohérence des zones d'activation face aux réalités cliniques, que ce soit pour le continuum d'atrophie de la maladie d'Alzheimer, l'identification de structures tumorales cérébrales spécifiques ou le ciblage de l'interligne articulaire dans l'ostéoarthrite du genou.

L'ensemble de ces résultats confirme la forte capacité de généralisation et la transversalité pathologique de notre architecture sur différents types d'examens radiologiques et d'imageries par résonance magnétique. Ce solide jalon scientifique étant posé, il ouvre la voie à la discussion des perspectives d'implémentation logicielle et à la conclusion générale de l'ensemble de notre étude.



Conclusion générale



Conclusion

Ce travail a permis de valider, à la fois sur le plan théorique et expérimental, la pertinence et l'efficacité d'un pipeline d'apprentissage profond conçu pour répondre au défi critique du déséquilibre des classes en imagerie médicale. En naviguant de la structuration des flux de données jusqu'à la confrontation clinique, les différents chapitres ont mis en lumière la capacité du modèle proposé à surmonter les défis complexes de l'imagerie contemporaine, à travers des expérimentations rigoureuses menées sur quatre jeux de données hétérogènes, comprenant des IRM cérébrales et des radiographies musculo-squelettiques.

La première contribution majeure de ce mémoire réside dans l'étude comparative standardisée de sept architectures CNN, établissant la supériorité d'EfficientNetB2 comme modèle de référence. Ce choix se justifie non seulement par sa performance statistique élevée, mais également par sa stabilité d'entraînement et sa robustesse face au surapprentissage, comparativement à des architectures plus lourdes telles que VGG19. Parallèlement, la conception d'une approche hybride originale combinant l'apprentissage en deux phases, le suréchantillonnage ciblé, l'augmentation des données et la pondération adaptative des classes, constitue un apport déterminant. En agissant simultanément sur la distribution des échantillons, le signal d'erreur rétropropagé et la dynamique du taux d'apprentissage, cette stratégie unifiée surpasse nettement les techniques appliquées isolément. Elle permet d'optimiser l'assimilation des représentations sémantiques même en présence d'une asymétrie inter-classe sévère, allant jusqu'à un ratio de $50:1$ pour la maladie d'Alzheimer. Avec une accuracy globale de 96,72 % et un rappel exceptionnel de 97,63 %, le pipeline proposé accroît significativement la sensibilité du système sur les catégories cliniques minoritaires, réduisant ainsi le risque de faux négatifs sur les cas pathologiques les plus déterminants.

Au-delà de ces performances, la validation de cette stratégie sur plusieurs jeux de données indépendants démontre que notre approche ne constitue pas une solution ad hoc dédiée à une seule pathologie, mais un cadre méthodologique transférable à d'autres modalités d'imagerie et d'autres régions anatomiques. Enfin, ce travail se distingue par la mise en place d'un protocole de validation de l'IA explicable (XAI) multiniveaux. Dépassant la simple inspection visuelle des cartes d'activation Grad-CAM, ce cadre rigoureux combine tests mathématiques, masquage anatomique (*Brain-Masking*) et protocoles d'occlusion causale pour écarter définitivement le risque de « faux positifs visuels ». Cette démarche trouve son aboutissement dans la validation qualitative menée par le Dr Sanhadji Sara, qui a confirmé la cohérence des zones d'activation avec les réalités cliniques, qu'il s'agisse du continuum d'atrophie de la maladie d'Alzheimer, de l'identification de structures tumorales cérébrales spécifiques ou du ciblage de l'interligne articulaire dans l'ostéoarthrite du genou.

Toutefois, ces résultats encourageants doivent être nuancés par certaines limites méthodologiques qu'il convient d'expliciter avec rigueur. Tout d'abord, la représentativité statistique reste limitée sur certaines classes ultra-minoritaires, à l'image de la catégorie *Moderate Demented* du jeu de données Alzheimer qui ne contient que six échantillons dans le sous-ensemble de test, ce qui impose d'interpréter les scores parfaits obtenus avec une relative prudence. De plus, l'évaluation s'est concentrée sur des corpus publics acquis dans des conditions contrôlées et standardisées ; les performances du pipeline pourraient donc fluctuer face à la forte hétérogénéité et à la variabilité inter-scanners de données hospitalières réelles. Sur le plan de l'explicabilité, environ 22 % des images analysées lors du test *Brain-Masked Grad-CAM* ont présenté une activation jugée non significative (verdict modéré), principalement en raison de la redondance spatiale des lésions corticales aux stades avancés de la maladie. Enfin, l'évaluation de la pertinence anatomique des cartes thermiques reste pour l'instant monocentrique, n'ayant été menée que par un seul médecin expert, ce qui mériterait d'être confronté à un panel élargi de cliniciens pour consolider la portée statistique de cette validation.

Perspectives futures

Les limites identifiées ouvrent des pistes de recherche prometteuses pour la suite de ces travaux :

- **Sur le plan de l'explicabilité** : l'intégration de méthodes complémentaires telles que Grad-CAM++, SHAP ou LIME permettrait d'obtenir des cartes d'attention encore plus fines et d'enrichir l'interprétation des décisions du modèle.
- **Sur le plan de la validation clinique** : le déploiement et l'évaluation de notre pipeline sur des données cliniques réelles, issues de flux hospitaliers de routine, constitueront une étape déterminante pour tester la robustesse du système face à la variabilité inter-constructeurs.
- **Sur le plan méthodologique** : l'extension de la méthodologie développée à d'autres pathologies complexes ou à l'intégration de données cliniques et biologiques complémentaires permettrait d'enrichir la compréhension globale du profil du patient.
- **Sur le plan applicatif** : la conception d'une interface logicielle d'aide à la décision, intégrant le modèle optimisé et ses cartes de saillance validées, faciliterait l'adoption et l'évaluation de cette technologie directement par les praticiens.

Enfin, afin de concrétiser la portée applicative de ce travail, une annexe présente Oasis Med — Plateforme d'Aide à la Décision Clinique, un outil web qui opérationnalise le pipeline d'intelligence artificielle développé dans ce mémoire. Cette plateforme constitue une première étape vers la traduction de nos contributions scientifiques en un outil concret, destiné à faciliter l'intégration de notre approche hybride et de notre protocole d'explicabilité dans un contexte clinique réel.

En somme, ce mémoire pose un jalon scientifique rigoureux qui concilie précision diagnostique et transparence décisionnelle. En répondant aux exigences réglementaires actuelles d'intégrité et d'explicabilité, les contributions de ce travail participent activement à l'effort de recherche visant à transformer les modèles d'apprentissage profond en outils de confiance, sécurisés et viables pour la routine médicale de demain.

Annexe

Oasis Med —Plateforme d'Aide à la Décision Clinique

1. Présentation Générale de la Plateforme

Oasis Med est une plateforme web d'aide à la décision clinique basée sur l'intelligence artificielle, conçue pour assister les radiologues et les cliniciens dans le diagnostic de pathologies neurologiques, oncologiques et musculosquelettiques. La plateforme intègre des modèles d'apprentissage profond avancés pour l'analyse d'images médicales (IRM et radiographies), permettant une détection précoce et précise de la maladie d'Alzheimer, des tumeurs cérébrales et de l'ostéoarthrite du genou.

Le système s'articule autour d'une interface clinique intuitive permettant la gestion des patients, la soumission de cas, l'analyse automatisée par IA, et la validation médicale des résultats. Les médecins disposent d'outils d'explicabilité visuelle basés sur la méthode Grad-CAM pour interpréter de manière transparente et valider rigoureusement les prédictions du modèle avant de finaliser les rapports cliniques.

2. Interfaces de la plateforme

Cette section présente les différentes interfaces graphiques développées pour la plateforme Oasis Med. À travers une série de captures d'écran (de la Figure A.1 à la Figure A.15), nous détaillons le parcours utilisateur du médecin praticien, depuis le tableau de bord initial jusqu'aux outils avancés de validation par intelligence artificielle et de configuration du système.

2.1. Tableau de bord principal

Le tableau de bord offre une vue d'ensemble complète de l'activité clinique du médecin. Comme illustré à la Figure A.1, il affiche les indicateurs clés de performance : le nombre total de cas soumis, les analyses automatisées par l'IA, les dossiers en attente de validation médicale, ainsi que les rapports finalisés. Une courbe de tendance hebdomadaire met en regard l'évolution des flux de patients face aux volumes traités par le modèle, complétée par un diagramme de distribution des types d'examens et un suivi de l'indice de fiabilité globale du système.

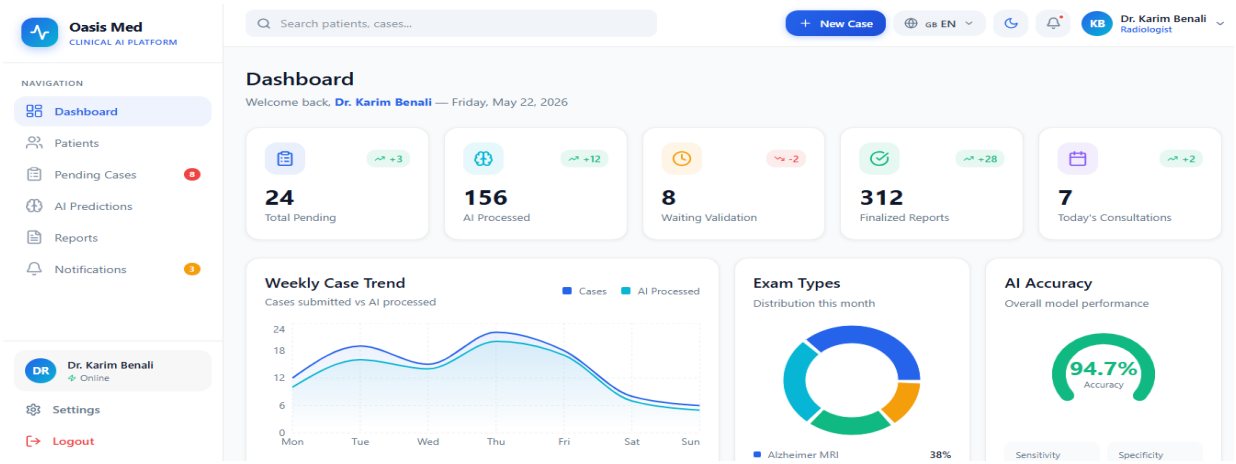


Figure A.1 *Tableau de bord principal avec statistiques cliniques et indicateurs de performance IA.*

2.2. Interface de gestion des patients

Ce module rassemble l'ensemble des patients enregistrés sous la responsabilité du médecin. Comme le montre la Figure A.2, chaque entrée affiche les informations essentielles : l'identifiant unique, le nom, l'âge, le sexe, le groupe sanguin, la date de la dernière visite, le nombre de cas associés et le statut clinique actuel. Un bouton de consultation « Review » permet un accès direct et immédiat au dossier complet du patient.

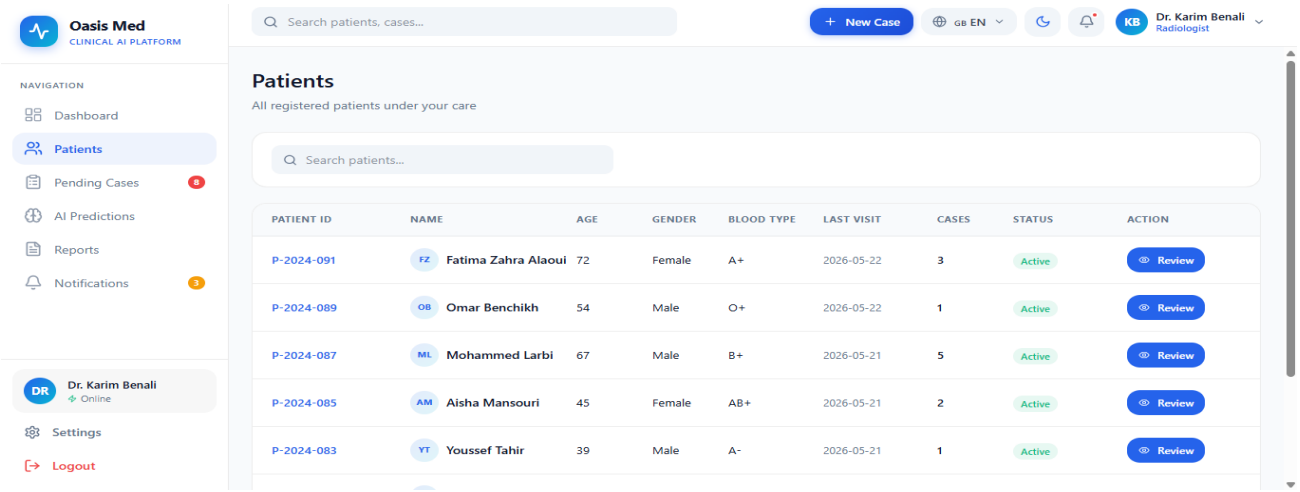


Figure A.2 *Interface de liste des patients avec informations démographiques et accès rapide aux dossiers.*

2.3. Gestion des examens en attente (Pending Cases)

Cette interface centralise l'ensemble des examens nécessitant une analyse ou une validation médicale. Comme l'illustre la Figure A.3, cette vue dispose de filtres dynamiques par état (Tous, En attente, En cours de traitement, Validés) et d'un système de tri par degré de priorité. Chaque ligne résume les métadonnées critiques du dossier : l'identifiant du patient, la modalité d'imagerie, le statut d'importation des volumes, l'état d'avancement de l'analyse automatisée par l'IA, le niveau d'urgence clinique (CRITIQUE ou ÉLEVÉ), la date de soumission, ainsi qu'un bouton d'accès rapide pour l'évaluation du cas.

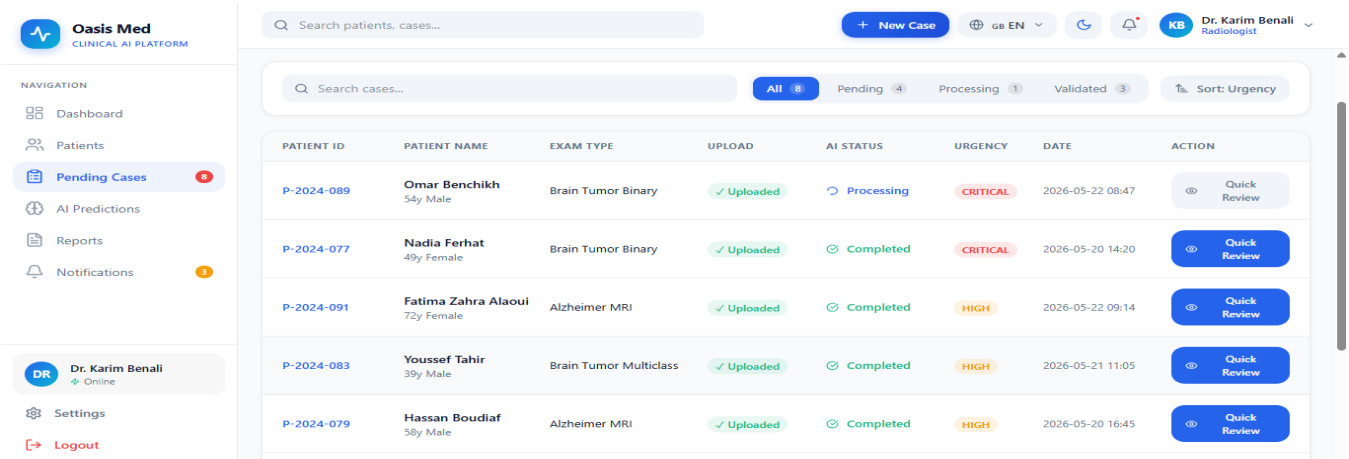


Figure A.3 *Interface des cas en attente avec filtres par statut et niveaux d'urgence.*

2.4. Interface de création d'un nouveau cas clinique

La création d'un nouveau cas suit un processus en trois étapes : (1) Dossier Patient, (2) Téléchargement d'Image, et (3) Analyse IA. La Figure A.4 illustre la première étape, où le médecin consulte les informations du patient (âge, sexe, date de naissance, groupe sanguin, médecin référent, service) et sélectionne le type d'examen IA approprié parmi les modèles disponibles : IRM Alzheimer (AlzNet-v2.4.1), Tumeur Cérébrale Binaire (TumorDet-v1.8.2), Tumeur Cérébrale Multiclasse (TumorNet-v3.1.0) et Ostéoarthrite du genou (OsteoNet-v1.2.0).

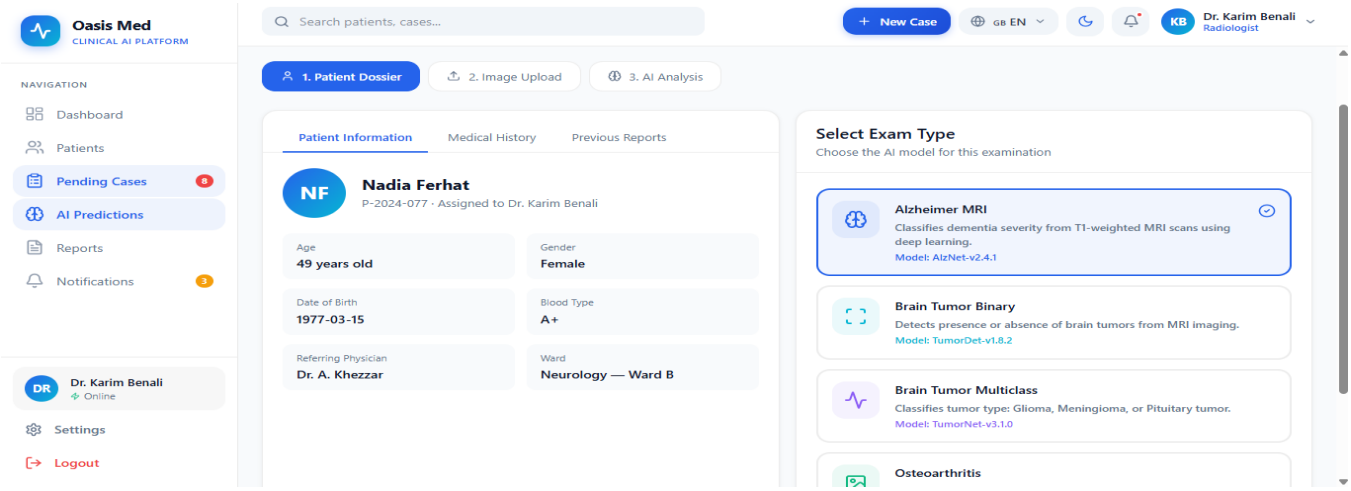


Figure A.4 *Interface de création de cas — étape 1 : dossier patient et sélection du type d'examen IA.*

2.5. Chargement de l'image médicale

La deuxième étape du processus est dédiée au chargement de l'image médicale pour l'analyse. Comme indiqué dans la Figure A.5, l'interface propose une zone de glisser-déposer acceptant les formats DICOM, PNG, JPG et TIFF. Un aperçu de l'image en niveaux de gris est affiché à droite, accompagné des métadonnées techniques : la modalité (IRM pondérée en T1), l'orientation (Axiale), l'épaisseur de coupe et la taille de la matrice. Le bouton « Lancer l'Analyse IA » déclenche l'exécution du pipeline d'inférence.

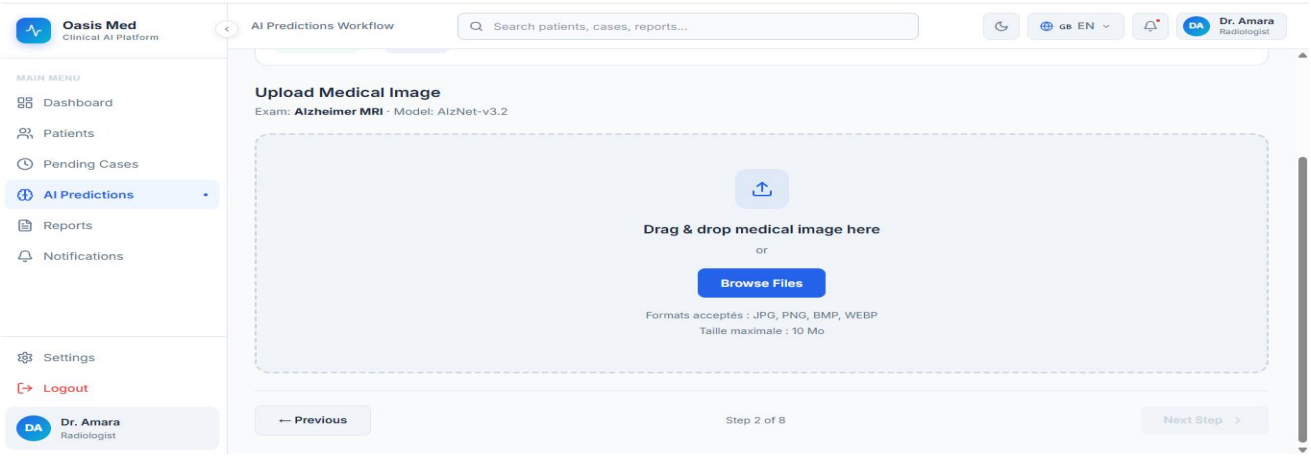


Figure A.5 *Interface de téléchargement d'image médicale avec prévisualisation et métadonnées DICOM.*

2.6. Résultats de l'analyse IA et visualisation Grad-CAM

Une fois l'analyse complétée, l'interface de la station de travail IA affiche les résultats de la prédiction. La Figure A.6 montre le résultat prédit (Démence Légère) avec son score de confiance et la distribution des probabilités par classe. La visualisation Grad-CAM met en évidence les régions clés activées par le modèle, avec une carte thermique identifiant les zones anatomiques les plus influentes dans la décision, telles que l'hippocampe, le cortex entorhinal et le lobe temporal.

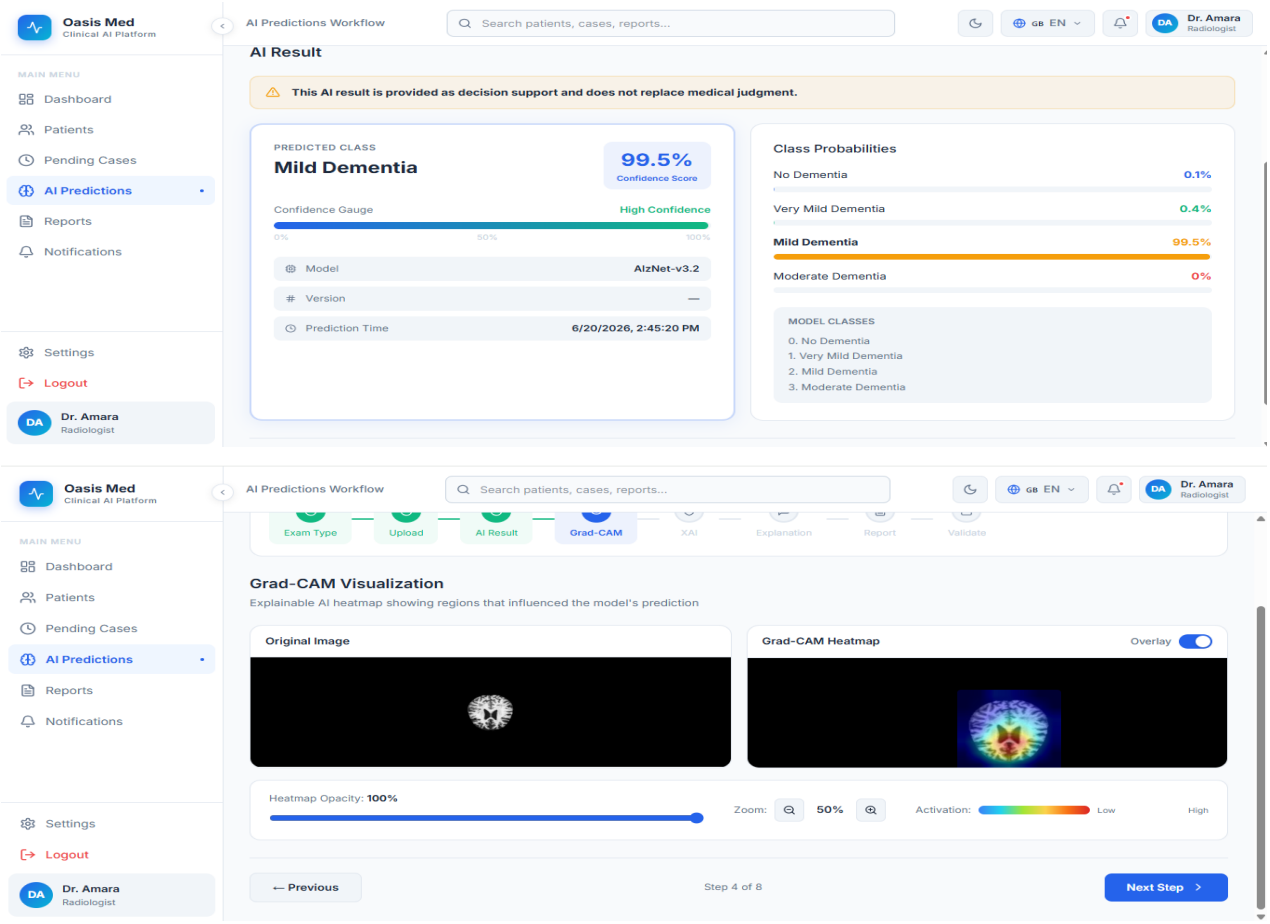


Figure A.6 *Station de travail IA — résultats de prédiction et visualisation Grad-CAM avec régions anatomiques clés.*

2.7. Validation XAI et explication clinique

La section de validation XAI (Explainable AI) fournit une évaluation rigoureuse de la fiabilité de la prédiction. La Figure A.7 présente les résultats des tests de validation : Stabilité de la Confiance (SUCCÈS), Qualité de la Localisation (SUCCÈS — focus primaire sur l'hippocampe, cliniquement plausible), Focus sur l'arrière-plan (SUCCÈS — aucune activation d'arrière-plan significative), Avertissement de faible confiance (SUCCÈS), Sensibilité à l'occlusion (ALERTE — sensibilité modérée à l'occlusion de la région temporale gauche) et Robustesse face aux attaques adverses (SUCCÈS).

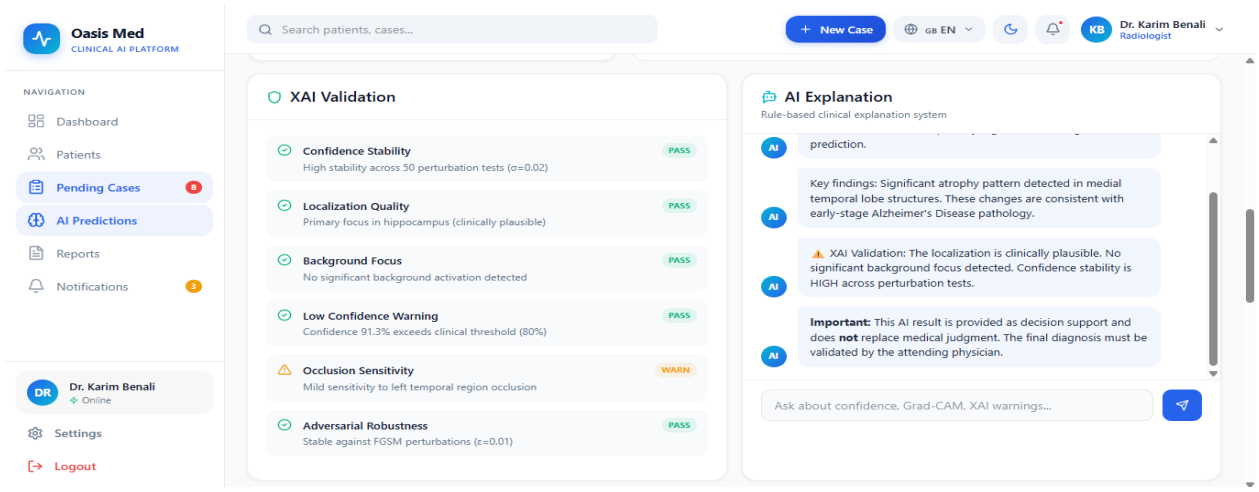


Figure A.7 *Interface de validation XAI avec résultats des tests de robustesse et explication clinique par IA.*

Le panneau d'explication IA (à droite) détaille les conclusions clés : atrophie significative des structures du lobe temporal médian, cohérentes avec une pathologie d'Alzheimer précoce, accompagnée d'une note importante rappelant que le résultat IA est fourni à titre d'aide à la décision et ne remplace pas le jugement médical.

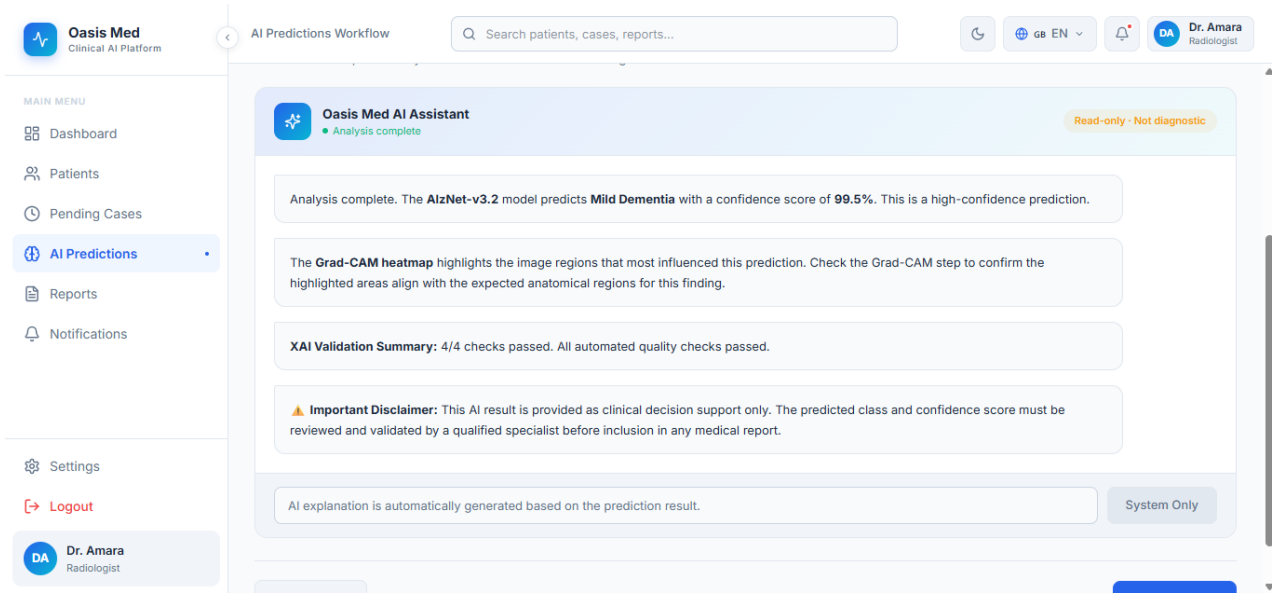


Figure A.8 *Vue détaillée du système d'explication clinique basé sur des règles (panneau droit).*

2.8. Rapport médical

L'interface de rapport médical permet au médecin de rédiger et de valider le compte-rendu final. La Figure A.9 illustre cette interface comprenant : un résumé automatique généré par l'IA (basé sur AlzNet-v2.4.1), les détails de l'examen (type, modèle, date de l'examen, établissement), un champ de diagnostic éditable, les commentaires du médecin et les recommandations cliniques (évaluation neuropsychologique, traitement par inhibiteurs de la cholinestérase, IRM de suivi dans 6 mois, ressources pour les aidants).

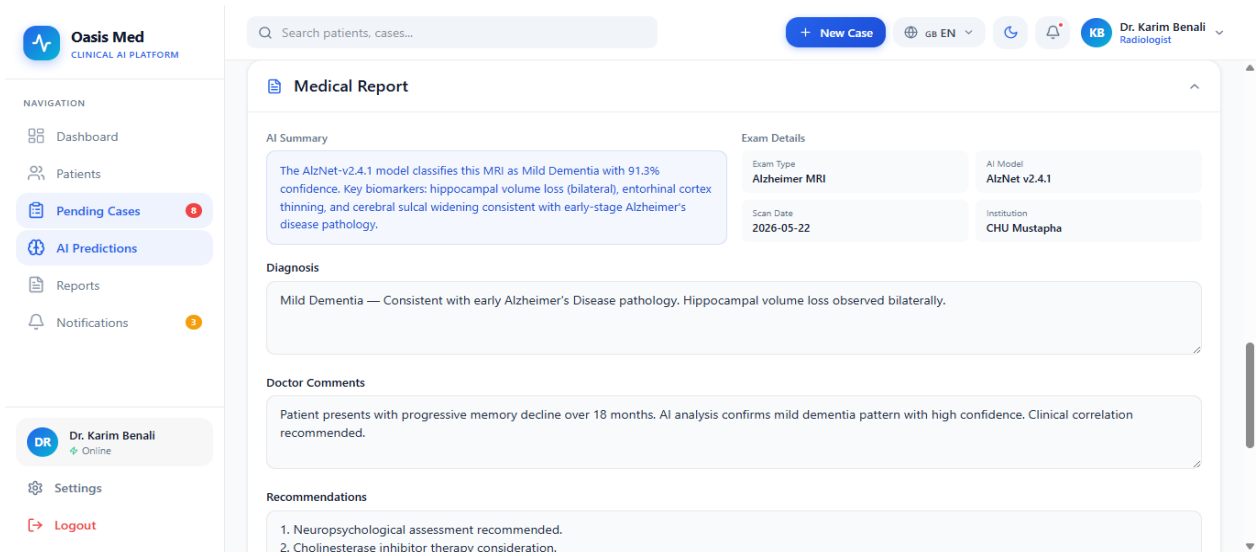


Figure A.9 *Interface de rapport médical avec résumé IA, diagnostic, commentaires et recommandations cliniques.*

2.9. Validation finale et signature du rapport

La section de validation finale permet au médecin d'approuver ou de modifier la prédiction de l'IA avant de finaliser le rapport. Comme le montre la Figure A.10, le médecin peut choisir entre deux actions : Approuver le Résultat IA (validation directe) ou Remplacer la Prédiction IA (saisie manuelle). Après cela, il finalise le rapport, génère le PDF et archive le cas. La signature numérique du médecin traitant (Dr. Karim Benali, Radiologue — CHU Mustapha, Licence #ALG-12847) est alors intégrée au document officiel.

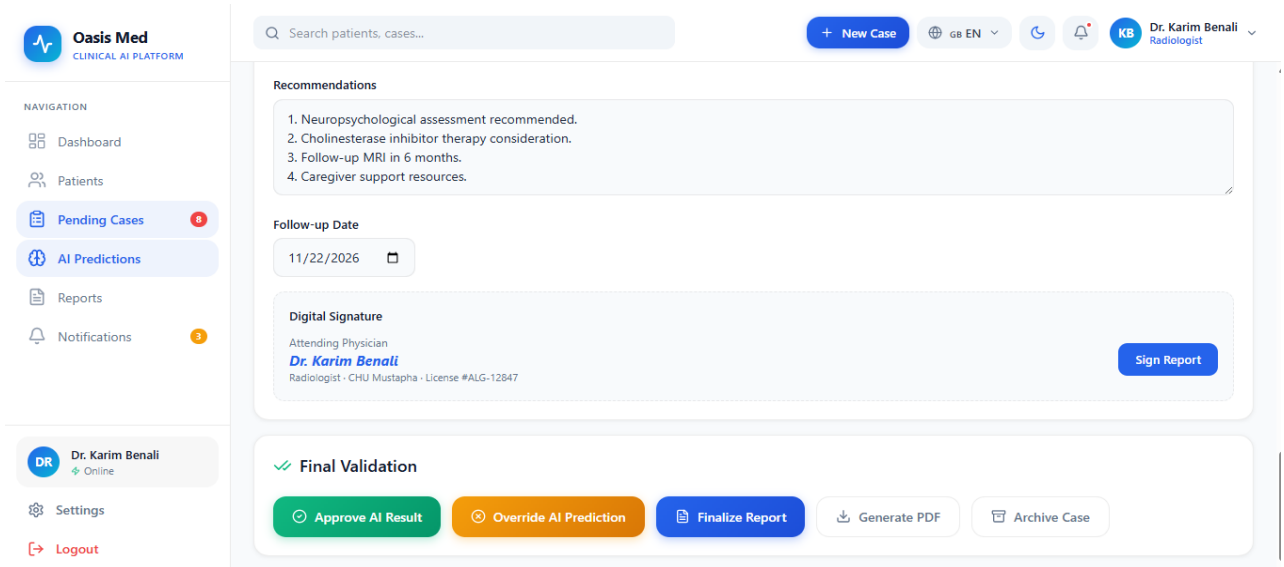


Figure A.10 *Interface de validation finale avec actions de décision clinique et signature numérique du médecin.*

2.10. Système de notifications

Le centre de notifications centralise toutes les alertes cliniques en temps réel. La Figure A.11 présente l'interface avec les 4 notifications non lues, classées par priorité : Nouveau Chargement IRM (ÉLEVÉ — traitement IA démarré pour la patiente Fatima Zahra), Validation Requise (ÉLEVÉ — résultat de tumeur cérébrale nécessitant une validation urgente), Rapport Prêt (MOYEN — rapport IA disponible pour révision) et Rappel de Suivi (MOYEN — IRM de suivi programmée pour demain).

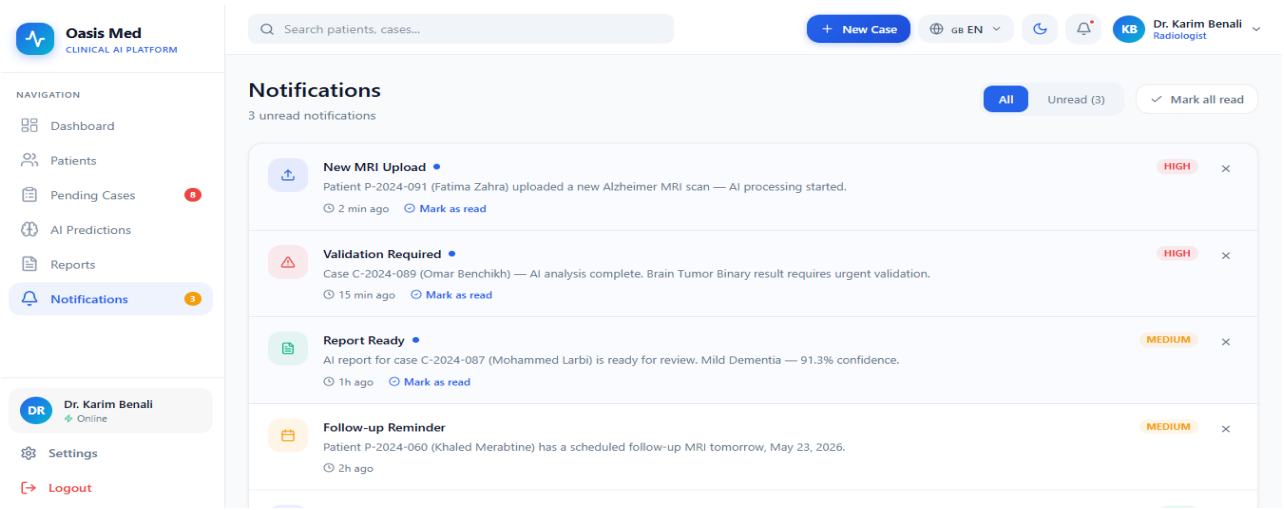


Figure A.11 *Centre de notifications avec alertes cliniques classées par priorité et statut de lecture.*

2.11. Rapports finalisés

L'interface des rapports offre une vue d'ensemble de tous les documents validés. La Figure A.12 présente les statistiques globales de ces comptes-rendus. Le tableau liste chaque élément par son identifiant, le nom du patient, le type d'examen, le diagnostic posé, le score de confiance et la date, avec un accès direct au rapport final et une option de téléchargement.

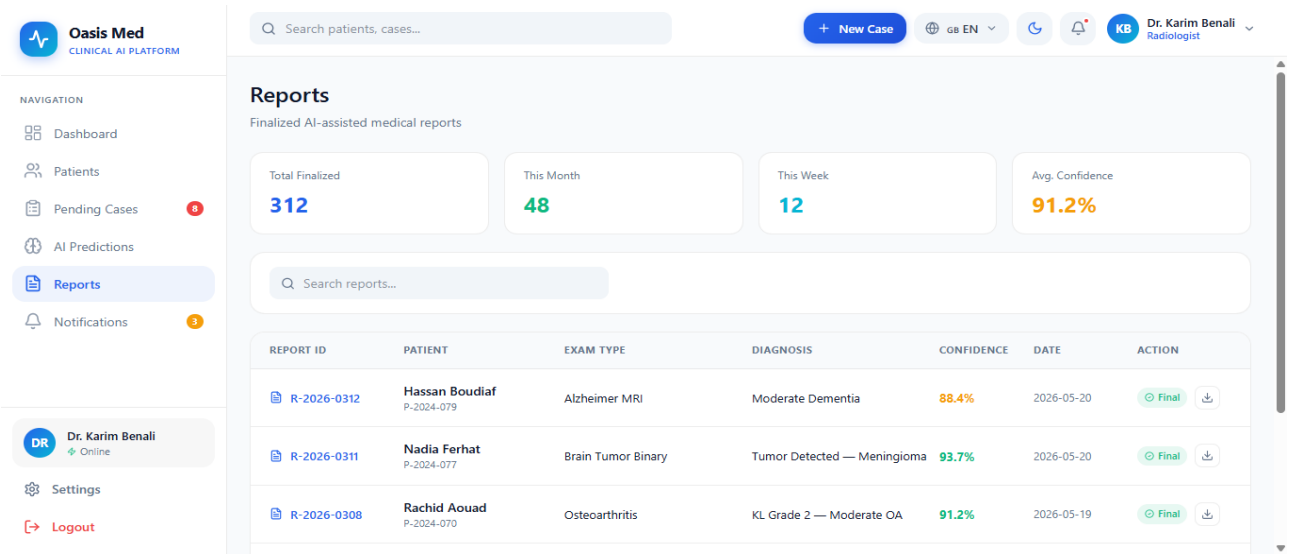


Figure A.12 *Interface des rapports finalisés avec statistiques globales et liste des comptes rendus.*

2.12. Paramètres (Settings)

L'interface des paramètres permet au médecin de personnaliser et de configurer son espace de travail sur la plateforme Oasis Med. Elle regroupe l'ensemble des options de gestion du compte, de la sécurité et des préférences cliniques.

Les paramètres disponibles incluent :

- **Informations du compte :** modification du profil médical, de la photo, de la spécialité et des coordonnées. La Figure A.13 montre l'organisation de cette interface de gestion du profil.

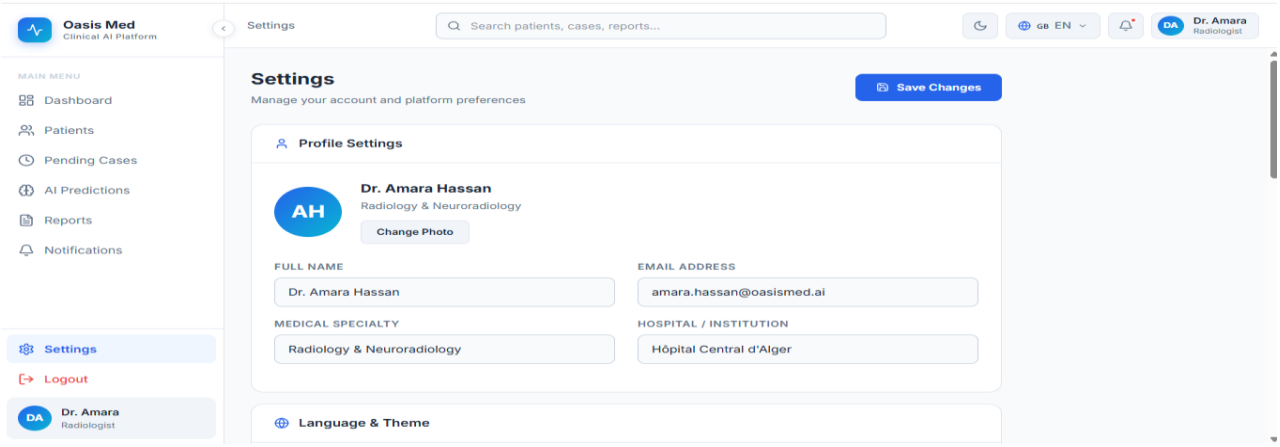


Figure A.13 *Interface des Informations du compte.*

- **Sécurité et accès :** gestion du mot de passe, de l'authentification à deux facteurs et des sessions actives. Cette section permet de sécuriser l'accès aux données de santé des patients (illustrée par la Figure A.14).

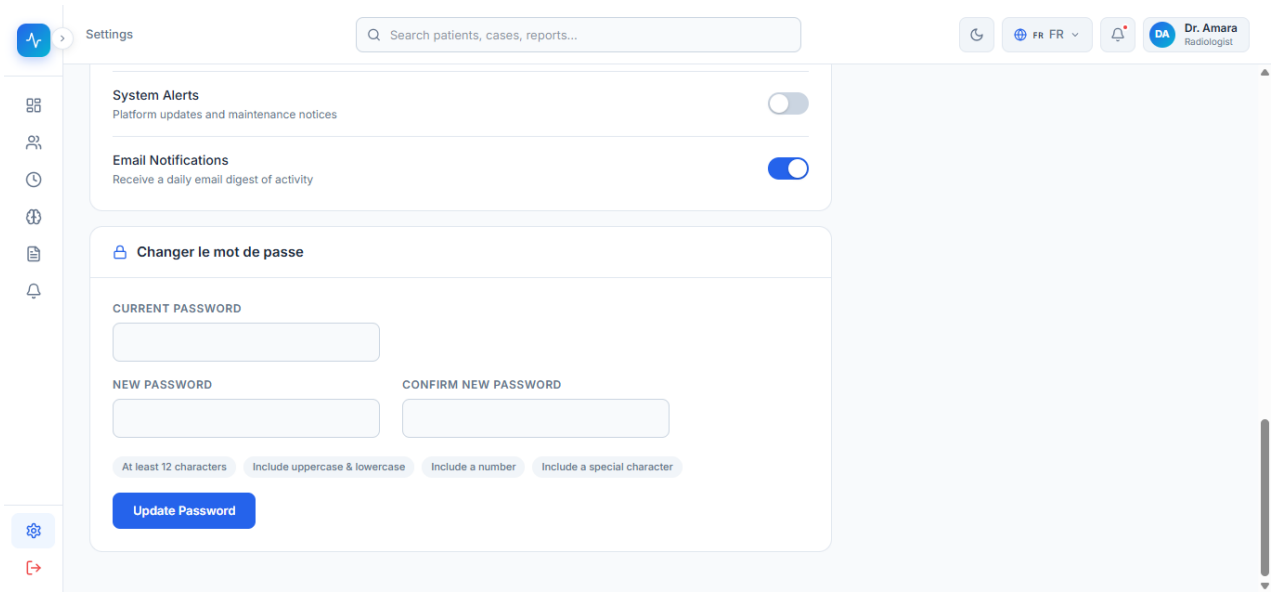


Figure A.14 : Interface de sécurité et accès

- **Notifications :** configuration des alertes cliniques, des rappels de suivi et des niveaux de priorité. La Figure A.15 présente les différentes options de personnalisation de ces alertes.

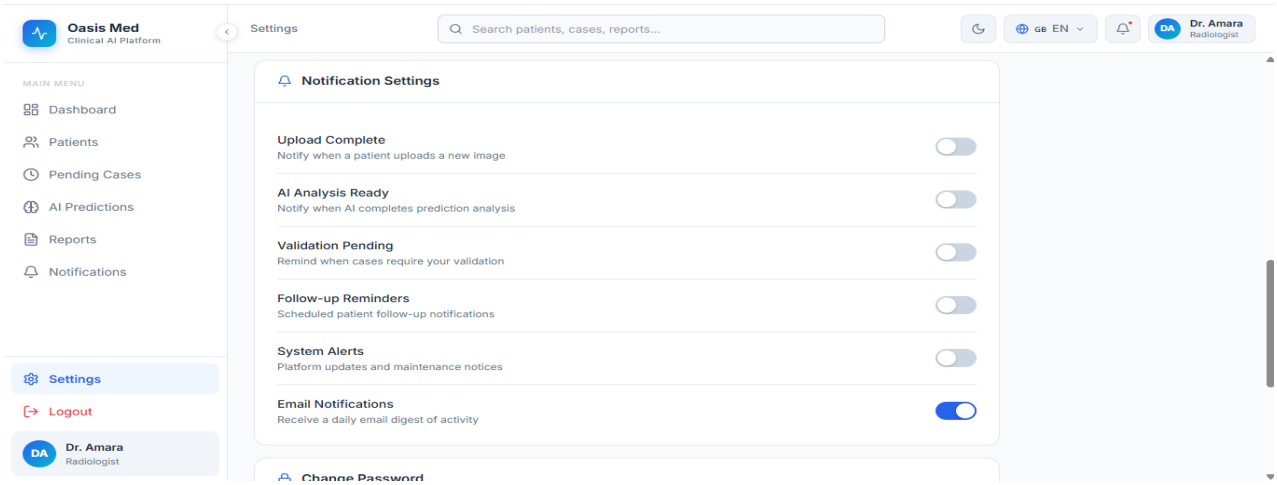


Figure A.15 *Interface des* Notifications.

- **Préférences d'affichage :** langue de l'interface, thème visuel et disposition des tableaux de bord. Comme le montre la Figure A.16, le médecin peut adapter l'aspect visuel de l'application selon ses besoins.

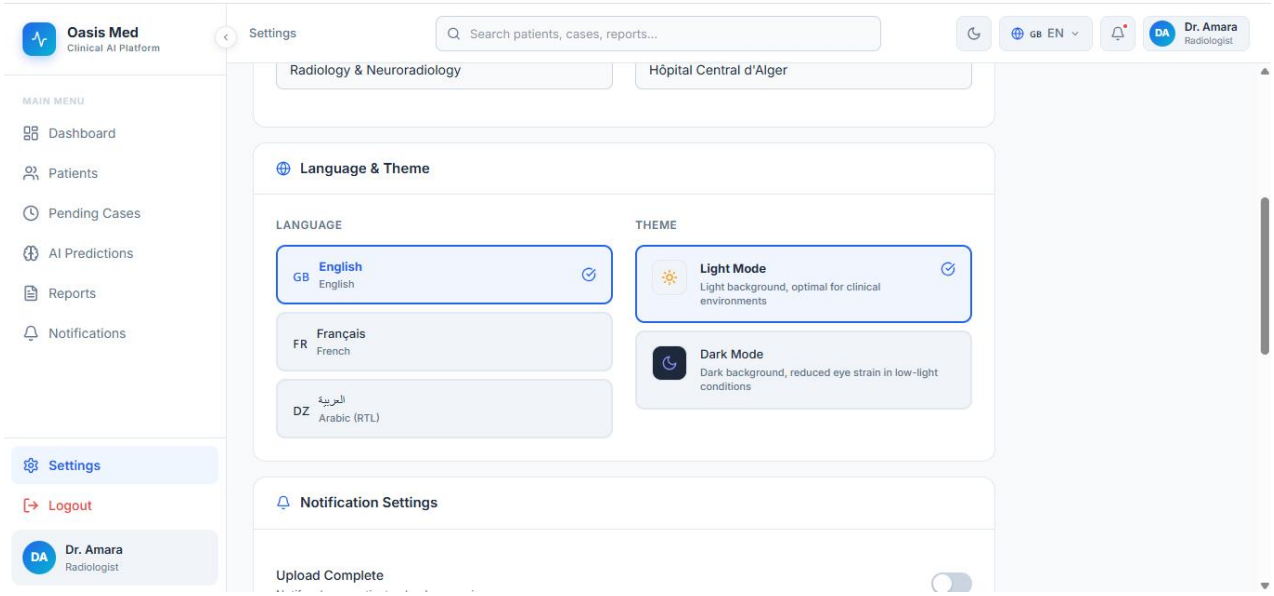


Figure A.16 *Interface des Préférences d'affichage.*

Références bibliographiques

Références bibliographiques

[1] M. Buda, A. Maki, et M. A. Mazurowski, « A systematic study of the class imbalance problem in convolutional neural networks », *Neural Netw.*, vol. 106, p. 249-259, oct. 2018, doi: 10.1016/j.neunet.2018.07.011.
[2] I. Araf, A. Idri, et I. Chairi, « Cost-sensitive learning for imbalanced medical data: a review », *Artif. Intell. Rev.*, vol. 57, nº 4, p. 80, mars 2024, doi: 10.1007/s10462-023-10652-8.
[3] B. Krawczyk, « Learning from imbalanced data: open challenges and future directions », *Prog. Artif. Intell.*, vol. 5, nº 4, p. 221-232, nov. 2016, doi: 10.1007/s13748-016-0094-0.
[4] G. Litjens *et al.*, « A survey on deep learning in medical image analysis », *Med. Image Anal.*, vol. 42, p. 60-88, déc. 2017, doi: 10.1016/j.media.2017.07.005.
[5] A. Esteva *et al.*, « A guide to deep learning in healthcare », *Nat. Med.*, vol. 25, nº 1, p. 24-29, janv. 2019, doi: 10.1038/s41591-018-0316-z.
[6] O. Ronneberger, P. Fischer, et T. Brox, « U-Net: Convolutional Networks for Biomedical Image Segmentation », 18 mai 2015, *arXiv*: arXiv:1505.04597. doi: 10.48550/arXiv.1505.04597.
[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, et D. Batra, « Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization », 3 décembre 2019. doi: 10.1007/s11263-019-01228-7.
[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, et W. P. Kegelmeyer, « SMOTE: Synthetic Minority Over-sampling Technique », *J. Artif. Intell. Res.*, vol. 16, p. 321-357, juin 2002, doi: 10.1613/jair.953.
[9] M. Salmi, D. Atif, D. Oliva, A. Abraham, et S. Ventura, « Handling imbalanced medical datasets: review of a decade of research », *Artif. Intell. Rev.*, vol. 57, nº 10, p. 273, sept. 2024, doi: 10.1007/s10462-024-10884-2.
[10] C. J. Hellín, A. A. Olmedo, A. Valledor, J. Gómez, M. López-Benítez, et A. Tayebi, « Unraveling the Impact of Class Imbalance on Deep-Learning Models for Medical Image Classification », *Appl. Sci.*, vol. 14, nº 8, p. 3419, janv. 2024, doi: 10.3390/app14083419.
[11] W. Fithian et T. Hastie, « Local case-control sampling: Efficient subsampling in imbalanced data sets », *Ann. Stat.*, vol. 42, nº 5, p. 1693-1724, oct. 2014, doi: 10.1214/14-AOS1220.
[12] V. Sampath, I. Maurtua, J. J. Aguilar Martín, et A. Gutierrez, « A survey on generative adversarial networks for imbalance problems in computer vision tasks », *J. Big Data*, vol. 8, nº 1, p. 27, janv. 2021, doi: 10.1186/s40537-021-00414-0.
[13] J. M. Johnson et T. M. Khoshgoftaar, « Survey on deep learning with class imbalance », *J. Big Data*, vol. 6, nº 1, p. 27, mars 2019, doi: 10.1186/s40537-019-0192-5.
[14] D. Chicco et G. Jurman, « The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation », *BMC Genomics*, vol. 21, nº 1, p. 6, janv. 2020, doi: 10.1186/s12864-019-6413-7.
[15] R. C. Prati, G. E. A. P. A. Batista, et M. C. Monard, « Class Imbalances versus Class Overlapping: An Analysis of a Learning System Behavior », in *MICAI 2004: Advances in Artificial Intelligence*, R. Monroy, G. Arroyo-Figueroa, L. E. Sucar, et H. Sossa, Éd., Berlin, Heidelberg: Springer, 2004, p. 312-321. doi: 10.1007/978-3-540-24694-7_32.
[16] S. BELAROUCI, « Traitement et classification des données médicales non équilibrées », Université Aboubakr Belkaïd – Tlemcen –.
[17] A. Esteva *et al.*, « Deep learning-enabled medical computer vision », *Npj Digit. Med.*, vol. 4, nº 1, p. 5, janv. 2021, doi: 10.1038/s41746-020-00376-2.

[18] T. Saito et M. Rehmsmeier, « The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets », *PLOS ONE*, vol. 10, nº 3, p. e0118432, mars 2015, doi: 10.1371/journal.pone.0118432.
[19] R. Blagus et L. Lusa, « SMOTE for high-dimensional class-imbalanced data », *BMC Bioinformatics*, vol. 14, nº 1, p. 106, mars 2013, doi: 10.1186/1471-2105-14-106.
[20] « Oversampling Techniques | Kaggle ». Consulté le: 5 juin 2026. [En ligne]. Disponible sur: https://www.kaggle.com/general/449169
[21] H. He, Y. Bai, E. A. Garcia, et S. Li, « ADASYN: Adaptive synthetic sampling approach for imbalanced learning », in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, Ieee, 2008, p. 1322-1328. Consulté le: 6 juin 2026. [En ligne]. Disponible sur: https://ieeexplore.ieee.org/abstract/document/4633969/
[22] G. E. Batista, A. L. Bazzan, et M. C. Monard, « Balancing training data for automated annotation of keywords: a case study. », *Wob*, vol. 3, p. 10-18, 2003.
[23] I. Tomek, « Two Modifications of CNN », *IEEE Trans. Syst. Man Cybern.*, vol. SMC-6, nº 11, p. 769-772, nov. 1976, doi: 10.1109/TSMC.1976.4309452.
[24] D. L. Wilson, « Asymptotic properties of nearest neighbor rules using edited data », *IEEE Trans. Syst. Man Cybern.*, nº 3, p. 408-421, 2007.
[25] C. Shorten et T. M. Khoshgoftaar, « A survey on Image Data Augmentation for Deep Learning », *J. Big Data*, vol. 6, nº 1, p. 60, juill. 2019, doi: 10.1186/s40537-019-0197-0.
[26] I. J. Goodfellow *et al.*, « Generative Adversarial Networks », 10 juin 2014, *arXiv*: arXiv:1406.2661. doi: 10.48550/arXiv.1406.2661.
[27] X. Yi, E. Walia, et P. Babyn, « Generative adversarial network in medical imaging: A review », *Med. Image Anal.*, vol. 58, p. 101552, 2019.
[28] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, et F. Herrera, *Learning from Imbalanced Data Sets*. Cham: Springer International Publishing, 2018. doi: 10.1007/978-3-319-98074-4.
[29] « Cross-Entropy Loss », AI Wiki. Consulté le: 6 juin 2026. [En ligne]. Disponible sur: https://aiwiki.ai/wiki/cross_entropy_loss
[30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, et P. Dollár, « Focal loss for dense object detection », in *Proceedings of the IEEE international conference on computer vision*, 2017, p. 2980-2988. Consulté le: 7 juin 2026. [En ligne]. Disponible sur: http://openaccess.thecvf.com/content_iccv_2017/html/Lin_Focal_Loss_for_ICCV_2017_paper.html
[31] C. N. Silla et A. A. Freitas, « A survey of hierarchical classification across different application domains », *Data Min. Knowl. Discov.*, vol. 22, nº 1-2, p. 31-72, janv. 2011, doi: 10.1007/s10618-010-0175-9.
[32] E. H. I. Eliwa, « Enhancing Skin Cancer Diagnosis Through Fine-Tuning of Pretrained Models: A Two-Phase Transfer Learning Approach », doi: 10.1155/ijbc/4362941.
[33] A. M. IESC, « Artificial Intelligence, Machine Learning, and Deep Learning: Same context, Different concepts », Master Intelligence Economique et Stratégies Compétitives. Consulté le: 7 juin 2026. [En ligne]. Disponible sur: https://master-iesc-angers.com/artificial-intelligence-machine-learning-and-deep-learning-same-context-different-concepts/
[34] F. Jiang *et al.*, « Artificial intelligence in healthcare: past, present and future », *Stroke Vasc. Neurol.*, vol. 2, nº 4, 2017, Consulté le: 7 juin 2026. [En ligne]. Disponible sur: https://svn.bmj.com/Content/2/4/230.abstract
[35] L. Alzubaidi *et al.*, « A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications », *J. Big Data*, vol. 10, nº 1, p. 46, avr. 2023, doi: 10.1186/s40537-023-00727-2.

[36] M. BENZIOUCHE et A. HAOUARI, « Détection et classification de l'arthrose par l'apprentissage profond », Université Aboubakr Belkaïd – Tlemcen –.
[37] K. Simonyan et A. Zisserman, « Very Deep Convolutional Networks for Large-Scale Image Recognition », 10 avril 2015, *arXiv*: arXiv:1409.1556. doi: 10.48550/arXiv.1409.1556.
[38] S. Ioffe et C. Szegedy, « Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift », 2 mars 2015, *arXiv*: arXiv:1502.03167. doi: 10.48550/arXiv.1502.03167.
[39] D. P. Kingma et J. Ba, « Adam: A Method for Stochastic Optimization », 30 janvier 2017, *arXiv*: arXiv:1412.6980. doi: 10.48550/arXiv.1412.6980.
[40] M. Tan et Q. V. Le, « EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks », 11 septembre 2020, *arXiv*: arXiv:1905.11946. doi: 10.48550/arXiv.1905.11946.
[41] K. He, X. Zhang, S. Ren, et J. Sun, « Deep residual learning for image recognition », in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, p. 770-778. Consulté le: 10 juin 2026. [En ligne]. Disponible sur: http://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html
[42] G. Huang, Z. Liu, L. Van Der Maaten, et K. Q. Weinberger, « Densely connected convolutional networks », in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, p. 4700-4708. Consulté le: 10 juin 2026. [En ligne]. Disponible sur: http://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.html
[43] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, et Z. Wojna, « Rethinking the inception architecture for computer vision », in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, p. 2818-2826. Consulté le: 10 juin 2026. [En ligne]. Disponible sur: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.html
[44] « Fig. 3. VGG19 architecture [24]. », ResearchGate. Consulté le: 10 juin 2026. [En ligne]. Disponible sur: https://www.researchgate.net/figure/GG19-architecture-24_fig3_376180056
[45] F. Chollet, « Xception: Deep learning with depthwise separable convolutions », in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, p. 1251-1258. Consulté le: 10 juin 2026. [En ligne]. Disponible sur: http://openaccess.thecvf.com/content_cvpr_2017/html/Chollet_Xception_Deep_Learning_CVPR_2017_paper.html
[46] « Architecture of the Xception deep CNN model », ResearchGate. Consulté le: 10 juin 2026. [En ligne]. Disponible sur: https://www.researchgate.net/figure/Architecture-of-the-Xception-deep-CNN-model_fig2_351371226
[47] Pitpitt, « Courbe précision-rappel », DataFranca. Consulté le: 15 juin 2026. [En ligne]. Disponible sur: https://datafranca.org/wiki/Courbe_pr%C3%A9cision-rappel
[48] « DCA-stratégie : Guide Pratique 10e Édition | PDF », Scribd. Consulté le: 15 juin 2026. [En ligne]. Disponible sur: https://fr.scribd.com/document/679743401/Dca-Version-10-3
[49] C. Rudin, « Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead », *Nat. Mach. Intell.*, vol. 1, n° 5, p. 206-215, 2019.
[50] N. A. Smuha, « Regulation 2024/1689 of the Eur. Parl. & Council of June 13, 2024 (eu artificial intelligence act) », *Int. Leg. Mater.*, vol. 64, n° 5, p. 1234-1381, 2025.
[51] S. Zhu, M. Gilbert, I. Chetty, et F. Siddiqui, « The 2021 landscape of FDA-approved artificial intelligence/machine learning-enabled medical devices: an analysis of the characteristics and intended use », *Int. J. Med. Inf.*, vol. 165, p. 104828, 2022.

[52] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, et E. K. Oermann, « Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study », *PLOS Med.*, vol. 15, nº 11, p. e1002683, nov. 2018, doi: 10.1371/journal.pmed.1002683.
[53] E. Tjoa et C. Guan, « A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI », *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, nº 11, p. 4793-4813, nov. 2021, doi: 10.1109/TNNLS.2020.3027314.
[54] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, et H. Müller, « Causability and explainability of artificial intelligence in medicine », *WIREs Data Min. Knowl. Discov.*, vol. 9, nº 4, p. e1312, juill. 2019, doi: 10.1002/widm.1312.
[55] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, et B. Kim, « Sanity Checks for Saliency Maps », in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2018. Consulté le: 10 juin 2026. [En ligne]. Disponible sur: https://proceedings.neurips.cc/paper_files/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html
[56] M. D. Zeiler et R. Fergus, « Visualizing and Understanding Convolutional Networks », in *Computer Vision – ECCV 2014*, vol. 8689, D. Fleet, T. Pajdla, B. Schiele, et T. Tuytelaars, Éd., in Lecture Notes in Computer Science, vol. 8689. , Cham: Springer International Publishing, 2014, p. 818-833. doi: 10.1007/978-3-319-10590-1_53.
[57] vera.damerjian.phd@gmail.com, « Dice Coefficient vs. IoU in Medical Image Segmentation - VeraDP ». Consulté le: 11 juin 2026. [En ligne]. Disponible sur: https://veradp-ai.com/iou-vs-dice/
[58] N. Arun *et al.*, « Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging », *Radiol. Artif. Intell.*, vol. 3, nº 6, p. e200267, nov. 2021, doi: 10.1148/ryai.2021200267.
[59] I. Bouzidi, R. Makhloufi, et M. Bouakkaz, « La Détection De La Maladie D'alzheimer Avec L'apprentissage Profond ». Consulté le: 15 juin 2026. [En ligne]. Disponible sur: https://www.theses-algerie.com
[60] H. Ghaffari, H. Tavakoli, et G. P. Jahromi, « Deep transfer learning–based fully automated detection and classification of Alzheimer's disease on brain MRI », *Br. J. Radiol.*, vol. 95, nº 1136, p. 20211253, juin 2022, doi: 10.1259/bjr.20211253.
[61] M. Femmam, S. Femmam, M. E. Fareh, O. A. Senni, et A. Ferhani, « Deep Learning-Based Classification and Diagnosis of Alzheimer's & Dementia Using Multi-scale Feature Extraction from Baseline MRI Scans », *J. Image Graph.*, vol. 12, nº 4, 2024, doi: 10.18178/joig.12.4.427-436.
[62] S. B. Shahid *et al.*, « Novel deep learning for multi-class classification of Alzheimer's in disability using MRI datasets », *Front. Bioinforma.*, vol. 5, p. 1567219, août 2025, doi: 10.3389/fbinf.2025.1567219.
[63] M. Saim, « Development of a diagnostic aid system for neurodegenerative pathologies: Application to the detection of Alzheimer's disease », Abou Bekr Belkaid University-Tlemcen-Faculty of Technology, Tlemcen, 2025.
[64] A. Loddo, S. Buttau, et C. Di Ruberto, « Deep learning based pipelines for Alzheimer's disease diagnosis: A comparative study and a novel deep-ensemble method », *Comput. Biol. Med.*, vol. 141, p. 105032, févr. 2022, doi: 10.1016/j.compbiomed.2021.105032.
[65] H. Tufail, A. Ahad, I. Puspitasari, I. Shayea, P. J. Coelho, et I. M. Pires, « Deep Learning in Smart Healthcare: A GAN-based Approach for Imbalanced Alzheimer's Disease Classification », *Procedia Comput. Sci.*, vol. 241, p. 146-153, janv. 2024, doi: 10.1016/j.procs.2024.08.021.

[66] M. Ryspayeva et O. Salykova, « Multi-domain synthetic medical image generation and dataset balancing with DGAN-WP-TL », *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.*, vol. 13, nº 1, p. 2556687, déc. 2025, doi: 10.1080/21681163.2025.2556687.
[67] M. Askarizade, « Efficient Alzheimer's disease classification using transfer learning with efficientNetV2S », *J. Emerg. Eng. Technol.*, vol. Volume 1, nº Issue 1, p. 1-12, janv. 2026.
[68] L. Lazli, F. Cheriet, et M. Boukadoum, « Multiclass prediction of Alzheimer's disease using balanced multimodal data and deep ensemble learning », *Biomed. Signal Process. Control*, vol. 114, p. 109026, avr. 2026, doi: 10.1016/j.bspc.2025.109026.
[69] la rédaction, « Définition | Python : qu'est-ce que c'est ? | Futura tech », Futura. Consulté le: 14 juin 2026. [En ligne]. Disponible sur: https://www.futura-sciences.com/tech/definitions/informatique-python-19349/
[70] « Qu'est-ce que Kaggle ? » Consulté le: 14 juin 2026. [En ligne]. Disponible sur: https://www.datacamp.com/blog/what-is-kaggle
[71] S. Kumar et S. Shastri, « Alzheimer MRI Preprocessed Dataset », vol. 1, juin 2025, doi: 10.17632/3r8hw8wmmk.1.
[72] W. Karim et M. Jerin, « Brain MRI Images (Tumor and No Tumor) », vol. 1, déc. 2025, doi: 10.17632/h5gjxxfmrs.1.
[73] M. I. K. Hira *et al.*, « Brain Tumor MRI Dataset (Glioma, Meningioma, Pituitary, No Tumor) », vol. 6, avr. 2026, doi: 10.17632/zwr4ntf94j.6.
[74] « Osteoarthritis Prediction ». Consulté le: 14 juin 2026. [En ligne]. Disponible sur: https://www.kaggle.com/datasets/farjanakabirsamanta/osteoarthritis-prediction