Classification d'images IRM cérébrales fondée sur un mécanisme d'attention invariant à l'échelle

par

Talía VÁZQUEZ ROMAGUERA

MÉMOIRE PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
COMME EXIGENCE PARTIELLE À L'OBTENTION DE LA MAÎTRISE
AVEC MÉMOIRE
M. Sc. A.

MONTRÉAL, LE 11 DÉCEMBRE 2025

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

PRÉSENTATION DU JURY

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE:

M. Matthew Toews, directeur de mémoire
Département de génie des systèmes, École de technologie supérieure

M. Jose Dolz, président du jury
Département de génie logiciel et des technologies de l'information, École de technologie supérieure

M. Christian Desrosiers, examinateur externe
Département de génie logiciel et des technologies de l'information, École de technologie supérieure

IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 27 NOVEMBRE 2025

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

AVANT-PROPOS

Ce mémoire est le fruit d'un parcours à la croisée de l'ingénierie, de l'intelligence artificielle et des sciences biomédicales. Il s'inscrit, dans le cadre de la classification fine d'images cérébrales, dans un effort visant à mieux relier la structure anatomique du cerveau aux représentations apprises par les réseaux de neurones profonds. Face à la complexité croissante des modèles, la neuroimagerie est aujourd'hui confrontée à un double défi : tirer parti de leur puissance prédictive tout en préservant une interprétation fidèle à la réalité biologique. Ce travail propose une voie intermédiaire entre les approches interprétables fondées sur des descripteurs locaux et les architectures profondes peu interprétables apprises de bout en bout, au moyen d'une architecture hybride alliant la transparence des premières et la performance des secondes.

REMERCIEMENTS

Je souhaite tout d'abord exprimer ma profonde gratitude à mon directeur de recherche, le Professeur Matthew Toews, pour m'avoir donné l'opportunité de poursuivre mes études de maîtrise au sein de son laboratoire et pour m'avoir guidée tout au long de ce travail. Je le remercie sincèrement pour sa patience, sa disponibilité, son dévouement, l'aide financière accordé dans le cadre de cette maîtrise, et tous les efforts qu'il a consacrés pour m'enseigner et m'accompagner dans cette recherche.

J'adresse également mes remerciements à l'ensemble de mes collègues du laboratoire LIVIA, tout particulièrement à Liam, pour sa bienveillance et le partage généreux de ses connaissances.

Enfin, j'exprime ma reconnaissance la plus profonde à ma famille : mon époux, mes parents, ma sœur, mon beau-frère et mes magnifiques nièces. Votre amour, votre soutien et votre patience indéfectible ont été une source inestimable de motivation, de force et d'inspiration tout au long de ces années d'études.

Classification d'images IRM cérébrales fondée sur un mécanisme d'attention invariant à l'échelle

Talía VÁZQUEZ ROMAGUERA

RÉSUMÉ

Ce mémoire aborde la classification fine des IRM structurelles du cerveau tout en préservant leur interprétabilité anatomique. Quatre familles de modèles complémentaires sont comparées : (1) un pipeline hiérarchique SIFT–MLP qui classe les points clés 3D et agrège les votes au niveau du sujet ; (2) des CNN 3D entraînés de bout en bout sur des volumes d'IRM pondérées en T1 ; (3) un CNN peu profond exploitant des cartes de descripteurs SIFT 3D ; et (4) un modèle hybride 3D-SIFT–Attention qui combine la morphologie locale invariante d'échelle et le contexte global au moyen d'une attention croisée. De plus, nous proposons une méthode d'Integrated-Gradients–Guided Keypoint Attribution (IG-KPA), qui aligne les attributions sur les points clés SIFT afin d'améliorer l'explicabilité. Le modèle hybride améliore la précision de classification dans tous les cas, par rapport aux réseaux standards ou aux points clés seuls, sur les jeux de données ADNI, HCP et OASIS, et pour les tâches de classification du sexe, de la maladie d'Alzheimer et de l'âge. Une exactitude de 94 % est obtenue pour la classification du sexe à partir de données rigidement recalées ; toutefois, les différences de taille entre hommes et femmes constituent un facteur de confusion majeur. Après correction des effets de taille par recalage déformable, la précision est limitée à environ 92 %, en accord avec la littérature, ce qui suggère que l'IRM d'environ 8 % des individus ressemble davantage à celle du sexe opposé. Les analyses d'interprétabilité localisent systématiquement les signaux dans des régions neuro-anatomiquement plausibles : les territoires thalamo-capsulaires profonds et périventriculaires pour la classification du sexe, et principalement des signatures ventriculaires pour l'âge et la maladie d'Alzheimer. En général, les résultats mettent en évidence l'intérêt des mécanismes d'attention qui intègrent des indices locaux invariants d'échelle avec un contexte spatial global pour les tâches de classification fine d'images cérébrales.

Mots-clés: IRM cérébrale, caractéristiques invariantes d'échelle (SIFT), mécanisme d'attention croisée, interprétabilité

Brain MRI Classification using a Scale-Invariant Keypoint Attention Mechanism

Talía VÁZQUEZ ROMAGUERA

ABSTRACT

This thesis addresses the fine-grained classification of structural brain MRI while maintaining anatomical interpretability. We compare four complementary model families : (1) a hierarchical SIFT–MLP pipeline that classifies 3D keypoints and aggregates subject-level votes ; (2) end-to-end 3D CNNs trained on native T1-weighted MRI volumes ; (3) a shallow CNN operating on 3D SIFT descriptor maps ; and (4) a hybrid 3D-SIFT–Attention model that fuses scale-invariant local morphology with global context via cross-attention. Furthermore, we propose an Integrated-Gradients–Guided Keypoint Attribution (IG-KPA) method, which aligns attributions with SIFT keypoints to enhance explainability. Our hybrid model improves classification accuracy in all cases, compared to standard networks or keypoints alone, for ADNI, HCP, OASIS datasets and sex, Alzheimer's and age classification. State-of-the-art sex classification accuracy of 94% is obtained from rigidly aligned data, however size differences between males and females is a significant confound. After removing size differences via deformable registration, sex classification accuracy appears limited to 92%, consistent with other work, indicating that the brain MRIs of 8% of people appear more similar to those of the opposite sex. Moreover, interpretability analyses consistently localize signal to neuroanatomically plausible regions, with bilateral deep thalamo-capsular and periventricular territories for sex classification, and predominantly ventricular signatures for age and Alzheimer's disease. Overall, the results underscore the advantage of attention mechanisms that merge stable, scale-invariant local cues with broader spatial context in fine-grained brain image classification tasks.

Keywords: Brain MRI, Scale-invariant features (SIFT), Cross-attention, Interpretability

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

IRM	Imagerie par résonance magnétique
TIV	Volume intracrânien total
DoG	Différences de Gaussiennes
SIFT	*Scale-Invariant Feature Transform*
HOG	Histogramme des gradients orientés
kNN	Méthode des k plus proches voisins
SVM	Machine à vecteurs de support
PCA	Analyse en composantes principales
FBM	Morphométrie fondée sur les caractéristiques
MLP	Perceptron multicouche
MAE	Erreur absolue moyenne
EER	Taux d'erreur égal
ROI	Région d'intérêt
SHAP	Explications additives de type Shapley
IG	Gradients intégrés
GradCAM	Cartographie d'activation pondérée par les gradients
HCP	Human Connectome Project
OASIS	Open Access Series of Imaging Studies
ADNI	Alzheimer's Disease Neuroimaging Initiative
AD	Maladie d'Alzheimer

INTRODUCTION

L'étude du cerveau humain repose aujourd'hui largement sur l'imagerie médicale, et en particulier sur l'Imagerie par Résonance Magnétique (IRM) grâce à sa capacité à générer des images de haute résolution entre les tissus mous, sans recourir à des radiations ionisantes (De Pietro *et al.*, 2024). En particulier, l'IRM permet de visualiser le cerveau selon plusieurs coupes (voir Figure 0.1) et d'étudier des régions anatomiques fines (hippocampe, cortex, noyaux profonds). Ce pouvoir de résolution anatomique et la richesse des signaux qu'elle fournit en font un support privilégié pour l'analyse quantitative et l'application de méthodes d'intelligence artificielle.

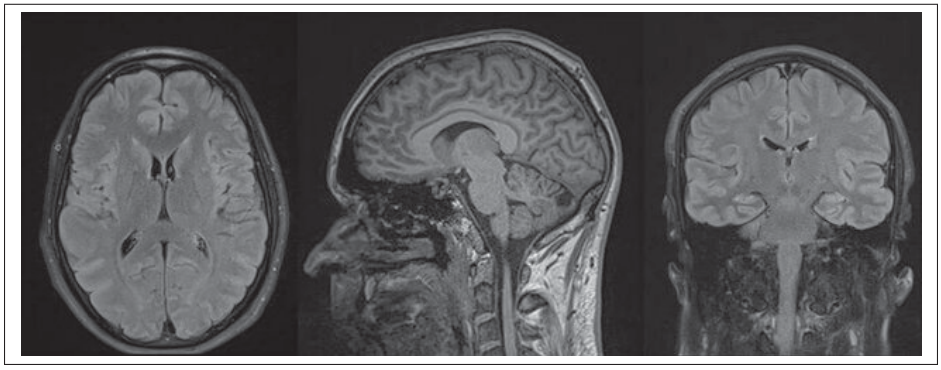

Figure 0.1 IRM cérébrale normale — coupes axiale, sagittale et coronale (de gauche à droite)
Tirée de Dimitrov *et al.*(2018)

Cependant, exploiter les images IRM cérébrales pour des tâches de classification est un défi de taille, notamment lorsqu'il s'agit de une classification à grain fin (*fine-grained image classification*). Dans une tâche classique, les classes à distinguer sont très différentes entre elles. Par exemple, distinguer un organe d'un autre, ou un tissu sain vs pathologique bien visible. En revanche, dans la classification à grain fin, les différences entre classes sont faibles, subtiles, souvent contextuelles, et peuvent être masquées par du bruit, des variations d'acquisition ou des différences anatomiques individuelles. Le modèle doit apprendre des signaux discriminants délicats (textures, variations locales, microstructures) tout en étant robuste aux variations globales

(Ren, Li, An, Zhang & Sun, 2024). Cela constitue une difficulté majeure dans les applications médicales, où les classes (stades de maladie, sous-types, différences intra-groupe) sont proches.

Un exemple représentatif de tâche fine-grained en neuroimagerie est l'analyse du dimorphisme sexuel cérébral — c'est-à-dire l'étude des différences structurelles entre les cerveaux d'individus de sexe masculin et féminin (McCarthy, Arnold, Ball, Blaustein & De Vries, 2012). Des différences globales sont régulièrement observées, par example un volume cérébral total en moyenne plus élevé chez les hommes (Ruigrok *et al.*, 2014) et, chez les femmes, une proportion plus importante de substance grise dans plusieurs régions corticales (Lotze *et al.*, 2019). Toutefois, une grande partie de ces effets s'explique par la taille globale et lorsqu'on controlle ce facteur, les différences restantes sont très modestes (Eliot, Ahmed, Khan & Patel, 2021) et leur détection exige des modèles sensibles, robustes et bien interprétables. Sur le plan neuroscientifique, comprendre ces différences, même faibles, peut éclairer les bases biologiques des variations cognitives ou de la vulnérabilité à certaines pathologies selon le sexe (Ritchie *et al.*, 2018). Sur le plan méthodologique, le sexe est une variable « de référence » stable que l'on peut tenter de prédire avant d'étendre les approches à des tâches plus délicates (prédiction de l'âge, détection de maladies) dans le même cadre. Toutefois, notre investigation se limite au sexe binaire auto-déclaré et ne tient pas compte des personnes transgenres ou intersexes, qui ne s'inscrivent pas nécessairement dans les catégories binaires homme/femme.

Les réseaux de neurones convolutionnels 3D (3D-CNN) offrent un potentiel considérable pour traiter directement les volumes IRM et extraire automatiquement des caractéristiques discriminantes. Ils peuvent capturer des motifs spatiaux complexes dans les trois dimensions, surpassant souvent les approches manuelles (Brueggeman, Thomas, Koomar, Hoskins & Michaelson, 2020). Toutefois, ils sont généralement perçus comme des boîtes noires : leurs décisions sont difficiles à interpréter, ce qui constitue un obstacle majeur à leur adoption en milieu clinique ou neuroscientifique (Borys *et al.*, 2023). En imagerie médicale, il est essentiel que les modèles

soient non seulement performants mais aussi explicables, afin que les experts puissent vérifier que les décisions reposent sur des régions anatomiquement plausibles, et non sur des artefacts ou des biais (van der Velden, Kuijf, Gilhuijs & Viergever, 2022).

La contribution de cette mémoire est de proposer un cadre méthodologique unifié pour la classification d'IRM cérébrales qui concilie interprétabilité et performance. Le focus principal sera la classification du sexe, mais nous envisageons aussi d'étendre cette approche à d'autres tâches comme la prédiction de l'âge ou la détection de pathologies cérébrales. La démarche adoptée suit une progression méthodologique claire : nous commençons par tester des méthodes basiques basées sur des descripteurs interprétables, puis nous passons à des approches pures avec des CNN sur les images brutes, en essayant d'interpréter ce que le réseau apprend. Enfin, nous proposons une approche hybride qui capitalise sur la nature interprétable des descripteurs manuels tout en bénéficiant de la puissance des CNN, avec une attention croisée ou une fusion intelligente pour obtenir un compromis interprétabilité / performance.

L'objectif global de ce mémoire est de concevoir et d'évaluer un cadre hybride d'apprentissage profond pour la classification fine-grained d'IRM cérébrales, avec une exigence forte d'explicabilité. Nous ambitionnons de (i) developer des approches fondées sur des descripteurs, (ii) analyser les limites des CNN volumétriques, et (iii) proposer une architecture hybride unifiant les deux paradigmes. Nous souhaitons démontrer que ce type d'architecture peut (a) offrir des performances compétitives sur des tâches de classification (sexe, âge, pathologie), et (b) fournir des explications plus transparentes et fiables, par l'attribution de contributions aux descripteurs et aux voxels ou régions cérébrales.

Ce mémoire s'organise de la façon suivante. Le Chapitre 1 Travaux connexes présente une revue des approches existantes : des méthodes manuelles aux CNN profonds, en passant par les stratégies d'explicabilité en imagerie médicale.Le Chapitre 2, intitulé Méthodologie, décrit les prétraitements, les méthodes mises en œuvre, ainsi que la technique d'explicabilité proposée. Le

Chapitre 3 Résultats et analyse expérimentale expose les résultats quantitatifs sur les différentes tâches, puis les analyses qualitatives. Enfin, la Conclusion synthétise les apports du mémoire et ouvre sur des développements ultérieurs.

CHAPITRE 1

TRAVAUX CONNEXES

L'étude de la classification fine en neuro-imagerie — notamment la différenciation du sexe, de l'âge et de pathologies neurodégénératives à partir d'IRM structurelles — a donné lieu à un large éventail d'approches. Celles-ci se répartissent en quatre familles : (i) les méthodes fondées sur des caractéristiques locales, (ii) les méthodes basées sur des caractéristiques apprises, (iii) les approches hybrides combinant caractéristiques locales et apprises, et (iv) les méthodes dédiées à l'explicabilité et à l'interprétabilité des modèles. Nous en proposons ci-après une synthèse, en soulignant les atouts et les limites propres à chacune.

1.1 Méthodes fondées sur des caractéristiques locales

Les méthodes fondées sur l'extraction de caractéristiques locales constituent une approche historique et fondamentale en vision par ordinateur et en analyse d'images médicales. Elles reposent sur l'idée que certaines régions d'une image, caractérisées par des variations marquées d'intensité ou de texture, contiennent une information particulièrement discriminante.

Une caractéristique locale comporte deux éléments : un détecteur et un descripteur (Li & Allinson, 2008). Le détecteur vise à localiser des points 3-D répétables, riches en structure de gradient. Les premiers travaux s'appuyaient sur des détecteurs de coins (Morevec, 1977). Harris, Stephens *et al.* (1988) ont ensuite amélioré la stabilité au voisinage des bords, avant que Lowe (2004) n'introduise *Scale-Invariant Feature Transform* (SIFT), une formulation en espace d'échelles (*scale-space*) assurant une détection cohérente malgré les variations d'échelle. Du côté des descripteurs, l'objectif est de résumer l'apparence locale en un vecteur compact, distinctif et robuste. Parmi les méthodes de référence figurent les histogrammes d'orientations de gradients (HOG), qui agrègent les orientations pondérées par sous-régions et offrent une invariance éprouvée à l'échelle, à la rotation et au contraste (Lowe, 2004; Dalal & Triggs, 2005). Plus précisément, SIFT détecte des extrémums dans un espace d'échelles construit par différence de gaussiennes (DoG), assigne une orientation dominante locale, puis encode chaque point via

des histogrammes d'orientations du gradient sur des sous-régions (Figure 1.1), produisant un descripteur de longueur fixe (généralement 128).

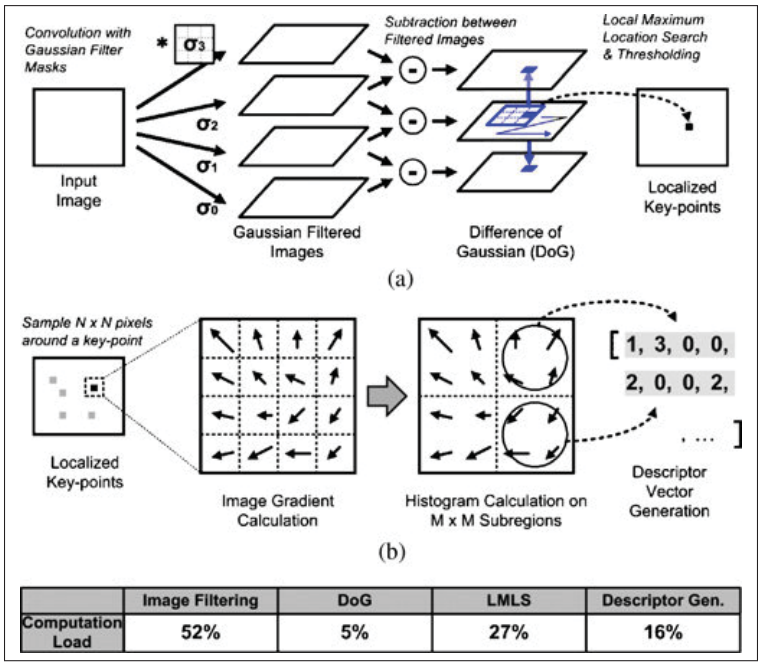

Figure 1.1 SIFT : (a) détection et localisation des points d'intérêt ; (b) description des points par agrégation des gradients en un descripteur de longueur fixe
Tirée de Kim *et al.*(2009)

L'extension de ces concepts à l'imagerie tridimensionnelle a permis d'appliquer le même principe aux volumes IRM, en généralisant la construction de l'espace multi-échelles et le calcul des histogrammes de gradients au domaine volumique (Toews & Wells III, 2013; Rister, Yi, Shivakumar & Rubin, 2017). Ces adaptations ouvrent la voie à l'utilisation de points clés dans l'analyse du cerveau humain à partir d'IRM structurelles.

Les pipelines fondés sur des caractéristiques locales incluent généralement un prétraitement (par exemple, extraction du crâne, recalage rigide ou déformable), suivi de la détection de points d'intérêt et du calcul de descripteurs. Deux voies de conception sont ensuite possibles : (i) encoder et agréger les descripteurs en une représentation de longueur fixe pour entraîner des modèles d'apprentissage automatique classiques ; ou (ii) recourir à l'appariement de caractéristiques,

où l'évaluation d'un sujet repose sur la quantité et la qualité des correspondances entre points d'intérêt, souvent résumées par des statistiques de rang.

Dans la première approche, Yuan, Chen, Zeng, Wang & Hu (2015) proposent le descripteur 3D-WHGO, fondé sur des histogrammes d'orientations de gradients pondérés par la magnitude. Après l'apprentissage d'un dictionnaire via *k*-means, une pyramide spatiale 3D multi-résolution permet d'agréger les occurrences encodées. Le vecteur final est ensuite classé par une machine à vecteurs de support (SVM). L'approche est évaluée sur des IRM T1 issues du *1000 Functional Connectomes Project* et atteint des précisions de l'ordre de 90 %, et jusqu'à 95 % selon les sites et paramètres.

À l'inverse, Chauvin, Kumar, Desrosiers, Wells & Toews (2021) illustrent la voie fondée sur l'appariement de caractéristiques (Figure 1.1). Ils extraient des points d'intérêt 3D SIFT-Rank, établissent des correspondances par *k*-plus proches voisins (kNN), puis évaluent une distance de Jaccard souple (*soft-Jaccard distance*) sur les ensembles correspondants. Testée sur HCP Q4, leur méthode atteint une AUC de 0,97 pour la classification du sexe en combinant apparence et géométrie, contre 0,93 pour l'apparence seule.

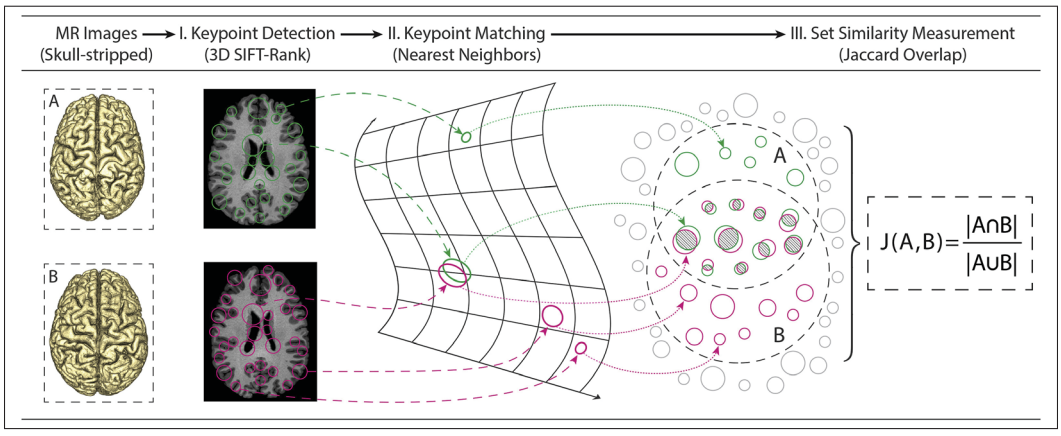

Figure 1.2 Schéma général d'un pipeline basé sur l'appariement de points d'intérêt pour la comparaison inter-sujets. .
Tirée de Chauvin *et al.*(2020)

Enfin, Toews, Romaguera, Wells & Makris (2025) prolongent ce paradigme d'appariement fondé sur des points SIFT et l'indice de Jaccard. Leur méthode repose sur un score continu de similarité inter-sujets mesurant le recouvrement des ensembles de caractéristiques. Appliquée à des IRM T1 après recalage rigide, elle atteint une exactitude d'environ 0,92 dans la cohorte HCP, avec un taux d'erreur d'environ 8 % selon le type de recalage (rigide ou non linéaire). L'analyse met en évidence une paire de caractéristiques thalamiques bilatérales, illustrée à la Figure 1.1, détectée dans 97% des sujets, mais souligne que la performance globale émerge principalement d'une constellation distribuée de traits faibles, plutôt que d'un petit nombre de régions dominantes. Ces résultats mettent en évidence que des descripteurs strictement locaux, contiennent une information suffisante pour capturer les différences sexuelles subtiles dans la morphologie cérébrale.

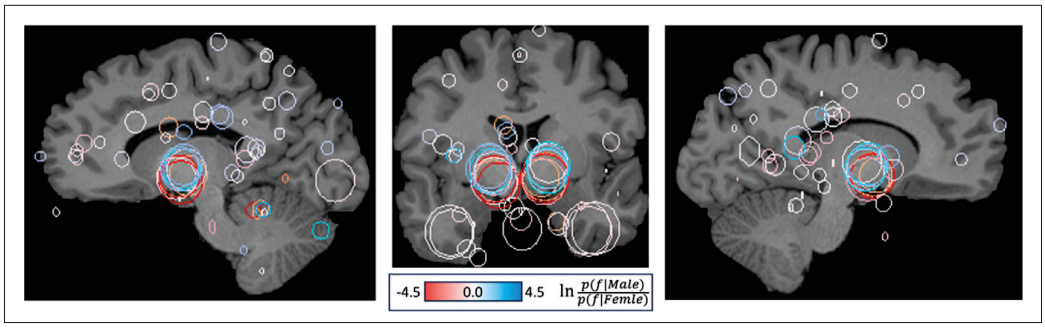

Figure 1.3 Visualisation de la variation des caractéristiques thalamiques pour les cinq femmes et les cinq hommes les plus nettement classés sur une IRM cérébrale représentative. Les cercles indiquent la localisation et l'échelle des caractéristiques, et la couleur encode la log-vraisemblance du sexe associé à chaque caractéristique : masculin (bleu) vs féminin (orange)
Tirée de Toews *et al.*(2025)

Au-delà de la classification du sexe, plusieurs travaux se sont orientés vers la prédiction de l'âge et la détection de maladies neurodégénératives à partir de caractéristiques locales. À titre d'exemple, Guo *et al.* (2024) proposent un pipeline fondé sur des descripteurs radiomiques extraits au niveau des des régions d'intérêt (ROI) sur des IRM pondérées T1, combiné à une recherche AutoML limitée à *XGBoost, Random Forest et ExtraTrees*, puis à une interprétabilité par *SHapley Additive exPlanations (SHAP)*. Leur analyse met en évidence l'hippocampe et le

gyrus parahippocampique comme régions les plus informatives pour la prédiction de l'âge, et atteint une faible erreur sur ADNI (XGBoost, MAE ≈ 1,51 an en validation croisée répétée). Ces résultats suggèrent que l'usage de descripteurs radiomiques locaux (au niveau des ROI), combiné à des attributions SHAP, renforce l'interprétabilité en reliant les prédictions du modèle à des structures anatomiques spécifiques.

Pour les applications centrées sur la maladie d'Alzheimer, Toews *et al.* (2010) proposent *Feature-Based Morphometry* (FBM), une méthode de classification fondée sur des caractéristiques locales invariantes d'échelle. Des points d'intérêt 3D SIFT, extraits sur de nombreux sujets, sont regroupés selon leur géométrie et leur apparence en caractéristiques modèles (*clusters*). Chaque caractéristique modèle reçoit ensuite un rapport de vraisemblance AD vs NC assorti d'un contrôle du taux de fausses découvertes, et la classification d'un nouveau sujet s'obtient en agrégeant ces preuves locales. Au-delà de performances compétitives (p. ex., EER ≈ 0,80 sur OASIS pour des AD modérés), l'atout majeur réside dans l'interprétabilité. Chaque indice discriminant possède une localisation 3D, une échelle et un patron d'intensité concrets, ce qui permet d'indiquer quelles structures, où, à quelle fréquence et avec quelle force statistique elles s'associent à l'AD.

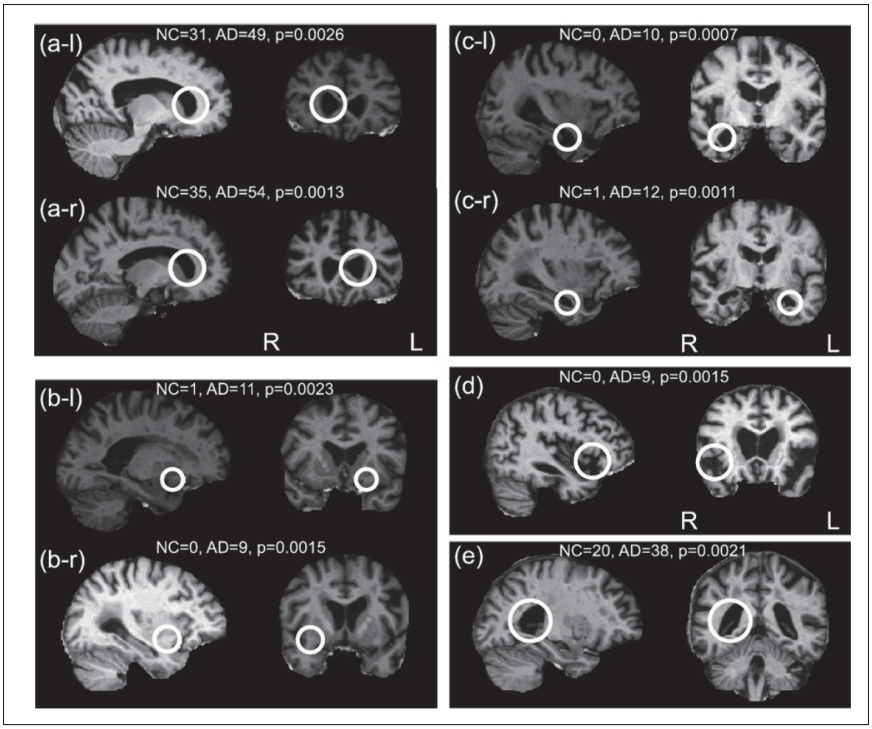

Figure 1.4 Caractéristiques significatives liées à la maladie d'Alzheimer (AD) identifiées par la FBM
Tirée de Toews *et al.*(2010)

Les Figures 1.4 et 1.5 l'illustrent clairement : les caractéristiques les plus significatives se concentrent dans des régions reconnues de l'AD—élargissement ventriculaire (y compris des cornes temporales), altérations temporales/sous-corticales et augmentation du LCR dans la scissure de Sylvius—tandis que les planches NC mettent en évidence des motifs complémentaires près du plan médian et un parenchyme préservé.

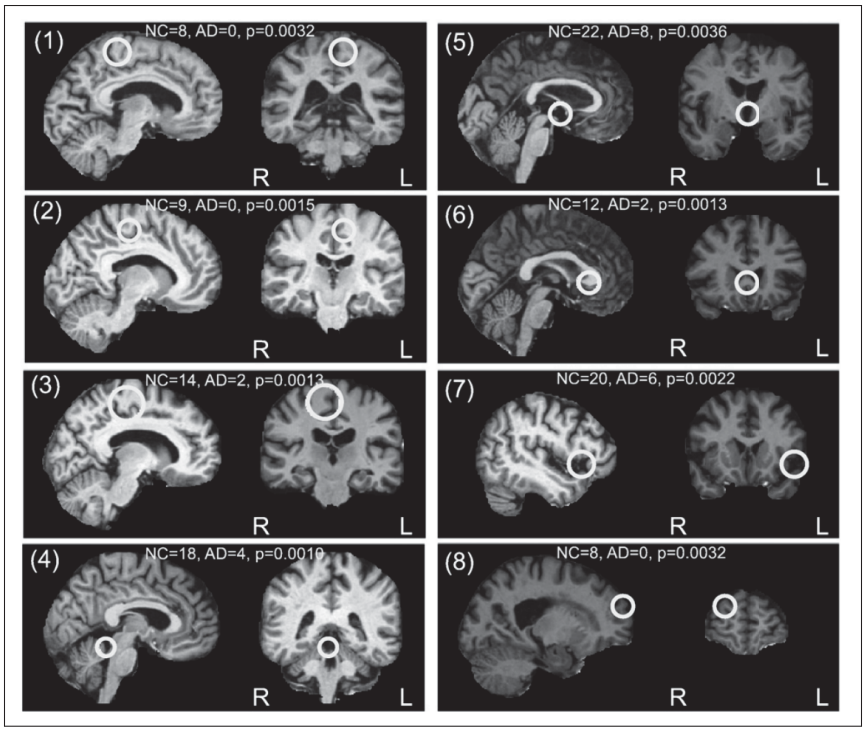

Figure 1.5 Caractéristiques significatives liées aux sujets témoins (NC) identifiées par la FBM
Tirée de Toews *et al.*(2010)

Dans la même veine, Li, Zhang, Song & Liu (2017) proposent une chaîne en deux étapes combinant la correspondance de points d'intérêt fondée sur SIFT et la modélisation de l'apparence locale via HOG. Ils détectent d'abord des points d'intérêt multi-échelles et retiennent un sous-ensemble représentatif. Des descripteurs HOG sont ensuite extraits autour de ces points ; des SVM spécifiques à chaque point sont entraînés, puis les scores sont agrégés pour classifier un nouveau sujet.La méthode est évaluée sur OASIS et PPMI et rapporte une exactitude maximale de 86,05%. Sur le plan de l'interprétabilité, les auteurs projettent les points à score élevé sur l'atlas AAL afin d'identifier les régions discriminantes, mettant en évidence des concentrations au niveau du thalamus, du lobe temporal, du cortex calcarin et du cingulum chez les sujets atteints d'AD.

Malgré les atouts des méthodes basées sur des caractéristiques locales — stabilité multi-échelle et traçabilité anatomique — les descripteurs restent conçus « à la main » et séparés du classifieur,

de sorte que leurs performances dépendent étroitement d'un prétraitement soigné ainsi que de la qualité et de la pertinence des caractéristiques extraites. Leur dépendance à une étape préalable d'ingénierie de variables limite en outre la modélisation de la complexité hiérarchique et du contexte global des images cérébrales, ce qui ouvre la voie aux approches d'apprentissage profond, capables de s'affranchir de cette contrainte en apprenant directement, de bout en bout, des représentations discriminantes à partir des données brutes.

1.2 Méthodes basées sur des caractéristiques apprises

L'avènement de l'apprentissage profond a transformé l'analyse d'images médicales en général, et l'imagerie cérébrale par IRM en particulier. Contrairement aux méthodes d'apprentissage automatique classiques, qui nécessitent l'extraction préalable de caractéristiques manuelles, les réseaux de neurones profonds sont capables d'apprendre directement des représentations discriminantes à partir des données brutes. Les réseaux de neurones convolutifs tridimensionnels (3D-CNN) représentent aujourd'hui l'architecture de référence pour l'analyse automatique de volumes IRM (Dorfner, Patel, Kalpathy-Cramer, Gerstner & Bridge, 2025). Ils étendent les réseaux 2D initialement développés pour la reconnaissance d'images (LeCun, Bottou, Bengio & Haffner, 2002) à des données volumétriques, en appliquant les convolutions dans les trois dimensions spatiales. Leur conception s'inspire directement des travaux pionniers de (Hubel & Wiesel, 1962), qui ont mis en évidence l'organisation hiérarchique du cortex visuel. Ces auteurs ont montré que les cellules simples répondent préférentiellement à des motifs élémentaires — tels que des bords ou des orientations spécifiques — tandis que les cellules complexes intègrent l'information de plusieurs cellules simples pour obtenir des réponses plus invariantes à la position (voir Figure 1.6). Ce principe de champs récepteurs hiérarchiques est au cœur du fonctionnement des CNN : chaque neurone traite une région locale de l'image afin d'extraire une caractéristique particulière. La taille du champ récepteur dépend du filtre et de la profondeur du réseau ; plus le réseau est profond, plus la zone d'intégration est étendue, permettant de combiner des motifs locaux en représentations globales.

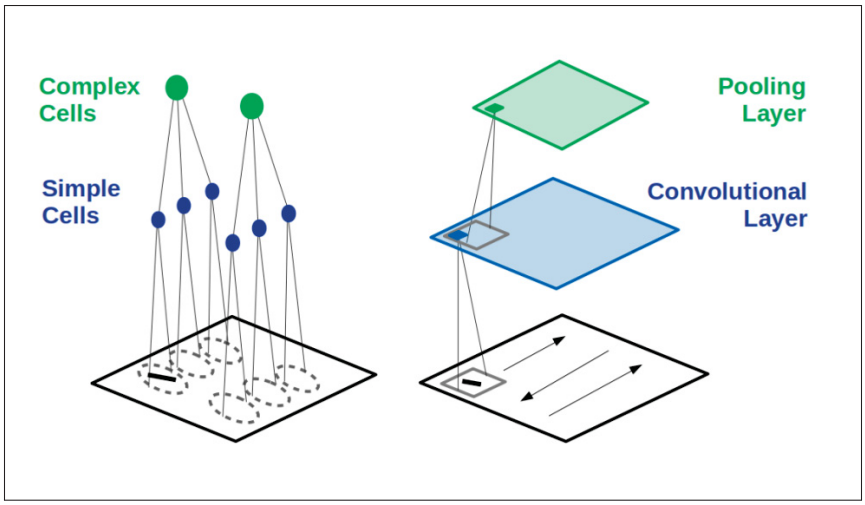

Figure 1.6 Relation entre cellules simples et complexes dans le cortex visuel et leur analogie avec les opérations de base d'un CNN (convolution et pooling).
Tirée de Lindsay(2021)

Grâce à cette hiérarchie de filtres appris, les 3D-CNN capturent aussi bien des motifs locaux (textures, contours, gradients régionaux) que des structures anatomiques de haut niveau (formes, volumes, asymétries). Brueggeman *et al.* (2020) ont montré que les CNN surpassent systématiquement les approches fondées sur des descripteurs manuels dans la classification du sexe à partir d'IRM cérébrales, confirmant que les représentations apprises automatiquement sont mieux adaptées pour capturer les différences subtiles liées au dimorphisme sexuel. Leurs performances élevées ont également motivé leur utilisation pour gérer la variabilité inter-sites. Par exemple, Yuan, Zeng, Shen, Liu & Hu (2018) ont proposé le *3D Deep Adding Network* (3D-DANet), intégrant plusieurs noyaux convolutifs de tailles différentes et une opération d'addition multi-échelle pour agréger l'information spatiale. Ce modèle a atteint une exactitude supérieure à 92 % sur plus de 6 000 IRM provenant de 61 centres, démontrant la faisabilité du deep learning sur de vastes ensembles multi-centres.

Cependant, ces modèles présentent aussi des limites. Leur dépendance à la taille des données pose problème dans le domaine médical, où les cohortes annotées sont souvent réduites, avec un risque de surapprentissage ou d'exploitation de biais, au détriment de l'apprentissage de caractéristiques réellement pertinentes pour la tâche. Dans la classification du sexe, un enjeu

majeur concerne le contrôle des facteurs confondants, en particulier le volume intracrânien total (TIV), dont l'effet peut masquer ou amplifier artificiellement des différences entre groupes. Ebel *et al.* (2023) ont montré que les performances de leur 3D-CNN (BraiNN) peuvent dépasser 91 % sur des données non appariées, mais chutent à environ 76 % lorsque les images sont appariées sur le volume intracrânien, révélant la sensibilité des modèles aux biais liés à la taille cérébrale. Par ailleurs, les réseaux profonds sont souvent perçus comme des « boîtes noires », leurs mécanismes internes restant difficiles à interpréter. Dans cette perspective, Dibaji, Ospel, Souza & Bento (2024) proposent une approche centrée sur la justice algorithmique et l'explicabilité. À partir de quatre ensembles de données (CC-359, OASIS-3, ADNI et CamCAN), ils entraînent un 3D-CNN avec un prétraitement minimal (recalage rigide et normalisation d'intensité) et atteignent 87 % d'exactitude pour la classification du sexe. Grâce à des cartes de saillance, ils identifient des contributions différenciées de régions supra- et infratentorielles et observent une réduction des biais lorsque les distributions de TIV se recouvrent davantage. Leur travail souligne la nécessité de modèles explicables et équitables, capables de maintenir des performances cohérentes entre sous-groupes démographiques.

Dans le cadre de la prédiction de l'âge, Cole *et al.* (2017) ont utilisé un CNN entraîné sur un large corpus d'IRM structurelles pour estimer l'« âge cérébral » d'un individu. Leur approche a montré que le modèle pouvait prédire l'âge chronologique avec une erreur moyenne inférieure à cinq ans, et que l'écart entre âge prédit et âge réel constituait un biomarqueur pertinent du vieillissement pathologique.

Les applications à la maladie d'Alzheimer et aux troubles neurodégénératifs ont également bénéficié de ces avancées. Basaia *et al.* (2019) ont proposé un CNN 3D entraîné sur la base ADNI, capable de discriminer patients Alzheimer et témoins avec une précision supérieure à 85 %, tout en identifiant des régions clés telles que l'hippocampe et le cortex temporal médian comme discriminantes. Korolev, Safiullin, Belyaev & Dodonova (2017) ont de leur côté comparé des architectures profondes avec des méthodes traditionnelles sur les mêmes données et ont confirmé la supériorité des CNN dès lors que des volumes suffisants sont disponibles pour l'entraînement.

Dans l'ensemble, les méthodes d'apprentissage profond représentent un saut qualitatif majeur par rapport aux approches traditionnelles, en permettant une extraction automatique et hiérarchique des représentations discriminantes. Toutefois, la nécessité de grands ensembles de données, la sensibilité aux biais et le manque d'interprétabilité limitent encore leur utilisation en pratique clinique. Ces constats ouvrent la voie à des recherches visant à rendre les modèles plus transparents, notamment via des architectures hybrides ou des modules d'explicabilité intrinsèque, qui seront abordés dans les sections suivantes.

1.3 Méthodologies hybrides combinant caractéristiques locales et apprises

Les méthodologies hybrides en IRM structurelle visent à combiner des descripteurs conçus manuellement avec des représentations apprises par réseaux de neurones profonds, afin d'exploiter l'information complémentaire qu'elles offrent pour la classification du sexe, de l'âge et des pathologies neurologiques. Ces approches intègrent généralement les deux types de caractéristiques par fusion précoce (concaténation), fusion tardive (ensemblage) ou, plus récemment, par des stratégies fondées sur l'attention.

Pour la détection de pathologies, Hasan, Jalab, Meziane, Kahtan & Al-Ahmad (2019) ont combiné des caractéristiques extraites par un CNN avec des descripteurs texturaux fondés sur une matrice de cooccurrence des niveaux de gris modifiée (MGLCM), appliqués à des IRM pondérées T1 et T2 issues d'un ensemble clinique local. Leur approche hybride a atteint une exactitude de 99,3 %, surpassant nettement les performances obtenues par les descripteurs isolés. De même, Bhadra, Maity & Sil (2024) ont proposé une méthode de fusion de caractéristiques appliquée aux IRM structurelles du jeu de données ADNI, combinant des descripteurs texturaux Gabor–GLCM et des caractéristiques apprises par CNN à l'aide d'un module de type *Feature Pyramid Network* (FPN), améliorant significativement la détection de la maladie d'Alzheimer par rapport aux modèles univariés.

Dans les tâches neurodégénératives, Akindele *et al.* (2025) ont présenté un modèle attentionnel basé sur un réseau ResNet–50 enrichi par des modules d'attention CBAM et *Multi-Head*

Self-Attention (MHSA), entraîné sur les ensembles *OASIS* et *MIRIAD*. Cette architecture a atteint des précisions supérieures à 98 %, démontrant l'intérêt des mécanismes d'attention pour la modélisation hiérarchique des dépendances locales et globales, bien qu'elle n'intègre pas de caractéristiques manuelles.

Dans d'autres domaines d'imagerie biomédicale, les fusions attentionnelles ont montré leur supériorité. En pathologie computationnelle, Goel, Kapse, Pati & Prasanna (2024) ont proposé le modèle *CoCa-MIL*, qui combine des descripteurs manuels et profonds via des modules de co- et de *cross-attention*. Cette approche a obtenu des performances supérieures à celles d'une simple concaténation des vecteurs de caractéristiques, démontrant la capacité de l'attention croisée à modéliser des interactions complexes entre modalités.

Cependant, de telles stratégies de fusion attentionnelle restent extrêmement rares dans les pipelines d'IRM cérébrales structurelles. Plus important encore, aucune étude recensée portant sur la classification du sexe, la prédiction de l'âge ou le diagnostic de la maladie d'Alzheimer n'a combiné des descripteurs SIFT tridimensionnels—connus pour leur invariance d'échelle et leur robustesse morphologique locale—avec des représentations apprises par apprentissage profond, que ce soit par concaténation ou par attention. Cette absence met en évidence un vide méthodologique et souligne la nouveauté de notre étude, qui vise à utiliser un mécanisme de *cross-attention* pour fusionner des caractéristiques SIFT volumétriques avec des représentations profondes tridimensionnelles, afin d'obtenir une classification interprétable et robuste à travers les dimensions démographiques et pathologiques.

1.4 Explicabilité et interprétabilité

L'adoption des réseaux neuronaux en neuroimagerie, malgré leurs performances remarquables, reste limitée par leur caractère opaque. Les cliniciens et chercheurs exigent non seulement des résultats fiables, mais également une compréhension des mécanismes sous-jacents aux prédictions. Cette exigence d'*explicabilité* est particulièrement cruciale en imagerie médicale, où la confiance dans les décisions algorithmiques doit être fondée sur leur alignement avec les

connaissances anatomiques et physiopathologiques établies. L'*interprétabilité* vise ainsi à relier les représentations internes d'un modèle aux structures biologiques pertinentes, garantissant que le réseau s'appuie sur des signaux plausibles plutôt que sur des artefacts d'acquisition ou des biais méthodologiques (Arun *et al.*, 2021; van der Velden *et al.*, 2022).

Les premières tentatives d'interprétation des réseaux convolutionnels se sont focalisées sur la compréhension des représentations internes à travers les réponses neuronales locales. La méthode de maximisation d'activation (*activation maximization*) (Erhan, Bengio, Courville & Vincent, 2009) consiste à optimiser un signal d'entrée afin de maximiser l'activation d'un neurone ou d'une couche donnée, fournissant ainsi une approximation des motifs auxquels le réseau est sensible. D'autres approches, telles que les DeconvNets (Zeiler & Fergus, 2014) ou la rétropropagation guidée (*guided backpropagation*) (Springenberg, Dosovitskiy, Brox & Riedmiller, 2014), visent à projeter les activations des couches intermédiaires vers l'espace d'entrée, mettant en évidence les régions de l'image responsables de l'activation.

Une autre famille de techniques repose sur la perturbation systématique de l'entrée. L'analyse de sensibilité par occlusion (*occlusion sensitivity analysis*) introduite par Zeiler & Fergus (2014) consiste à masquer sélectivement des régions de l'image et à mesurer l'impact de cette suppression sur la sortie du modèle, de maniere que les régions dont l'occlusion réduit fortement la confiance du modèle sont considérées comme critiques pour la prédiction. Ce type d'analyse est intuitif, mais particulièrement coûteux en calcul, car il nécessite un grand nombre de passes avant pour évaluer chaque région masquée. Dans le cas des volumes 3D, le nombre de régions possibles croît rapidement avec la taille du volume, ce qui rend la méthode difficilement applicable à grande échelle.

Plus récemment, les techniques d'attribution cherchent à identifier les régions de l'image qui influencent le plus la décision du modèle. Parmi elles, *Grad-CAM (Gradient-weighted Class Activation Mapping)* s'est imposée comme l'une des plus utilisées en imagerie médicale (Selvaraju *et al.*, 2017). Elle calcule les gradients de la sortie du modèle par rapport aux cartes de caractéristiques de la dernière couche convolutionnelle, produisant une carte de chaleur grossière

localisant les régions discriminantes. Grad-CAM a notamment été appliqué à la classification du sexe (Qiu, Zhou & Lin, 2023), permettant de vérifier que les régions mises en évidence correspondaient à des structures cérébrales connues pour leur dimorphisme.

D'autres méthodes étendent ce principe, par example *Integrated Gradients* (IG) (Sundararajan, Taly & Yan, 2017) quantifie la contribution de chaque voxel en intégrant les gradients le long d'un chemin reliant une entrée de référence à l'image étudiée. *SmoothGrad* (Smilkov, Thorat, Kim, Viégas & Wattenberg, 2017) améliore la stabilité des cartes de saillance en ajoutant du bruit gaussien et en moyennant les gradients obtenus. En parallèle, des méthodes d'approximation locale telles que *LIME* (Ribeiro, Singh & Guestrin, 2016) et *SHAP* (Lundberg & Lee, 2017) tentent de reproduire localement le comportement du modèle par des classificateurs plus simples, attribuant un poids d'importance à chaque caractéristique ou région.

1.5 Synthèse et motivations

Les travaux examinés mettent en évidence l'évolution progressive des approches de classification fine en neuro-imagerie, depuis les descripteurs locaux conçus manuellement jusqu'aux représentations hiérarchiques apprises automatiquement, en passant par des tentatives récentes de fusion et d'explicabilité. Chaque famille de méthodes présente des forces complémentaires mais également des limites structurelles qui motivent la recherche d'un cadre unifié conciliant robustesse morphologique, expressivité hiérarchique et transparence anatomique.

Les méthodes fondées sur des caractéristiques locales — telles que SIFT, HOG, ou FBM — ont démontré leur capacité à détecter des motifs anatomiques stables et interprétables, tout en offrant une invariance robuste à l'échelle et à la rotation. Elles permettent d'ancrer les différences inter-sujets dans des régions cérébrales précises, rendant la classification anatomiquement traçable. Toutefois, leur conception manuelle et leur séparation du processus d'apprentissage limitent leur adaptabilité.

À l'inverse, les approches basées sur des caractéristiques apprises (3D-CNN, ResNet, etc.) ont révolutionné la modélisation de l'information en apprenant automatiquement des hiérarchies de

représentations depuis les données brutes. Elles capturent simultanément des motifs locaux et globaux, atteignant des performances remarquables pour la classification du sexe, l'estimation de l'âge ou le diagnostic de la maladie d'Alzheimer. Cependant, ces modèles sont fortement dépendants du volume et de la diversité des données, vulnérables aux biais (p. ex. volume intracrânien), et souvent perçus comme des « boîtes noires » difficilement interprétables.

Les méthodologies hybrides proposent une voie intermédiaire prometteuse. En combinant la robustesse et l'interprétabilité des descripteurs explicites avec la puissance des modèles profonds, elles visent à concilier performance et transparence. Alors que de telles fusions ont démontré leur efficacité dans d'autres domaines biomédicaux, elles restent quasi inexistantes dans le contexte de l'IRM cérébrale tridimensionnelle. En particulier, aucune étude recensée n'a, à ce jour, exploré la combinaison explicite de descripteurs SIFT volumétriques avec des représentations profondes apprises.

Enfin, la question de l'*explicabilité* traverse l'ensemble de ces approches. Les méthodes post-hoc, telles que Grad-CAM, permettent de relier les décisions à des structures anatomiques plausibles, mais sont souvent globaux ou diffus. Les approches intrinsèquement explicables et les modèles hybrides apparaissent ainsi comme des perspectives particulièrement pertinentes pour dépasser ces limites.

C'est précisément dans ce contexte que s'inscrit ce mémoire, dont la méthodologie présentée au Chapitre 2 se propose de explorer systématiquement des approches fondées sur des descripteurs locaux, des modèles profonds *end-to-end* et des architectures hybrides.

CHAPITRE 2

MÉTHODOLOGIE

Dans ce chapitre, nous présentons en détail les méthodes mises en œuvre pour la classification fine des IRM cérébrales tridimensionnelles. Nous suivons une progression allant des méthodes les plus simples et explicites vers des modèles plus complexes, combinant représentations manuelles et apprentissage profond. Nous décrivons successivement : (i) une méthode basée sur descripteurs SIFT et un perceptron multicouche (MLP), (ii) des réseaux CNN appliqués directement aux images IRM, (iii) une approche CNN utilisant des volumes SIFT comme entrée, et enfin (iv) une architecture hybride de type 3D–SIFT–Attention.

2.1 Classification hiérarchique par caractéristiques SIFT et classifieur MLP

Cette section présente une approche de classification hiérarchique fondée sur les descripteurs locaux SIFT et un perceptron multicouche (MLP) pour effectuer une classification à l'échelle de l'image. Une vue d'ensemble du flux complet est illustrée dans la Figure 2.1. Premièrement, les points-clés SIFT sont extraits de l'image en utilisant l'algorithme 3D SIFT Rank décrit en Section 2.1.1. Ils sont ensuite filtrés lors d'une étape de prétraitement afin d'écarter les caractéristiques non pertinentes. Deuxièmement, chaque point-clé retenu est classé individuellement à l'aide d'un MLP. Enfin, les prédictions locales issues de tous les points-clés sont agrégées pour produire la décision finale au niveau de l'image. Cette organisation en pipeline permet de relier explicitement les caractéristiques locales aux décisions globales, ce qui renforce l'interprétabilité du modèle.

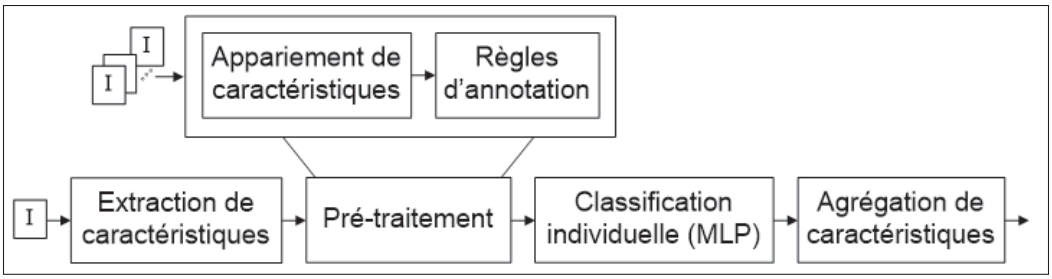

Figure 2.1 Chaîne hiérarchique SIFT+MLP. Le pipeline comporte quatre étapes principales : extraction des descripteurs SIFT, prétraitement et filtrage, classification des points-clés individuels par MLP, puis agrégation des prédictions locales pour obtenir la classification finale de l'image.

2.1.1 Extraction de caractéristiques avec 3D SIFT–Rank

L'algorithme *3D SIFT–Rank*, proposé par Toews & Wells (2009), constitue une extension tridimensionnelle du SIFT classique (Lowe, 2004), initialement conçu pour les images 2D. L'objectif est de détecter des points clés stables dans un volume IRM et de les décrire par des signatures locales invariantes à l'échelle et à la rotation, avant de les transformer en une représentation compacte par classement ordinal.

Pour chaque image cérébrale $I(\mathbf{x})$, en position $\mathbf{x} = (x, y, z)$, un espace d'échelle gaussien est construit :

$$L(\mathbf{x}, \sigma) = (G_\sigma * I)(\mathbf{x}), \tag{2.1}$$

où G_σ est un noyau gaussien de variance σ^2.

La différence de gaussiennes (DoG) est ensuite calculée :

$$D(\mathbf{x}, \sigma) = L(\mathbf{x}, k\sigma) - L(\mathbf{x}, \sigma), \tag{2.2}$$

avec $k > 1$ comme pas interéchelle. Les points clés $\{(\mathbf{x}_i, \sigma_i)\}$ sont définis comme les extrémas locaux de $|D(\mathbf{x}, \sigma)|$ en position et en échelle :

$$\{(\mathbf{x}_i, \sigma_i)\} = \operatorname*{local\,argmax}_{\mathbf{x},\sigma} |D(\mathbf{x}, \sigma)|. \tag{2.3}$$

Autour de chaque point clé, un voisinage normalisé par l'échelle est extrait, et les gradients y sont regroupés dans une grille $2 \times 2 \times 2$ de cellules spatiales. Dans chacune des huit cellules, les orientations de gradient sont quantifiées en 8 classes uniformes, ce qui conduit à un vecteur de 64 dimensions $\mathbf{d}_i \in \mathbb{R}^{64}$

Pour réduire la sensibilité aux variations monotones d'intensité entre scanners ou sujets, chacune des 64 valeurs est soumise à une transformation par rang (*ordinal normalization*). Les composantes du descripteur sont remplacées par leur ordre relatif, ce qui améliore la comparabilité inter-sujets en supprimant l'effet d'échelle absolue des intensités. Le résultat est un descripteur local robuste, pouvant être ensuite agrégé pour obtenir une signature globale du volume.

2.1.2 Stratégie de prétraitement

Le pipeline de prétraitement prépare les points-clés SIFT pour l'entraînement du MLP. Il a pour objectif de transformer l'ensemble brut des caractéristiques extraites en un sous-ensemble de descripteurs stables, discriminants et informatifs, afin d'améliorer la qualité de l'apprentissage supervisé.

Pour chaque image I, on extrait un ensemble de points-clés décrivant les structures locales les plus saillantes :

$$\mathcal{F}(I) = \{f_i\}_{i=1}^{n}, \qquad f_i = (x_i, y_i, z_i, s_i, \theta_i, \mathbf{d}_i),$$

où (x_i, y_i, z_i) représentent la position spatiale du point-clé, s_i son échelle, θ_i son orientation dominante, et $\mathbf{d}_i \in \mathbb{R}^{64}$ le descripteur normalisé des gradients locaux. Ces descripteurs

constituent les entrées de la chaîne de prétraitement, qui vise à identifier et conserver uniquement les caractéristiques les plus cohérentes entre sujets et les plus pertinentes pour la tâche de classification.

Les étapes de prétraitement sont les suivantes :

1. **Filtrage par échelle.** On conserve uniquement les caractéristiques dont l'échelle dépasse un seuil minimal $s_{\min}$, afin de réduire le bruit contextuel :

$$\mathcal{F}'(I) = \{f_i \in \mathcal{F}(I) \mid s_i > s_{\min}\}.$$

   Cette étape élimine les structures instables et détermine les points-clés utilisés lors de la mise en correspondance.

2. **Mise en correspondance inter-images.** Pour chaque point $f_i \in \mathcal{F}'(I_p)$ d'une image de référence I_p, on recherche les descripteurs similaires dans les $N-1$ autres images $\{I_q\}_{q \neq p}$. La similarité est mesurée par la distance euclidienne :

$$D(f_i^p, f_j^q) = \|\mathbf{d}_i^p - \mathbf{d}_j^q\|_2.$$

   Le résultat est un graphe de correspondances inter-sujets reliant chaque point-clé à ses homologues les plus proches.

3. **Filtrage géométrique.** Cette étape écarte les correspondances incohérentes sur le plan spatial. Un seuil géométrique T est appliqué pour rejeter les correspondances incohérentes. Seules les K meilleures correspondances respectant ce critère sont conservées.

4. **Étiquetage des caractéristiques par vote** Chaque caractéristique est ensuite étiquetée selon la distribution de ses correspondances avec les classes cibles (par exemple, sexe masculin ou féminin).

   Soit $M(f_i)$ le nombre de correspondances vers des images masculines et $F(f_i)$ vers des images féminines, l'étiquette est attribuée par une règle de vote majoritaire avec facteur de

dominance 2 :

$$\begin{cases} M(f_i) \geq 2 \times F(f_i) & \Rightarrow \text{Étiqueter comme masculin} \\ F(f_i) \geq 2 \times M(f_i) & \Rightarrow \text{Étiqueter comme féminin} \\ \text{sinon} & \Rightarrow \text{Non informatif.} \end{cases}$$

Les caractéristiques non dominantes ou incohérentes avec la vérité terrain de l'image source sont éliminées. Cette étape produit, pour chaque sujet, un sous-ensemble cohérent de caractéristiques locaux, étiquetés et validés, qui servira à la constitution des ensembles d'entraînement, de validation et de test employés lors de l'apprentissage supervisé du MLP

Algorithme 2.1 Prétraitement global et génération d'annotations locales

Algorithme : Génération du jeu d'entraînement (points-clés → {M,F})

Input : Images $\mathcal{I} = \{I^{(j)}\}_{j=1}^{N}$ avec étiquettes $y^{(j)} \in \{M, F\}$; seuil d'échelle $s_{\min}$; seuil géométrique T ; top-K ; ratio ρ

Output : Jeu annoté uniforme $\mathcal{T} = \{(f_i, \ell_i)\}$ pour l'apprentissage du MLP, où chaque $f_i = (x_i, y_i, z_i, s_i, \theta_i, \mathbf{d}_i)$ et $\ell_i \in \{M, F\}$

for *chaque image* $I^{(j)} \in \mathcal{I}$ **do**
    Extraire les points-clés et descripteurs $F \leftarrow$ détection SIFT (3D) sur $I^{(j)}$;
    Appliquer un **filtrage d'échelle** : $F \leftarrow \{f \in F \mid s_f \geq s_{\min}\}$;
    for *chaque point-clé* $f \in F$ **do**
        $(f^\star, \ell^\star) \leftarrow$ AnnoterPointClé(f, T, K, ρ);
        if $(f^\star, \ell^\star) \neq \varnothing$ **then**
            Ajouter $(f^\star, \ell^\star)$ à $\mathcal{T}$;
        end if
    end for
end for
return $\mathcal{T}$;

Algorithme 2.2 Annotation d'un point-clé par votes de correspondances

Algorithme : AnnoterPointClé(f, T, K, ρ)

Input : Point-clé source $f = (x, y, z, s, \theta, d)$ issu de I_p ;
ensemble des autres images $\{I_q\}_{q \neq p}$; seuil géométrique T_g ;
top-K K ; facteur de dominance ρ

Output : Paire annotée $(f^\star, \ell^\star)$ avec $f^\star = (x, y, z, s, \theta, d)$ et $\ell^\star \in \{M, F\}$,
ou $\varnothing$ si rejet

Récupérer des correspondances candidates de f dans d'autres images;
Classer les correspondances par proximité de descripteurs (croissante);
Écarter les paires qui violent la cohérence géométrique (erreur $> T_g$);
Conserver les K meilleures paires restantes;
if *aucune paire retenue* **then**
    return $\varnothing$
end if
Compter c_M et c_F = nombre de correspondances vers sujets M et F;
Récupérer l'étiquette globale du sujet de f : gt $\in \{M, F\}$;
if $c_M \geq \rho\, c_F$ ***et*** $gt = M$ **then**
    return $\big([x_f, y_f, z_f, s_f, \theta_f, d_f],\ M\big)$
end if
else if $c_F \geq \rho\, c_M$ ***et*** $gt = F$ **then**
    return $\big([x_f, y_f, z_f, s_f, \theta_f, d_f],\ F\big)$
end if
else
    return $\varnothing$ (non-informatif exclu)
end if

2.1.3 Classification individuelle et agrégation probabiliste

La dernière étape du pipeline repose sur un classifieur neuronal de type *Multi-Layer Perceptron* (MLP), chargé d'apprendre la relation entre les caractéristiques SIFT locaux et la variable de sexe biologique. Chaque descripteur annoté $(\mathbf{z}_i,\ l_i)$, issu de l'ensemble $\mathcal{T}$, constitue une

observation d'entrée. Le réseau estime pour chaque vecteur $\mathbf{z}_i$ une distribution de probabilité conditionnelle sur les classes $y \in \{M, F\}$, notée $p_\theta(y \mid \mathbf{z}_i)$.

Le modèle MLP est composé de quatre couches entièrement connectées, séparées par des activations non linéaires et des normalisations de lot (*Batch Normalization*) pour stabiliser l'entraînement. Sa configuration, illustrée à la figure 2.2, comprend une couche d'entrée de 80 neurones correspondant à la dimension du descripteur SIFT, suivie de trois couches cachées entièrement connectées. La première couche cachée compte 200 neurones, accompagnée d'une activation ReLU et d'une normalisation de lot. Les deuxième et troisième couches cachées comportent chacune 100 neurones et appliquent la même combinaison ReLU + normalisation de lot. Enfin, la couche de sortie se compose de deux neurones linéaires correspondant aux classes *Homme* et *Femme*.

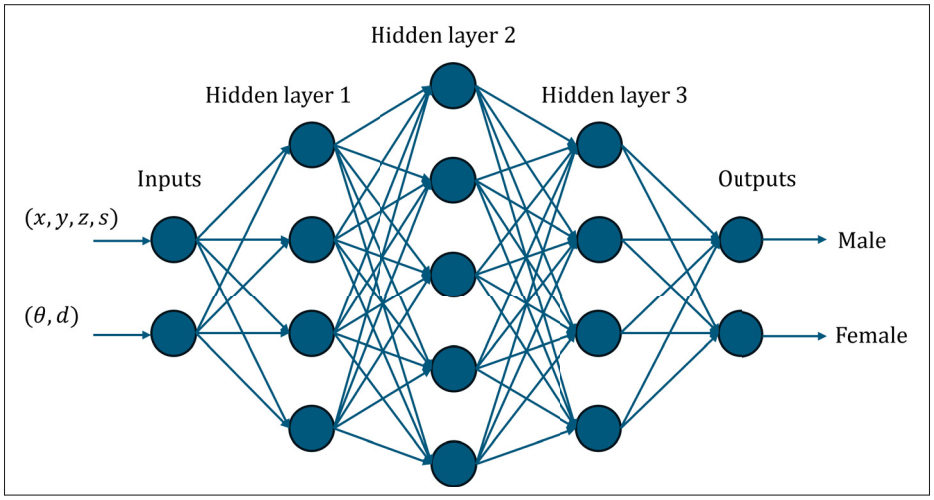

Figure 2.2 Architecture du MLP entraîné sur les caractéristiques SIFT

Le modèle est entraîné à l'aide de la perte d'entropie croisée, en optimisant les paramètres θ par descente de gradient stochastique à l'aide de l'algorithme Adam (Kingma, 2014).

Pour chaque descripteur d'entrée $\mathbf{z}_i$, le réseau produit un vecteur de *logits* $\mathbf{s}_i = f_\theta(\mathbf{z}_i)$, transformé en probabilités par la fonction *softmax* :

$$p_\theta(y = c \mid \mathbf{z}_i) = \frac{\exp(s_{i,c})}{\sum_{k\in\{M,F\}} \exp(s_{i,k})}, \quad c \in \{M, F\}.$$

La classe locale est prédite par la règle du maximum a posteriori :

$$\hat{y}_i = \arg\max_c p_\theta(y = c \mid \mathbf{z}_i).$$

2.1.4 Agrégation probabiliste au niveau du sujet.

Chaque sujet S est représenté par un ensemble de points-clés $\mathcal{F}_S = \{\mathbf{z}_i\}_{i=1}^{n_S}$ extraits de son volume IRM. Afin d'obtenir une décision globale cohérente, les probabilités individuelles sont moyennées pour construire une distribution agrégée :

$$\bar{p}_\theta(y = c \mid S) = \frac{1}{n_S} \sum_{\mathbf{z}_i \in \mathcal{F}_S} p_\theta(y = c \mid \mathbf{z}_i).$$

La prédiction finale du sujet est alors donnée par

$$\hat{y}_S = \arg\max_c \bar{p}_\theta(y = c \mid S).$$

Cette agrégation probabiliste exploite la redondance des observations locales tout en conservant une interprétation bayésienne : la probabilité moyenne $\bar{p}_\theta(y = c \mid S)$ représente la confiance estimée du modèle dans la classe c à l'échelle du sujet. Par ailleurs, les probabilités locales $p_\theta(y = c \mid \mathbf{z}_i)$ sont réutilisées comme poids continus dans les cartes de contribution tridimensionnelles, établissant un lien direct entre la confiance prédictive et l'intensité visuelle observée dans les figures de contributions.

2.2 Classification par réseaux de neurones convolutifs (CNN)

Les réseaux de neurones convolutifs tridimensionnels (3D CNN) ont démontré une capacité remarquable à extraire des représentations hiérarchiques directement à partir des volumes d'imagerie médicale. Dans ce travail, plusieurs architectures de CNN 3D ont été mises en œuvre pour la tâche de classification binaire à partir d'IRM cérébrales, où les étiquettes sont définies comme 0 et 1.

Afin de couvrir différents niveaux de complexité, trois types d'architectures ont été analysés :

(i) un **CNN 3D de base** (*Vanilla CNN*) servant de référence minimale ;

(ii) des **réseaux résiduels** (ResNet-10, ResNet-18 et ResNet-50) de profondeurs variables afin d'évaluer l'effet de la capacité du modèle ;

(iii) un **modèle hybride à double branche** combinant le volume IRM et le volume SIFT 3D à l'aide d'un module d'attention croisée (*cross-attention*) à tête unique.

Dans tous les cas, le problème consiste à apprendre une fonction $f_\theta(X)$ qui prédit la probabilité $\hat{y} \in [0, 1]$ associée à la classe $y \in \{0, 1\}$, en minimisant la perte d'entropie croisée :

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \right], \tag{2.4}$$

2.2.0.1 Modalités d'entrée et formulation du problème

La définition du tenseur d'entrée X varie selon la représentation utilisée :

2.2.0.1.1 CNN basé sur l'IRM.

$$X = I(\mathbf{x}) \in \mathbb{R}^{1 \times H \times W \times D},$$

où $I(\mathbf{x})$ correspond au volume IRM prétraité, normalisé par sujet et structuré en format *canal-premier*.

2.2.0.1.2 CNN basé sur SIFT.

$$X = S(\mathbf{x}) \in \mathbb{R}^{C_s \times H \times W \times D},$$

où $S(\mathbf{x})$ représente un volume de descripteurs 3D SIFT-Rank à 64 canaux ($C_s = 64$). Chaque canal encode la distribution locale des orientations de gradient dans une région anatomique spécifique, offrant ainsi une invariance intégrée à l'échelle et à la rotation (Lowe, 2004; Toews & Wells, 2009).

2.2.0.1.3 Modèle hybride (IRM + SIFT).

$$X = \{I(\mathbf{x}), S(\mathbf{x})\},$$

pour lequel deux fonctions d'encodage sont apprises :

$$z_I = f_I(I; \theta_I), \quad z_S = f_S(S; \theta_S),$$

et fusionnées via un bloc d'attention croisée à tête unique :

$$F_{\text{fusion}} = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) V + z_I, \tag{2.5}$$

suivi d'une tête de classification $f_{\text{cls}}(F_{\text{fusion}}; \theta_{\text{cls}})$. Cette conception favorise l'apprentissage de correspondances spatiales entre les caractéristiques issues de l'IRM (requêtes Q) et celles issues du volume SIFT (clés K, valeurs V), tout en limitant la complexité et en préservant l'interprétabilité.

Cette formulation unifiée permet de comparer directement les modèles mono-entrée et double-entrée sous le même objectif d'apprentissage, tout en tirant parti des propriétés géométriques et invariantes de SIFT.

2.2.0.2 Architecture CNN 3D de base

Le modèle de base est un 3D-CNN composé de six blocs séquentiels Conv3D–BatchNorm–ReLU–MaxPool, comme illustré à la Figure 2.3. Le nombre de filtres augmente à chaque bloc (32, 64, 128, 256, 512, 512).Le réseau se termine par une couche de *global average pooling* et une couche pleinement connectée à deux sorties.

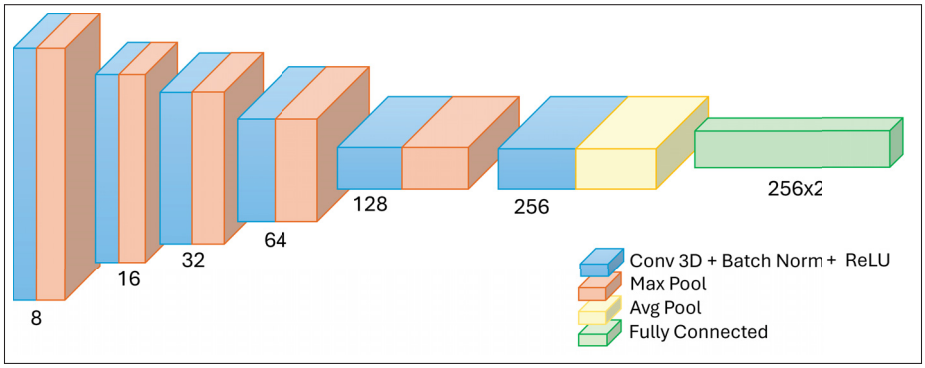

Figure 2.3 Architecture du CNN 3D de base. Chaque bloc comprend une convolution 3D, une normalisation de lot et une activation ReLU suivie d'un sous-échantillonnage.

Ce design minimaliste vise la lisibilité et la légèreté computationnelle, tout en servant de point de référence pour évaluer les bénéfices des connexions résiduelles et des mécanismes de fusion. Des noyaux $3 \times 3 \times 3$ sont utilisés afin d'équilibrer la taille du champ réceptif et le nombre de paramètres, adapté à des volumes de 128^3 voxels.

2.2.0.3 Réseaux résiduels (ResNets)

Les réseaux résiduels (He, Zhang, Ren & Sun, 2016) résolvent le problème de dégradation observé dans les réseaux profonds en introduisant des connexions identitaires facilitant la

rétropropagation :

$$H(x) = F(x) + x, \tag{2.6}$$

où $F(x)$ est la fonction résiduelle apprise par un ou plusieurs blocs convolutifs.

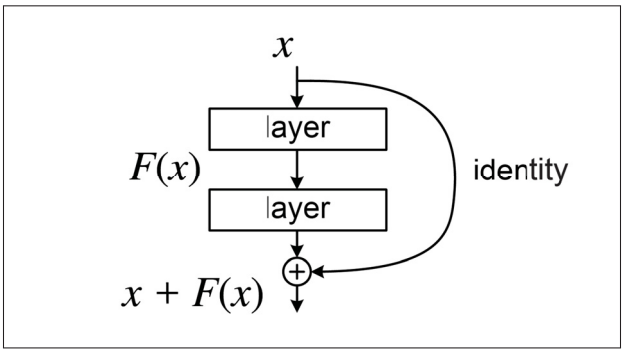

Figure 2.4 Illustration d'un bloc résiduel 3D avec connexion identitaire permettant un apprentissage stable dans les architectures profondes.
Tirée de He *et al.*(2016)

Cette approche encourage le raffinement progressif des représentations plutôt que leur réapprentissage complet, ce qui améliore la stabilité de l'optimisation et la généralisation.

- **ResNet-10/18** utilisent des blocs basiques constitués de deux convolutions $3 \times 3 \times 3$ suivies d'une normalisation de lot (BN) et d'une activation ReLU (Figures 2.5 et 2.6) ;
- **ResNet-50** repose sur des blocs *bottleneck* composés de convolutions 1×1, 3×3 et 1×1, réduisant la complexité tout en approfondissant la hiérarchie des caractéristiques (Figure 2.7).

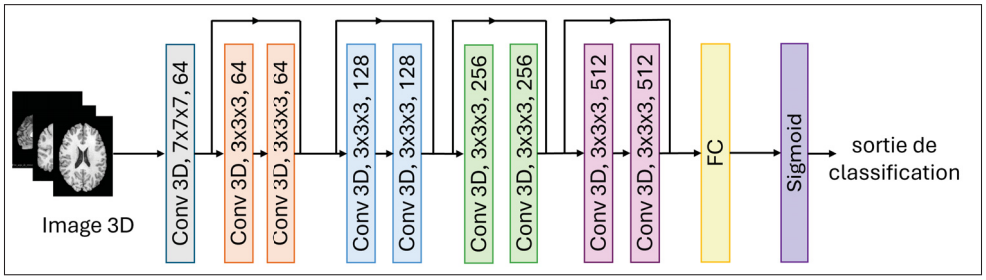

Figure 2.5 Architecture ResNet-10

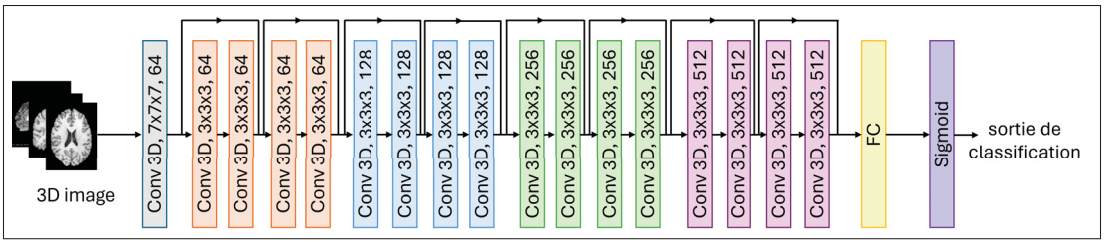

Figure 2.6 Architecture ResNet-18

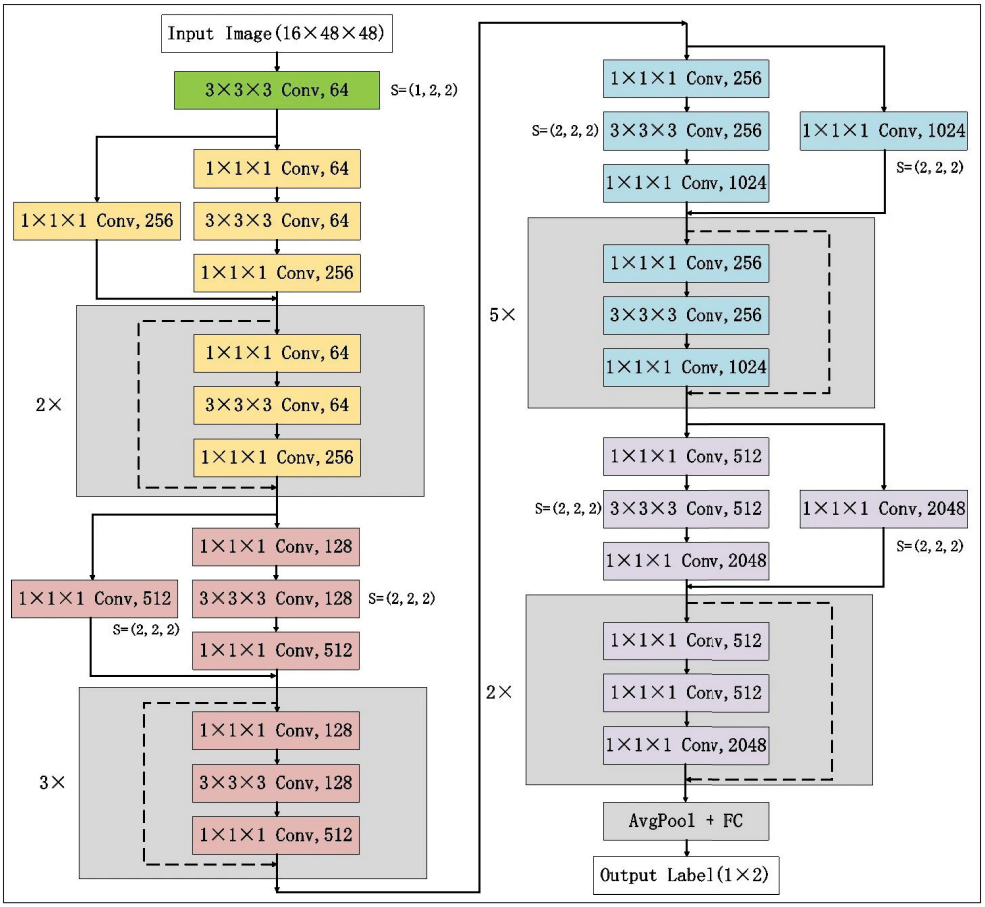

Figure 2.7 Architecture ResNet-50
Tirée de Yu *et al.*(2021)

Les connexions résiduelles sont particulièrement bénéfiques pour les données volumétriques, car elles atténuent la disparition du gradient et favorisent une meilleure convergence lorsque le nombre de paramètres augmente.

2.2.0.4 CNN basé sur les volumes SIFT

L'architecture **SIFT-CNN** établit un lien entre les descripteurs construits à la main et l'apprentissage profond. Le volume d'entrée est un tenseur $S \in \mathbb{R}^{64\times H\times W\times D}$ où chaque canal correspond à une dimension du descripteur SIFT-Rank extrait aux positions clés du cerveau.

Le réseau est constitué de quatre blocs convolutifs Conv3D–BN–ReLU suivis d'un *global average pooling* et d'une couche pleinement connectée pour la classification binaire. Ce design permet d'exploiter les invariances géométriques et photométriques déjà encodées dans les descripteurs SIFT et concentre l'apprentissage sur les régions anatomiques les plus informatives.

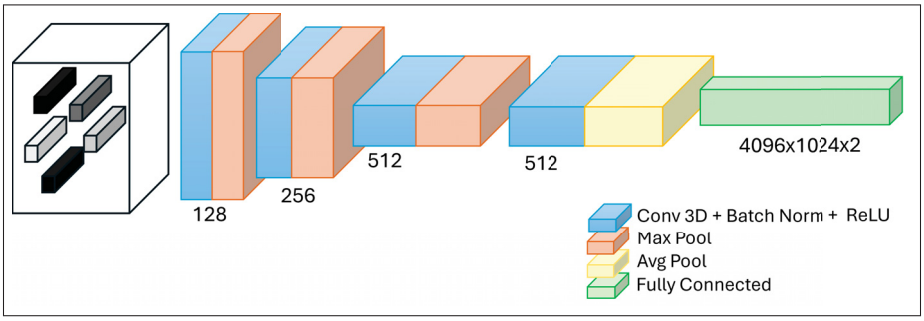

Figure 2.8 Architecture du SIFT-CNN : un réseau 3D peu profond traitant le volume SIFT pour extraire des représentations discriminantes invariantes à l'échelle et à la rotation.

2.2.0.5 Modèle hybride proposé (3D–SIFT–Attention)

Afin de combiner le contexte anatomique global des volumes IRM avec les indices locaux de texture issus de SIFT, un modèle hybride à deux branches a été conçu, dont l'architecture est illustrée à la Figure 2.9.

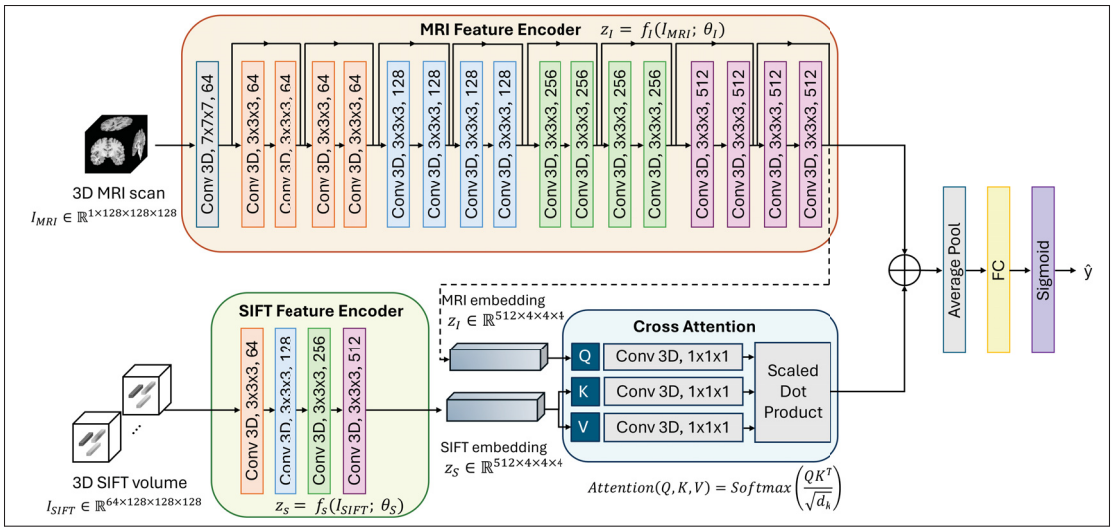

Figure 2.9 Architecture du modèle 3D–SIFT–Attention. La branche IRM encode les intensités natives au moyen d'un ResNet 3D ; la branche SIFT encode une carte de descripteurs SIFT-Rank 3D à 64 canaux. Un module de cross-attention fusionne les deux flux avant la tête de classification.

- **Branche IRM :** un encodeur ResNet–18 3D extrait un tenseur de caractéristiques $F_{\text{IRM}} \in \mathbb{R}^{512\times4\times4\times4}$ représentant la structure anatomique globale.
- **Branche SIFT :** un CNN 3D léger encode les volumes de descripteurs SIFT et produit un tenseur F_{SIFT} de dimensions compatibles, capturant des motifs géométriques locaux et invariants d'échelle.
- **Bloc d'attention croisée :** un module à tête unique établit des correspondances spatiales entre F_{IRM} (requêtes) et F_{SIFT} (clés et valeurs). Il applique le mécanisme d'attention par produit scalaire normalisé :

$$\text{Attn}(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \tag{2.7}$$

où $Q = W_Q F_{\text{IRM}}$, $K = W_K F_{\text{SIFT}}$ et $V = W_V F_{\text{SIFT}}$. La sortie est ensuite fusionnée aux caractéristiques IRM par une connexion résiduelle :

$$F_{\text{fusion}} = \text{Attn}(Q,K,V) + F_{\text{IRM}}. \tag{2.8}$$

Ce mécanisme de fusion met en correspondance les structures géométriques locales issues de SIFT avec les indices d'intensité globaux de l'IRM, améliorant ainsi la complémentarité des représentations. L'utilisation d'une seule tête d'attention garantit un compromis entre expressivité, interprétabilité et coût computationnel maîtrisé.

L'ensemble du code permettant de reproduire le prétraitement des données, l'entraînement et l'évaluation du modèle proposé est mis à disposition dans le dépôt GitHub associé [1]

2.2.0.6 Stratégie d'entraînement

Toutes les architectures CNN ont été entraînées sur la même tâche de classification binaire et les mêmes partitions de données afin de garantir la comparabilité. L'entraînement a été effectué selon un protocole de validation croisée stratifiée à 5 plis, préservant la distribution des classes dans chaque sous-ensemble.

L'optimisation a été réalisée avec l'optimiseur Adam et la perte d'entropie croisée binaire, avec des lots de 32 échantillons. Tous les modèles résiduels ont été adaptés à des entrées tridimensionnelles via des convolutions 3D et initialisés avec des poids pré-entraînés sur *ImageNet* (Deng *et al.*, 2009), puis ajustés sur des volumes IRM à canal unique. Les taux d'apprentissage ont été fixés à 10^{-3} pour ResNet-10/18 et à 10^{-2} pour ResNet-50.

Un ordonnancement de type *reduce-on-plateau* a été appliqué pour réduire dynamiquement le taux d'apprentissage lorsque la perte de validation stagnait, et un *early stopping* a été employé pour prévenir le surapprentissage. Les données d'entraînement ont été augmentées par retournement sagittal aléatoire. La normalisation et l'équilibrage des classes ont été systématiquement appliqués pour stabiliser la convergence et permettre une évaluation équitable entre architectures.

[1] https://github.com/taliav26/3DSIFTAttention

2.2.1 Explicabilité (IG–KPA)

Nous proposons l'approche *Integrated–Gradients–Guided Keypoint Attribution* (IG–KPA), illustrée à la Figure 2.10, qui convertit des attributions denses voxel par voxel en un ensemble compact et vérifiable de points clés SIFT appariés, représentant les éléments qui influencent le plus la décision du modèle.

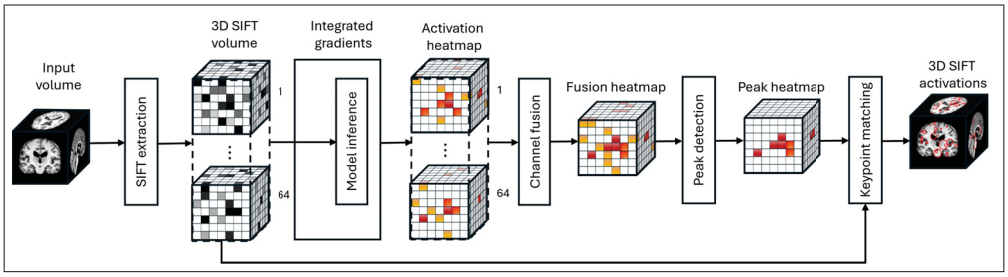

Figure 2.10 Schéma du pipeline de la méthode IG-KPA (Integrated Gradients–Keypoint Attribution). À partir d'un volume IRM tridimensionnel, des descripteurs SIFT sont extraits et organisés en un volume 3D-SIFT. Lors de l'inférence du modèle, la méthode des gradients intégrés (IG) calcule des attributions voxel-par-voxel, produisant des cartes d'activation pour chaque canal. Une fusion de canaux agrège ces cartes en une carte de chaleur unique, à partir de laquelle les pics sont détectés. La carte de pics est ensuite mise en correspondance avec les points clés d'origine, générant des activations 3D-SIFT localisées qui mettent en évidence les régions anatomiques les plus influentes dans la décision du modèle

Soit $f(V, S)$ le logit du réseau pour la classe prédite, étant donné un volume IRM pondéré T1 $V \in \mathbb{R}^{X\times Y\times Z}$ et un volume SIFT 3D $S \in \mathbb{R}^{C\times X\times Y\times Z}$ (canaux normalisés par min–max de $[0, 63]$ vers $[0, 1]$). Soit $\mathcal{F} = \{(x_i, y_i, z_i, \text{scale}_i)\}_{i=1}^{N}$ l'ensemble des points clés SIFT du sujet dans la grille voxel native (x, y, z).

Nous calculons les attributions pour la branche SIFT à l'aide des IG le long du chemin rectiligne reliant des références nulles $\tilde{V} = \mathbf{0}$, $\tilde{S} = \mathbf{0}$ à l'entrée :

$$\text{IG}_S[c, i] = \left(S_{c,i} - \tilde{S}_{c,i}\right) \int_0^1 \frac{\partial f(V_\alpha, S_\alpha)}{\partial S_{c,i}} \, \mathrm{d}\alpha, \qquad (V_\alpha, S_\alpha) = (\tilde{V}, \tilde{S}) + \alpha \, (V - \tilde{V}, S - \tilde{S}), \quad (2.9)$$

où c indexe les canaux et i les positions voxel. Nous utilisons n_{steps} valeurs de α régulièrement espacées et contrôlons la stabilité numérique via l'indicateur de convergence de Captum

(Kokhlikyan *et al.*, 2020)ainsi que le *taux de complétude* des IG (somme des attributions par rapport à la différence de logit).

Ensuite, nous agrégeons les attributions par canal via la règle ℓ_1 afin d'obtenir un champ d'attribution 3D unique :

$$H[i] \;=\; \sum_{c=1}^{C} \left|\mathrm{IG}_S[c,i]\right|. \tag{2.10}$$

Nous rééchantillonnons H sur la grille d'image native (X, Y, Z) par interpolation trilinéaire, puis appliquons une normalisation min–max vers $[0, 1]$ pour une sélection de pics robuste. Sauf mention contraire, toutes les coordonnées sont exprimées dans le repère voxel d'origine (x, y, z).

Finalement, nous détectons les maxima de saillance spatiale par suppression de non-maxima 3D, avec deux garde-fous : (i) un plafond de cardinalité $K \leq \min(K_{\max}, N)$, évitant plus de pics que de points clés disponibles ; et (ii) un seuil d'amplitude imposant $|H[p]| \geq \overline{|H|}$ (moyenne du champ fusionné). Une séparation minimale $d_{\min}$ (en voxels) prévient les doublons regroupés, ce qui produit un ensemble de positions de pics $\mathcal{P} = \{(x_p, y_p, z_p)\}_{p=1}^{K}$.

Chaque pic est apparié au point clé SIFT le plus proche à l'aide d'un arbre KD avec rayon euclidien r (valeur par défaut 1,5 voxels), en imposant des correspondances bijectives pour éviter les duplications. Les pics sans point clé à distance $\leq r$ sont écartés. L'explication finale est l'ensemble apparié

$$\mathcal{M} \;=\; \left\{\, (x_j, y_j, z_j, \mathrm{scale}_j,\; H[x_j, y_j, z_j]) \,\right\}, \tag{2.11}$$

trié par H. Nous exportons $\mathcal{M}$ sous forme de tableau (CSV) et de surimpression 3D (VTK) pour la visualisation avec 3D-Slicer (Pieper, Halle & Kikinis, 2004).

CHAPITRE 3

RÉSULTATS ET ANALYSE EXPÉRIMENTALE

Ce chapitre présente et analyse les résultats expérimentaux obtenus à partir des approches méthodologiques décrites au Chapitre 2. L'objectif principal est d'évaluer les performances de classification ainsi que le degré d'interprétabilité des modèles développés, afin d'examiner dans quelle mesure les représentations apprises reflètent des différences morphologiques réelles entre groupes.

La tâche principale étudiée est la classification du sexe à partir d'images IRM cérébrales. Cette tâche constitue un banc d'essai de référence permettant de comparer différentes architectures et d'analyser la dimorphie sexuelle du cerveau à travers les cartes de saillance et les points clés identifiés par les modèles. Nous nous intéressons particulièrement à la capacité des approches à apprendre des caractéristiques dimorphiques, c'est-à-dire des formes anatomiques similaires mais présentant des différences structurelles cohérentes selon le sexe.

En complément, le modèle le plus performant sur la tâche principale a été évalué sur deux tâches de validation supplémentaires : la classification par âge et la classification de la maladie d'Alzheimer, afin d'examiner la capacité de généralisation et la cohérence des mécanismes d'interprétation sur d'autres dimensions biologiques. Ces expériences complémentaires, menées sur des jeux de données indépendants, permettent de vérifier la robustesse du modèle hybride proposé et de confirmer la plausibilité biologique des régions d'activation mises en évidence.

3.1 Jeux de données

Les expériences présentées dans ce mémoire s'appuient sur plusieurs bases de données d'imagerie cérébrale tridimensionnelle librement accessibles. Plus précisément, le Human Connectome Project (HCP) (Van Essen *et al.*, 2013) a été utilisé pour la classification du sexe, tandis que le Open Access Series of Imaging Studies (OASIS) (Marcus *et al.*, 2007) a servi pour la classification de l'âge ainsi que pour une première analyse de la maladie d'Alzheimer. Enfin, le

Alzheimer's Disease Neuroimaging Initiative (Jack Jr *et al.*, 2008) a été utilisé pour approfondir l'étude de la classification d'Alzheimer sur une cohorte clinique plus vaste.

Le jeu de données HCP Q4 comprend des IRM 3D pondérées en T1 à haute résolution (0,7 mm isotropique, $260 \times 311 \times 260$ voxels) acquises chez des adultes âgés de 22 à 36 ans dont la moyenne d'âge est d'environ 29 ans (Van Essen *et al.*, 2012). Dans nos expériences fondées sur des caractéristiques apprises avec des CNN, nous avons analysé un ensemble de 1 040 sujets issus de cette base de données, dont 570 femmes et 470 hommes, après extraction du crâne (*skull stripping*) afin de restreindre l'analyse aux tissus cérébraux. Pour les méthodes fondées uniquement sur des caractéristiques locales SIFT, nous avons constitué un sous-ensemble équilibré de 211 femmes et 211 hommes tirés aléatoirement de 422 familles distinctes, de manière à limiter les biais de parenté. Dans tous les cas, les étiquettes de sexe proviennent de l'auto-déclaration et, conformément aux métadonnées HCP, les sujets s'auto-déclarant femmes et fournissant un historique menstruel complet (dix champs) sont considérés comme femmes, les autres comme hommes.

Le jeu de données OASIS-1 comprend des IRM pondérées T1 provenant de 416 sujets âgés de 18 à 96 ans, incluant des individus en bonne santé et des patients présentant une probable maladie d'Alzheimer. Les acquisitions ont été harmonisées et segmentées selon des protocoles standardisés, avec une résolution isotropique de 1 mm. Cette base permet d'aborder deux types de tâches : (i) la classification par âge, définie à partir d'un seuil médian pour distinguer les groupes jeunes et âgés, et (ii) la classification Alzheimer / sujets sains, fondée sur le score clinique CDR (*Clinical Dementia Rating*).

Le jeu de données ADNI rassemble des IRM T1 multi-sites acquises sur des scanners 1,5 T et 3 T avec résolution typiquement $\approx$ 1,0–1,2 mm isotropique. La cohorte inclut des témoins cognitivement normaux (CN), des troubles cognitifs légers (MCI) et des patients Alzheimer (AD), assortis d'évaluations cliniques standardisées (p. ex., CDR). Dans ce travail, nous utilisons les sous-ensembles T1 d'ADNI pour la classification Alzheimer / sujets sains et pour évaluer la robustesse inter-sites des modèles.

Dans le cadre des expériences, les volumes ont été rééchantillonnés à une taille uniforme de $128 \times 128 \times 128$ voxels afin de réduire la charge computationnelle. Un prétraitement standard a été appliqué, comprenant une normalisation en z-score de l'intensité, ainsi qu'un enregistrement rigide ou déformable sur un espace de référence commun, selon la nature de l'expérience considérée.

3.2 Résultats de classification

L'évaluation des performances de classification repose sur trois métriques principales : l'exactitude, la précision et le rappel. L'**exactitude** (*accuracy*) mesure la proportion globale de prédictions correctes sur l'ensemble des sujets, toutes classes confondues. La **précision** (*precision*) quantifie la fiabilité des prédictions positives, c'est-à-dire la fraction d'exemples réellement positifs parmi ceux prédits comme tels. Le **rappel** (*recall*) évalue la sensibilité du modèle, soit la capacité à identifier correctement les exemples appartenant à une classe donnée. Ces métriques, calculées pour chaque classe et moyennées selon une validation croisée à cinq plis, permettent d'apprécier à la fois la performance globale et l'équilibre du modèle entre détection correcte et fausses classifications. L'ensemble des résultats reportés est accompagné de l'écart-type inter-plis afin de rendre compte de la stabilité des modèles étudiés.

3.2.1 Classification du sexe sur HCP et comparaison avec les modèles invariants d'échelle représentatifs

Le Tableau 3.1 résume les performances obtenues par plusieurs architectures pour la classification du sexe à partir d'IRM cérébrales tridimensionnelles issues du jeu de données HCP. Les expériences mettent en évidence des différences nettes entre les modèles peu profonds et les architectures profondes, ainsi qu'entre les approches reposant sur des caractéristiques locales construites à la main et celles apprenant directement des représentations globales. Ces résultats peuvent être interprétés à travers la manière dont chaque modèle encode le contexte spatial, l'invariance d'échelle et la correspondance morphologique entre sujets.

Tableau 3.1 Performances de classification du sexe sur HCP (validation croisée à 5 plis)

Données	Modèle	Exactitude	Précision	Rappel	Paramètres (M)
Rigide	SIFT-MLP (caractéristiques / sujets)	0.70 / 0.84	0.70 / 0.84	0.70 / 0.84	–
Rigide	Vanilla CNN	0.700 ± 0.186	0.695 ± 0.181	0.984 ± 0.021	1.18
Rigide	3D-SIFT ResNet18	0.836 ± 0.041	0.840 ± 0.039	0.836 ± 0.041	34.5
Déformable	3D-SIFT (Toews *et al.*, 2025)	85.1 ± 0.5	–	–	–
Rigide	3D-SIFT CNN	0.870 ± 0.033	0.880 ± 0.035	0.870 ± 0.030	4.76
Rigide	ResNet-50	0.928 ± 0.026	0.930 ± 0.028	0.930 ± 0.028	46.2
Rigide	3D-SIFT (Toews *et al.*, 2025)	0.933 ± 0.0	–	–	–
Rigide	ResNet-18	0.936 ± 0.014	0.936 ± 0.014	0.934 ± 0.014	33.1
Rigide	**3D-SIFT-Attention**	**0.940 ± 0.020**	0.939 ± 0.022	0.939 ± 0.016	38.9
Déformable	ResNet-50	0.898 ± 0.015	0.898 ± 0.017	0.896 ± 0.017	46.2
Déformable	ResNet-18	0.914 ± 0.023	0.914 ± 0.023	0.916 ± 0.022	33.1
Déformable	**3D-SIFT-Attention**	**0.924 ± 0.016**	0.935 ± 0.033	0.928 ± 0.031	38.9

L'approche SIFT-MLP, fondée sur des descripteurs tridimensionnels SIFT suivis d'un perceptron multicouche, a atteint une exactitude de 70 % au niveau des caractéristiques locales et jusqu'à 84 % après agrégation des prédictions au niveau du sujet. À titre de comparaison, le Vanilla CNN, un réseau convolutif à six couches entraîné directement sur les intensités brutes des voxels, n'a atteint qu'une moyenne de 71 % et a montré une variabilité élevée entre les plis de validation. Ce contraste souligne la capacité des descripteurs invariants locaux à surpasser un modèle convolutif entraînable lorsque les données sont limitées. Les descripteurs SIFT encodent l'orientation des gradients dans de petits voisinages tridimensionnels, capturant ainsi la forme et la courbure locales d'une manière invariante aux changements d'échelle et de rotation. Le MLP, bien que simple, profite de la stabilité de ces entrées et de la redondance de centaines de points-clés par sujet. En agrégeant les prédictions locales, il parvient à lisser le bruit individuel tout en renforçant la cohérence des statistiques de texture et de forme. D'un côté, la performance relativement élevée du SIFT–MLP peut s'expliquer par sa stratégie de prétraitement fondée sur les correspondances inter-sujets, conceptuellement proche de celle proposée par Toews *et al.* (2025). Dans cette étude, les auteurs ont atteint une précision de 93.3 % en utilisant un modèle

non paramétrique basé exclusivement sur la comparaison d'ensembles de points-clés SIFT tridimensionnels. Leur méthode mesure une similarité de type Jaccard adoucie entre l'ensemble de points-clés d'un sujet test et des ensembles de référence masculins ou féminins, intégrant directement les forces de correspondance dans une fonction de décision à l'échelle du sujet. Notre pipeline adopte une logique comparable, mais la transpose dans un cadre d'apprentissage supervisé. Concrètement, les points-clés extraits sont d'abord filtrés selon leur échelle afin de ne conserver que les structures stables et de grande taille. Les descripteurs sont ensuite mis en correspondance entre sujets à l'aide de la distance euclidienne, les paires incohérentes sont écartées par filtrage géométrique, et chaque caractéristique restante est étiquetée par un vote majoritaire en fonction de la distribution de ses correspondances entre classes. Seules les caractéristiques cohérentes et informatives sont conservées pour l'entraînement du MLP, produisant ainsi un ensemble d'apprentissage épuré et discriminant. Ce processus reproduit partiellement la mise en correspondance inter-sujets décrite par Toews *et al.* (2025), mais remplace leur pondération continue par une étape de quantification et d'étiquetage dur, à partir de laquelle les caractéristiques sont traitées comme des échantillons indépendants. Bien que cette simplification permette un apprentissage plus scalable (linéaire plutôt que quadratique en nombre de sujets), elle supprime les nuances des correspondances faibles que le modèle de Toews *et al.* (2025) intègre directement dans sa mesure de similarité. Cette différence méthodologique explique l'écart de performance entre 84 % et 93.3 %, soulignant l'intérêt de conserver les poids de correspondance lors de la décision finale. Néanmoins, notre approche supervisée présente plusieurs avantages : elle apprend un pondérage discriminant des dimensions du descripteur, fournit des sorties probabilistes interprétables comme des mesures de confiance et offre une complexité computationnelle adaptée à de grands ensembles de données.

D'un autre côté, la performance plus faible du réseau convolutif suggère que l'information discriminante du sexe est davantage concentrée dans des structures morphologiques localisées que dans des motifs d'intensité globaux. De plus, la précision moyenne inférieure et la forte variance inter-plis observées pour le vanilla CNN reflètent l'instabilité du réseau à apprendre

des représentations invariantes, probablement en raison de la taille restreinte du jeu de données et de l'absence de pré-entraînement.

Comme attendu, les architectures convolutives profondes intégrant des modules résiduels appliquées directement aux IRM brutes ont obtenu des performances nettement supérieures. Les modèles ResNet10, ResNet18 et ResNet50 ont respectivement atteint 88 %, 93 % et 92 % d'exactitude. L'amélioration par rapport aux réseaux plus simples confirme que la profondeur introduit une capacité hiérarchique de représentation, permettant aux réseaux résiduels d'intégrer progressivement l'information locale issue des gradients et contours dans des abstractions spatiales de plus haut niveau. Cependant, le gain n'est pas monotone avec la profondeur : la ResNet18 surpasse légèrement la ResNet50. Ce résultat peut s'expliquer par la combinaison de plusieurs facteurs. D'une part, la ResNet50 comporte un nombre de paramètres beaucoup plus élevé, augmentant le risque de surapprentissage dans un contexte de données limitées. D'autre part, la redondance de caractéristiques devient importante une fois que les hiérarchies spatiales pertinentes (bords, surfaces, configurations anatomiques) ont été apprises, rendant les couches supplémentaires moins informatives. Enfin, la propagation du gradient sur de longues chaînes peut nuire à la stabilité de l'optimisation et freiner la convergence, même avec des connexions résiduelles. Ces observations indiquent que la profondeur optimale dépend non seulement de la taille des données mais aussi de l'échelle d'information intrinsèque à la tâche. La ResNet18 semble constituer un compromis efficace : suffisamment profonde pour modéliser des dépendances spatiales non locales, mais pas au point d'introduire une complexité inutile.

Le modèle 3D-SIFT-CNN, qui traite les volumes de descripteurs SIFT à l'aide de quatre couches convolutives, a atteint une précision moyenne de 87 %, surpassant le 3D-SIFT-ResNet18 (83.6 %). Cette inversion de tendance montre qu'une convolution appliquée à des caractéristiques déjà invariantes peut être très efficace même avec un réseau peu profond. Les descripteurs SIFT encodent déjà des gradients d'orientation, des contrastes locaux et des motifs de courbure de manière normalisée en échelle et en rotation, c'est-à-dire qu'ils remplissent en grande partie le rôle des premières couches convolutives d'un réseau appris sur intensités brutes. Dès lors, l'intégration d'une architecture résiduelle profonde comme ResNet18 ne garantit pas l'extraction d'abstractions

réellement nouvelles, car plusieurs filtres appris peuvent s'avérer redondants par rapport aux structures déjà encodées par SIFT. Au lieu d'apporter une information complémentaire, ces couches supplémentaires risquent de augmenter la complexité sans bénéfice en termes de diversité représentationnelle. Cette sur-paramétrisation peut conduire à un surapprentissage et à une moindre capacité de généralisation. En revanche, un réseau peu profond tel que le SIFT–CNN dispose d'une capacité suffisante pour apprendre uniquement les relations spatiales entre des descripteurs déjà stables, sans risquer de dupliquer des invariances existantes. Ces observations suggèrent que, lorsqu'une partie du processus d'invariance est déjà prise en charge par la représentation d'entrée, la profondeur optimale du réseau doit être adaptée en conséquence pour éviter la redondance structurelle et favoriser la complémentarité entre filtres appris et caractéristiques préalables.

Le modèle 3D-SIFT-Attention a obtenu la meilleure performance globale avec 94 % d'exactitude sur les données enregistrées de manière rigide. Cette architecture combine la stabilité locale des caractéristiques SIFT avec la modélisation contextuelle globale d'une ResNet18 via un bloc de *cross-attention* à une seule tête. Le mécanisme d'attention apprend à pondérer les activations du réseau résiduel selon leur corrélation avec la carte de caractéristiques SIFT, mettant ainsi en évidence les régions où les deux modalités s'accordent. Cette fusion dynamique offre une sensibilité spatiale complémentaire : les SIFT capturent les détails à haute fréquence (bords, plis corticaux), tandis que la ResNet encode des structures plus globales et abstraites (symétries, volumes, enveloppes anatomiques). L'attention permet d'aligner ces deux échelles de manière adaptative, renforçant les régions informatives sans dépendre exclusivement de l'enregistrement spatial explicite. Le modèle hybride bénéficie ainsi d'une meilleure robustesse et d'une variabilité réduite entre plis, illustrant l'intérêt d'une représentation multi-échelle combinant caractéristiques locales et globales.

Enfin, la comparaison entre enregistrements rigides et déformables révèle un schéma récurrent. Pour toutes les architectures, l'enregistrement déformable conduit à une légère diminution des performances (par exemple, la ResNet18 chute de 93.6 % à 91.4 %). Ce résultat corrobore l'hypothèse selon laquelle les réseaux profonds peuvent exploiter, parfois involontairement, des

indices liés à la taille ou à la forme du cerveau lorsque les volumes ne sont pas normalisés. L'enregistrement rigide conserve les différences de taille interindividuelles, qui peuvent servir de raccourcis décisionnels pour la classification. Une fois ces différences supprimées par l'alignement déformable, le réseau doit s'appuyer sur des indices de texture et de forme plus subtils, plus difficiles à apprendre.

Dans l'ensemble, la combinaison d'un prétraitement fondé sur la correspondance, d'un apprentissage supervisé des descripteurs et d'une fusion multi-échelle permet de relier le paradigme invariant d'échelle de Toews *et al.* (2025) aux architectures profondes résiduelles modernes. Cette continuité méthodologique suggère que, même après correction de la taille globale, l'information discriminante entre sexes demeure portée par des configurations locales et distribuées de gradients morphologiques subtils, plutôt que par des différences volumétriques marquées. Les modèles qui préservent explicitement les correspondances inter-sujets, comme celui de Toews *et al.* (2025), ou qui apprennent à les pondérer de manière adaptative à travers des mécanismes d'attention croisée, comme notre architecture hybride SIFT–Attention, atteignent ainsi le meilleur compromis entre pouvoir discriminant, invariance et interprétabilité.

3.2.2 Classification de l'âge (OASIS1)

Le meilleur modèle obtenu pour la classification du sexe, à savoir l'architecture hybride 3D-SIFT-Attention, a ensuite été évalué sur une tâche distincte et à partir d'un autre jeu de données, OASIS1. Cette expérience visait la classification par âge, définie à partir d'un seuil médian séparant les groupes de sujets *jeunes* et *âgés*. La distribution des âges est illustrée à la Figure 3.1, où l'on observe une répartition fortement bimodale, avec un premier pic de sujets jeunes adultes (environ 20–30 ans) et un second pic de sujets âgés (environ 70–85 ans), tandis que les tranches d'âges intermédiaires sont très peu représentées, laissant peu de sujets à proximité du seuil de décision (60 ans).

Les résultats correspondants sont présentés dans le Tableau 3.2.

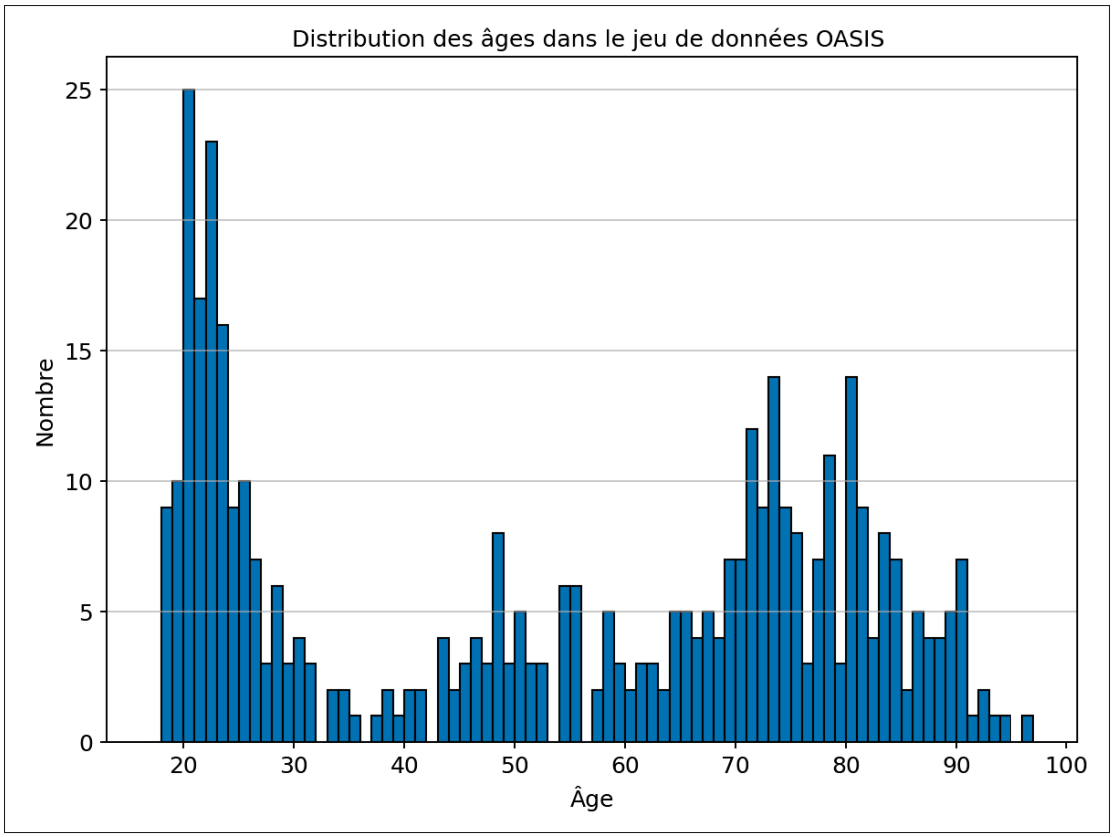

Figure 3.1 Distribution des âges dans OASIS1, utilisée pour définir les groupes *jeunes* (<60 ans) et *âgés* (60 ans)

Tableau 3.2 Performances de classification de l'âge (jeunes vs âgés) sur OASIS1 (validation croisée à 5 plis). Toutes les expériences ont été réalisées avec enregistrement rigide.

Modèle	**Exactitude**	**Précision**	**Rappel**	**Paramètres (M)**
ResNet-18	0.928 ± 0.027	0.931 ± 0.028	0.928 ± 0.029	33.1
3D-SIFT-Attention	**0.930 ± 0.023**	0.935 ± 0.022	0.931 ± 0.021	38.9

Le modèle démontre une performance de classification robuste et équilibrée, caractérisée par une précision et un rappel élevés pour les deux classes, ainsi qu'une faible variabilité inter-plis. Ces résultats traduisent une bonne capacité de généralisation du modèle pour distinguer de

manière fiable les individus jeunes et âgés. L'exactitude globale est pratiquement identique pour les deux architectures, indiquant que chacune d'elles classe correctement environ 93 % des sujets en moyenne. L'analyse statistique appariée réalisée sur les cinq plis de validation (test t apparié, $p = 0.93$, $d_z = 0.04$) confirme que la différence observée entre la *ResNet-18* et le modèle *3D–SIFT–Attention* n'est pas statistiquement significative. L'écart moyen d'exactitude (≈ 0.2 %) demeure négligeable, et l'intervalle de confiance à 95 % pour la différence moyenne [–0.0598, 0.0558] inclut zéro, ce qui indique que les deux architectures présentent des performances équivalentes pour la classification âge jeunes / âgés. Ainsi, le modèle interprétable proposé reproduit les performances prédictives d'une *ResNet-18* de référence, tout en offrant un avantage conceptuel en termes d'explicabilité — grâce à la génération de cartes d'attention spécifiques à chaque classe et à l'attribution de caractéristiques locales dérivées des descripteurs SIFT — comme il sera détaillé dans la section 3.3.

3.2.3 Classification d'Alzheimer sur des cohortes multiples

Pour finir, nous avons évalué le modèle hybride *3D–SIFT–Attention* sur la tâche de classification d'Alzheimer en exploitant des données issues de cohortes distinctes. Trois configurations ont été considérées : OASIS-1 ($n = 235$; 135 sains, 100 atteints), ADNI ($n = 372$; 201 sains, 171 atteints) et une configuration combinée fusionnant les deux jeux ($n = 607$). Comme le montre le Tableau 3.3, l'hybride surpasse systématiquement le *ResNet18* dans les trois cadres. Les performances sont les plus modestes sur OASIS, s'améliorent sur ADNI et atteignent leur maximum dans la configuration combinée, où l'exactitude et le score F1 sont les plus élevés.

On observe en outre qu'au moment de regrouper les cohortes, les performances du modèle de base *se dégradent*, tandis que celles de l'hybride *s'améliorent* ; cette divergence suggère que l'intégration de descripteurs SIFT, couplée à un mécanisme d'attention, confère une robustesse accrue face aux décalages inter-cohortes et à la variabilité de pose/registration, facteurs susceptibles de perturber des représentations uniquement convolutionnelles.

Tableau 3.3 Performances de classification de la maladie d'Alzheimer sur OASIS1 et ADNI (validation croisée à 5 plis)

Modèle (jeu de données)	Exactitude	Précision	Rappel	F1
ResNet18 (OASIS)	0.646 ± 0.067	0.694 ± 0.091	0.614 ± 0.086	0.586 ± 0.116
3D–SIFT–Attn (OASIS)	**0.688 ± 0.067**	**0.714 ± 0.070**	**0.690 ± 0.076**	**0.676 ± 0.074**
ResNet18 (ADNI)	0.644 ± 0.101	0.602 ± 0.204	0.624 ± 0.114	0.580 ± 0.168
3D–SIFT–Attn (ADNI)	**0.782 ± 0.057**	**0.794 ± 0.051**	**0.790 ± 0.059**	**0.780 ± 0.058**
ResNet18 (ADNI+OASIS)	0.632 ± 0.147	0.698 ± 0.035	0.624 ± 0.075	0.574 ± 0.170
3D–SIFT–Attn (ADNI+OASIS)	**0.826 ± 0.113**	**0.830 ± 0.115**	**0.830 ± 0.101**	**0.820 ± 0.119**

Enfin, ces résultats ne sont pas directement comparables à l'état de l'art car notre protocole d'entraînement est délibérément contraint (volume de données limité). Les approches de référence, qu'elles soient fondées sur des méthodes classiques ou sur l'apprentissage profond, rapportent en effet des performances très élevées pour la classification binaire Alzheimer versus témoins sains sur ADNI et OASIS. Par exemple, Desikan *et al.* (2009) exploitent des mesures régionales de volume et d'épaisseur corticale et obtiennent une discrimination pratiquement parfaite (AUC = 1,0) entre patients AD et témoins dans ADNI. Plus récemment, Shaffi, Viswan & Mahmud (2024) évaluent un ensemble de *vision transformers* pré-entraînés sur ImageNet ——ViT (Dosovitskiy, 2020), Swin (Liu *et al.*, 2021), DeiT (Touvron *et al.*, 2021), BEiT (Bao, Dong, Piao & Wei, 2021)– sur de grands jeux de coupes 2D dérivées d'OASIS et d'ADNI, et rapportent des exactitudes dans la plage haute des 90 %, avec notamment DeiT à 98,43 % sur OASIS et BEiT à 100 % sur ADNI, en s'appuyant sur plusieurs milliers de coupes par base de données. Un cadre plus proche du nôtre, en termes d'échelle de données et de choix architecturaux, est celui du réseau 3D CNN de Yagis *et al.* (2020), qui opère également sur des volumes IRM structurels 3D et est entraîné sur 100 examens AD et 100 témoins par jeu de données. Ce modèle atteint 73,4 % d'exactitude sur ADNI et 69,9 % sur OASIS avec une validation croisée à 5 plis, des valeurs bien plus comparables à nos résultats de classification AD. D'autres travaux, tels que le ResNet-18 enrichi d'attention, de convolutions *depthwise* et de modules *squeeze-and-excitation* proposé par Francis & Prakash Verma (2025), rapportent environ 93 % d'exactitude sur ADNI en s'appuyant sur un transfert d'apprentissage à partir de réseaux pré-entraînés sur des images naturelles, sur un

très grand nombre de coupes 2D et sur des cohortes de formation plus larges, ce qui représente un cadre nettement plus favorable que le nôtre. Dans notre cas, le modèle 3D–SIFT–Attn opère sur des volumes T1 3D complets, est initialisé à partir de poids pré-entraînés sur ImageNet puis entièrement affiné sur des cohortes volontairement limitées (235 volumes OASIS et 372 volumes T1 d'ADNI), et est utilisé comme une architecture unique pour plusieurs tâches, sans pré-entraînement spécifique au domaine sur de larges corpus d'IRM cérébrales 3D. Dans ces contraintes, il atteint une exactitude de 0,688 sur OASIS, 0,782 sur ADNI et 0,826 sur le jeu combiné ADNI+OASIS, ce qui est globalement cohérent avec le 3D CNN de Yagis *et al.* (2020) tout en répondant à un régime d'apprentissage plus exigeant. Étant donné que nombre de méthodes à l'état de l'art reposent sur des milliers de coupes 2D par jeu de données, sur un transfert d'apprentissage massif et parfois sur des pré-entraînements supplémentaires sur des images médicales, ainsi que sur un réglage fin spécifique à chaque base, une comparaison directe des exactitudes brutes avec nos modèles 3D compacts serait intrinsèquement inéquitable. Notre contribution doit plutôt être comprise comme exploratoire, montrant que l'intégration de descripteurs SIFT 3D, initialement développés pour l'appariement robuste et l'indexation à grande échelle d'images médicales volumétriques, permet de produire des modèles hybrides qui demeurent compétitifs tout en offrant, dans plusieurs cas, des explications plus localisées et neuro-anatomiquement pertinentes que les CNN 3D standards. Des gains supplémentaires sont donc attendus en (i) augmentant l'échelle des données (davantage de sujets, de sites et de protocoles) et/ou (ii) recourant à un pré-entraînement spécifique au domaine (par exemple auto-supervisé sur de larges corpus d'IRM cérébrales 3D), avant l'affinage supervisé sur OASIS, ADNI et l'ensemble combiné.

3.2.4 Résultats quantitatifs de l'analyse IG–KPA

Afin d'évaluer quantitativement les performances et la stabilité de la méthode proposée IG–KPA, nous avons calculé, pour chaque sujet, le nombre de pics d'attribution détectés par les gradients intégrés, le nombre de caractéristiques SIFT correspondantes, ainsi que les temps d'inférence et de calcul des attributions. Les pics représentent les maxima locaux de saillance dans les cartes

IG, tandis que les correspondances désignent les points situés à moins de 1,5 voxel d'un point clé SIFT.

Tableau 3.4 Résumé quantitatif des performances de la méthode IG–KPA. Les valeurs sont indiquées en moyenne ± écart type.

Groupe	Enregistrement	Inférence (ms)	IG (ms)	Pics	Caract. SIFT	Corrélés	Taux (%)
Hommes	Rigide	44 ± 114	3918 ± 1064	168 ± 14	5924 ± 573	140 ± 14	83 ± 5
Femmes	Rigide	123 ± 99	3484 ± 22	122 ± 15	4334 ± 397	101 ± 13	82 ± 6
Hommes	Déformable	67 ± 139	3395 ± 31	163 ± 16	6060 ± 341	136 ± 15	84 ± 3
Femmes	Déformable	54 ± 97	3400 ± 25	165 ± 4	5541 ± 373	139 ± 3	84 ± 2
Jeunes	Rigide	50 ± 90	3799 ± 26	200 ± 0	2515 ± 108	162 ± 7	81 ± 3
Âgés	Rigide	66 ± 136	3788 ± 25	200 ± 0	3620 ± 220	166 ± 4	83 ± 2
Sains	Rigide	67 ± 139	3397 ± 23	200 ± 0	3006 ± 407	184 ± 1	92 ± 1
Alzheimer	Rigide	160 ± 346	3778 ± 18	200 ± 0	3421 ± 255	182 ± 4	91 ± 2

Pour l'ensemble des jeux de données, le calcul IG–KPA nécessite environ 3,4 à 3,9 s par volume, démontrant une charge de calcul stable et prévisible, indépendante de la variabilité anatomique. Le temps d'inférence reste inférieur à 0,2 s par sujet, confirmant que le coût est dominé par le calcul des attributions. Le nombre de pics détectés demeure constant (100 à 200 par sujet), illustrant la robustesse du critère de saillance.

Sous enregistrement rigide, les deux sexes présentent des taux de correspondance comparables entre pics IG et points SIFT (82–83 %), montrant que les régions mises en évidence reflètent des structures locales cohérentes. Chez les hommes, les cartes de saillance sont légèrement plus denses (~168 pics contre 122 chez les femmes), tandis que l'enregistrement déformable améliore la cohérence spatiale dans les deux groupes (jusqu'à 84 %). Cette augmentation suggère qu'en compensant les différences morphologiques interindividuelles, le modèle identifie une proportion plus élevée de caractéristiques SIFT pertinentes pour la décision. Une tendance similaire est observée dans la tâche Alzheimer, ce qui soutient l'idée que, lorsque la classification devient plus complexe, le modèle hybride s'appuie davantage sur l'information structurale issue de SIFT pour renforcer la performance discriminante.

Dans la tâche de prédiction de l'âge, environ 200 pics de saillance sont identifiés par sujet. Les participants âgés présentent un taux de correspondance légèrement supérieur (83 %) et un plus grand nombre de points SIFT, traduisant probablement une plus forte hétérogénéité morphologique liée au vieillissement.

Dans la tâche Alzheimer, IG–KPA atteint la correspondance la plus élevée, avec plus de 91 % des pics IG coïncidant avec des points SIFT, aussi bien chez les sujets sains que malades. Cette forte cohérence soutient l'amélioration des performances de classification du modèle hybride *3D–SIFT–Attn* décrite aux Sections 3.2.3 , suggérant que le réseau exploite davantage les indices structuraux encodés par SIFT pour différencier les cerveaux Alzheimer des cerveaux sains. Combinés aux résultats obtenus avec l'enregistrement déformable, ces constats confirment que l'amélioration de la classification découle d'une plus forte sensibilité du modèle aux caractéristiques locales de structure.

3.3 Résultats d'interprétabilité

Cette section examine comment chaque modèle encode l'information morphologique liée au sexe à partir d'IRM cérébrales tridimensionnelles, afin de déterminer si les représentations internes correspondent à des régions anatomiquement pertinentes plutôt qu'à des biais globaux tels que la taille du cerveau ou les artefacts d'enregistrement. Chaque architecture est interprétée selon une approche adaptée à sa nature.

3.3.1 SIFT–MLP

La figure Figure 3.2 présente des cartes de contribution agrégées construites en déposant, à chaque caractéristique, un point flou tridimensionnel dont l'intensité est pondérée par la confiance et dont la largeur est ajustée à l'échelle. Les surimpressions bleues indiquent l'évidence en faveur du sexe masculin et les rouges celle du sexe féminin. Elles sont calculées à partir des caractéristiques extraites chez des sujets correctement classés du jeu de test. Les colonnes (de gauche à droite) représentent les caractéristiques soutenant une décision correcte “homme”, les

caractéristiques féminines mal classées comme “homme”, les caractéristiques masculines mal classées comme “femme”, et les caractéristiques soutenant une décision correcte “femme”.

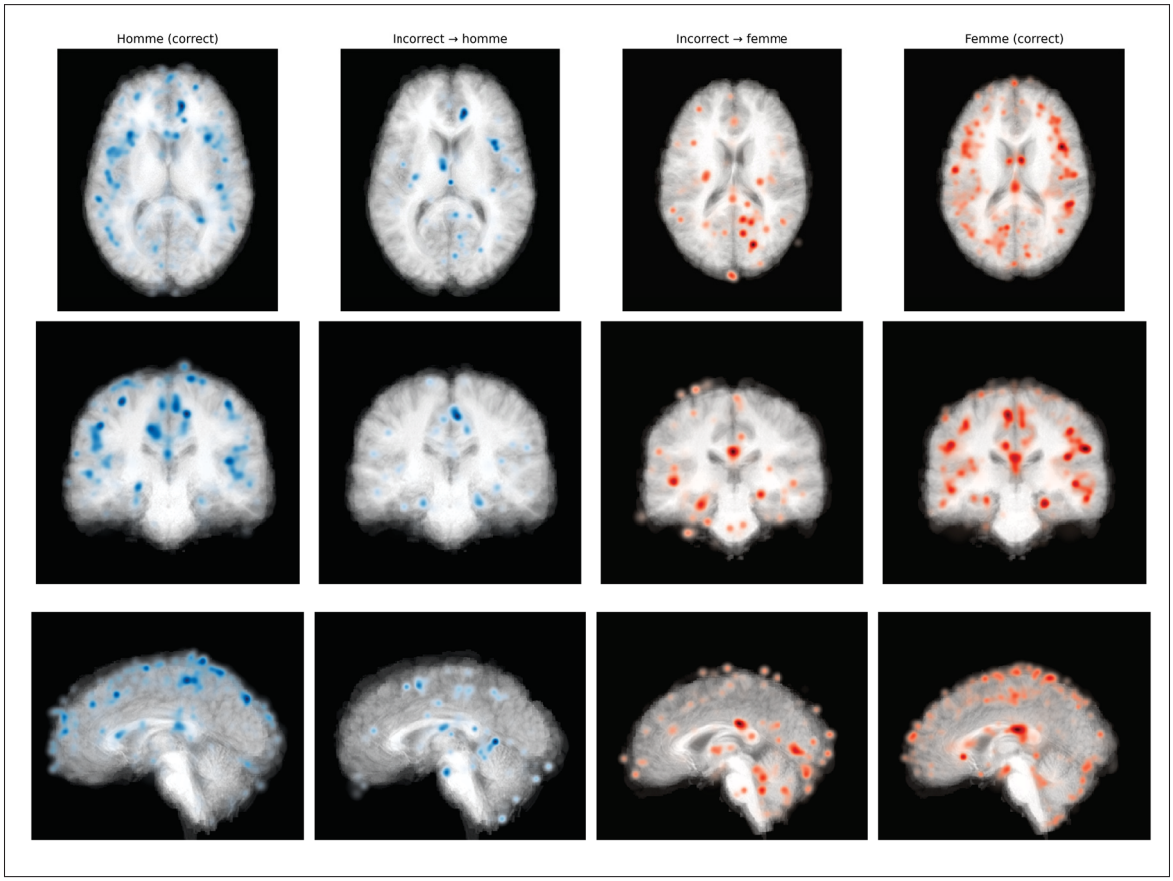

Figure 3.2 Cartes de vote locales (niveau caractéristique). Visualisation des classifications des caractéristiques SIFT au sein de sujets correctement classés. Les colonnes 1 et 4 montrent les caractéristiques correctement classées ; les colonnes 2 et 3 montrent les caractéristiques mal classées, agrégées sur des sujets correctement classés des deux sexes

Les prédictions correctes révèlent un motif bilatéral central, symétrique autour de la ligne médiane et au voisinage des ventricules. Le même patron est observé chez les femmes correctement classées ; cependant, la distribution y est plus étendue, avec des contributions supplémentaires, quoique moins intenses, au cervelet et dans les régions sous-corticales. À l’inverse, les erreurs de classification présentent des motifs plus dispersés et moins cohérents, avec des foyers orientés vers le tronc cérébral et le cervelet, ainsi que des activations isolées en surface corticale, sans la signature centrale nette observée dans les prédictions correctes.

La Figure 3.3 montre la répartition spatiale des 20% de caractéristiques les plus informatives par sujet, illustrée pour les cinq meilleurs cas classés dans chacune des quatre catégories : hommes correctement classés, femmes classées à tort comme hommes, hommes classés à tort comme femmes et femmes correctement classées. Chaque cercle correspond à une caractéristique tridimensionnelle détectée, dont la couleur représente le degré de contribution à la classe prédite et la taille, l'échelle spatiale de cette caractéristique.

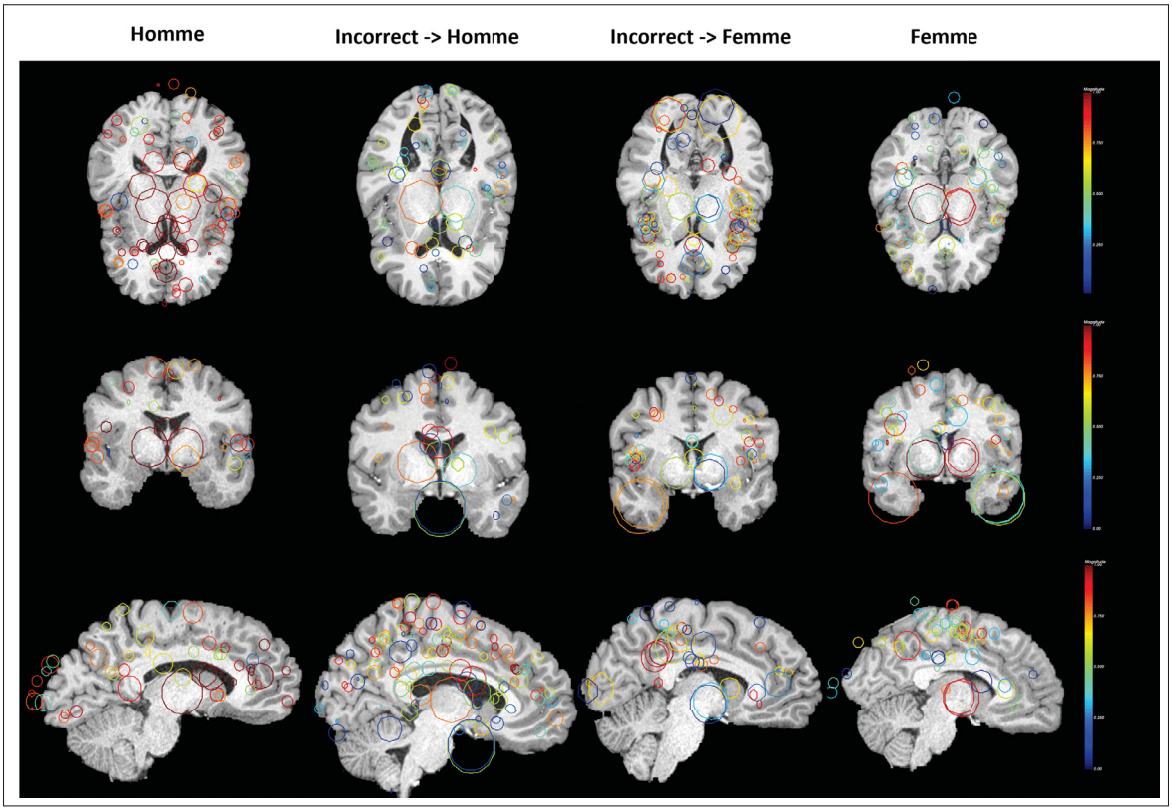

Figure 3.3 Caractéristiques tridimensionnelles les plus informatives pour les cinq sujets les mieux classés dans chacune des quatre catégories : hommes correctement classés, femmes classées à tort comme hommes, hommes classés à tort comme femmes et femmes correctement classées

On observe un schéma récurrent sous la forme de deux regroupements bilatéraux situés dans les régions thalamiques gauche et droite, présents de manière systématique chez la quasi-totalité des sujets. Chez les sujets correctement classés, ces caractéristiques thalamiques présentent une évidence plus forte, tandis que chez les sujets mal classés, leur contribution apparaît plus

faible. Cette observation rejoint l'interprétation de (Toews *et al.*, 2025), qui ont identifié des caractéristiques thalamiques bilatérales analogues, capables de prédire le sexe avec une précision d'environ 0,74 et détectées chez 97% des individus. Ces regroupements semblent donc constituer une signature anatomique commune et robuste reflétant la variabilité morphologique liée au sexe dans les images IRM tridimensionnelles.

3.3.2 CNN

Dans cette sous-section, nous interprétons les régions qui pilotent la décision du modèle de base ResNet-18 entraîné sur des IRM 3D. L'analyse porte d'abord sur la tâche principale de classification du sexe, puis sur deux tâches de validation (classification de l'âge et de la maladie d'Alzheimer). Nous utilisons des cartes Grad-CAM extraites à trois profondeurs (couche initiale, intermédiaire, finale) afin d'obtenir des indices spatiaux class-discriminants sur lesquels le réseau s'appuie.

Pour la tâche de classification du sexe, les figures 3.4 et 3.5 illustrent qualitativement les régions cérébrales les plus contributives aux décisions du modèle, respectivement pour les pipelines rigidement et déformablement enregistrés. Dans le premier cas, les couches initiales révèlent des activations fines et fragmentées le long des contours (p. ex. interfaces CSF–parenchyme), conformément à la hiérarchie de représentation typique des CNN où les premières couches captent des bords/gradients avant de composer des textures et des patrons plus globaux aux étages profonds (Zeiler & Fergus, 2014). Les couches finales concentrent les activations dans une zone bilatérale profonde, centrée autour des noyaux gris et des ventricules latéraux.

Dans le cas des images déformablement enregistrées, les premières couches présentent des activations périphériques marquées, probablement liées à des effets de bord ou d'interpolation. Les couches intermédiaires se caractérisent par des activations plus granulaires et hétérogènes, tandis que la dernière couche conserve le motif central bilatéral, accompagné d'un halo périphérique résiduel. Ces observations suggèrent que le réseau encode des représentations structurelles globales relativement stables

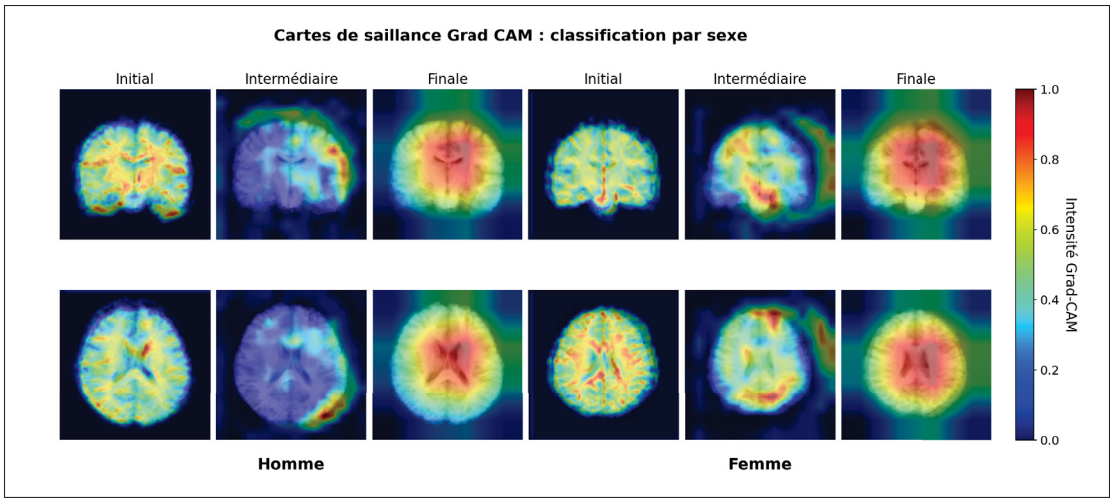

Figure 3.4 Cartes de saillance Grad-CAM moyennées pour la tâche de classification du sexe, obtenues avec le modèle de base ResNet-18 sur des IRM 3D rigidement enregistrées. Chaque colonne correspond à une profondeur du réseau (initiale, intermédiaire, finale) et chaque ligne à une vue anatomique (coronale, axiale) ; les cartes synthétisent les régions les plus contributives aux prédictions pour les groupes masculin et féminin.

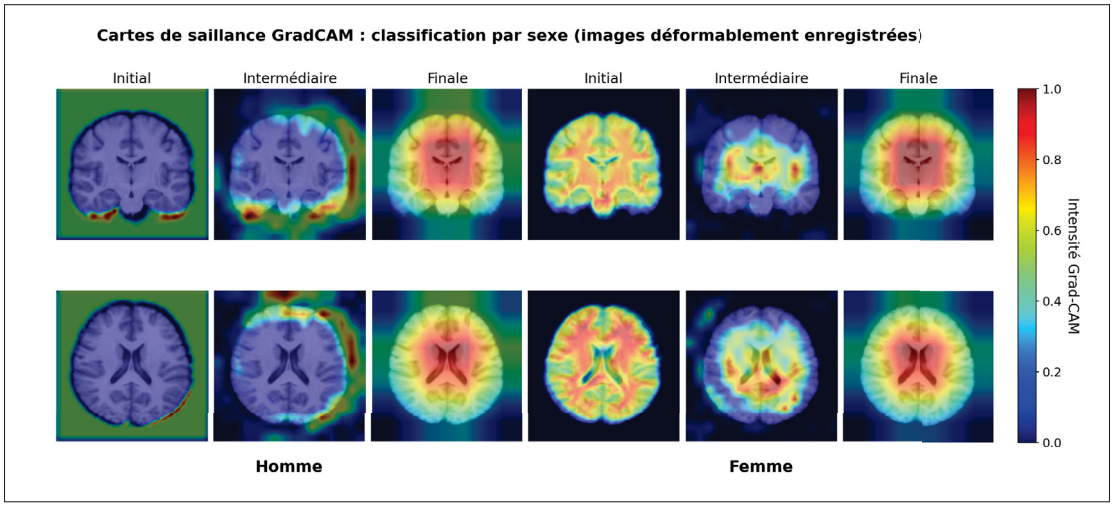

Figure 3.5 Cartes de saillance Grad-CAM moyennées pour la tâche de classification du sexe, obtenues avec le modèle de base ResNet-18 sur des IRM 3D déformablement enregistrées. Chaque colonne correspond à une profondeur du réseau (initiale, intermédiaire, finale) et chaque ligne à une vue anatomique (coronale, axiale) ; les cartes synthétisent les régions les plus contributives aux prédictions pour les groupes masculin et féminin

Les cartes Grad-CAM de la tâche de classification de l'âge (Figure 3.6) montrent que le modèle ResNet-18 distingue nettement les cerveaux jeunes et âgés. Chez les sujets jeunes, les activations précoces suivent les contours corticaux, puis s'élargissent en couches intermédiaires pour former une distribution diffuse couvrant le cortex et le cervelet. Chez les sujets âgés, les activations se concentrent dans les régions centrales profondes, traduisant la sensibilité du modèle à des marqueurs dégénératifs tels que l'élargissement ventriculaire et l'atrophie centrale. Globalement, le réseau semble reconnaître le vieillissement à travers des indices localisés de dégradation, tandis que la jeunesse est caractérisée par la préservation morphologique globale du cortex, conformément aux marqueurs IRM établis du vieillissement normal.

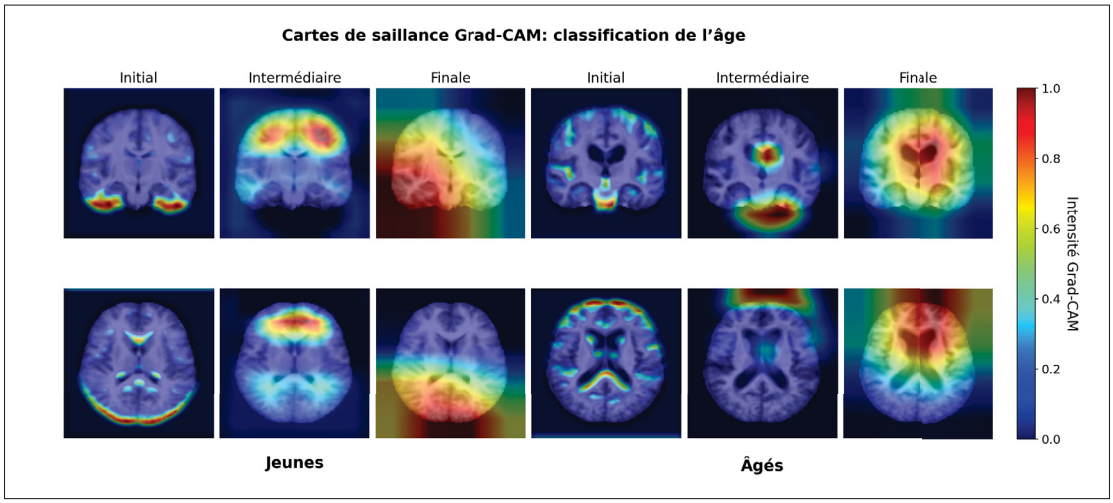

Figure 3.6 Cartes de saillance Grad-CAM moyennées pour la classification de l'âge (jeunes vs âgés), obtenues avec le modèle de base ResNet-18 sur des IRM 3D. Les colonnes représentent les profondeurs du réseau (initiale, intermédiaire, finale) et les lignes les vues anatomiques (coronale, axiale). Les cartes résument les régions les plus contributives aux décisions pour chaque groupe d'âge.

Pour la maladie d'Alzheimer (Figure 3.7), les couches initiales des sujets sains et atteints présentent des activations périphériques complémentaires. Dans les couches intermédiaires, la saillance se resserre progressivement dans les régions temporales médiales chez les patients, indiquant que le réseau commence à détecter des motifs de désorganisation structurelle caractéristiques de la pathologie, tandis que chez les sujets sains, l'activation adopte la forme d'un halo diffus, traduisant une organisation corticale homogène. Enfin, dans les couches finales, les deux classes

présentent une concentration bilatérale de la saillance autour des ventricules, sans qu'il soit possible d'identifier des sous-structures locales clairement différenciées.

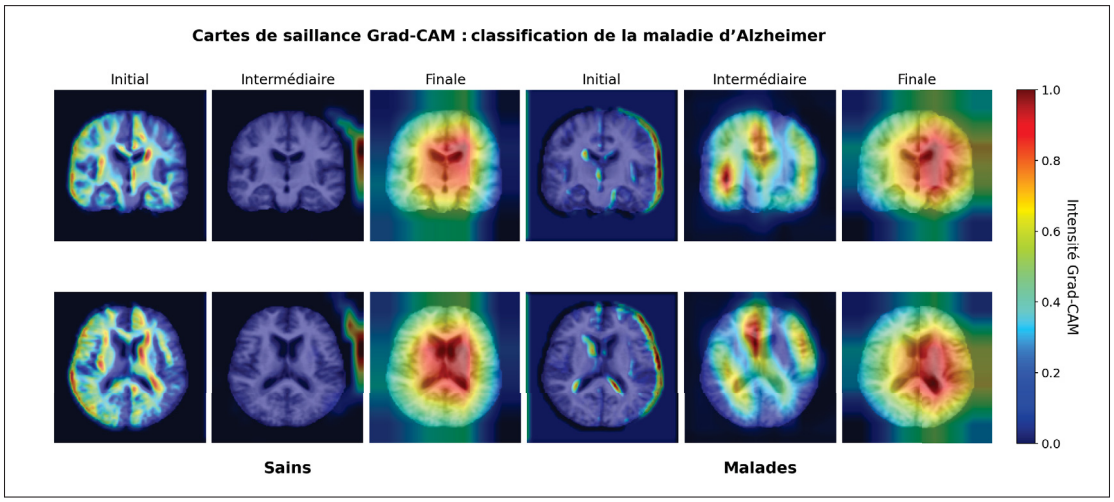

Figure 3.7 Cartes de saillance Grad-CAM moyennées pour la classification de la maladie d'Alzheimer (sujets sains vs patients atteints), obtenues avec le modèle de base ResNet-18 sur des IRM 3D. Chaque colonne correspond à une profondeur du réseau (initiale, intermédiaire, finale) et chaque ligne à une vue anatomique (coronale, axiale) ; les cartes mettent en évidence les régions les plus contributives aux décisions pour chaque groupe

De manière générale, ces résultats confirment que les activations Grad-CAM illustrent une hiérarchie de représentation typique des réseaux convolutifs : les couches précoces captent des caractéristiques locales de bas niveau (bords, contours), tandis que les couches profondes intègrent des informations morphologiques globales. Toutefois, la diffusion spatiale des cartes aux étages finaux reflète une limite connue de la méthode : bien qu'efficace pour une lecture qualitative, Grad-CAM tend à produire des cartes larges et peu spécifiques, sans garantie de granularité fine. Ces cartes demeurent néanmoins informatives et cohérentes avec les régions anatomiques pertinentes, ce qui valide qualitativement la capacité du modèle à exploiter des indices neuroanatomiques plausibles.

3.3.3 3D–SIFT–Attention

Après l'analyse post hoc par Grad-CAM du modèle de base ResNet-18, nous introduisons un modèle hybride 3D–SIFT–Attention visant une interprétabilité plus intrinsèque tout en

préservant les visualisations Grad-CAM sur la branche ResNet. La motivation est de rapprocher le processus décisionnel d'entités anatomiques locales et stables : des points d'intérêt 3D SIFT décrivant la morphologie fine, sur lesquels un mécanisme d'attention croisée apprend à allouer des poids explicites en fonction du contexte global de l'image. Pour renforcer cette lecture locale, nous appliquons sur la branche SIFT la méthode des IG-KPA (voir section 2.2.1), utilisée ici pour identifier et conserver les points clés les plus contributifs à la prédiction.

La Figure 3.8 présente les cartes Grad-CAM calculées sur la branche ResNet du modèle proposé 3D–SIFT–Attention à différentes profondeurs du réseau. Sous enregistrement rigide, les couches initiales (et, dans une moindre mesure, intermédiaires) révèlent un biais marqué de taille et contour global. Après enregistrement déformable, qui normalise la taille du cerveau et réduit la variabilité de forme inter-sujets, cette prédominance globale s'atténue fortement. À la couche finale, les entrées rigidement enregistrées présentent des foyers dépendants de la classe, tandis que, sous enregistrement déformable, la saillance converge pour les deux sexes vers des régions centrales, indiquant une dépendance à des indices structurels fins plutôt qu'à des artefacts superficiels ou globaux. Ce basculement confirme que le contrôle de la taille crânienne via l'alignement déformable incite le réseau à capturer des variations morphologiques subtiles de type texture, fournissant une base d'interprétation plus pertinente.

Les Figures 3.9 et 3.10 présentent (a) les cartes d'attention croisée et (b) les attributions IG-KPA pour la classification du sexe dans le modèle hybride 3D–SIFT–Attention. Dans les deux conditions, l'attention se concentre dans des régions profondes bilatérales — principalement périventriculaires et cingulo-callosales, avec des extensions thalamo-capsulaires — et le motif spatial est très similaire chez les hommes et les femmes, ce qui indique un recours à des indices locaux partagés plutôt qu'à des localisations anatomiques distinctes (cohérent avec l'invariance d'échelle des descripteurs SIFT).

Par rapport au rigide, l'enregistrement déformable produit des cartes d'attention plus symétriques et une constellation plus dense de points clés IG-KPA contributifs dans le thalamus. En co-enregistrant l'anatomie homologue et en normalisant la taille du cerveau, l'alignement déformable

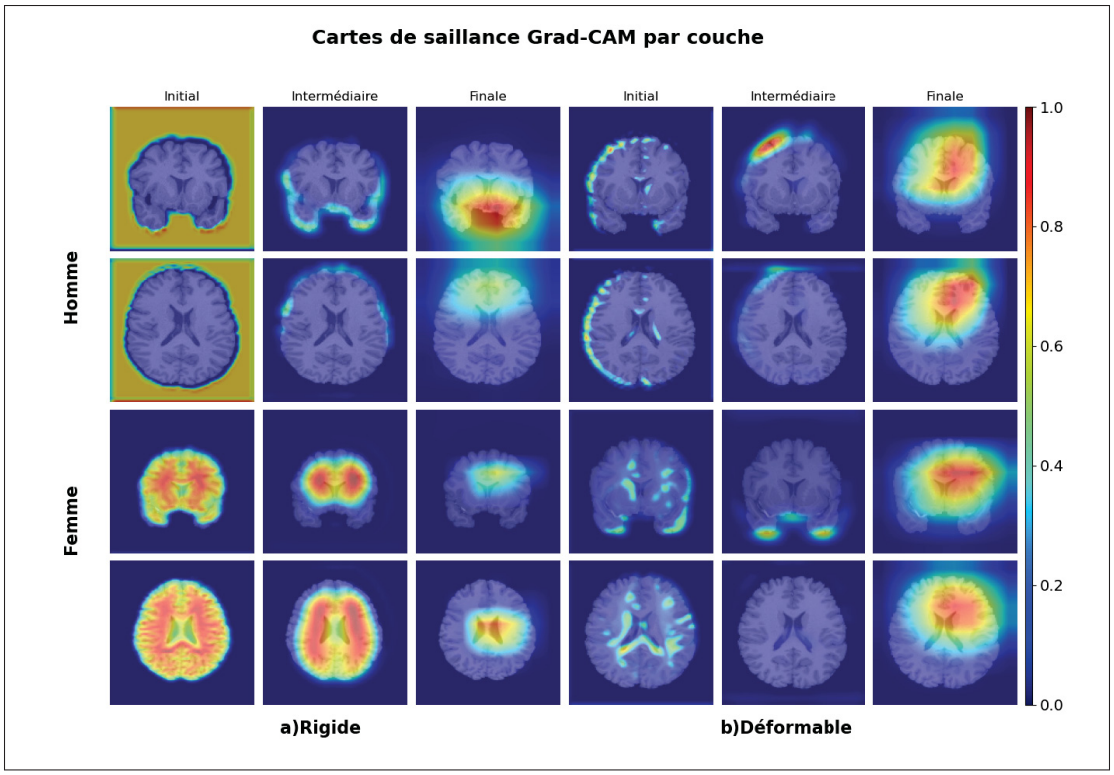

Figure 3.8 Cartes de saillance Grad-CAM de la branche ResNet du modèle hybride 3D SIFT-Attention, par couche et type d'enregistrement

réduit les confusions liées au contour et redistribue l'attribution des contours globaux vers des caractéristiques co-localisées autour du thalamus, rapprochant encore les cartes homme–femme et mettant l'accent sur des indices fins de texture plutôt que sur des signaux géométriques grossiers.

Pour la tâche de classification de l'âge, le modèle présente une signature centrée sur les ventricules à travers les différentes méthodes d'interprétabilité. Les couches initiales de Grad-CAM semblent capter la taille globale, tandis que les couches tardives s'élargissent et culminent autour des ventricules latéraux. Les cartes d'attention croisée confirment une mise en évidence périventriculaire chez les sujets âgés. IG-KPA, invariant d'échelle au niveau des caractéristiques, se concentre dans des régions profondes de la ligne médiane (incluant des indices thalamiques), reproduit une activité le long des parois ventriculaires et, dans certains cas âgés, fait apparaître un foyer hippocampique focal. Les sujets jeunes présentent plutôt des motifs plus compacts

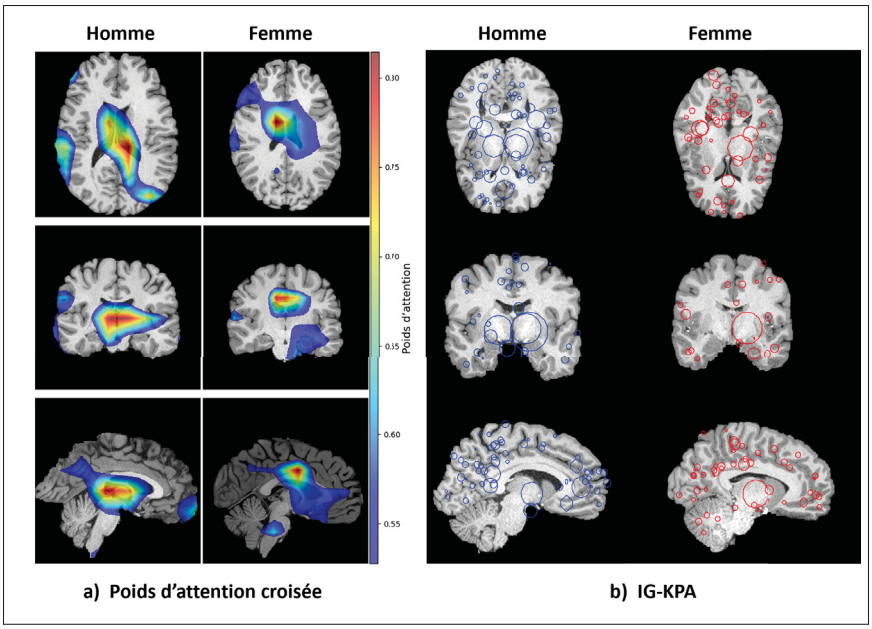

Figure 3.9 Interprétations sous enregistrement rigide pour la tâche sexe avec le modèle hybride 3D SIFT-Attention (a) Poids d'attention croisée (b) IG-KPA

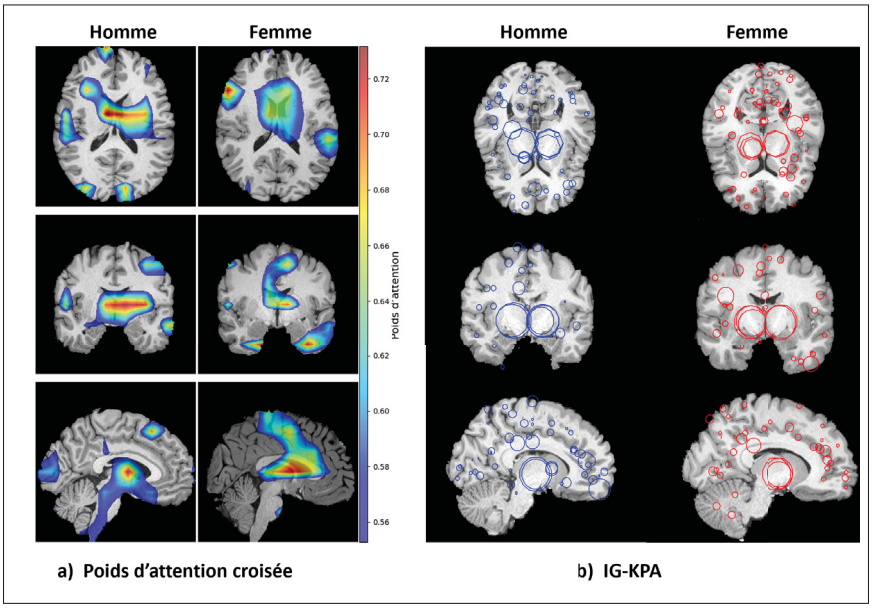

Figure 3.10 Interprétations sous enregistrement déformable pour la tâche sexe avec 3D SIFT-Attention. (a) Poids d'attention croisée (b) IG-KPA

et centraux, avec une diffusion périphérique plus faible. Ces observations convergentes sont plausibles biologiquement : l'élargissement ventriculaire et des modifications morphométriques diffuses accompagnent le vieillissement sain, et le volume cérébral diminue avec l'âge. Toutefois, puisque les saillances précoces reflètent clairement la taille et que nous ne l'avons pas corrigée, nous évitons des affirmations régionales fortes au-delà de la robuste signature centrée sur les ventricules.

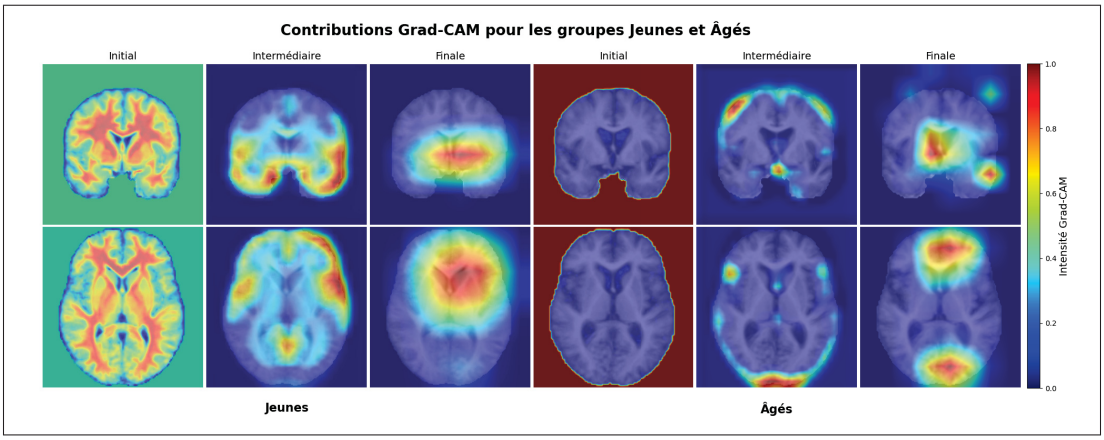

Figure 3.11 Cartes de saillance Grad-CAM de la branche ResNet du modèle hybride 3D SIFT-Attention pour la tâche de classification d'âge

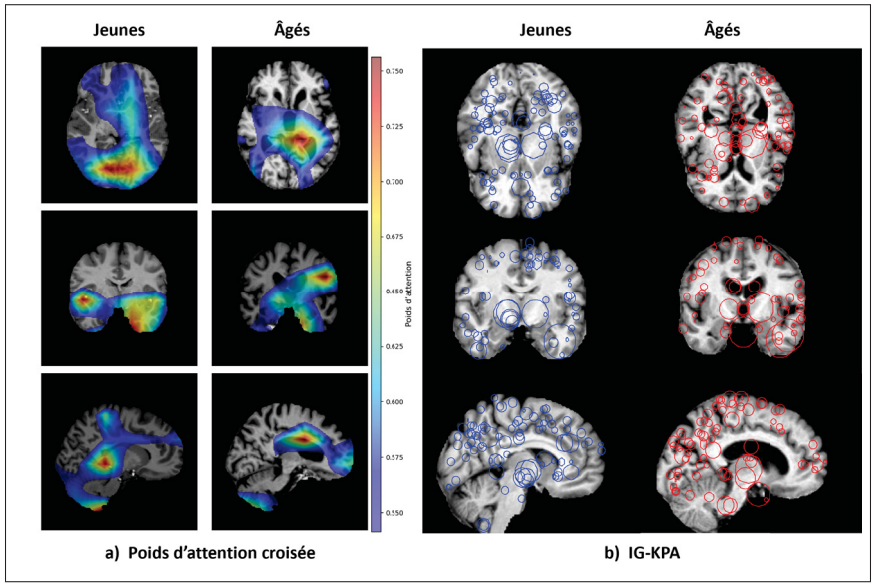

Figure 3.12 Interprétations pour la tâche âge (Jeunes vs Âgés) avec 3D SIFT-Attention. (a) Poids d'attention croisée (b) IG-KPA

De même, pour la classification d'Alzheimer, les Figures 3.12–3.13 révèlent des motifs morphométriques compatibles avec la maladie. Grad-CAM (Fig. 3.12) met en évidence de larges régions profondes autour des ventricules, signe d'une atrophie globale. À l'inverse, l'attention croisée (Fig. 3.13a) montre des réponses plus nettes et de plus forte amplitude le long des parois ventriculaires, au sein des structures sous-corticales (substance grise profonde) et au niveau des sillons corticaux—prononcées chez les patients et minimales chez les témoins. En complément, IG-KPA (Fig. 3.13b) concentre des grappes de points clés le long des bords entre les cavités remplies de liquide et le tissu cérébral—incluant les ventricules, les zones sous-corticales adjacentes et les limites médio-temporales—capturant des indices fins de texture et de bord caractéristiques de la perte tissulaire.

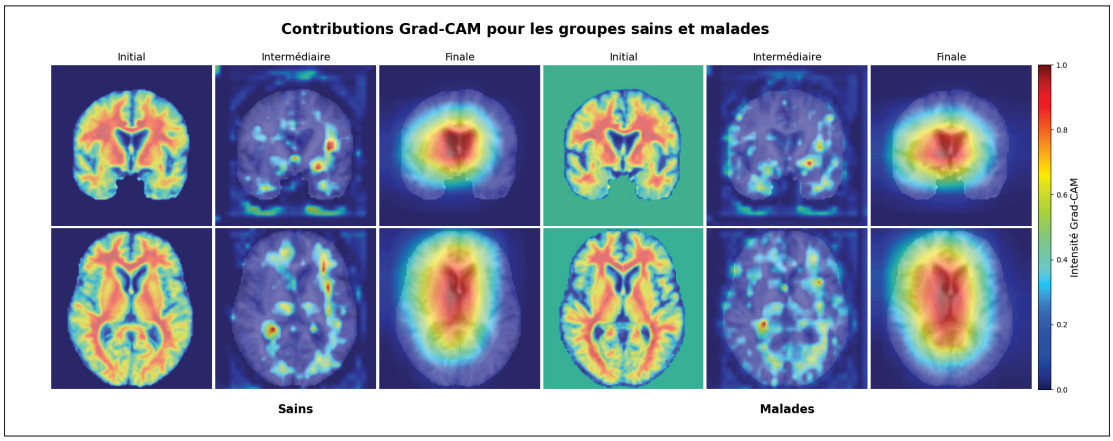

Figure 3.13 Cartes de saillance Grad-CAM de la branche ResNet du modèle hybride 3D SIFT-Attention pour la tâche de classification Alzheimer

3.4 Discussion et limites

Ce chapitre a présenté les résultats quantitatifs et qualitatifs de différentes stratégies de modélisation pour la classification à partir d'IRM structurelles, allant de modèles interprétables peu profonds à des architectures profondes. Le modèle hiérarchique SIFT-MLP, bien que dépassé en exactitude par des réseaux plus profonds, a établi une solide base interprétable. Il produit des schémas de saillance localisables, fondés sur des points clés, en accord avec l'anatomie discriminante, offrant ainsi une lecture spatiale fine. À l'inverse, les architectures ResNet

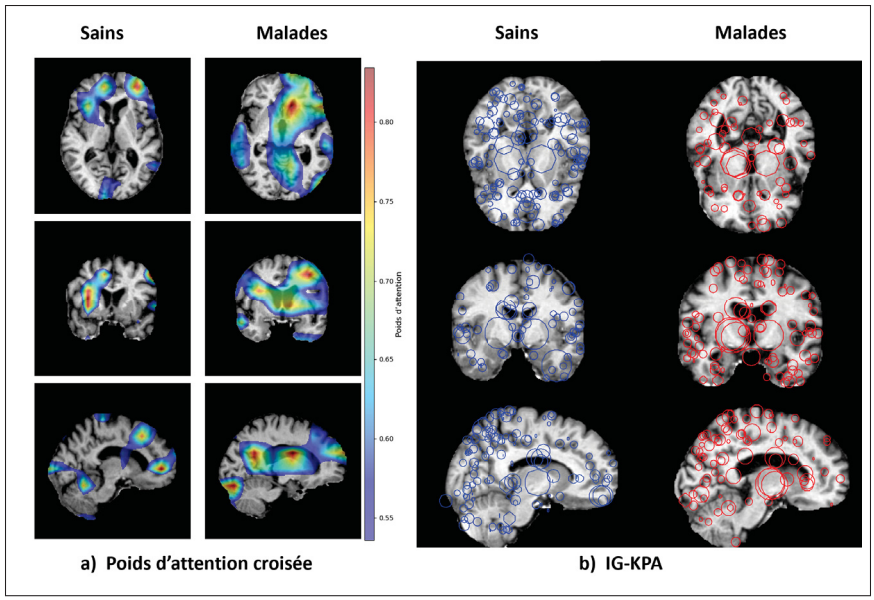

Figure 3.14 Interprétations pour la classification Alzheimer (Sains vs Malades) avec 3D SIFT-Attention. (a) Poids d'attention croisée (b) IG-KPA

atteignent des performances supérieures mais avec une explicabilité plus limitée. L'analyse Grad-CAM du ResNet de base met en évidence une progression typique des premières couches (sensibles aux bords) vers des activations tardives larges et diffuses. Malgré ces différences, Figure 3.15 montre une forte concordance des sorties modèles : la corrélation de Spearman entre ResNet-18 et SIFT-MLP atteint $\rho = 0{,}879$, indiquant que ces modèles—bien qu'opérant sur des représentations distinctes—ordonnent les sujets de manière similaire. Cette concordance monotone justifie le recours à des méthodes fondées sur SIFT comme appui interprétatif des décisions des CNN dans des tâches de neuro-imagerie fines comme la classification du sexe. Sans prétendre à une causalité, l'alignement avec les prédictions du ResNet en fait un proxy fiable pour repérer les régions probablement déterminantes.

Dans cette logique, nous proposons une architecture hybride multi-branche combinant la puissance discriminante des ResNet profonds et la précision spatiale de SIFT. Ce modèle montre des performances améliorées ou comparables selon les tâches, tout en mettant en avant des régions neuroanatomiques plausibles. En plus, la Figure 3.16 met en évidence un contraste clair : le modèle de base ResNet diffuse sa saillance et inclut des régions hors cerveau, tandis que, dans

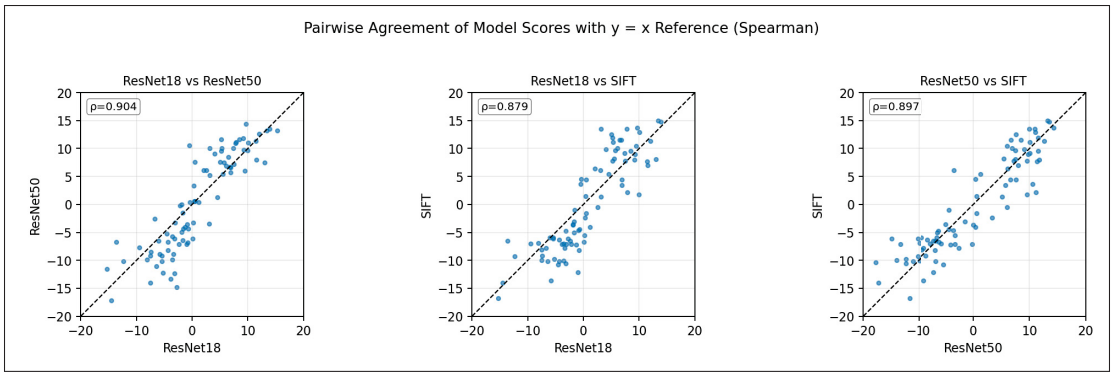

Figure 3.15 Accord par paires des scores des modèles par rapport à la référence $y = x$ (en tirets). Les corrélations de Spearman sont indiquées dans chaque panneau : $\rho = 0{,}904$ (ResNet18 vs ResNet50), $\rho = 0{,}879$ (ResNet18 vs SIFT) et $\rho = 0{,}897$ (ResNet50 vs SIFT).

l'architecture hybride, la branche ResNet présente des activations plus précises, confinées au tissu cérébral, ce qui est cohérent avec une meilleure contrainte anatomique et suggère que la branche SIFT régularise l'attention du CNN vers des structures morphologiques pertinentes.

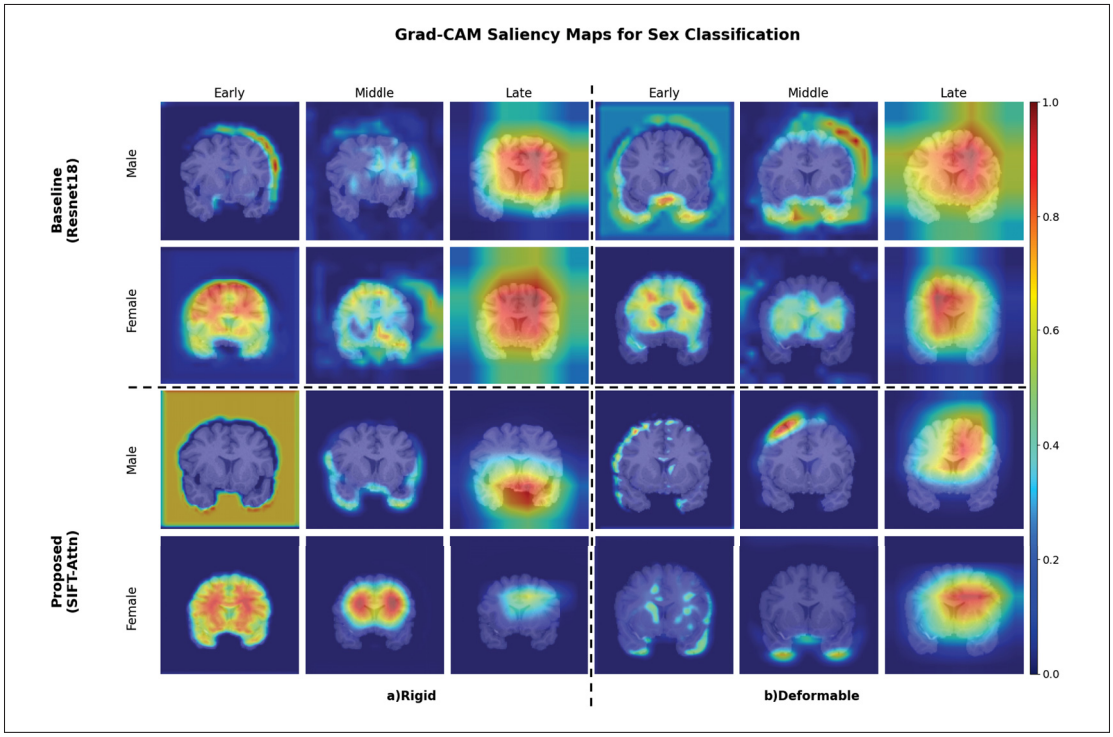

Figure 3.16 Saillance Grad-CAM pour la classification du sexe — ResNet-18 de référence vs modèle proposé 3D-SIFT–Attention

Cependant, ces gains en interprétabilité ont toutefois un coût en complexité. L'ajout de la branche SIFT introduit des couches et paramètres supplémentaires, augmentant l'empreinte mémoire et le calcul. Pour y remédier, des CNN clairsemés (*Sparse CNN (Liu, Wang, Foroosh, Tappen & Pensky, 2015)*) dans la branche SIFT permettrait de traiter uniquement les zones actives du volume, préservant ainsi la résolution spatiale locale tout en réduisant le nombre d'opérations et de paramètres effectifs.

Sur le plan anatomique, les modèles identifient de façon cohérente des régions pertinentes pour chaque tâche. Pour la classification du sexe, les activations ciblent fréquemment les structures thalamiques bilatérales profondes, les zones périventriculaires, ainsi que les cortex occipital et temporal. Ces observations concordent avec des études morphométriques décrivant des différences liées au sexe dans le cervelet, le thalamus et les régions postérieures. Le caractère mosaïque des marqueurs liés au sexe—distribués et souvent partagés par des sous-ensembles de sujets—apparaît également dans nos résultats. Les modèles révèlent des corrélations parcimonieuses mais cohérentes à l'échelle du cerveau, ce qui renforce l'hypothèse de la mosaïque (Joel, 2021) et rejoint les observations de (Ebel *et al.*, 2023) et (Toews *et al.*, 2025), selon lesquelles les traits informatifs ne se concentrent pas en une seule région mais se répartissent sur des zones anatomiques diverses, avec une variabilité inter-individuelle.

Pour l'âge et Alzheimer, les modèles s'attachent principalement aux régions ventriculaires et périventriculaires, en accord avec des marqueurs établis de vieillissement et de neurodégénérescence (élargissement des ventricules latéraux, atrophie des structures adjacentes). Les cartes Grad-CAM et les attributions IG-KPA de l'hybride se localisent de façon constante dans ces zones, ce qui confère une validité de façade au raisonnement du modèle. Dans le contexte inter-cohortes ADNI+OASIS, l'hybride généralise mieux, ce qui est probablement dû à à l'invariance d'échelle des descripteurs SIFT et à leur résistance aux biais spécifiques de cohorte.

CONCLUSION ET RECOMMANDATIONS

Cette thèse explore des approches de classification de phénotypes neuroanatomiques fins à partir d'IRM structurelles tout en préservant l'interprétabilité. Le chapitre 1 situe le travail dans le paysage actuel de l'imagerie cérébrale et de l'apprentissage automatique, en retraçant la transition des descripteurs locaux vers des architectures profondes de bout en bout. Il met en évidence la tension récurrente entre performances et interprétabilité, ainsi que l'exigence d'explications fiables dans des contextes à forts enjeux comme l'imagerie médicale. Cette revue motive l'idée directrice de la thèse : combiner une morphologie locale stable et invariante d'échelle avec un modèle contextuel global, puis évaluer les explications à l'aide d'outils qui rattachent la saillance à des structures anatomiques vérifiables.

Le chapitre 2 présente une méthodologie comparative couvrant quatre familles de modèles et un flux d'explicabilité unifié. L'objectif est de concilier la précision des CNN avec la localité et la reproductibilité des indices. D'abord, un pipeline hiérarchique SIFT-MLP classe les points d'intérêt SIFT 3D et agrège les votes locaux au niveau du sujet, établissant un socle interprétable ancré dans l'anatomie. Ensuite, des CNN 3D (ResNet-18/50) sont entraînés de bout en bout sur les volumes T1 bruts ; en parallèle, un SIFT-CNN consomme des volumes de descripteurs SIFT 3D pour évaluer l'apport de convolutions apprises sur des caractéristiques déjà invariantes. Enfin, une architecture hybride 3D–SIFT–Attention fusionne le contexte global appris par le CNN avec la morphologie locale stable via de l'attention croisée, afin d'exploiter explicitement des échelles spatiales complémentaires. Pour rendre les explications auditables, le chapitre introduit IG–KPA, qui transforme la saillance issue des Integrated Gradients en ensembles discrets de points SIFT appariés, directement interprétables d'un point de vue anatomique.

Le Chapitre 3 montre que le modèle hybride atteint des performances de pointe pour la classification du sexe sur HCP en enregistrement rigide, avec une exactitude maximale de 94

Le modèle hybride égale par ailleurs ResNet-18 pour la classification de l'âge (OASIS), tout en offrant des explications plus riches, et surpasse la base pour la maladie d'Alzheimer sur OASIS, ADNI et leur combinaison, avec des gains maximaux en contexte multi-cohortes. L'accord élevé entre les sorties SIFT et ResNet justifie l'usage d'attributions fondées sur SIFT comme proxy interprétatif lorsque les cartes Grad-CAM du CNN sont diffuses. De façon cohérente, les Grad-CAM de la branche ResNet dans l'hybride sont plus localisées et confinées au cerveau que dans le modèle de base, suggérant une régularisation guidée par SIFT vers des régions morphologiquement pertinentes. Selon les tâches, l'attention croisée et l'analyse IG–KPA mettent en évidence des motifs en accord avec la littérature : territoires bilatéraux profonds thalamo-capsulaires et périventriculaires pour le sexe, et signatures centrées sur les ventricules pour l'âge et Alzheimer.

Pour la suite, nous recommandons de renforcer la fidélité des explications par des sanity checks et des contre-factuels/ablations autour des amas SIFT dominants et de tirer parti d'un pré-entraînement auto-supervisé spécifique à l'IRM 3D.

BIBLIOGRAPHIE

Akindele, R. G., Cen, S., Yu, M., Aribilola, I., Xinrang, T. & Liu, Y. (2025). A hybrid attention-based deep learning framework for precise early diagnosis of Alzheimer's disease. *Discover Applied Sciences*, 7(8), 856.

Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J. & Gidwani, M. (2021). Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology : Artificial Intelligence*, 3(6), e200267.

Bao, H., Dong, L., Piao, S. & Wei, F. (2021). Beit : Bert pre-training of image transformers. *arXiv preprint arXiv :2106.08254*.

Basaia, S., Agosta, F., Wagner, L., Canu, E., Magnani, G., Santangelo, R. & Filippi, M. (2019). Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage : Clinical*, 21, 101645.

Bau, D., Zhou, B., Khosla, A., Oliva, A. & Torralba, A. (2017). Network dissection : Quantifying interpretability of deep visual representations. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549.

Bhadra, S. R., Maity, S. P. & Sil, J. (2024). Enhancement of Alzheimer's Disease Detection Technique by Fusing Features of Deep Learning and Handcrafted Methods on sMRI Images. *2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, pp. 1079–1090.

Borys, K., Schmitt, Y. A., Nauta, M., Seifert, C., Krämer, N., Friedrich, C. M. & Nensa, F. (2023). Explainable AI in medical imaging : An overview for clinical practitioners–Beyond saliency-based XAI approaches. *European journal of radiology*, 162, 110786.

Brueggeman, L., Thomas, T. R., Koomar, T., Hoskins, B. & Michaelson, J. J. (2020). Sex differences in the brain : Divergent results from traditional machine learning and convolutional networks. *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 899–903.

Chauvin, L., Kumar, K., Desrosiers, C. & Toews, M. (2020). Neuroimage signature from salient keypoints is highly discriminative. *NeuroImage*, 205, 116301.

Chauvin, L., Kumar, K., Desrosiers, C., Wells, W. & Toews, M. (2021). Efficient pairwise neuroimage analysis using the soft jaccard index and 3d keypoint sets. *IEEE transactions on medical imaging*, 41(4), 836–845.

Choudhary, S. & Kesswani, N. (2020). Analysis of KDD-Cup'99, NSL-KDD and UNSW-NB15 datasets using deep learning in IoT. *Procedia Computer Science*, 167, 1561–1573.

Cole, J. H., Poudel, R. S., Tsagkrasoulis, D., Caan, M. W., Steves, C., Spector, T. D. & Montana, G. (2017). Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, 163, 115–124.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M. O., Chupin, M., Benali, H. & Colliot, O. (2011). Automatic classification of patients with Alzheimer's disease from structural MRI : a comparison of ten methods using the ADNI database. *NeuroImage*, 56(2), 766–781.

Dalal, N. & Triggs, B. (2005). Histograms of oriented gradients for human detection. *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, 1, 886–893.

De Pietro, S., Di Martino, G., Caroprese, M., Barillaro, A., Cocozza, S., Pacelli, R., Cuocolo, R., Ugga, L., Briganti, F., Brunetti, A. et al. (2024). The role of MRI in radiotherapy planning : a narrative review "from head to toe". *Insights into Imaging*, 15(1), 255.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009). Imagenet : A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255.

Desikan, R. S., Cabral, H. J., Hess, C. P., Dillon, W. P., Glastonbury, C. M., Weiner, M. W., Schmansky, N. J., Greve, D. N., Salat, D. H., Buckner, R. L. et al. (2009). Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. *Brain*, 132(8), 2048–2057.

Dibaji, M., Ospel, J., Souza, R. & Bento, M. (2024). Sex differences in brain MRI using deep learning toward fairer healthcare outcomes. *Frontiers in Computational Neuroscience*, 18, 1452457.

Dimitrov, I., Atanasova, S., Kaprelyan, A., Ivanov, B., Nestorova, V., Drenska, K., Chuperkova, Z. & Aleksandrov, I. (2018). Gerstmann syndrome in a young man : a case report. *Trakia Journal of Sciences*, 16(3), 239.

Diogo, V. S., Ferreira, H. A. & Prata, D. (2022). Early diagnosis of Alzheimer's disease using machine learning : a multi-diagnostic, generalizable approach. *Alzheimer's Research & Therapy*, 14(1), 107.

Dorfner, F. J., Patel, J. B., Kalpathy-Cramer, J., Gerstner, E. R. & Bridge, C. P. (2025). A review of deep learning for brain tumor analysis in MRI. *NPJ Precision Oncology*, 9(1), 2.

Dosovitskiy, A. (2020). An image is worth 16x16 words : Transformers for image recognition at scale. *arXiv preprint arXiv :2010.11929*.

Ebel, M., Domin, M., Neumann, N., Schmidt, C. O., Lotze, M. & Stanke, M. (2023). Classifying sex with volume-matched brain MRI. *NeuroImage : Reports*, 3(1), 100181.

Eliot, L., Ahmed, A., Khan, H. & Patel, J. (2021). Dump the "dimorphism" : Comprehensive synthesis of human brain studies reveals few male-female differences beyond size. *Neuroscience & Biobehavioral Reviews*, 125, 667–697.

Erhan, D., Bengio, Y., Courville, A. & Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal*.

Feis, D. L., Brodersen, K. H., von Cramon, D. Y., Luders, E. & Tittgemeyer, M. (2013). Decoding gender dimorphism of the human brain using multimodal anatomical and diffusion MRI data. *NeuroImage*, 70, 250–257.

Francis, S. B. & Prakash Verma, J. (2025). Deep CNN ResNet-18 based model with attention and transfer learning for Alzheimer's disease detection. *Frontiers in Neuroinformatics*, 18, 1507217.

Goel, P., Kapse, S., Pati, P. & Prasanna, P. (2024). Coca-Mil : Attention-Based Handcrafted-Deep Feature Fusion in Computational Pathology. *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5.

Guo, X., Ding, Y., Xu, W., Wang, D., Yu, H., Lin, Y., Chang, S., Zhang, Q. & Zhang, Y. (2024). Predicting brain age gap with radiomics and automl : A Promising approach for age-Related brain degeneration biomarkers. *Journal of Neuroradiology*, 51(3), 265–273.

Harris, C., Stephens, M. et al. (1988). A combined corner and edge detector. *Alvey vision conference*, 15(50), 10–5244.

Hasan, A. M., Jalab, H. A., Meziane, F., Kahtan, H. & Al-Ahmad, A. S. (2019). Combining deep and handcrafted image features for MRI brain scan classification. *IEEE Access*, 7, 79959–79967.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hubel, D. H. & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1), 106.

Huff, L., Papanastasiou, G., Soares, J., Salehi, M. & Frangi, A. F. (2021). Advances in explainable deep learning in medical imaging. *Frontiers in Artificial Intelligence*, 4, 668781.

Jack Jr, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L. Whitwell, J., Ward, C. et al. (2008). The Alzheimer's disease neuroimaging initiative (ADNI) : MRI methods. *Journal of Magnetic Resonance Imaging : An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4), 685–691.

Joel, D. (2021). Beyond the binary : Rethinking sex and the brain. *Neuroscience & Biobehavioral Reviews*, 122, 165–175.

Kim, J.-Y., Kim, D., Lee, S., Kim, K. & Yoo, H.-J. (2009). Visual image processing ram : memory architecture with 2-D data location search and data consistency management for a multicore object recognition processor. *IEEE transactions on circuits and systems for video technology*, 20(4), 485–495.

Kingma, D. P. (2014). Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980*.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S. et al. (2020). Captum : A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv :2009.07896*.

Korolev, S., Safiullin, A., Belyaev, M. & Dodonova, Y. (2017). Residual and plain convolutional neural networks for 3D brain MRI classification. *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI)*, pp. 835–838.

Lancaster, J., Lorenz, R., Leech, R. & Cole, J. H. (2018). Bayesian optimization for neuroimaging pre-processing in brain age classification and prediction. *Frontiers in Aging Neuroscience*, 10, 28.

LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. (2002). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.

Li, H., Zhang, K., Song, L. & Liu, Y. (2017). Classification of brain disease in magnetic resonance images using two-stage local feature fusion. *PLoS ONE*, 12(2), e0171749.

Li, J. & Allinson, N. M. (2008). A comprehensive review of current local features for computer vision. *Neurocomputing*, 71(10-12), 1771–1787.

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system : Past, present, and future. *Journal of cognitive neuroscience*, 33(10), 2017–2031.

Liu, B., Wang, M., Foroosh, H., Tappen, M. & Pensky, M. (2015). Sparse convolutional neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 806–814.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. & Guo, B. (2021). Swin transformer : Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022.

Lotze, M., Domin, M., Gerlach, F. H., Gaser, C., Lueders, E., Schmidt, C. O. & Neumann, N. (2019). Novel findings from 2,838 adult brains on sex differences in gray matter brain volume. *Scientific reports*, 9(1), 1671.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 91–110.

Lundberg, S. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, pp. 4765–4774.

Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C. & Buckner, R. L. (2007). Open Access Series of Imaging Studies (OASIS) : cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9), 1498–1507.

McCarthy, M. M., Arnold, A. P., Ball, G. F., Blaustein, J. D. & De Vries, G. J. (2012). Sex differences in the brain : the not so inconvenient truth. *Journal of Neuroscience*, 32(7), 2241–2247.

Meier, T. B., Desphande, A. S., Vergun, S., Nair, V. A., Song, J., Biswal, B. B., Meyerand, M. E., Birn, R. M. & Prabhakaran, V. (2012). Support vector machine classification and characterization of age-related reorganization of functional brain networks. *NeuroImage*, 60(1), 601–613.

Morevec, H. P. (1977). Towards automatic visual obstacle avoidance. *Proceedings of the 5th international joint conference on Artificial intelligence-Volume 2*, pp. 584–584.

Nasser, F., Pereira, C. & Esteva, A. (2024). Utilizing radiomic features for automated MRI keypoint detection. *Proceedings of BIOSTEC 2024*.

Pieper, S., Halle, M. & Kikinis, R. (2004). 3D Slicer. *2004 2nd IEEE international symposium on biomedical imaging : nano to macro (IEEE Cat No. 04EX821)*, pp. 632–635.

Qiu, A., Zhou, Y. & Lin, W. (2023). Interpretation of convolutional neural networks for sex classification of brain MRI. *NeuroImage*, 265, 119741.

Ren, J., Li, C., An, Y., Zhang, W. & Sun, C. (2024). Few-shot fine-grained image classification : A comprehensive review. *AI*, 5(1), 405–425.

Ribeiro, M. T., Singh, S. & Guestrin, C. (2016). Why should i trust you ? Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.

Rister, B., Yi, D., Shivakumar, K. & Rubin, D. L. (2017). Volumetric image registration from invariant keypoints. *IEEE Transactions on Image Processing*, 26(10), 4900–4910.

Ritchie, S. J., Cox, S. R., Shen, X., Lombardo, M. V., Reus, L. M., Alloza, C., Harris, M. A., Alderson, H. L., Hunter, S., Neilson, E. et al. (2018). Sex differences in the adult human brain : evidence from 5216 UK biobank participants. *Cerebral cortex*, 28(8), 2959–2975.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.

Ruigrok, A. N., Salimi-Khorshidi, G., Lai, M.-C., Baron-Cohen, S., Lombardo, M. V., Tait, R. J. & Suckling, J. (2014). A meta-analysis of sex differences in human brain structure. *Neuroscience & Biobehavioral Reviews*, 39, 34–50.

Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S. Q., Nguyen, C. D., Ngo, V. D., Seekins, J., Blankenberg, F. G., Ng, A. Y. et al. (2022). Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 4(10), 867–878.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. (2017). Grad-cam : Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Serre, T. (2019). Deep learning : the good, the bad, and the ugly. *Annual Review of Vision Science*, 5, 399–426.

Shaffi, N., Viswan, V. & Mahmud, M. (2024). Ensemble of vision transformer architectures for efficient Alzheimer's Disease classification. *Brain Informatics*, 11(1), 25.

Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. (2017). Smoothgrad : removing noise by adding noise. *arXiv preprint arXiv :1706.03825*.

Springenberg, J. T., Dosovitskiy, A., Brox, T. & Riedmiller, M. (2014). Striving for simplicity : The all convolutional net. *arXiv preprint arXiv :1412.6806*.

Sundararajan, M., Taly, A. & Yan, Q. (2017). Axiomatic attribution for deep networks. *International Conference on Machine Learning*, pp. 3319–3328.

Toews, M. & Wells, W. (2009). Sift-rank : Ordinal description for invariant feature correspondence. *2009 ieee conference on computer vision and pattern recognition*, pp. 172–177.

Toews, M. & Wells III, W. M. (2013). Efficient and robust model-to-image alignment using 3D scale-invariant features. *Medical image analysis*, 17(3), 271–282.

Toews, M., Wells III, W., Collins, D. L. & Arbel, T. (2010). Feature-based morphometry : Discovering group-related anatomical patterns. *NeuroImage*, 49(3), 2318–2327.

Toews, M., Romaguera, T. V., Wells, W. & Makris, N. (2025). Representative scale-invariant characteristics of male and female brains in magnetic resonance images. *NeuroImage : Reports*, 5(1), 100267.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. *International conference on machine learning*, pp. 10347–10357.

van der Velden, B. H. M., Kuijf, H. J., Gilhuijs, K. G. & Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, 79, 102470.

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S. W. et al. (2012). The Human Connectome Project : a data acquisition perspective. *Neuroimage*, 62(4), 2222–2231.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H. et al. (2013). The WU-Minn human connectome project : an overview. *Neuroimage*, 80, 62–79.

Wachter, S., Mittelstadt, B. & Russell, C. (2017). Counterfactual explanations without opening the black box : Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 841–887.

Wang, L., Shen, H., Tang, F., Zang, Y. & Hu, D. (2012). Combined structural and resting-state functional MRI analysis of sexual dimorphism in the young adult human brain : an MVPA approach. *NeuroImage*, 61(4), 931–940.

Wei, X.-S., Song, Y.-Z., Mac Aodha, O., Wu, J., Peng, Y., Tang, J., Yang, J. & Belongie, S. (2021). Fine-grained image analysis with deep learning : A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(12), 8927–8948.

Yagis, E., Citi, L., Diciotti, S., Marzi, C., Atnafu, S. W. & De Herrera, A. G. S. (2020). 3d Convolutional neural networks for diagnosis of alzheimer's disease via structural mri. *2020 IEEE 33rd international symposium on computer-based medical systems (CBMS)*, pp. 65–70.

Yu, H., Li, J., Zhang, L., Cao, Y., Yu, X. & Sun, J. (2021). Design of lung nodules segmentation and recognition algorithm based on deep learning. *BMC bioinformatics*, 22(Suppl 5), 314.

Yuan, L., Chen, F., Zeng, L. L., Wang, L. & Hu, D. (2015). Gender identification of human brain image with a novel 3D descriptor. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(2), 551–561.

Yuan, L., Zeng, L., Shen, H., Liu, L. & Hu, D. (2018). Multi-source domain adaptation for gender classification of structural brain MRI. *IEEE Transactions on Medical Imaging*, 37(8), 1719–1731.

Zeiler, M. D. & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. *ECCV*.