République Algérienne Démocratique et Populaire

Ministère de l'Enseignement supérieur et de la Recherche scientifique

Universités de 20 Août 1955 Skikda

Faculté de Technologie

Département de Génie Electrique

Mémoire

Présenté en vue de l'obtention du diplôme de **Master**

Filière : Génie Biomédicale

Spécialité : Instrumentation Biomédicale

Détection automatique des tumeurs cérébrales par Deep Learning à partir d'images IRM

Par:

- **BOUTAGHANE Aroua**
- **CHENIKI Sirine**

Soutenu publiquement le : 22 Juin 2026

Devant le jury composé de :

Président	BENABBAS Mohamed Tahar	MRB	Centre de recherche en sciences pharmaceutiques – Constantine
Encadrant	DAOUDI Chouaib	MCB	Université 20 Aout 1955 Skikda
Examinateur	SEKHANE Djalal	MCB	Université 20 Aout 1955 Skikda

Année Universitaire 2025-2026

Remerciements

Avant toute chose, nous adressons nos louanges et nos remerciements les plus sincères à ALLAH, Le Tout-Puissant, pour la force, la patience et la volonté qu'Il nous a accordées tout au long de ce travail. C'est grâce à Sa miséricorde que nous avons pu franchir les étapes qui ont mené à l'aboutissement de ce mémoire.

Nous exprimons notre profonde gratitude à notre encadrante, Monsieur Daoudi chouaib, pour son accompagnement bienveillant, la qualité de ses orientations, ainsi que pour ses précieux conseils et remarques, qui ont grandement enrichi notre réflexion.

Nos remerciements s'étendent à toutes les personnes, de près ou de loin, qui ont contribué à la réalisation de ce projet, que ce soit par leur aide matérielle, morale ou intellectuelle.

Nous tenons à remercier nos parents, pour leurs sacrifices inestimables, leur amour inconditionnel et leur soutien indéfectible. Grâce à eux, nous avons pu persévérer même dans les moments les plus difficiles.

À nos frères, sœurs, famille et amis, qui nous ont entourés de leur affection, de leur patience et de leurs encouragements constants tout au long de ce parcours.

Enfn, nos remerciements vont également à Messieurs les membres du jury, pour avoir accepté d'évaluer ce travail. Nous espérons qu'il saura répondre à leurs attentes et témoigner du sérieux et de l'engagement investis tout au long de cette recherche.

Dédicace

Ce mémoire est le fruit d'un long parcours universitaire parcouru main dans la main.

Je le dédie à ma chère binôme et amie Aroua, avec qui j'ai partagé toutes les années de la fac, les joies, les difficultés, les révisions tardives et les moments inoubliables. Sans elle, ce travail n'aurait pas la même saveur.

À mes parents, pour leur amour, leur sacrifice et leur soutien sans faille. À eux je dois tout.

À mes grands-parents, que Dieu les bénisse, pour l'héritage et les valeurs qu'ils m'ont transmis.

À toute ma famille, grande et unie. Je tiens à citer tout particulièrement mon oncle Adel, l'une des personnes les plus chères à mon cœur.

À l'âme de mon grand-père Ahcen Mefrouche, que Dieu lui accorde Sa miséricorde.
J'aurais aimé qu'il soit là pour voir ce jour. Que ce travail soit une prière pour lui.

Cheniki sirine

Dédicace

Je dédie ce travail, avant tout, à Dieu, le Tout-Puissant, donné la force, la patience et la foi nécessaires pour surmonter toutes les épreuves et atteindre cet objectif.

À mon binôme et chère amie Sirine, pour sa présence à mes côtés, sa sincère amitié, sa patience, son soutien et tous les efforts partagés tout au long de la réalisation de ce travail. Merci pour tous les moments de collaboration et de complicité que nous avons vécus ensemble.

À ma chère grand-mère Aldjia, dont les prières et la tendresse ont toujours été une source de force et de réconfort. Que Dieu la préserve et la comble de Ses bienfaits.

À mes très chers parents, pour tous les sacrifices qu'ils ont consentis, pour leur amour inconditionnel, leur patience, leur soutien et leurs encouragements tout au long de mon parcours. Que Dieu leur accorde une bonne santé, une longue vie remplie de bonheur et de sérénité.

À mon cher frère Ahmed, pour sa présence, son affection et son soutien constant.

À mon cher oncle Bilel et sa femme, pour leurs prières et leurs encouragements.

À toute ma famille, pour leur amour, leur confiance et leurs encouragements.

À mes chères amies khouloud et cherifa, pour tous les beaux moments partagés, leur écoute et leur fidélité.

Boutaghane aroua

الملخص

تُعَدُّ أورام الدماغ من أكثر الأمراض العصبية تعقيدًا بسبب تنوعها المورفولوجي وتأثيرها المحتمل على الوظائف الحيوية للدماغ. ويُعتبر التشخيص المبكر عاملًا أساسيًا في تحسين فرص العلاج ورفع معدلات البقاء على قيد الحياة. ومع التطور المتسارع في مجال الذكاء الاصطناعي، أصبحت تقنيات التعلم العميق من الأدوات الواعدة التي يمكن أن تُساهم في أتمتة تحليل الصور الطبية ودعم الأطباء في اتخاذ القرار التشخيصي.

يهدف هذا العمل إلى تطوير ودراسة نظام ذكي للتصنيف الآلي لأورام الدماغ اعتمادًا على صور التصوير بالرنين المغناطيسي(MRI) ، وذلك من خلال مقارنة أداء عدة معماريات للشبكات العصبية الالتفافية .(CNN) شملت الدراسة أربع فئات رئيسية هي: الورم الدبقي(Glioma) ، والورم السحائي(Meningioma) ، وورم الغدة النخامية (Pituitary Adenoma)، بالإضافة إلى فئة الصور السليمة الخالية من الأورام.(No Tumor)

تم في هذا البحث تصميم وتقييم ثلاثة نماذج مختلفة، وهي: شبكة عصبية التفافية مخصصة(CNN) ، ومعمارية VGG16المعتمدة على التعلم بالنقل(Transfer Learning) ، ومعمارية ResNet50 المدعومة بتقنية الضبط الدقيق .(Fine-Tuning)وقد اعتمدت المنهجية المقترحة على مرحلتين من التقييم، حيث استُخدمت قاعدة بيانات Kaggle الخاصة بـ Masoud Nickparvar في التدريب والتحقق(Validation) ، بينما استُخدمت قاعدة بيانات مستقلة من Mendeley Dataلاختبار قدرة النماذج على التعميم (Generalization) في ظروف أقرب إلى الواقع السريري.

شملت مراحل المعالجة المسبقة إعادة تحجيم الصور، وتطبيع قيم البكسلات، وترميز الفئات، بالإضافة إلى تطبيق تقنيات زيادة البيانات (Data Augmentation) بهدف تحسين تنوع العينات وتقليل ظاهرة فرط التعلّم .(Overfitting) كما تم تقييم أداء النماذج باستخدام عدة مؤشرات إحصائية تشمل الدقة(Accuracy) ، والاستدعاء(Recall) ، والدقة النوعية(Precision) ، ودرجةF1 ، ومصفوفات الالتباس(Confusion Matrices) ، ومنحنيات.ROC-AUC

أظهرت النتائج التجريبية أن نموذج ResNet50 حقق أفضل أداء بين النماذج المدروسة، حيث بلغت الدقة 94.9٪ على مجموعة التحقق و91.0٪ على مجموعة الاختبار الخارجية المستقلة، مما يدل على قدرة عالية على التعميم واستقرار الأداء. كما أبرزت الدراسة أهمية استخدام قواعد بيانات مستقلة في تقييم أنظمة التعلم العميق، إذ كشفت عن وجود مؤشرات على فرط التعلّم في بعض النماذج الأخرى مثل VGG16 رغم تحقيقها نتائج مرتفعة أثناء مرحلة التحقق.

تؤكد النتائج المتحصل عليها الإمكانات الكبيرة لتقنيات التعلم العميق في مجال التشخيص المساعد للأورام الدماغية، كما تُبرز أهمية الدمج بين المعماريات العميقة المتقدمة واستراتيجيات التحقق الصارمة من أجل تطوير أنظمة موثوقة وقابلة للاستخدام في التطبيقات الطبية المستقبلية. ومع ذلك، تبقى الحاجة قائمة لإجراء دراسات سريرية متعددة المراكز والتحقق من فعالية هذه النماذج على قواعد بيانات أوسع وأكثر تنوعًا قبل اعتمادها في الممارسة الطبية اليومية.

الكلمات المفتاحية

أورام الدماغ، التصوير بالرنين المغناطيسي، الذكاء الاصطناعي، التعلم العميق، الشبكات العصبية الالتفافية، VGG16، ResNet50، التعلم بالنقل، التصنيف الطبي، التشخيص بمساعدة الحاسوب.

Abstract

Brain tumors are among the most challenging neurological diseases due to their morphological diversity and their potential impact on vital brain functions. Early diagnosis plays a crucial role in improving therapeutic management and patient prognosis. In this context, Deep Learning techniques have emerged as powerful tools for automating medical image analysis.

This thesis presents a comparative study of several convolutional neural network architectures for the automatic classification of brain tumors from MRI images. Four classes were considered: glioma, meningioma, pituitary adenoma, and no tumor. Three models were developed and evaluated: a custom CNN, VGG16 with Transfer Learning, and ResNet50 with Fine-Tuning.

The proposed methodology relies on a rigorous external validation strategy using two independent datasets. The Kaggle dataset provided by Masoud Nickparvar was used for training and validation, while the Mendeley Data dataset was used as an independent external test set to assess model generalization. Performance was evaluated using multiple complementary metrics, including accuracy, precision, recall, F1-score, confusion matrices, and ROC-AUC curves.

Experimental results demonstrated that ResNet50 achieved the best performance, reaching 94.9% accuracy on the validation set and 91.0% accuracy on the independent external test set. The study also highlighted the overfitting limitations of VGG16 and emphasized the importance of external validation for obtaining reliable estimates of model performance in medical imaging applications.

The findings confirm the potential of deep neural networks as decision-support tools for brain tumor classification. However, further multicenter and clinical validations remain necessary before deployment in real-world healthcare environments.

Keywords: Brain Tumor, MRI, Deep Learning, CNN, VGG16, ResNet50, Transfer Learning, Medical Image Classification.

Résumé

Les tumeurs cérébrales représentent l'une des pathologies neurologiques les plus complexes en raison de leur diversité morphologique et de leur impact potentiel sur les fonctions vitales du cerveau. Le diagnostic précoce constitue un facteur déterminant pour améliorer la prise en charge thérapeutique et le pronostic des patients. Dans ce contexte, les techniques de Deep Learning offrent des perspectives prometteuses pour l'automatisation de l'analyse des images médicales.

Ce mémoire présente une étude comparative de plusieurs architectures de réseaux de neurones convolutifs appliquées à la classification automatique des tumeurs cérébrales à partir d'images IRM. L'étude porte sur quatre classes : gliome, méningiome, adénome hypophysaire et absence de tumeur. Trois modèles ont été développés et évalués : un CNN personnalisé, VGG16 avec Transfer Learning et ResNet50 avec Fine-Tuning.

La méthodologie repose sur une validation externe rigoureuse utilisant deux sources de données indépendantes. Le dataset Kaggle de Masoud Nickparvar a été utilisé pour l'entraînement et la validation, tandis que le dataset Mendeley Data a servi à l'évaluation finale de la capacité de généralisation des modèles. Les performances ont été évaluées à l'aide de plusieurs métriques complémentaires : accuracy, précision, rappel, F1-score, matrices de confusion et courbes ROC-AUC.

Les résultats montrent que ResNet50 constitue l'architecture la plus performante avec une accuracy de 94,9 % sur l'ensemble de validation et de 91,0 % sur le test externe indépendant. L'étude met également en évidence les limites du surapprentissage observées avec VGG16 et souligne l'importance d'une validation externe pour l'évaluation fiable des systèmes de Deep Learning en imagerie médicale.

Les résultats obtenus confirment le potentiel des architectures profondes pour assister les spécialistes dans la classification des tumeurs cérébrales. Toutefois, des validations multicentriques et cliniques demeurent nécessaires avant toute intégration dans la pratique hospitalière.

Mots-clés : Tumeurs cérébrales, IRM, Deep Learning, CNN, VGG16, ResNet50, Transfer Learning, Classification médicale.

Table des matières

Chapitre 2 Matériel et Méthodes

Chapitre 3 : Résultats et Discussion

LES FIGURES

LISTE DES TABLEAUX

Introduction générale

Introduction générale

Les tumeurs cérébrales figurent parmi les pathologies neurologiques les plus complexes et les plus préoccupantes en raison de leur impact direct sur les fonctions vitales du cerveau. Elles regroupent un ensemble hétérogène de lésions pouvant être bénignes ou malignes, dont l'évolution peut entraîner des déficits neurologiques sévères, une altération de la qualité de vie, voire le décès du patient. Selon leur localisation, leur taille et leur nature histologique, ces tumeurs peuvent affecter différentes régions cérébrales et provoquer des symptômes variés tels que des céphalées persistantes, des troubles moteurs, des déficits cognitifs ou encore des crises épileptiques.

Le diagnostic précoce constitue un facteur déterminant dans la prise en charge thérapeutique et l'amélioration du pronostic des patients. Parmi les différentes modalités d'imagerie médicale disponibles, l'Imagerie par Résonance Magnétique (IRM) représente aujourd'hui la technique de référence pour l'exploration des structures cérébrales. Grâce à son excellente résolution des tissus mous et à son caractère non invasif, l'IRM permet d'obtenir des informations détaillées sur la morphologie et l'étendue des lésions tumorales.

Cependant, l'interprétation des images IRM demeure une tâche complexe qui exige une expertise médicale importante. La quantité croissante de données générées par les systèmes d'imagerie modernes, associée à la variabilité des caractéristiques tumorales, peut rendre l'analyse longue et parfois sujette à des variations d'interprétation entre spécialistes. Dans ce contexte, le développement d'outils d'aide au diagnostic capables d'assister les radiologues dans l'identification et la classification des tumeurs cérébrales représente un enjeu majeur pour la médecine moderne.

Les réseaux de neurones convolutifs (CNN) sont au cœur de la vision par ordinateur moderne. Contrairement aux réseaux denses, ils préservent la topologie spatiale des images grâce à trois types de couches : les couches de convolution, qui appliquent des filtres pour extraire des cartes de caractéristiques (des contours simples vers des motifs complexes) ; les couches de pooling, qui réduisent la dimension tout en conservant l'information saillante ; et les couches denses finales, qui réalisent la classification. Cette architecture hiérarchique rend les CNN particulièrement adaptés à l'analyse d'images médicales. [56] discriminantes directement à partir des images leur confère un avantage considérable par rapport aux approches classiques fondées sur l'extraction manuelle de caractéristiques.

Malgré les performances remarquables rapportées dans la littérature scientifique, plusieurs défis demeurent. De nombreuses études reposent sur des bases de données limitées ou utilisent uniquement des protocoles de validation interne susceptibles de surestimer les

performances réelles des modèles. La capacité de généralisation des systèmes développés à des données provenant de sources indépendantes constitue ainsi l'un des principaux défis actuels de l'intelligence artificielle appliquée à l'imagerie médicale.

C'est dans ce contexte que s'inscrit le présent mémoire intitulé :

« Détection et Classification Automatisée des Tumeurs Cérébrales par Deep Learning : Optimisation des Architectures CNN et Aide au Diagnostic Clinique ».

L'objectif principal de ce travail est de développer et d'évaluer un système intelligent capable de classifier automatiquement différentes catégories de tumeurs cérébrales à partir d'images IRM. Une attention particulière est portée à l'étude comparative de plusieurs architectures de Deep Learning afin d'identifier la solution offrant le meilleur compromis entre précision, robustesse et capacité de généralisation.

Pour atteindre cet objectif, trois modèles ont été étudiés : un réseau de neurones convolutif personnalisé (CNN), l'architecture VGG16 exploitant le transfert d'apprentissage (Transfer Learning) et l'architecture ResNet50 associée à une stratégie de Fine-Tuning. Les modèles ont été entraînés à partir d'un dataset public issu de Kaggle, puis évalués sur une base de données totalement indépendante provenant de Mendeley Data afin de mesurer objectivement leur capacité de généralisation. Les performances ont été analysées à l'aide de plusieurs indicateurs tels que l'accuracy, la précision, le rappel, le score F1, les matrices de confusion et les courbes ROC-AUC.

Les principales contributions de ce travail peuvent être résumées comme suit :

- Étude comparative de trois architectures de Deep Learning appliquées à la classification multiclasses des tumeurs cérébrales ;
- Mise en œuvre des techniques de Transfer Learning et de Fine-Tuning pour améliorer les performances des modèles
- Évaluation rigoureuse de la capacité de généralisation à l'aide d'un dataset externe indépendant ;
- Analyse comparative des performances et des limites des architectures étudiées ;
- Contribution au développement d'un système d'aide au diagnostic susceptible d'assister les professionnels de santé dans l'interprétation des examens IRM.

Afin de présenter de manière structurée l'ensemble des travaux réalisés, ce mémoire est organisé en trois chapitres :

Le premier chapitre présente les notions fondamentales relatives aux tumeurs cérébrales, aux techniques d'imagerie médicale, à l'intelligence artificielle et aux architectures de Deep Learning utilisées dans le domaine de la classification des images médicales. Une analyse critique des travaux récents permet également de situer le positionnement scientifique du projet.

Le deuxième chapitre est consacré à la méthodologie adoptée. Il décrit les bases de données utilisées, les différentes étapes de prétraitement, les techniques d'augmentation des données ainsi que la conception et l'entraînement des architectures CNN, VGG16 et ResNet50.

Le troisième chapitre présente les résultats expérimentaux obtenus, leur analyse comparative, l'évaluation de la capacité de généralisation des modèles ainsi qu'une discussion critique des performances observées.

Enfin, une conclusion générale synthétise les principaux résultats obtenus, met en évidence les contributions du travail réalisé et propose plusieurs perspectives de recherche pour l'amélioration future des systèmes de classification automatisée des tumeurs cérébrales.

Chapitre 1
État de l'art

1. Introduction

Le cerveau humain constitue le centre de régulation des fonctions vitales, cognitives et motrices. Au sein de ce système complexe, l'émergence de pathologies tumorales représente l'un des défis majeurs de l'oncologie moderne. En raison de leur hétérogénéité biologique et de leur caractère infiltrant, ces lésions exigent des protocoles diagnostiques de haute précision afin d'orienter au mieux les stratégies thérapeutiques. Dans ce contexte, l'Imagerie par Résonance Magnétique (IRM) s'est imposée comme l'examen de référence, permettant une évaluation morphologique détaillée de l'encéphale sans exposition aux rayonnements ionisants.

Cependant, l'interprétation clinique des examens IRM repose encore largement sur l'analyse visuelle du radiologue. Cette tâche, particulièrement chronophage, est sujette à une variabilité inter-observatrice et peut être complexifiée par la charge de travail croissante des services de radiologie. L'intégration de l'intelligence artificielle, et plus spécifiquement de l'apprentissage profond (Deep Learning), offre aujourd'hui des perspectives prometteuses. En s'appuyant sur des architectures neuronales capables d'extraire des caractéristiques complexes, ces technologies visent à fournir des outils d'aide à la décision capable d'automatiser certains segments de l'analyse iconographique et de renforcer la reproductibilité de la classification des lésions.

Ce premier chapitre pose le cadre théorique de l'étude : aspects cliniques des tumeurs cérébrales, techniques d'imagerie, principes de l'apprentissage profond, revue des travaux antérieurs et positionnement de notre approche.

1.1. Généralités sur les tumeurs cérébrales

1.1.1. Définition des tumeurs cérébrales

Une tumeur cérébrale est une masse de cellules anormales se développant dans le tissu cérébral ou, plus largement, dans le système nerveux central (cerveau et moelle épinière). Sa croissance est incontrôlée et peut être de nature bénigne ou maligne.

Même bénigne, une tumeur peut être dangereuse si elle comprime des structures vitales ou gêne la circulation du liquide céphalo-rachidien ; sa localisation et sa vitesse de croissance comptent donc autant que sa nature histologique.

La classification OMS 2021 a fait évoluer les critères, intégrant désormais les marqueurs moléculaires aux données histologiques pour un typage plus précis des tumeurs. [3-5]

En pratique clinique, ce terme désigne un ensemble très hétérogène de lésions. Certaines sont primaires et se développent directement dans le cerveau, tandis que d'autres sont secondaires, résultant de la propagation de cellules cancéreuses provenant d'autres organes. Cependant, ces dernières ne sont pas abordées ici en détail. Les différentes sources consultées montrent que la définition d'une tumeur cérébrale dépasse l'idée d'une simple masse dans le cerveau : il s'agit d'une prolifération anormale de cellules du système nerveux central dont les propriétés cellulaires, moléculaires, la localisation et la dynamique de croissance déterminent le comportement clinique. [3]

1.1.2 Classification générale : tumeurs bénignes et malignes

La distinction entre « bénin » et « malin » reste utile, mais elle s'avère souvent trop simpliste lorsqu'il s'agit de tumeurs cérébrales. Les classifications récentes de l'OMS ne se fondent plus exclusivement sur l'histologie ; elles intègrent désormais des paramètres moléculaires pour définir les grandes catégories tumorales. Cette approche a conduit à une révision profonde de plusieurs groupes, notamment les gliomes, les épendymomes et les médulloblastomes. [4] [5]

Sur le plan pratique, une tumeur bénigne est généralement non cancéreuse, se développe plus lentement et est moins invasive. Toutefois, cela ne signifie pas qu'elle est sans gravité : dans le contexte cérébral, même une tumeur bénigne peut poser problème en raison de sa localisation et de l'effet de masse qu'elle exerce sur des structures particulièrement sensibles. Les travaux consultés soulignent que la classification clinique est souvent doublée d'une évaluation en termes de bas grade ou haut grade, ce qui reflète plus fidèlement le continuum du comportement biologique. [6] [7]

À l'opposé, une tumeur maligne est cancéreuse, se développe plus rapidement et tend davantage à infiltrer les tissus adjacents. De manière succincte, les synthèses explorées expliquent que les tumeurs bénignes « croissent lentement et se propagent rarement », alors que les tumeurs malignes « croissent beaucoup plus rapidement » et présentent une infiltration plus prononcée des tissus cérébraux environnants. [7]

Prenons l'exemple des méningiomes pour illustrer les limites d'une distinction aussi binaire. Les méningiomes sont souvent définis comme des tumeurs intracrâniennes généralement bénignes. Cependant, l'OMS les classe en trois grades : le grade I pour les formes bénignes, le grade II pour les formes atypiques, et le grade III pour celles reconnues comme malignes. Même lorsqu'ils sont histologiquement bénins, les méningiomes peuvent récidiver. Une étude rapporte notamment un taux de récidive de 10 à 20 % après une résection complète. [8] [9]

En résumé, le terme « bénin » renvoie principalement à une croissance plus lente et un comportement moins agressif, tandis que « malin » décrit une tumeur cancéreuse, plus invasive et présentant un risque accru d'évolution défavorable. Toutefois, dans le cas des tumeurs cérébrales, les critères moléculaires et le grade OMS jouent un rôle aussi déterminant que cette classification binaire, soulignant ainsi l'importance des nouvelles approches modernes en matière de classification. [4,5,6]

1.1.3 Présentation des principales tumeurs étudiées

Un premier examen des publications récentes met en lumière trois types de tumeurs intracrâniennes distinctes, chacune se différenciant par leur biologie, leur agressivité et leur approche thérapeutique. Les gliomes, les méningiomes et les adénomes hypophysaires figurent parmi les plus fréquents, mais ils s'inscrivent dans des cadres différents, tant du point de vue clinique que du degré de malignité. Les gliomes se démarquent par une évolution souvent sévère, tandis que les méningiomes et les adénomes hypophysaires sont, dans la majorité des cas, bénins, même si certaines formes peuvent présenter des caractéristiques cliniques significatives. [10]

1. Gliome :

Les gliomes, tumeurs primitives les plus fréquentes du système nerveux central, naissent des cellules gliales et présentent une agressivité variable (de bas grade à glioblastome). Leur diagnostic précoce conditionne fortement le pronostic. [10-13]

La Figure 1.1 illustre un exemple de gliome de bas grade observé en IRM. Ce type de tumeur présente généralement une croissance lente et une infiltration progressive des tissus cérébraux adjacents. L'analyse radiologique de ces lésions constitue une étape essentielle dans le diagnostic et la planification thérapeutique [11-19].

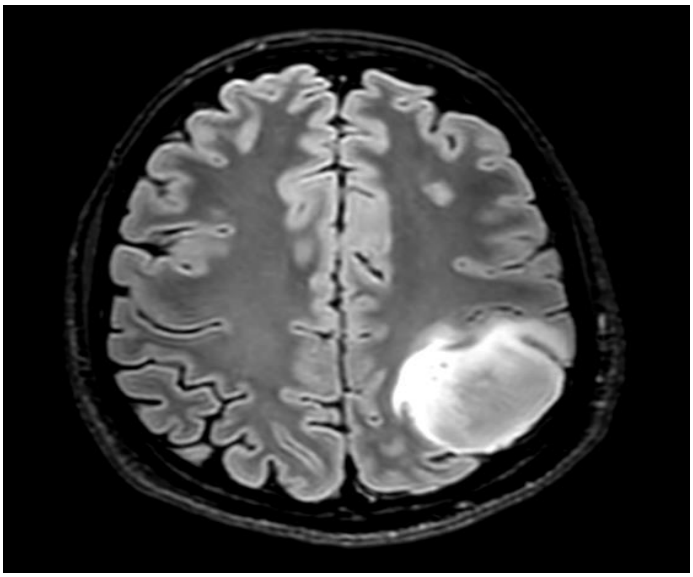


Figure 1.1 : un gliome de bas grade

2. Méningiome :

En ce qui concerne les méningiomes, ceux-ci sont actuellement reconnus comme les tumeurs primitives intracrâniennes les plus communes. La majorité d'entre eux sont des tumeurs de bas grade à croissance lente, souvent diagnostiquées par hasard. La prise en charge repose principalement sur une chirurgie visant une résection complète, souvent curative si l'emplacement le permet. Lorsque la résection est incomplète ou lorsque les lésions sont non opérables ou de grade supérieur [20].

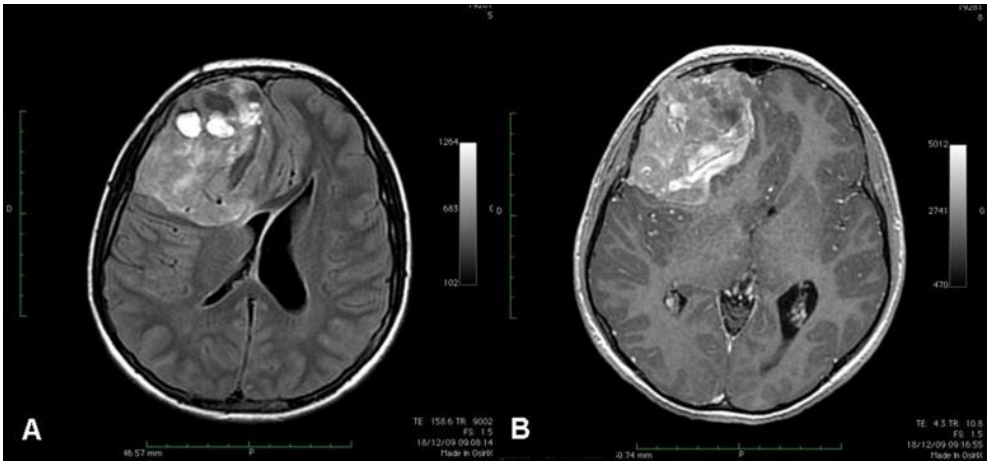


Figure 1.2 : Méningiome papillaire : une forme rarissime et agressive de méningiome

La radiothérapie constitue une alternative thérapeutique. Les recherches récentes ont mis en évidence une évolution de leur classification : l'approche purement histologique tend à être complétée par des analyses génomiques, des profils de méthylation ADN ainsi que

d'autres marqueurs épigénomiques afin de mieux appréhender le risque de récidive et le comportement biologique des tumeurs [22].

Cependant, les auteurs s'accordent à souligner l'insuffisance des essais contrôlés rigoureux, justifiant en partie des disparités dans les stratégies thérapeutiques employées [23].

3. Adénomes hypophysaires :

Quant aux adénomes hypophysaires (également appelés PitNETs depuis l'introduction d'une nouvelle nomenclature), ils sont généralement considérés comme des tumeurs bénignes à croissance lente issues de l'hypophyse antérieure. Leur présentation clinique est particulièrement variée : tandis que certaines lésions se manifestent par des syndromes d'hypersécrétion hormonale, d'autres provoquent des symptômes liés à un effet de masse, tels que des céphalées, des troubles visuels ou un hypopituitarisme. Les avancées en matière de classification s'appuient désormais largement sur l'immunohistochimie des hormones hypophysaires et l'analyse des facteurs de transcription, ce qui renforce la précision du diagnostic, bien que l'appellation PitNET demeure l'objet de débats au sein de la communauté scientifique [24].

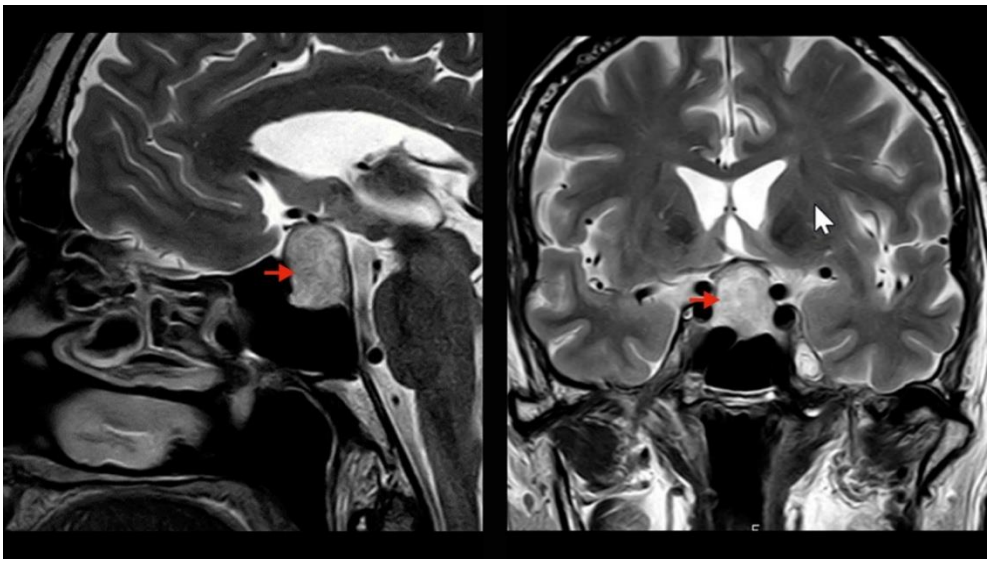


Figure 1.3 : Adénome hypophysaire [25].

Sur le plan thérapeutique, la chirurgie transsphénoïdale reste généralement le traitement privilégié pour ces lésions. Toutefois, les traitements médicaux occupent également une place importante selon le sous-type d'adénome concerné : les agonistes dopaminergiques pour les prolactinomes, les analogues de la somatostatine pour les acromégalies modérées et la radiothérapie pour les formes résiduelles ou récidivantes [26].

Les caractéristiques biologiques et cliniques des différentes tumeurs cérébrales influencent directement leur prise en charge thérapeutique et leur pronostic. Dans ce contexte, la rapidité du diagnostic constitue un facteur déterminant pour améliorer les chances de traitement et réduire les complications neurologiques associées. Il est donc nécessaire d'examiner l'importance du diagnostic précoce dans la prise en charge des patients atteints de tumeurs cérébrales.

1.2 Importance du diagnostic précoce des tumeurs cérébrales

Le diagnostic précoce des tumeurs cérébrales constitue un enjeu majeur en neuro-oncologie (Tab 1.1). Une détection rapide permet non seulement d'identifier la présence d'une lésion intracrânienne, mais également de caractériser son type, son extension et son degré d'agressivité afin d'orienter la stratégie thérapeutique la plus appropriée. La précocité du diagnostic influence directement le pronostic des patients, les possibilités thérapeutiques ainsi que la qualité de vie à long terme [10-12, 20, 22-24, 26, 27].

Dans le cas des gliomes, qui représentent les tumeurs cérébrales primaires les plus agressives, une identification précoce favorise la réalisation d'une résection chirurgicale maximale tout en préservant les fonctions neurologiques essentielles. [16-17-18]

Elle permet également d'accélérer l'accès aux analyses histopathologiques et moléculaires nécessaires à la classification selon les recommandations de l'Organisation Mondiale de la Santé (OMS). À l'inverse, un retard diagnostique peut conduire à une progression tumorale importante, réduire les possibilités thérapeutiques et détériorer significativement le pronostic [10-12, 20, 22-24, 26, 27].

Pour les méningiomes, généralement caractérisés par une croissance lente, le diagnostic précoce permet d'intervenir avant l'apparition de complications liées à l'effet de masse exercé sur les structures cérébrales adjacentes. Une prise en charge précoce améliore les chances d'une résection complète et limite les risques de récidive ou de déficits neurologiques permanents [10-12, 20, 22-24, 26, 27].

Les adénomes hypophysaires nécessitent également une détection rapide afin d'éviter les complications endocriniennes et ophtalmologiques pouvant résulter de l'hypersécrétion hormonale ou de la compression des structures voisines. Une intervention précoce permet de réduire les conséquences fonctionnelles et d'améliorer l'efficacité des traitements médicaux ou chirurgicaux [10-12, 20, 22-24, 26, 27].

Malgré les progrès considérables de l'imagerie médicale, l'interprétation des examens radiologiques repose encore largement sur l'expertise humaine. Cette analyse peut être

influencée par plusieurs facteurs tels que la complexité des images, la variabilité inter-observateurs et l'augmentation constante du volume de données à traiter. Ces limitations peuvent entraîner des retards diagnostiques ou des divergences d'interprétation susceptibles d'affecter la prise en charge clinique [10-12, 20, 22-24, 26, 27].

Dans ce contexte, l'intégration de techniques d'intelligence artificielle et d'apprentissage profond apparaît comme une solution prometteuse pour assister les radiologues dans la détection et la classification des tumeurs cérébrales. Les systèmes automatisés basés sur les réseaux de neurones convolutifs permettent d'extraire des caractéristiques complexes à partir des images IRM et de fournir une aide à la décision rapide, reproductible et potentiellement plus précise. Ces avancées justifient l'intérêt croissant porté aux approches de Deep Learning dans le domaine de la neuro-oncologie moderne [10-12, 20, 22-24, 26, 27].

Tableau 1.1 : Comparaison des différentes tumeurs et leurs risques

Type tumoral	Risque d'un diagnostic tardif	Impact clinique principal
Gliome	Progression rapide	Diminution de la survie
Méningiome	Compression cérébrale	Déficits neurologiques
Adénome hypophysaire	Troubles endocriniens	Atteinte hormonale et visuelle

1.3 Techniques d'imagerie médicale

L'arsenal diagnostique en neuro-oncologie repose sur diverses modalités d'imagerie, essentielles pour la détection, la caractérisation morphologique et le suivi thérapeutique des tumeurs. Ces techniques permettent d'obtenir des données structurelles et métaboliques cruciales sur les lésions. Bien que la Tomodensitométrie (CT) et la Tomographie par Émission de Positons (PET) soient utilisés pour des indications spécifiques, l'Imagerie par Résonance Magnétique (IRM) demeure la modalité de référence en raison de sa polyvalence et de sa précision [28].

L'imagerie par Résonance Magnétique (IRM) est la modalité de référence pour l'exploration du système nerveux central. Contrairement aux techniques basées sur les rayons X, l'IRM exploite les propriétés magnétiques des noyaux d'hydrogène (protons), omniprésents dans l'eau et les graisses des tissus biologiques. Sous l'effet d'un champ magnétique puissant (généralement de 1,5 ou 3 Tesla) et d'impulsions de radiofréquence, ces protons émettent des signaux captés et transformés en images de haute résolution. [29]

Figure 1.4 : Résonance Magnétique (IRM) [30]

L'intérêt majeur de l'IRM en neuro-oncologie réside dans sa **multimodalité**. Une seule séance permet d'obtenir différentes "pondérations" (séquences), chacune révélant des propriétés tissulaires spécifiques :

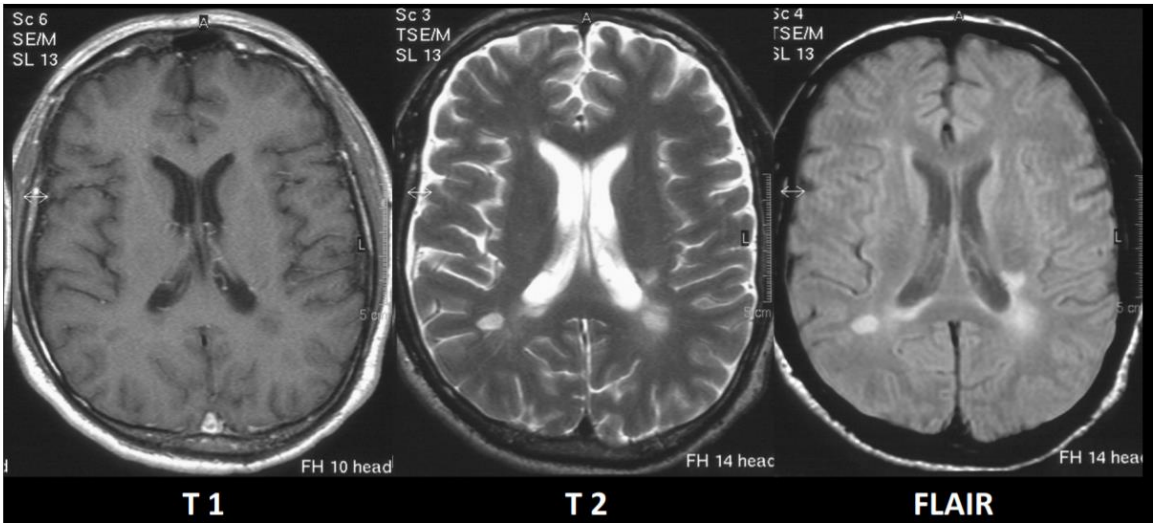


Figure 1.5 : Différentes séquences IRM utilisées dans l'analyse des tumeurs cérébrales [31].

A. Séquences Morphologiques Fondamentales

- **Pondération T1 (Temps de relaxation longitudinal) :** Cette séquence offre la meilleure résolution anatomique. Elle permet de distinguer avec précision la substance grise de la substance blanche. En pathologie tumorale, elle est souvent couplée à l'injection d'un produit de contraste (Gadolinium). La **T1 avec contraste (T1c)** est indispensable pour visualiser la rupture de la barrière hémato-encéphalique, signe d'une tumeur active ou de haut grade. [32]

- **Pondération T2 (Temps de relaxation transversal) :** Très sensible à l'eau, le T2 est l'outil de choix pour détecter les anomalies tissulaires. Les processus pathologiques (tumeurs, inflammations) apparaissent généralement en **hypersignal** (blanc). Elle permet de délimiter l'étendue globale de la lésion, incluant la masse tumorale et les zones de remaniement tissulaire. [32]

B. Séquences de Caractérisation Avancée

- **Séquence FLAIR (Fluid Attenuated Inversion Recovery) :** Il s'agit d'une séquence T2 où le signal du liquide céphalo-rachidien (LCR) est supprimé (apparaissant noir). Cette technique est cruciale car elle permet de différencier l'œdème vasogénique (brillant) des cavités ventriculaires, facilitant ainsi la détection des lésions périventriculaires et la segmentation précise de l'œdème. [32]

Dans les applications de Deep Learning, chaque séquence IRM apporte des informations complémentaires sur la tumeur et son environnement. Les séquences T1 permettent une visualisation détaillée de l'anatomie cérébrale, tandis que les séquences T2 et FLAIR mettent davantage en évidence les zones d'œdème et les anomalies tissulaires. L'exploitation combinée de ces séquences améliore significativement la capacité des modèles d'apprentissage profond à distinguer les différentes catégories de tumeurs cérébrales et à extraire des caractéristiques discriminantes pertinentes pour la classification automatique.

C. Importance pour l'Apprentissage Profond (Deep Learning)

Pour les modèles de classification et de segmentation, ces séquences ne sont pas de simples images, mais des **canaux d'information complémentaires**. La fusion de ces modalités (Input multi-canal) permet aux réseaux de neurones (CNN) d'intégrer à la fois l'anatomie (T1), l'activité biologique (T1c), et l'étendue de l'œdème (FLAIR/T2), garantissant ainsi une précision bien supérieure à une analyse basée sur une seule séquence. [32]

1.3.2 Tomodensitométrie (CT)

La tomodensitométrie (CT-scan), fondée sur l'absorption différentielle des rayons X, offre une résolution moindre que l'IRM pour les tissus mous mais reste complémentaire : rapide, elle est l'examen de choix en urgence (hémorragies, effet de masse) et visualise bien les calcifications et l'os. Ses limites sont l'irradiation et une faible distinction entre noyau tumoral et œdème. En IA, ses données (unités Hounsfield) peuvent être fusionnées avec l'IRM pour enrichir l'analyse tissulaire. [33]

Le scanner (Figure 1.6) complète l'IRM en urgence, mais expose aux rayonnements ionisants et distingue mal le noyau tumoral de l'œdème. En IA, ses données peuvent être fusionnées avec l'IRM. [33]

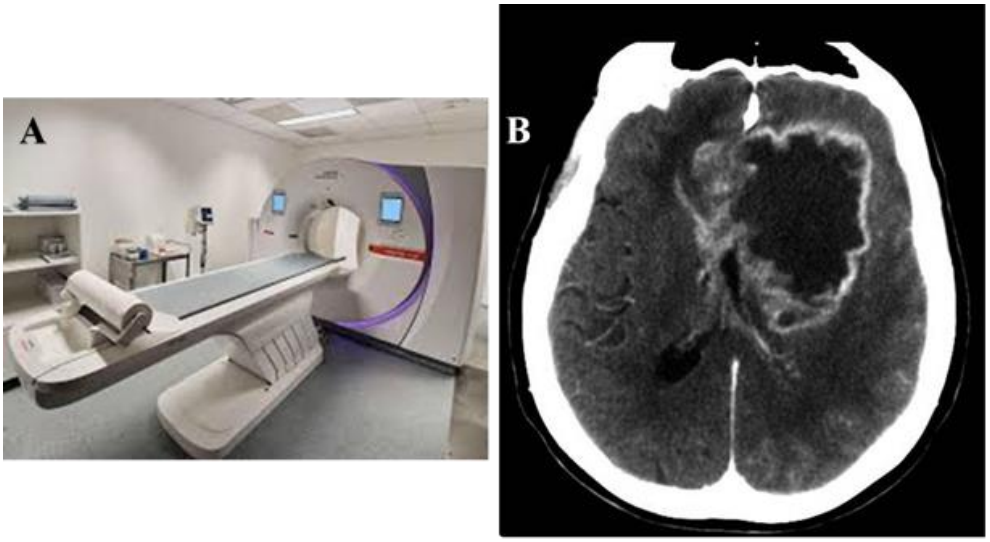


Figure 1.6 : (A) Scanner et (B) Scanner injecté, montrant une volumineuse tumeur avec une nécrose centrale, prédominant en région frontale gauche mais également étendue vers le corps calleux et la région frontale droite. Il existe un œdème péritumoral (hypodensité

1.3.3. Tomographie par émission de positons (PET)

La tomographie par émission de positons (PET) est une imagerie fonctionnelle : par injection d'un traceur (le plus souvent le 18F-FDG, analogue du glucose), elle révèle l'activité métabolique des tissus. Les cellules tumorales malignes, fortement consommatrices de glucose, apparaissent ainsi nettement, ce qui aide au grading et à la détection précoce, en complément de l'information morphologique de l'IRM. [34]

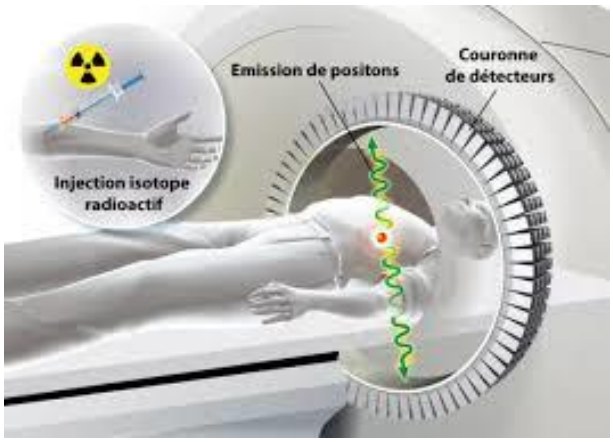


Figure 1.7 : PET-scan [36]

En neuro-oncologie, cette technique s'avère particulièrement discriminante pour différencier une récidive tumorale active d'une radionécrose (cicatrice post-radiothérapie), une distinction souvent complexe à établir avec une IRM standard. De plus, l'utilisation de traceurs spécifiques aux acides aminés permet de mieux délimiter les contours des gliomes infiltrants, ce qui optimise la planification chirurgicale et la délimitation des volumes cibles en radiothérapie. Pour les algorithmes de Deep Learning, l'intégration des données métaboliques du PET-scan apporte une dimension biologique cruciale, permettant de raffiner la classification des grades tumoraux et de prédire l'agressivité de la lésion avec une fiabilité accrue [37].

1.3.4. Comparaison et complémentarité des modalités d'imagerie

Le choix d'une modalité d'imagerie résulte d'un compromis entre résolution spatiale, spécificité biologique et contraintes cliniques. L'IRM reste le pilier du diagnostic grâce à son contraste des tissus mous, le scanner et le PET apportant des informations complémentaires (urgence, métabolisme).

Sur le plan de l'intelligence artificielle, cette complémentarité est exploitée à travers le concept de **fusion de données multimodales**. En intégrant les caractéristiques structurelles (IRM/CT) et métaboliques (PET), les modèles de Deep Learning peuvent pallier les limites intrinsèques de chaque modalité prise isolément. Cette approche holistique permet non seulement d'améliorer la précision de la classification des grades de malignité, mais aussi de fournir une cartographie lésionnelle plus robuste, essentielle pour la précision de la neuronavigation chirurgicale et de la radiothérapie conformationnelle. [38]

Tableau 1.2 : Comparaison des modalités d'imagerie

Modalité	Information principale	Avantages	Limites	Intérêt pour l'IA
IRM [39]	Anatomie, contraste tissulaire, œdème	Excellent contraste des tissus mous, absence d'irradiation	Coût, durée, variabilité des protocoles	Modalité principale pour la classification CNN
CT [40]	Densité, calcifications, urgence	Rapide, disponible, utile en urgence	Rayonnements ionisants, faible contraste des tissus mous	Complément possible
PET [41]	Métabolisme tumoral	Information fonctionnelle et biologique	Coût, irradiation, disponibilité limitée	Fusion multimodale possible

1.3.5. Bases de données d'images IRM utilisées en Deep Learning

Le développement des méthodes d'intelligence artificielle appliquées à l'imagerie médicale repose sur la disponibilité de bases de données volumineuses et correctement annotées. Dans le domaine de la détection et de la classification des tumeurs cérébrales, plusieurs bases de données publiques sont largement utilisées pour l'entraînement, la validation et l'évaluation des modèles d'apprentissage profond.

Parmi les ensembles de données les plus connus figurent la base Brain Tumor Segmentation Challenge (BraTS), qui regroupe des examens IRM multimodaux comprenant généralement les séquences T1, T1 avec injection de contraste (T1c), T2 et FLAIR. Cette base est devenue une référence internationale pour les travaux de segmentation et de classification des gliomes.

D'autres bases de données publiques, disponibles sur des plateformes telles que Kaggle, Figshare ou Mendeley Data, proposent des images IRM classées selon différents types de tumeurs cérébrales, notamment les gliomes, les méningiomes et les adénomes hypophysaires. Ces ensembles de données sont fréquemment utilisés dans les études portant sur la classification automatique à l'aide des réseaux de neurones convolutifs (CNN).

L'utilisation de ces bases de données présente plusieurs avantages, notamment la disponibilité d'un grand nombre d'images annotées, la possibilité de comparer les performances des modèles proposés avec celles de la littérature scientifique, ainsi que la reproductibilité des résultats expérimentaux. Toutefois, certaines limitations subsistent, telles que l'hétérogénéité des protocoles d'acquisition, les déséquilibres entre les classes et les variations liées aux équipements utilisés dans différents centres hospitaliers.

Dans ce mémoire, les expérimentations reposent sur une base de données publique contenant des images IRM appartenant à plusieurs catégories tumorales. Ce choix permet d'évaluer les performances des architectures de Deep Learning dans des conditions proches de celles rencontrées dans les travaux récents de la littérature scientifique [42].

1.4. Intelligence artificielle et Deep Learning

En neuro-oncologie, l'intelligence artificielle améliore la précision diagnostique et la détection précoce en identifiant des phénotypes complexes difficilement perceptibles à l'œil. Elle s'impose progressivement comme une aide à la décision, sans se substituer à l'expertise du radiologue.

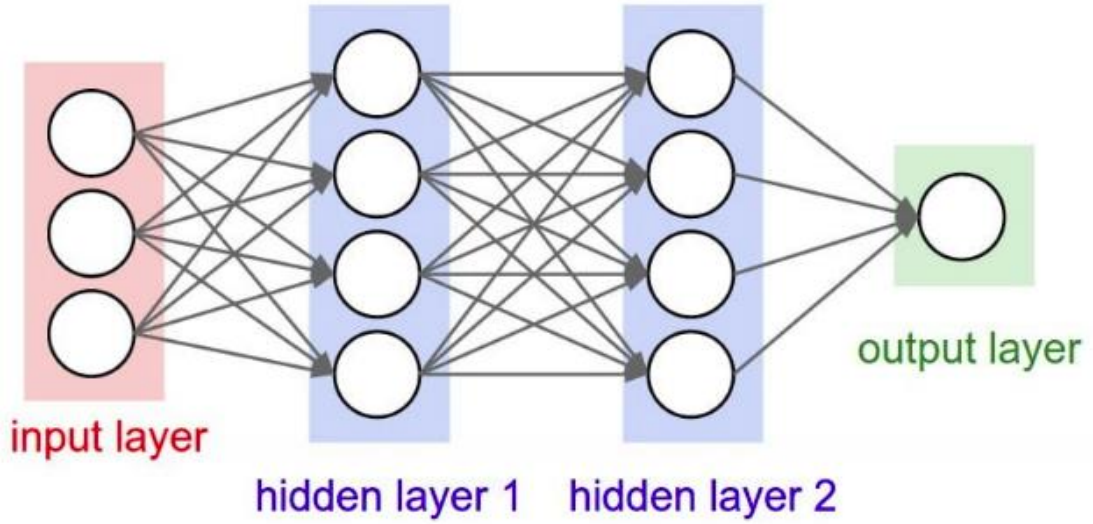


Figure 1.8 : Principe de fonctionnement de l'apprentissage profond [52]

- Input layer : couche d'entrée, qui reçoit l'information et la transfère aux nœuds sous-jacents.
- Hidden layer : les couches cachées sont celles où les calculs ont lieu.
- Output layer : couche de sortie où les résultats des calculs apparaissent.

Sur le plan technique, l'apprentissage automatique se divise en approches supervisée (données étiquetées), non supervisée (structures latentes) et par renforcement (décisions séquentielles). La classification d'images médicales relève du paradigme supervisé.

Tableau 1.3 : Principaux paradigmes du Machine Learning

Type d'apprentissage	Principe	Exemple médical
Supervisé	Données étiquetées	Classification de tumeurs
Non supervisé	Données non étiquetées	Clustering de patients
Renforcement	Récompenses/pénalités	Optimisation thérapeutique

Parmi ces différentes approches, l'apprentissage supervisé est le plus largement utilisé dans les applications de classification d'images médicales. Dans ce cadre, le modèle apprend à associer une image à une classe prédéfinie à partir d'exemples annotés par des experts. Les performances du système dépendent fortement de la qualité des données d'entraînement ainsi que de la capacité du modèle à généraliser ses connaissances sur de nouvelles données.

1.4.1. Réseaux de neurones convolutifs (CNN)

Les réseaux de neurones convolutifs, ou **CNN** (Convolutional Neural Networks), constituent la pierre angulaire des avancées actuelles en vision par ordinateur, particulièrement dans le domaine de la neuro-imagerie. Contrairement aux réseaux de

neurones denses classiques, les CNN sont spécifiquement architecturés pour préserver la topologie spatiale des images. [56]

Ils opèrent par une succession de transformations mathématiques réparties sur trois types de couches fondamentales :

- **Les couches de convolution :** Véritable cœur du réseau, elles utilisent des filtres (noyaux de convolution) qui parcourent l'image pour extraire des cartes de caractéristiques (feature maps). Dans les premières couches, le réseau détecte des motifs rudimentaires tels que les contours et les gradients d'intensité. Au fur et à mesure de la profondeur, les filtres parviennent à encoder des structures plus complexes, comme la texture d'une masse tumorale ou l'irrégularité de ses contours. [56]
- **Les couches de pooling (mise en commun) :** Ces couches assurent une réduction de la dimensionnalité spatiale des données (souvent par Max-Pooling). Leur rôle est double : réduire la charge computationnelle du modèle et conférer au réseau une certaine invariance aux petites translations et distorsions, garantissant que la détection d'une tumeur ne dépende pas de sa position exacte dans le cliché IRM. [56]
- **Les couches totalement connectées (Fully Connected) :** Situées à la fin de l'architecture, ces couches agrègent les caractéristiques de haut niveau extraites précédemment pour effectuer la classification finale. Elles calculent un score de probabilité permettant d'attribuer l'image à une catégorie spécifique (par exemple : Gliome, Méningiome ou Adénome). [56]

L'avantage déterminant des CNN en neuro-oncologie réside dans leur capacité à capturer des dépendances locales et globales au sein des images IRM. Ils surpassent les techniques conventionnelles en s'adaptant à l'hétérogénéité visuelle des tumeurs, permettant non seulement une détection précise, mais aussi une segmentation rigoureuse des sous-régions tumorales (noyau nécrotique, zone de rehaussement, œdème). [56]

1.4.2. Principe du Transfer Learning

L'apprentissage par transfert (transfer learning) réutilise un modèle pré-entraîné sur un grand corpus source (typiquement ImageNet) pour une tâche cible aux données limitées. Il exploite le fait que les couches initiales apprennent des caractéristiques génériques (contours, textures) transférables, seules les couches finales étant ré-adaptées. En neuro-imagerie, où les IRM annotées sont rares, cette stratégie est devenue incontournable pour atteindre de bonnes performances sans entraînement complet. [57]

Le processus repose sur l'idée que les couches initiales d'un CNN apprennent des caractéristiques génériques — telles que les courbes, les textures et les contrastes — qui sont universelles à toute image numérique. Au lieu d'initialiser les poids du réseau de manière aléatoire, ce qui nécessiterait une quantité massive de données et de temps de calcul pour converger, on utilise les poids déjà optimisés d'architectures performantes (telles que **VGG**, **ResNet** ou **Inception**). Le processus se décline généralement en deux étapes :

- **L'extraction de caractéristiques (Feature Extraction) :** On gèle les premières couches du réseau pour conserver les détecteurs de formes basiques, tout en remplaçant uniquement la couche de sortie pour l'adapter au nombre de classes de tumeurs à identifier.
- **Le réglage fin (Fine-Tuning) :** On débloque certaines couches profondes pour les ré-entraîner avec un taux d'apprentissage très faible. Cela permet au modèle de s'ajuster aux spécificités morphologiques des tumeurs cérébrales, affinant ainsi sa sensibilité diagnostique.

L'adoption du Transfer Learning offre des avantages substantiels : elle accélère considérablement la phase de convergence, réduit les risques de sur-apprentissage (overfitting) sur des petits échantillons et permet d'atteindre des niveaux de précision élevés, même avec des ressources computationnelles limitées. [57]

1.4.3. Architectures de référence en classification d'images

Le CNN simple : constitue l'architecture de base en apprentissage profond pour la classification d'images. Les sources consultées le décrivent comme un réseau convolutif séquentiel composé de couches de convolution, de pooling et de couches entièrement connectées. Il est principalement utilisé pour illustrer le fonctionnement général des modèles de classification ou comme point de départ pour des comparaisons. De nombreuses revues et études d'application soulignent qu'il s'agit d'une structure largement utilisée en vision par ordinateur, facile à mettre en œuvre et adaptée aux problèmes de faible complexité, mais peu performante face aux architectures plus avancées. [58]

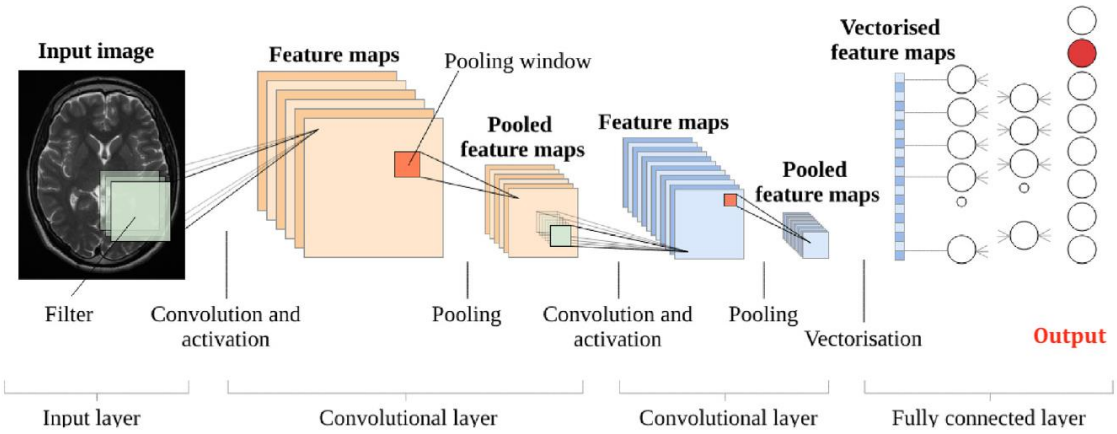


Figure 1.9 : Exemple d'un réseau de neurones convolutifs [59]

VGG16 : représente une étape plus avancée, étant une architecture plus profonde qui est devenue une référence incontournable dans le domaine de la classification d'images. Les études examinées insistent sur sa nature profondément hiérarchique et sa capacité à extraire des caractéristiques complexes, grâce à ses 16 couches standardisées. En pratique, cette architecture est souvent sélectionnée pour son équilibre entre profondeur suffisante pour capturer des motifs visuels riches et simplicité d'utilisation, notamment pour des tâches de fine-tuning par apprentissage par transfert. Cependant, son inconvénient majeur reste un coût calculatoire supérieur à celui du CNN simple. [60]

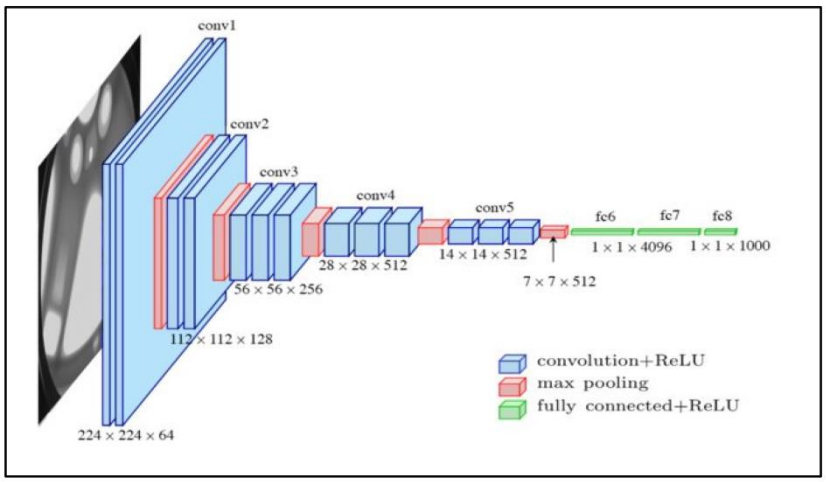


Figure 1.10 : Analyse des tumeurs cérébrales par Deep Learning et approches d'apprentissage par ensemble basées sur VGG-16. [61]

ResNet : propose une approche novatrice en introduisant le concept d'apprentissage résiduel. Présenté dans l'étude fondatrice de He et al. ce cadre permet de "faciliter l'entraînement" de réseaux bien plus profonds que ceux précédemment utilisés. Son innovation clé repose sur l'ajout de connexions résiduelles (skip connections), qui stabilisent l'entraînement même pour des architectures très profondes. Les chercheurs ont testé des réseaux allant jusqu'à 152 couches, surpassant la profondeur de VGG tout en maintenant une complexité moindre. Ils

ont également rapporté une nette amélioration des performances sur le dataset ImageNet par rapport aux approches antérieures. Malgré ses atouts, certains travaux ont mis en lumière des risques potentiels de surapprentissage selon les paramètres choisis. [62]

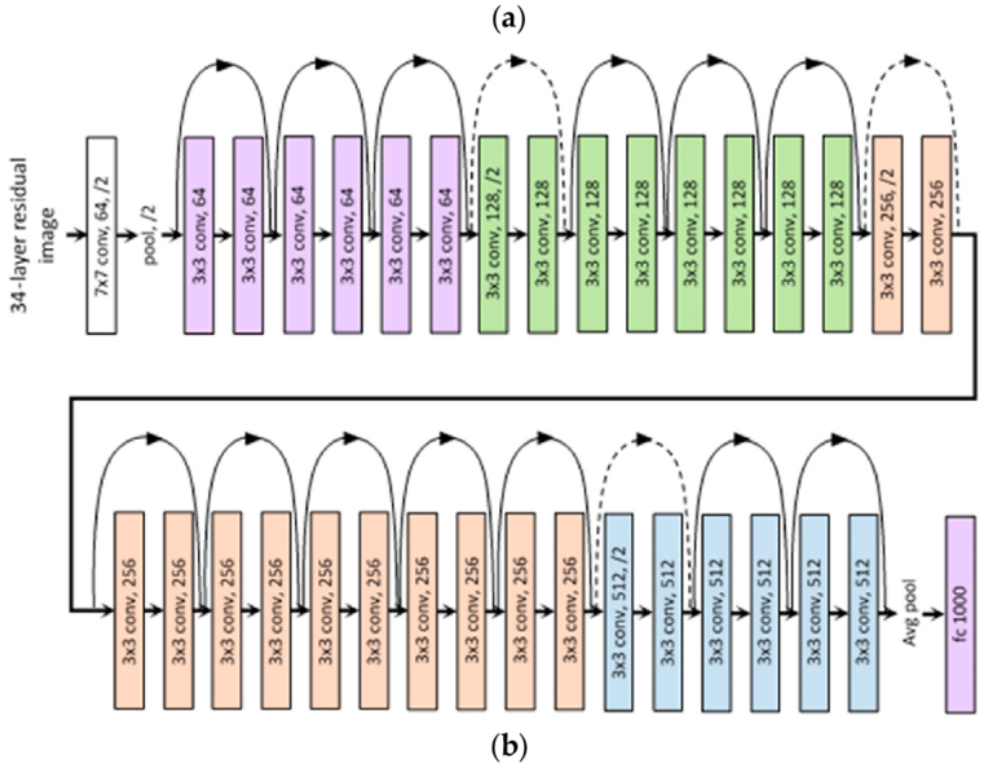


Figure 1.11 : Architecture de ResNet50. [63]

Pour résumer en une phrase comparative : le CNN simple agit comme point de départ, VGG16 incarne une référence classique plus profonde, et ResNet représente la solution moderne pour les réseaux très profonds grâce à ses connexions résiduelles qui garantissent la stabilité de l'entraînement. En termes pédagogiques, ces architectures suivent une progression logique : simplicité initiale, profondeur standardisée, et enfin profondeur optimisée par des innovations structurelles. [58]

Dans les architectures classiques dédiées à la classification d'images, un CNN simple constitue généralement une baseline ; VGG16 se distingue par son architecture profonde et uniforme basée sur des blocs convolutionnels homogènes ; tandis que ResNet innove en intégrant des connexions résiduelles pour optimiser efficacement les réseaux très profonds. [64]

1.5. Revue des travaux antérieurs

L'état de l'art actuel témoigne d'une mutation technologique majeure dans le domaine de la neuro-imagerie, où la littérature scientifique a été marquée par une convergence massive vers l'automatisation du diagnostic des pathologies cérébrales. L'avènement de l'apprentissage

profond (Deep Learning) a agi comme un catalyseur, permettant de transcender les limites des approches classiques par le développement de méthodologies robustes capables de traiter la haute dimensionnalité des images IRM. [65]

Ces travaux de recherche ne visent pas uniquement à maximiser la précision diagnostique, mais aspirent également à réduire la subjectivité inhérente à l'interprétation humaine et à optimiser les délais de prise en charge clinique. Dans cette section, nous dressons une synthèse analytique des avancées récentes les plus significatives, en mettant en lumière les architectures neuronales prédominantes, les bases de données de référence ayant servi de socle à ces études, ainsi que les performances de pointe rapportées par la communauté scientifique.

1.5.1. Études récentes sur la classification des tumeurs cérébrales par IRM

L'état de l'art actuel (2023-2025) montre que la classification automatique des pathologies cérébrales a franchi une étape décisive grâce à la convergence entre le **Transfer Learning** et les architectures de réseaux profonds hybrides. Les chercheurs ne se limitent plus à l'utilisation de modèles standards, mais explorent des stratégies d'optimisation complexes pour répondre aux défis de l'imagerie médicale.

Récemment, les travaux de **Boubdellaha et al. (2025)** ont mis en évidence l'efficacité de l'apprentissage ensembliste (Ensemble Learning). En combinant les architectures **VGG16** et **InceptionV3**, ils ont démontré qu'il est possible de stabiliser les prédictions et d'atteindre des niveaux de précision supérieurs à **99 %** sur des bases de données multiclasses. [66]

Dans une approche complémentaire, **Fki et al. (2024)** ont introduit l'algorithme **LuNet** couplé à des réseaux convolutifs profonds (**DCNN**), soulignant que l'innovation algorithmique, associée à une segmentation rigoureuse, reste la clé pour discriminer des tissus tumoraux aux caractéristiques visuelles très proches. [67]

Par ailleurs, l'importance du prétraitement spatial a été réaffirmée par **Dhakshnamurthy et al. (2024)**, qui ont intégré des filtres de **Wiener** à une architecture **ResNet-50**, prouvant que la réduction du bruit thermique des IRM est un facteur déterminant pour la performance finale du classifieur. [68]

Toutefois, comme le soulignent **Sultan et al. (2024)**, la performance brute (l'accuracy) doit être analysée avec une prudence méthodologique. Leurs recherches sur les modèles **Xception** rappellent que la capacité de généralisation d'un modèle dépend intrinsèquement de la gestion du déséquilibre des classes et de l'étanchéité des protocoles de validation afin d'éviter tout risque de fuite de données (Data Leakage). [69]

Enfin, les travaux de **Chakrabarty et al. (2023)** sur les **CNN 3D** apportent une perspective volumétrique essentielle, bien que plus exigeante en ressources calculatoires, montrant la complexité de la classification lorsque le nombre de types tumoraux augmente. [70]

1.5.2. Datasets utilisés dans la littérature

La performance et la capacité de généralisation d'un modèle de Deep Learning dépendent intrinsèquement de la qualité, de la diversité et du volume des données d'entraînement. Dans la littérature consacrée à la classification des tumeurs cérébrales, deux types de sources prédominent : les bases de données issues de challenges internationaux et les dépôts publics de données cliniques.

A. Les bases de données de référence (Challenges)

Le dataset **BraTS (Brain Tumor Segmentation)** est sans doute la référence la plus citée, notamment dans les travaux de **Fki et al. (2024)** et **Chakrabarty et al. (2023)**. Ce dataset se distingue par sa nature multimodale (incluant les séquences T1, T1c, T2 et FLAIR) et par la précision de ses annotations réalisées par des experts. Bien qu'il soit initialement conçu pour la segmentation, il est fréquemment utilisé pour l'entraînement de classifieurs robustes en raison de sa grande variabilité pathologique. [71]

B. Les dépôts publics et collaboratifs

Pour les études axées sur la classification multi-classes simplifiée, les chercheurs se tournent majoritairement vers des plateformes comme **Kaggle** ou **Figshare**.

- **Le Dataset de Nickparvar (Kaggle) :** Utilisé massivement dans les études récentes (ex: **Boubdellaha et al., 2025** ; **Sultan et al., 2024**), ce jeu de données regroupe plus de 7 000 images IRM réparties en quatre classes (Gliome, Méningiome, Adénome hypophysaire et absence de tumeur). Sa popularité s'explique par son prétraitement initial qui facilite la convergence des modèles de Transfer Learning. [72]
- **Le Dataset de Cheng (Figshare) :** Plus ancien mais toujours pertinent, il contient 3 064 coupes d'IRM de 233 patients. Il est souvent utilisé pour valider la robustesse des modèles face à des résolutions d'image variables. [72]

C. Limites et critiques des datasets publics

L'analyse critique de ces sources révèle plusieurs défis méthodologiques que l'on doit prendre en compte lors de l'évaluation des résultats :

1. **Biais de domaine :** La plupart de ces images proviennent de protocoles d'acquisition standardisés qui ne reflètent pas toujours la "réalité bruitée" des milieux hospitaliers. [73]
2. **Absence de métadonnées cliniques :** Le manque d'informations sur l'âge, le sexe ou le stade de la tumeur limite l'interprétation biologique des prédictions du modèle. [73]
3. **Problématique du partitionnement :** Comme souligné précédemment, l'utilisation de ces datasets publics nécessite une vigilance particulière concernant la séparation des données par patient afin d'éviter le **Data Leakage**, une faille souvent omise dans les publications affichant des taux d'exactitude frôlant la perfection. [73]

1.5.3. Analyse des méthodes et performances rapportées

Cette section examine les approches techniques adoptées dans la littérature récente pour optimiser la classification des tumeurs cérébrales. Les méthodologies recensées se distinguent par l'exploitation de trois leviers technologiques majeurs :

A. Le Transfer Learning et le Fine-tuning

La méthode la plus répandue consiste à réutiliser des architectures pré-entraînées sur de grandes bases de données d'images générales (ImageNet) pour les adapter au domaine médical. Les chercheurs comme **Sultan et al. (2024)** utilisent le **Fine-tuning**, qui consiste à geler les premières couches du réseau (qui extraient les formes simples) et à réentraîner les couches profondes pour apprendre les caractéristiques spécifiques des tissus tumoraux. Cette approche permet de pallier le manque de données médicales massives tout en bénéficiant de la puissance d'extraction de modèles éprouvés comme **Xception** ou **InceptionV3**. [74]

B. Les architectures hybrides et l'Ensemble Learning

Une tendance forte consiste à combiner plusieurs modèles pour compenser les faiblesses individuelles de chaque architecture. **Boubdellaha et al. (2025)** proposent une méthode d'**Ensemble Learning** basée sur le vote majoritaire ou la concaténation de vecteurs de caractéristiques (Feature Fusion). Par exemple, l'association de **VGG16** (robuste pour les textures) et d'**Inception** (performant pour les détails multi-échelles) permet d'obtenir une représentation plus riche de l'image IRM, se traduisant par des performances supérieures à **99%**. [75]

C. Algorithmes personnalisés et mécanismes de régularisation

Certaines études introduisent des modifications structurelles au sein même des réseaux. C'est le cas de **Fki et al. (2024)** avec l'algorithme **LuNet**, qui optimise la manière dont les neurones s'activent face aux bords flous des tumeurs. De plus, pour stabiliser l'apprentissage, des méthodes de **régularisation** sont systématiquement intégrées :

- **Le Dropout :** Utilisé pour désactiver aléatoirement des neurones et forcer le réseau à ne pas dépendre d'un seul chemin de décision. [76]
- **L'Augmentation de données (Data Augmentation) :** Application de rotations, de zooms et de symétries pour diversifier artificiellement le dataset et améliorer la robustesse du modèle. [76]

D. Analyse des performances et limites méthodologiques

Bien que les taux d'**Accuracy** rapportés soient extrêmement élevés, leur fiabilité dépend étroitement du protocole expérimental. Les méthodes affichant les meilleurs scores (ex: **99,50 %**) sont souvent celles qui intègrent un prétraitement lourd, comme le **filtrage de Wiener** pour la réduction du bruit (**Dhakshnamurthy et al., 2024**). Il est important de souligner que la performance brute doit toujours être mise en perspective avec la complexité de l'architecture et sa capacité à généraliser sur des données qu'elle n'a jamais rencontrées lors de l'entraînement. [77]

1.5.4. Tableau Comparatif des principales études de la littérature

Tableau 1.4 : Comparatif des principales études de la littérature

Auteurs	Dataset	Classes	Architecture	**Accuracy**	Limites & Critiques
Boubdellaha et al. (2025). **[66]**	Kaggle (Nickparvar)	7 023 images / 4 classes	Modèle hybride : **Ens-VGG16 + InceptionV3**	99.12%	Complexité de calcul
Dhakshnamurthy et al. (2024). **[68]**	Kaggle & Figshare	3 064 images / 3 classes	**CNN-ResNet50** avec optimisation Adam	98.50%	Sensibilité au bruit
Fki et al. (2024). **[67]**	BraTS 2021	2 000+ volumes / Multi-classes	DCNN + Algorithme LuNet	99.50%	Dépendance au Dataset
Sultan et al. (2024). **[69]**	Kaggle	Multi-sources / 4 classes	Transfer Learning (Xception)	97.20%	Dépendance au Dataset
Chakrabarty et al. (2023). **[70]**	Multi-institutionnel	2 105 examens / 6 classes	**3D CNN** (volumes complets)	93.35%	Besoin en RAM élevé

1.6. Analyse critique

Malgré les performances prometteuses rapportées par les travaux récents en classification des tumeurs cérébrales, une analyse rigoureuse de la littérature permet d'identifier plusieurs verrous méthodologiques et opérationnels. Ces limitations freinent encore l'adoption massive de ces systèmes en routine clinique et peuvent être synthétisées selon trois axes majeurs :

- **Dépendance et hétérogénéité des données :** La majorité des modèles performants sont entraînés sur des jeux de données publics (tels que Figshare ou BraTS) qui, bien que précieux, sont souvent standardisés et prétraités. En contexte clinique réel, les images IRM présentent une grande variabilité due aux différents constructeurs d'aimants (Siemens, GE, Philips) et aux protocoles d'acquisition, ce qui peut dégrader significativement la précision d'un modèle non robuste à ces variations. [78]
- **Déficit de généralisation et risque de sur-apprentissage :** L'utilisation d'architectures extrêmement profondes sur des bases de données médicales relativement restreintes expose les modèles au phénomène de **sur-apprentissage** (overfitting). Si les taux d'exactitude frôlent la perfection en laboratoire, leur capacité de généralisation face à des cas pathologiques atypiques ou des artefacts d'image reste souvent limitée. [79]
- **Boîte noire et interprétabilité :** L'un des obstacles majeurs reste l'aspect "boîte noire" des réseaux de neurones profonds. En neuro-oncologie, la décision médicale nécessite une justification spatiale (où se situe la lésion ?) et sémantique (quels traits ont conduit à ce diagnostic ?). Rares sont les travaux qui intègrent des mécanismes d'explicabilité permettant au radiologue de valider visuellement le raisonnement de l'IA. [80]

L'ensemble de ces constats souligne la nécessité de concevoir des systèmes de classification qui ne se contentent pas d'une haute précision brute, mais qui intègrent des stratégies de prétraitement robustes et des architectures optimisées. C'est précisément dans cette optique que s'articule notre contribution. Le chapitre suivant détaillera la méthodologie mise en œuvre pour relever ces défis, en présentant les outils technologiques et les choix de conception retenus pour notre système de diagnostic assisté par ordinateur.

1.6.1. Limites liées à la dimensionnalité des jeux de données

L'analyse distingue deux problématiques selon la taille des données. Pour les petits datasets, les défis sont le surapprentissage, les erreurs de généralisation et le compromis biais-variance, avec une validation croisée imbriquée plus robuste que le K-fold classique même jusqu'à 1000 échantillons. Pour les grands datasets, l'augmentation de la taille finit par plafonner les gains de précision, et les limites deviennent alors la qualité des données comme le bruit excessif ou une faible séparabilité, sans oublier que les tailles d'échantillon en oncologie sont souvent inférieures aux seuils requis. [81-82]

1.6.2. Déséquilibre des classes (Class Imbalance)

En imagerie médicale, les classes sont souvent déséquilibrées : les tumeurs fréquentes dominent les sous-types rares. Sans correction (pondération, augmentation ciblée), le modèle tend à favoriser les classes majoritaires au détriment des minoritaires, pourtant parfois les plus critiques.

Ce déséquilibre numérique induit un biais algorithmique majeur : le modèle tend à minimiser la fonction de perte globale en privilégiant la reconnaissance des classes les plus représentées. Par conséquent, la **sensibilité** de détection pour les classes minoritaires s'en trouve dégradée, générant un taux élevé de faux négatifs. Pourtant, l'identification de ces cas rares revêt souvent un intérêt pronostique crucial, car ils peuvent correspondre à des pathologies à progression rapide nécessitant une intervention immédiate. Sans une stratégie de compensation rigoureuse, le modèle risque de perdre sa pertinence clinique au profit d'une exactitude globale trompeuse. [83]

1.6.3. Risque de surapprentissage (Overfitting)

Le surapprentissage (overfitting) survient quand la complexité du modèle dépasse la richesse des données : le réseau mémorise le bruit et les spécificités du jeu d'entraînement au lieu d'apprendre des caractéristiques généralisables. Il se traduit par un écart croissant entre performances en entraînement et en validation.

Ce manque de flexibilité crée un écart critique entre performances de laboratoire et efficacité clinique : un modèle en surapprentissage affiche d'excellents scores sur ses données mais se dégrade fortement sur des images nouvelles.

1.6.4. Problématique de l'explicabilité (Black-Box)

L'opacité des modèles profonds, dits « boîtes noires », freine leur adoption clinique : malgré d'excellentes performances, la difficulté à expliquer leurs décisions pose des problèmes éthiques, légaux et de confiance. Des méthodes comme Grad-CAM visent à rendre ces décisions plus interprétables.

En médecine, la compréhension du « pourquoi » est aussi cruciale que la précision du résultat lui-même. Un radiologue ou un neurochirurgien ne peut valider une stratégie thérapeutique lourde sur la seule base d'une probabilité statistique ; il nécessite une justification spatiale permettant de corréler la décision de l'IA avec des signes radiologiques tangibles, tels que l'œdème péritumoral ou la prise de contraste. Cette absence de transparence limite la confiance du corps médical et souligne l'importance émergente de l'intelligence

artificielle explicable (**XAI**), visant à transformer ces « boîtes noires » en systèmes transparents et auditables. [85]

1.6.5. Défis de la généralisation en milieu clinique

Le passage du laboratoire à la routine hospitalière reste difficile. Ce « fossé de généralisation » provient surtout de la variabilité réelle des IRM : différences d'équipements (constructeurs, champ 1,5 T contre 3 T), hétérogénéité des protocoles d'acquisition et diversité des populations de patients. Un modèle performant sur un jeu de données peut ainsi se dégrader fortement sur des images d'une autre source — ce qui justifie précisément notre choix d'une validation externe. [62]

Plusieurs facteurs techniques et biologiques entrent en jeu :

- **Variabilité des équipements :** Les différences technologiques entre les constructeurs (Siemens, GE, Philips) et l'intensité du champ magnétique (1.5T vs 3T) produisent des variations significatives dans le rapport signal/bruit et le contraste des tissus.
- **Hétérogénéité des protocoles :** Les paramètres d'acquisition (temps d'écho, temps de répétition) varient d'un centre radiologique à un autre, créant des disparités visuelles que les modèles entraînés sur des bases publiques standardisées peinent à compenser.
- **Diversité démographique et pathologique :** La diversité des profils de patients et la variabilité de l'apparence des tumeurs à différents stades de progression introduisent un décalage entre les résultats expérimentaux et l'efficacité clinique réelle.

Cette absence de robustesse face à des données « hors distribution » (*Out-of-Distribution*) souligne l'impérieuse nécessité de développer des algorithmes capables de s'adapter à la diversité des données cliniques, garantissant ainsi un diagnostic fiable et reproductible, quel que soit l'environnement hospitalier. [86]

1.6.6. Solutions méthodologiques : pondération des classes, focal loss, validation croisée, séparation au niveau patient, test externe, Grad-CAM :

Pour répondre aux limites méthodologiques de la littérature, plusieurs choix ont été retenus : une pondération adaptée et l'utilisation de la focal loss pour gérer le déséquilibre des classes ; une validation croisée pour une estimation plus robuste des performances ; une séparation des données au niveau patient pour éviter les fuites d'information ; un test externe pour évaluer la généralisation sur des données indépendantes ; et l'intégration de Grad-CAM pour ajouter une dimension interprétable aux prédictions du modèle. [87,88-89,90]

1.7. Positionnement de l'étude

Face à l'évolution fulgurante des technologies d'intelligence artificielle et aux défis persistants de la neuro-oncologie, la classification automatisée des néoplasmes cérébraux s'impose désormais comme un champ de recherche prioritaire à la convergence de la médecine et de l'ingénierie.

Ce travail s'inscrit dans cette dynamique : il propose une méthodologie rigoureuse fondée sur le deep learning, avec une attention particulière à la validation externe pour garantir la généralisation des modèles.

1.7.1. Justification scientifique du travail proposé

La détection précoce des tumeurs cérébrales sur IRM est un enjeu clinique majeur, aujourd'hui limité par une interprétation manuelle chronophage et variable. Les réseaux de neurones convolutifs (CNN) y répondent en apprenant automatiquement des représentations discriminantes. Ce mémoire développe et évalue une chaîne complète de classification en quatre classes (gliome, méningiome, adénome hypophysaire, absence de tumeur), comparant un CNN simple, VGG16 et ResNet50. Sa contribution principale est méthodologique : recourir à une validation externe sur une source indépendante afin de vérifier que les performances se maintiennent sur des données réellement nouvelles — condition indispensable à toute application clinique.

1.7.2. Contribution attendue du mémoire

1.7.3. Intérêt académique et clinique de l'approche choisie

- **Intérêt académique :** Sur le plan académique, l'approche CNN apporte une contribution méthodologique en comparant systématiquement l'impact des architectures (CNN, VGG, ResNet) et des choix d'entraînement sur la performance diagnostique, au-delà du simple score. La valeur scientifique réside dans l'analyse des limites (dépendance aux données annotées, surapprentissage, difficulté de généralisation). Le mémoire peut ainsi cadrer sa contribution comme une démarche d'évaluation complète incluant une définition claire des classes, un protocole entraînement/validation/test, et des stratégies anti-surapprentissage pour garantir que les gains proviennent de la généralisation et non du surajustement.

- Sur le plan clinique, l'approche peut réduire le temps d'interprétation, aider à la décision et standardiser le tri des cas — sous réserve d'une validation prospective, l'outil restant une aide et non un substitut au radiologue.

1.8. Conclusion

Ce chapitre a présenté les fondements théoriques nécessaires à la compréhension du problème de classification automatisée des tumeurs cérébrales. Après avoir introduit les principales catégories tumorales ainsi que l'importance du diagnostic précoce, les différentes modalités d'imagerie médicale utilisées en neuro-oncologie ont été étudiées, avec un intérêt particulier pour l'IRM, qui constitue la principale source de données exploitée dans ce travail.

Les concepts fondamentaux de l'intelligence artificielle, du Machine Learning et du Deep Learning ont ensuite été présentés, suivis d'une étude détaillée des réseaux de neurones convolutifs et des architectures avancées VGG16 et ResNet50. Les techniques d'optimisation telles que le transfert d'apprentissage, le Fine-Tuning et l'augmentation des données ont également été abordées.

Enfin, une analyse des travaux récents a permis d'identifier les principales avancées ainsi que les limites actuelles des systèmes de classification des tumeurs cérébrales. Cette étude bibliographique a servi à définir le positionnement scientifique du présent mémoire. Le chapitre suivant sera consacré à la méthodologie adoptée, à la présentation de la base de données utilisée ainsi qu'à la conception des modèles développés pour la classification automatique des tumeurs cérébrales.

Chapitre 2

Matériel et Méthodes

1. Introduction

Ce chapitre décrit les moyens techniques et méthodologiques adoptés pour la détection et la classification automatique des tumeurs cérébrales à partir d'images IRM à l'aide de modèles de deep learning. Il présente les différentes étapes du processus, depuis la sélection des données jusqu'à l'évaluation des performances des modèles. Une attention particulière est accordée au prétraitement des images et à la conception des architectures neuronales. Afin de garantir la fiabilité, la transparence et la reproductibilité des résultats, un protocole expérimental rigoureux a été mis en place. La validation repose sur deux sources de données indépendantes : un jeu de données Kaggle utilisé pour l'apprentissage et la validation, et un jeu de données Mendeley Data réservé aux tests finaux. Cette approche permet d'évaluer la capacité de généralisation des modèles dans des conditions proches d'une utilisation réelle.

2.1. Démarche générale de l'étude

2.1.1. Schéma global du pipeline de traitement

La Figure 2.1 illustre l'ensemble du pipeline de traitement, depuis la collecte des images IRM jusqu'à l'évaluation finale des modèles :

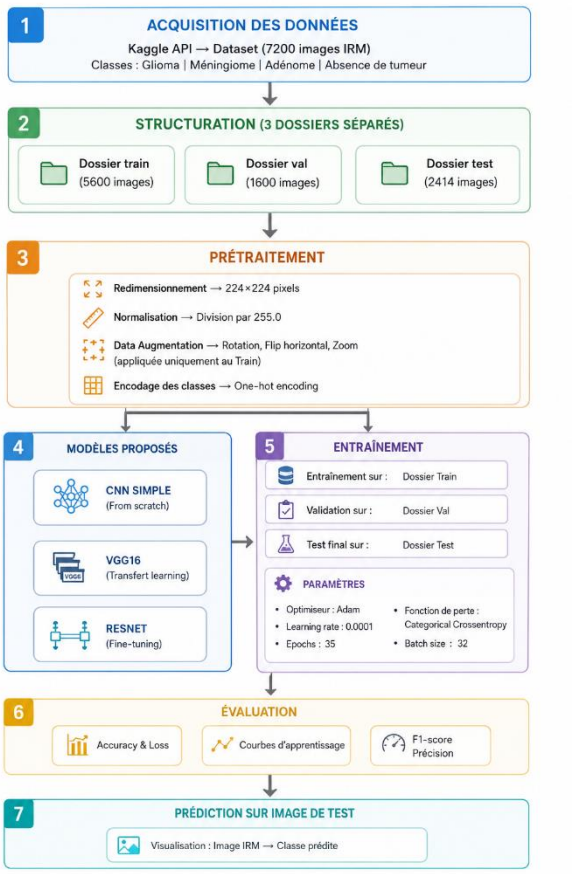


Figure 2.1 : Pipeline global de la méthodologie, de l'acquisition à l'évaluation

2.1.2 Étapes principales de la méthodologie

La méthodologie mise en œuvre dans cette étude s'articule autour de cinq grandes étapes:

1. **Acquisition et organisation des données :** Les images IRM proviennent du dataset Kaggle (7 200 images au total). Elles sont réparties en trois dossiers distincts : un dossier d'apprentissage (Training) contenant 5 600 images, un dossier de validation (Validation) avec 1600 images, et un dossier de test (Testing) comprenant 2414 images.
2. **Prétraitement des images :** Toutes les images sont uniformisées à une résolution de 224×224 pixels. Les valeurs des pixels sont normalisées par division par 255,0. Des techniques d'augmentation de données (rotations, retournements horizontaux, zooms) sont appliquées **uniquement** sur l'ensemble d'apprentissage, afin d'enrichir la diversité des exemples vus par le modèle.
3. **Entraînement des modèles :** Trois architectures de réseaux de neurones convolutifs sont comparées : un réseau simple construit sur mesure, VGG16 avec transfer learning, et ResNet-50 avec fine-tuning. L'entraînement est réalisé sur le dossier Training, tandis que le dossier Validation est utilisé pour le réglage des hyperparamètres et la détection précoce du surapprentissage.
4. **Évaluation des performances :** Les modèles sont évalués à l'aide des métriques suivantes : accuracy, précision, rappel, F1-score et loss. Des courbes d'apprentissage sont générées pour visualiser l'évolution de ces indicateurs au fil des époques.
5. **Prédiction finale sur données inédites :** Le modèle retenu est soumis à des images IRM jamais vues auparavant, issues du dossier Testing. La classe prédite est affichée sur chaque image, permettant une appréciation qualitative et quantitative de la capacité de généralisation du modèle.

2.2. Description du dataset

2.2.1. Source des données

L'étude utilise une validation externe avec deux sources de données indépendantes : une pour l'apprentissage/validation, l'autre pour le test. Cette approche, plus rigoureuse qu'un simple partitionnement aléatoire, évalue mieux la capacité du modèle à généraliser sur des images provenant d'appareils, protocoles et centres différents, se rapprochant ainsi des conditions cliniques réelles.

- **Ensemble d'apprentissage et de validation – Kaggle (Masoud Nickparvar)**

Les données d'apprentissage et de validation proviennent du jeu de données public « **Brain Tumor MRI Dataset** », disponible sur Kaggle et publié par **Masoud Nickparvar**. Selon la documentation officielle, ce jeu de données est une **collection composite** issue de trois bases ouvertes largement utilisées dans la littérature :

- **Figshare** : proposé par Cheng et al. ce jeu contient des IRM de tumeurs cérébrales (gliomes, méningiomes, tumeurs hypophysaires) et constitue une référence récurrente en classification.
- **SARTAJ** : introduit par Bhuvaji et al. il est structuré selon quatre classes tumorales et représente un benchmark standard.
- **Br35H** : composé d'IRM cérébrales avec ou sans tumeur (classification binaire), ses images de sujets sains alimentent la classe « No tumor » de notre étude.

La fusion de trois sources crée un corpus unifié et équilibré. Leur hétérogénéité (différents appareils et paramètres) renforce la robustesse du modèle. L'organisation en deux dossiers (Training et Validation) a été conservée pour garantir la reproductibilité et faciliter les comparaisons avec les travaux antérieurs.

- **Ensemble de test – Mendeley Data**

L'évaluation finale des performances repose sur un ensemble de test provenant d'une source **totalement indépendante** : le dépôt **Mendeley Data**. Ce choix présente plusieurs avantages décisifs :

- **Absence de contamination** : Les données d'entraînement/validation (Kaggle) et les données de test (Mendeley) ne se chevauchent pas, ce qui élimine tout risque de fuite d'information. Les métriques obtenues sur le test reflètent donc une **évaluation objective, non biaisée**.
- **Variabilité des conditions d'acquisition** : Les IRM issues de Mendeley Data ont été acquises selon des protocoles distincts (résolutions, champs magnétiques, paramètres techniques différents). Cette hétérogénéité est **délibérément recherchée** pour éprouver la robustesse du modèle face à la diversité du monde réel.
- **Validation externe stricte** : Dans le domaine de l'imagerie médicale, évaluer un modèle sur des données provenant d'une source autre que celle de l'entraînement constitue la **référence méthodologique** pour démontrer la généralisabilité d'une approche.

Le dossier Testing/, téléchargé depuis Mendeley Data, contient des images IRM réparties exactement selon les **quatre mêmes classes** que le jeu d'entraînement (glioma, méningiome, tumeur hypophysaire, absence de tumeur). Cette homogénéité sémantique des étiquettes permet une évaluation cohérente, sans nécessiter de réétiquetage ni d'adaptation supplémentaire.

Le dataset de Masoud Nickparvar a été choisi pour trois raisons : il est largement utilisé dans la littérature récente, ce qui facilite les comparaisons ; il présente une répartition équilibrée entre les classes tumorales, réduisant les biais ; et son volume important est adapté au transfert d'apprentissage avec des architectures profondes comme VGG16 et ResNet50.

2.2.2. Description des classes étudiées

Notre étude vise la classification de quatre catégories distinctes d'images IRM cérébrales : trois types tumoraux (gliome, méningiome, adénome hypophysaire) et une classe témoin correspondant à l'absence de tumeur.

1. Gliome

Les gliomes constituent la classe tumorale la plus agressive du dataset. Leur aspect radiologique est souvent hétérogène, avec présence possible d'œdème et de zones nécrotiques, ce qui en fait une catégorie particulièrement intéressante pour l'évaluation des capacités discriminantes des modèles de Deep Learning [91-92].

2. Méningiome

Le méningiome, tumeur extra-axiale issue des cellules arachnoïdiennes des méninges, constitue la tumeur cérébrale primitive la plus fréquente chez l'adulte, représentant environ 35 % des cas. Généralement bénin (grade I OMS) et à croissance lente, il se caractérise sur le plan radiologique par une masse homogène, bien limitée, fréquemment associée au « signe de la queue de la dure-mère » **[94]**. Bien que sa nature soit bénigne, sa localisation peut induire une symptomatologie clinique par compression des structures cérébrales adjacentes. Pour les besoins de l'étude, l'imagerie de cette entité pathologique est issue des bases de données Figshare, SARTAJ et Mendeley Data **[93]**.

3. Adénome hypophysaire

L'adénome hypophysaire est une tumeur bénigne de la région sellaire issue de l'hypophyse, représentant 10 à 15 % des tumeurs cérébrales pour une prévalence d'environ 1 cas sur 1 000. Sur le plan radiologique, l'IRM révèle une masse bien limitée présentant un rehaussement post-injection généralement moins intense que le parenchyme hypophysaire sain, permettant de distinguer les microadénomes (< 1 cm) des macroadénomes (≥ 1 cm), ces derniers pouvant comprimer le chiasma optique et causer des déficits visuels. Cliniquement, ces néoformations se divisent en formes sécrétantes et non sécrétantes, à l'origine de divers syndromes endocriniens **[95]**. L'échantillonnage de cette classe tumorale au sein de l'étude est extrait des cohortes Figshare, SARTAJ et Mendeley Data **[93]**.

4. Absence de tumeur (No tumor)

La classe « absence de tumeur » (ou *No tumor*) regroupe les examens IRM normaux, exempts de toute masse, œdème ou rehaussement pathologique **[92]**. Cette catégorie s'avère méthodologiquement cruciale pour évaluer la spécificité des modèles de *Deep Learning*, garantissant leur aptitude à exclure les faux positifs afin d'épargner aux patients sains des examens invasifs et une anxiété délétère. Sur le plan clinique, l'équilibre entre la détection des faux positifs et des faux négatifs conditionne directement la sécurité thérapeutique en évitant tout retard de prise en charge. Pour cette étude, l'iconographie des cerveaux sains est issue de la cohorte Br35H pour la phase d'entraînement et de validation, et du dépôt Mendeley Data pour la constitution de l'ensemble de test indépendant **[96]**.

2.2.3. Nombre d'images par classe

Afin d'assurer un apprentissage équilibré et une évaluation fiable, nous détaillons ci-dessous la répartition exacte des images pour chaque classe dans les trois ensembles (Training, Validation, Test).

Tableau 2.1 : Répartition des images par classe dans les trois ensembles (Entraînement et Validation : Kaggle ; Test : Mendeley Data)

Classe	Entraînement	Validation	Test
Gliome	1400	400	755
Méningiome	1400	400	626
Adénome hypophysaire	1400	400	546
Absence de tumeur	1400	400	487
Total	5600	1600	2414

Le dataset d'entraînement a été volontairement équilibré afin de garantir une représentation identique des quatre classes. Cette stratégie limite le risque qu'un modèle favorise artificiellement une classe majoritaire au détriment des autres catégories. À l'inverse, l'ensemble de test conserve la distribution naturelle des données fournies par Mendeley Data, permettant une évaluation plus réaliste des performances de généralisation.

L'analyse de la répartition des données révèle une asymétrie structurale délibérée entre les phases du projet : les ensembles d'apprentissage et de validation affichent un équilibre parfait avec 1 400 images par classe, tandis que l'ensemble de test présente un déséquilibre naturel où le gliome constitue la classe majoritaire (755 images) et l'absence de tumeur la classe minoritaire (487 images), soit un ratio de 1,55:1. Ce choix méthodologique se justifie par des impératifs de robustesse algorithmique : si un entraînement et une validation équilibrés neutralisent les biais de prédiction et garantissent un ajustement impartial des hyperparamètres, un jeu de test déséquilibré s'avère scientifiquement plus pertinent. En

simulant l'hétérogénéité des distributions de prévalence de la pratique clinique réelle, ce dispositif permet d'éprouver la robustesse et la valeur prédictive concrète du modèle face à des conditions d'utilisation non contrôlées.

2.2.4 Répartition des données

Nos données sont organisées en trois ensembles indépendants et disjoints : un ensemble d'apprentissage (Training), un ensemble de validation (Validation) et un ensemble de test (Test). Cette répartition à trois niveaux est essentielle pour garantir une évaluation non biaisée des performances de nos modèles.

Un risque méthodologique fréquent en Deep Learning médical est la fuite d'information (Data Leakage). Ce phénomène survient lorsqu'une même information, directement ou indirectement, est présente à la fois dans les ensembles d'entraînement et de test. Dans notre étude, l'utilisation d'une source de données totalement indépendante pour l'évaluation finale réduit fortement ce risque et constitue un élément important de la validité expérimentale du protocole.

2.2.4.1. Ensemble d'apprentissage (Training)

L'ensemble d'apprentissage provient de Kaggle (dataset de Masoud Nickparvar), contient 5600 images équilibrées (1400 par classe). Il sert à entraîner les trois architectures (CNN simple, VGG16, ResNet50). La data augmentation (rotation, flip, zoom) est appliquée uniquement sur cet ensemble pour enrichir la diversité des données et éviter les biais.

2.2.4.2. Ensemble de validation (Validation)

L'ensemble de validation contient 1600 images équilibrées (400 par classe) provenant de la même source Kaggle. Il est utilisé pour ajuster les hyperparamètres (learning rate, batch size), surveiller le surapprentissage en comparant les performances avec l'entraînement, et décider l'arrêt précoce (early stopping) lorsque la performance cesse de s'améliorer.

2.2.4.3. Ensemble de test (Test)

L'ensemble de test provient de Mendeley Data (source indépendante de Kaggle), contient 2414 images avec une répartition déséquilibrée par classe. Il est utilisé une seule fois après la fin de l'entraînement pour évaluer la généralisation réelle des modèles, comparer objectivement les trois architectures (CNN, VGG16, ResNet50), et estimer les performances en conditions réelles grâce à la source indépendante.

2.2.5. Limites et biais potentiels du dataset

Les limites mentionnées ci-dessous sont communes à la plupart des études utilisant des datasets publics en imagerie médicale. Elles n'altèrent en rien la validité de notre approche ni la qualité de nos résultats, mais leur reconnaissance honnête est une exigence de rigueur scientifique.

2.2.1.1. Biais liés à la source des données

L'utilisation de données publiques provenant de Kaggle et Mendeley ne constitue pas un problème majeur, car elle assure la reproductibilité de l'étude, un avantage scientifique important. De plus, la validation externe mise en place compense en partie cette limite. Par ailleurs, le fait que l'ensemble de test (Mendeley) contienne exclusivement des IRM de tumeurs cérébrales, avec une classe "No tumor" issue du même dataset, est en réalité un point fort pour la tâche de classification, car le modèle est évalué sur des données parfaitement adaptées à son objectif clinique.

2.2.1.2. Biais liés à la répartition des classes

L'utilisation d'un ensemble d'entraînement équilibré (1400 images par classe) est un choix méthodologique délibéré qui ne reflète pas les prévalences réelles des tumeurs, mais cela évite les biais de classification et permet au modèle d'apprendre aussi bien les classes rares que les classes fréquentes. Quant à l'ensemble de test déséquilibré (répartition 755/626/546/487 propre à Mendeley), il constitue un test plus réaliste qu'un test artificiellement équilibré, car le modèle est évalué sur une distribution proche de ce qu'il rencontrerait en pratique.

2.2.1.3. Biais liés à l'annotation et à la qualité

La qualité variable des images due à la fusion de trois bases (Figshare, SARTAJ, Br35H) n'est pas rédhibitoire car elle constitue une force : le modèle apprend sur des images diverses (différents appareils et protocoles), ce qui améliore sa généralisation. Les annotations non vérifiées par des experts ne posent pas de problème majeur car ces bases sont largement utilisées et validées par la communauté scientifique, et leurs labels sont considérés comme fiables. Enfin, l'absence d'informations cliniques (âge, sexe, symptômes) n'est pas un obstacle puisque l'objectif est la classification basée uniquement sur l'image, ce qui est le cas standard en deep learning médical, l'ajout de données cliniques constituant une extension future intéressante.

2.2.1.4. Biais liés à la validation externe

L'utilisation d'une seule source de test (Mendeley) reste valide car c'est déjà un progrès significatif par rapport aux études qui divisent un unique dataset, même si une validation sur plusieurs sources serait idéale. Quant au fait que la source Mendeley ne soit pas entièrement

documentée, cela n'affecte pas la comparaison équitable entre nos trois modèles (CNN, VGG16, ResNet50) évalués sur la même base de test.

2.2.1.5. Absences de données (non des biais)

L'absence d'autres pathologies (métastases, abcès, maladies inflammatoires) constitue une limite de périmètre car notre étude se concentre uniquement sur les tumeurs primitives, mais des travaux futurs pourraient étendre le modèle à d'autres lésions. L'absence d'images avec artefacts (mouvements ou artefacts métalliques) s'explique par le fait que les datasets publics sont généralement de bonne qualité, et une étape supplémentaire de test sur images bruitées serait intéressante à explorer. Enfin, l'utilisation exclusive de données IRM est cohérente avec l'usage clinique standard pour les tumeurs cérébrales, le modèle n'étant pas destiné à d'autres modalités comme le scanner ou le PET.

2.2.1.6. Limites méthodologiques générales

La taille modérée du dataset (9614 images) n'est pas un problème car la plupart des études en imagerie médicale utilisent des datasets de cette taille, ce qui rend les résultats statistiquement significatifs. L'absence de validation croisée n'est pas grave puisque la séparation Training/Validation/Test est claire et standard, et la cross-validation aurait alourdi les calculs sans bénéfice certain. Enfin, l'utilisation de données 2D uniquement est acceptable car la plupart des travaux en classification d'IRM cérébrales utilisent des coupes 2D, l'ajout de la 3D constituant une piste d'amélioration future.

Bien que l'utilisation d'un dataset externe pour le test améliore significativement la crédibilité des résultats, certaines menaces à la validité externe subsistent. Les données utilisées restent issues de bases publiques et ne reflètent pas nécessairement toute la diversité des situations cliniques rencontrées en pratique hospitalière. Des validations multicentriques incluant différents équipements d'imagerie et populations de patients seraient nécessaires avant toute application clinique.

2.3. Prétraitement des données

Toutes les images ont été redimensionnées à une résolution uniforme de 224 × 224 pixels afin d'assurer leur compatibilité avec les architectures VGG16 et ResNet50 pré-entraînées sur ImageNet. Cette résolution représente un compromis entre la conservation des informations anatomiques pertinentes et la réduction de la charge computationnelle. L'utilisation d'une taille d'entrée homogène garantit également la cohérence des traitements appliqués aux différents modèles étudiés. [97-98]

2.3.2. Normalisation

Les intensités des pixels sont normalisées dans l'intervalle [0, 1] (division par 255), ce qui stabilise le calcul des gradients et accélère la convergence. Pour VGG16 et ResNet50, on applique la normalisation spécifique d'ImageNet via preprocess_input (moyennes et écarts-types d'ImageNet). [99-100]

2.3.3. Data Augmentation

Le recours aux techniques d'augmentation de données est motivé par la nécessité d'optimiser l'invariance spatiale et morphologique du modèle face aux variabilités de l'imagerie clinique. L'application du retournement (*flip*) horizontal prévient le biais de mémorisation de la latéralisation en simulant la survenue bilatérale des tumeurs au sein des hémisphères cérébraux **[101, 103]**. Parallèlement, l'introduction de rotations modérées (10 %) compense les fluctuations d'orientation positionnelle des patients lors de l'acquisition **[104]**, tandis que les variations de zoom (10 %) modélisent l'hétérogénéité d'échelle inhérente à l'utilisation de différents appareils IRM **[102]**. Ce processus d'augmentation améliore la robustesse globale et la capacité de généralisation du réseau de neurones.

L'augmentation des données est appliquée uniquement sur l'ensemble d'entraînement, et non sur la validation ou le test. Sur le training, elle crée des variantes artificielles des images, ce qui augmente la taille effective du dataset et améliore la capacité de généralisation du modèle. En revanche, sur les ensembles de validation et de test, l'augmentation n'est pas appliquée car ces ensembles doivent refléter la distribution réelle des données pour permettre une évaluation fidèle des performances.

Tableau 2.2 : Techniques d'augmentation et leurs paramètres associés

Technique	Paramètre	Rôle
RandomFlip	"horizontal"	Retourne aléatoirement l'image horizontalement (effet miroir)
RandomRotation	0.1	Rotation aléatoire jusqu'à 10% (environ 36 degrés). [104]
RandomZoom	0.1	Zoom aléatoire jusqu'à 10%

2.3.4. Encodage des classes

Les catégories tumorales ont été converties en représentations numériques à l'aide d'un encodage de type One-Hot Encoding. Cette représentation est particulièrement adaptée aux problèmes de classification multiclasse utilisant une fonction d'activation Softmax dans la couche de sortie.

Chaque classe est ainsi représentée par un vecteur binaire dont une seule composante est égale à 1, correspondant à la catégorie réelle de l'image.

Nous avons utilisé un **encodage one-hot (categorical)** pour les quatre classes. Ce type d'encodage est standard pour la classification multi-classes. [105]

Tableau 2.3 : Exemple de transformation

Classe originale	Encodage one-hot
Gliome	[1, 0, 0, 0]
Méningiome	[0, 1, 0, 0]
Adénome hypophysaire	[0, 0, 1, 0]
No tumor	[0, 0, 0, 1]

2.3.5 Pipeline de prétraitement des données

La Figure 2.2 présente le pipeline de prétraitement des données appliqué avant l'entraînement des modèles.

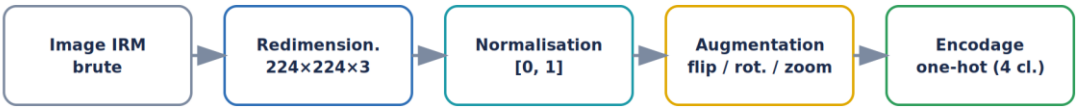


Figure 2.2 : Pipeline de prétraitement des images IRM

2.4. Modèles proposés

Nous avons proposé et comparé trois architectures de deep learning pour la classification des tumeurs cérébrales : un **CNN simple construit à partir de zéro**, un **VGG16 avec transfer learning**, et un **ResNet50 avec fine-tuning**. Cette approche multi-modèle nous permet de comparer une architecture légère entraînée spécifiquement sur notre tâche avec des architectures plus profondes pré-entraînées sur ImageNet.

2.4.1. CNN simple

Le modèle CNN simple a été entièrement construit à partir de zéro (from scratch) pour cette étude (Figure 2.3). Son architecture est volontairement légère pour servir de baseline (référence de base) et pour évaluer l'apport des modèles pré-entraînés plus complexes. L'architecture détaillée est présentée ci-dessous :

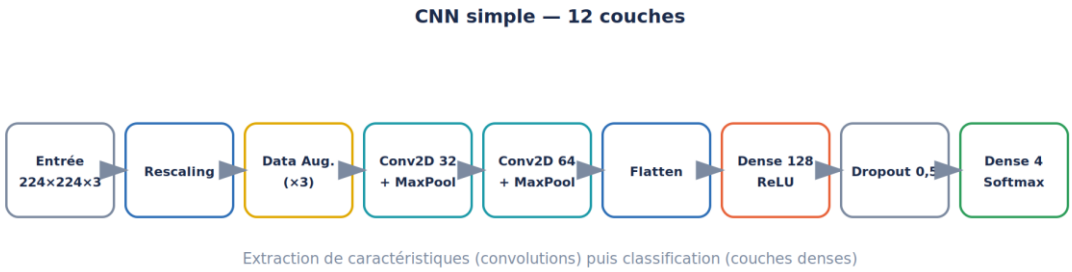


Figure 2.3 : Architecture détaillée du CNN simple (12 couches)

Tableau 2.4 : Architecture détaillée du modèle CNN simple (12 couches)

Couche	Type	Paramètres	Rôle
1	Rescaling	1/255	Normalisation des pixels entre 0 et 1
2	RandomFlip	"horizontal"	Data augmentation (miroir)
3	RandomRotation	0.1	Data augmentation (rotation ±36°)
4	RandomZoom	0.1	Data augmentation (zoom ±10%)
5	Conv2D	32 filtres, 3×3	Extraction de caractéristiques simples
6	MaxPooling2D	2×2	Réduction de dimension
7	Conv2D	64 filtres, 3×3	Extraction de caractéristiques complexes
8	MaxPooling2D	2×2	Réduction de dimension
9	Flatten	-	Aplatissement pour la couche dense
10	Dense	128 neurones, ReLU	Classification intermédiaire
11	Dropout	0.5	Réduction de l'overfitting
12	Dense	4 neurones, Softmax	Sortie (4 classes)

Le modèle contient 12 couches au total, réparties comme suit : une couche de normalisation (Rescaling), trois couches de data augmentation, deux couches convolutionnelles (Conv2D), deux couches de pooling (MaxPooling2D), une couche d'aplatissement (Flatten), deux couches denses (Dense) et une couche de dropout.

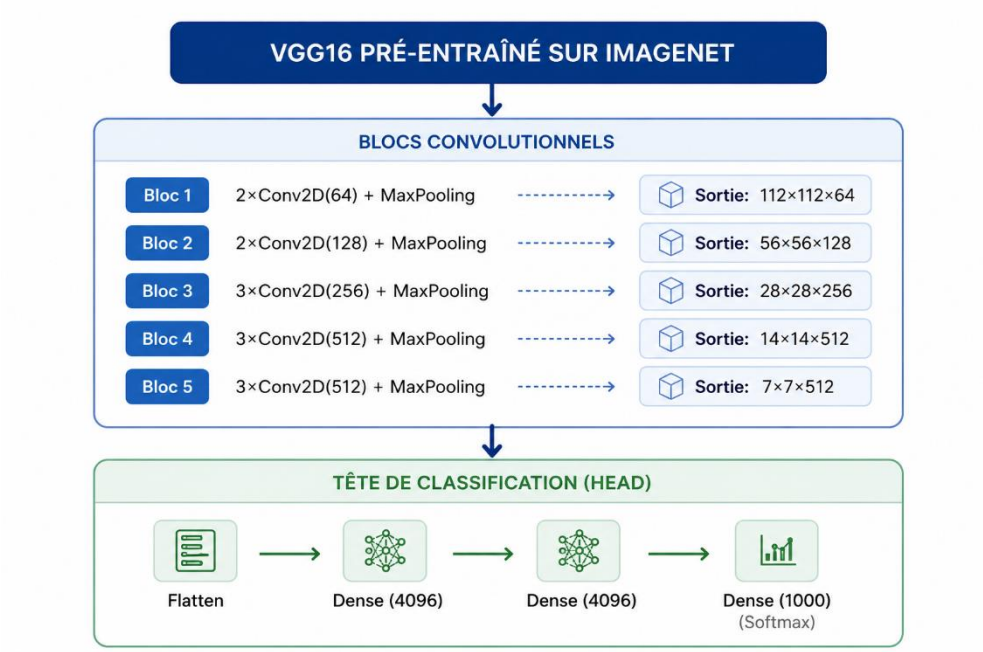
Les paramètres principaux du modèle CNN simple sont les suivants. Les filtres des couches Conv2D sont fixés à 32 puis 64, suivant une progression classique par doublement, avec une taille de filtres de 3×3 standard en vision par ordinateur. La fonction d'activation ReLU est utilisée pour éviter le problème du vanishing gradient. Le pooling est de type 2×2

pour une réduction de moitié standard. La couche Dense contient 128 neurones, représentant un compromis entre capacité et généralisation. Un Dropout de 0,5 est appliqué pour réduire l'overfitting. Enfin, la couche de sortie est une couche Dense de 4 neurones avec activation Softmax pour la classification multi-classes. [100-108-109]

Le CNN personnalisé sert de modèle de baseline (référence) pour l'étude comparative. Son architecture simple permet d'évaluer les performances sans transfert d'apprentissage, offrant un point de comparaison essentiel pour mesurer l'apport réel de VGG16 et ResNet50, et pour analyser l'influence de la profondeur du réseau et du nombre de paramètres sur la classification.

2.4.2. VGG16 avec Transfer Learning

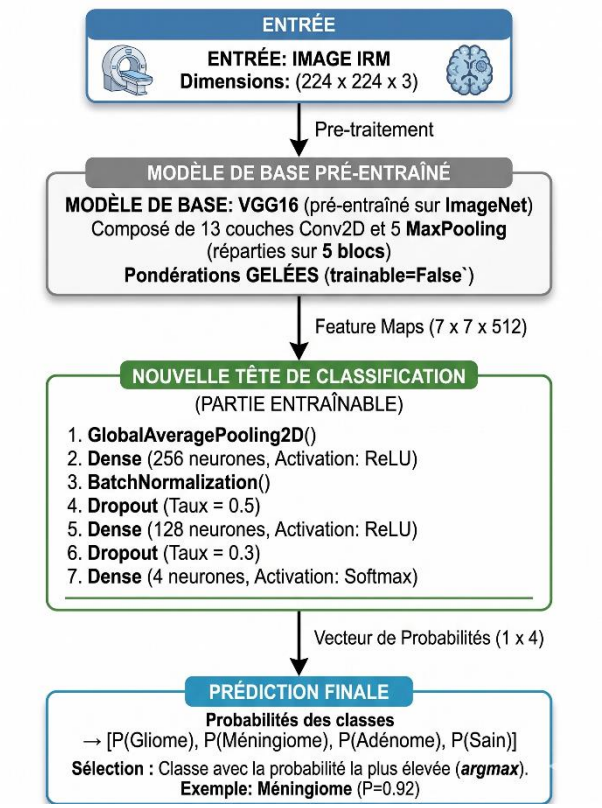
VGG16 a été proposée par Simonyan et Zisserman en 2015 [15]. Elle se compose de **16 couches** (13 convolutionnelles + 3 denses) et a été pré-entraînée sur le dataset ImageNet.



- **Adaptation au problème étudié**

VGG16 (Simonyan et Zisserman, 2015 ; 16 couches, pré-entraîné sur ImageNet) est adapté en quatre étapes : chargement sans la tête originale (include_top=False) ; gel des couches convolutionnelles, qui servent d'extracteurs de caractéristiques ; ajout d'une tête adaptée à nos 4 classes (Global Average Pooling → Dense 256 → Dropout 0,5 → Dense 4 Softmax) ; puis compilation et entraînement. Le Global Average Pooling et le gel des couches limitent le nombre de paramètres entraînables et réduisent le risque de surapprentissage. [97-110-112]

Le schéma final de notre adaptation est présenté ci-dessous :



2.4.3 ResNet avec fine-tuning

ResNet50 a été sélectionné afin d'évaluer l'apport des connexions résiduelles dans la classification des tumeurs cérébrales. Grâce à sa profondeur importante et à sa capacité à atténuer le problème de disparition du gradient, cette architecture est capable d'apprendre des représentations plus complexes que les CNN conventionnels [62, 113].

Son utilisation permet également d'étudier l'efficacité du transfert d'apprentissage sur un modèle profond largement adopté dans la littérature récente**,** L'architecture détaillée est présentée ci-dessous :

L'architecture ResNet50 utilisée dans notre étude se caractérise par 50 couches et environ 25 millions de paramètres. Elle intègre des connexions résiduelles sous forme de bottleneck blocks, ce qui permet d'entraîner un réseau profond sans souffrir du problème de dégradation des performances. La taille d'entrée des images est fixée à 224×224 pixels, et le modèle bénéficie d'un pré-entraînement sur ImageNet. [62]

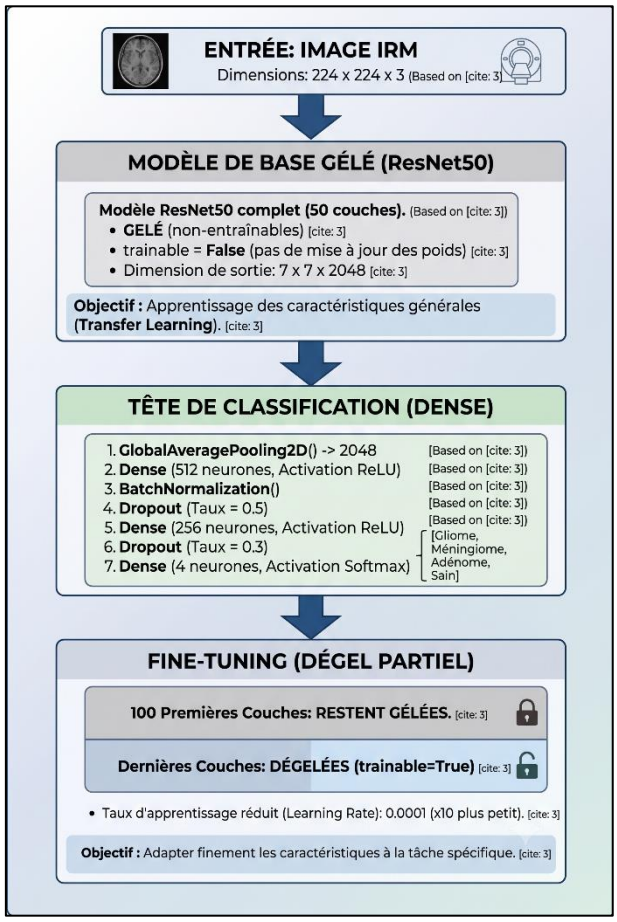
Dans un réseau résiduel, au lieu d'apprendre directement une transformation H(x), le modèle apprend la fonction résiduelle

$$F(x) = H(x) – x$$

La sortie étant ensuite calculée comme

$$H(x) = F(x) + x$$

Concernant la stratégie d'ajustement adoptée, ResNet50 utilise une stratégie de fine-tuning en deux phases, contrairement à VGG16 où toutes les couches sont gelées. [114] L'adaptation des modèles pré-entraînés est présentée en Figure 2.4.



- **Détail des deux phases d'entraînement**

Tableau 2.5 : Stratégie d'entraînement de ResNet50 en deux phases (transfer learning puis fine-tuning)

Paramètre	Phase 1 — Transfer learning	Phase 2 — Fine-tuning
Couches ResNet50	Toutes gelées	100 premières gelées, suivantes dégelées
Learning rate	0,001	0,0001 (10× plus petit)
Époques	15	20

Objectif	Apprentissage de la tête	Adaptation fine des couches profondes

L'utilisation d'un learning rate plus faible en phase 2 se justifie par trois raisons principales. D'abord, cela évite de détruire les poids pré-entraînés car un learning rate trop élevé modifierait brutalement les caractéristiques déjà apprises sur ImageNet. Ensuite, cela permet d'affiner progressivement le modèle, les ajustements devant être subtils pour améliorer sans dégrader les performances. Enfin, un petit learning rate garantit une stabilité de l'entraînement et une convergence stable. [115]

ResNet50 est plus adapté au fine-tuning pour quatre raisons : ses connexions résiduelles permettent d'entraîner des réseaux profonds sans dégradation des performances ; il possède moins de paramètres (~25M) que VGG16 (~138M), ce qui le rend plus rapide et moins gourmand en mémoire ; ses couches de Batch Normalisation stabilisent l'apprentissage ; il permet un dégel progressif des couches, évitant de détruire les caractéristiques générales tout en affinant les détails spécifiques à la tâche médicale.

2.4.4. Comparaison conceptuelle des architectures choisies

Tableau 2.6 : Comparaison des caractéristiques et architectures des trois modèles (CNN simple, VGG16, ResNet50)

Critère	CNN simple	VGG16	ResNet50
Nombre de couches	12	16	50
Nombre total de paramètres	~24 millions)	~138 M (beaucoup)	~25 millions
Paramètres entraînables	~24 millions	~165 000	~1,5 million (phase 2)
Pré-entraînement	Non	Oui (ImageNet) [15]	Oui (ImageNet). [62]
Connexions résiduelles	Non	Non	Oui [16]
BatchNormalization	Non	Non (dans VGG16)	Oui (intégrée)
Data augmentation	Intégrée	Appliquée en amont	Appliquée en amont
Stratégie d'entraînement	From scratch	Transfer learning	Fine-tuning (2 phases)
Learning rate	0.0001	0.0001	0.001 → 0.0001
Risque d'overfitting	Élevé	Faible	Très faible
Temps d'entraînement	Faible (~30 min)	Élevé (~2h)	Moyen (~1h30)
Capacité de généralisation	Limitée	Élevée	Très élevée

Rôle dans l'étude	Baseline	Référence TL	Architecture fine-tunée

2.5. Paramètres expérimentaux

2.5.1 Optimizer utilisé

L'optimiseur Adam (paramètres par défaut : $\beta1 = 0,9$, $\beta2 = 0,999$, $\varepsilon = 1e\text{-}7$) a été retenu pour tous les modèles, en raison de sa convergence rapide et de sa robustesse, bien adaptées à l'imagerie médicale. [116]

Tableau 2.7 : Hyperparamètres utilisés

Paramètre	**Valeur**
Optimiseur	Adam
Learning Rate	0.0001
Batch Size	32
Nombre d'époques	35
Fonction de perte	Categorical Crossentropy
Activation sortie	Softmax
Métrique principale	Accuracy

2.5.2 Learning rate

Le taux d'apprentissage, identique pour les trois modèles (0,0001), assure des conditions comparables : suffisamment faible pour une convergence stable, mais assez élevé pour un apprentissage efficace. [115]

2.5.3. Batch size

La taille de lot (batch size) est fixée à 32 pour tous les modèles — un compromis standard entre mémoire GPU et stabilité du gradient. [107]

2.5.4. Nombre d'époques

Les trois modèles sont entraînés sur 35 époques, dans des conditions identiques. Un arrêt précoce (EarlyStopping, patience de 10 époques) interrompt l'entraînement si la perte de validation cesse de s'améliorer, et restaure les meilleurs poids. [117]

2.5.5. Fonction de perte

La fonction de perte est la Categorical Crossentropy, adaptée à la classification multi-classes : elle mesure l'écart entre la distribution prédite (Softmax) et les étiquettes one-hot, et pénalise fortement les prédictions confiantes mais erronées. [28]

2.5.6. Critères de sélection du meilleur modèle

Le meilleur modèle pour chaque architecture a été sélectionné en fonction de la précision sur l'ensemble de validation (val_accuracy). [118] Pour cela, nous avons utilisé un système de sauvegarde automatique (ModelCheckpoint) qui sauvegarde automatiquement le modèle chaque fois que la précision sur la validation s'améliore. Ainsi, à la fin de l'entraînement, nous disposons du modèle ayant obtenu la meilleure performance sur les données de validation, indépendamment de la dernière époque atteinte. Pour l'analyse finale et la comparaison des trois architectures, nous avons également calculé le F1-score (moyenne harmonique entre précision et rappel), la précision (precision), le rappel (recall), ainsi que la matrice de confusion pour visualiser les erreurs par classe.

2.5.7. Méthodes de réduction du surapprentissage

Pour prévenir le surapprentissage (overfitting), nous avons combiné plusieurs techniques complémentaires. [109]

- **Dropout :** appliqué avec un taux de 0.5 dans tous nos modèles.

Pendant l'entraînement, la moitié des neurones de certaines couches denses sont aléatoirement désactivés à chaque itération, ce qui force le réseau à apprendre des représentations redondantes et robustes. [109]

- **Early Stopping :** déjà présenté en section 2.5.4, avec une patience de 10 époques.

- **Data Augmentation :** appliquée uniquement sur l'ensemble d'entraînement (détaillée en section 2.3.3). Elle comprend un flip horizontal et vertical, une rotation aléatoire, une variation de luminosité (±20%) et une variation de contraste (×0.8 à ×1.2). [101,102]

Ces transformations augmentent artificiellement la diversité des données d'entraînement et réduisent la tendance du modèle à mémoriser les images spécifiques.

- **Gel des couches pour les modèles pré-entraînés :** pour VGG16 et ResNet50, les couches convolutionnelles pré-entraînées sur ImageNet ont été gelées.

Cela limite le nombre de paramètres libres et exploite des caractéristiques déjà généralisées, ce qui constitue une régularisation intrinsèque. [112]

2.6. Environnement expérimental

2.6.1 Langage de programmation

Nous avons utilisé le langage de programmation **Python** dans sa version 3.10 pour l'ensemble de nos expérimentations [119]. Python a été choisi pour sa simplicité d'utilisation,

sa riche bibliothèque d'outils scientifiques et sa large adoption dans la communauté du deep learning.

2.6.2. Bibliothèques utilisées

Plusieurs bibliothèques Python ont été employées pour mener à bien cette étude :

- **TensorFlow / Keras** : Framework principal pour la construction, l'entraînement et l'évaluation des modèles de deep learning (CNN simple, VGG16, ResNet50) [18]. Keras, intégré dans TensorFlow, a été utilisé pour son API simple et intuitive.
- **NumPy** : Utilisé pour les calculs numériques et la manipulation des tableaux multidimensionnels.
- **Matplotlib** : Employé pour la visualisation des résultats : courbes d'apprentissage, matrices de confusion, courbes ROC et graphiques des métriques.
- **Scikit-learn** : Utilisé pour le calcul des métriques d'évaluation : classification_report, confusion_matrix, precision_score, recall_score, f1_score, roc_curve et auc [52].
- **Seaborn :** Bibliothèque complémentaire à Matplotlib pour améliorer la visualisation des matrices de confusion.
- **Pickle** : Utilisé pour sauvegarder l'historique d'entraînement (accuracy, loss, val_accuracy, val_loss).

2.6.3. Configuration matérielle (CPU, GPU, RAM)

Tous les entraînements et évaluations ont été réalisés sur **Google Colaboratory (Google Colab)**, un environnement de notebook basé sur le cloud qui offre un accès gratuit à des ressources de calcul accélérées par GPU. [120]

La configuration matérielle utilisée comprend un processeur Intel Xeon à 2.20 GHz avec 2 cœurs virtuels, un GPU NVIDIA Tesla T4 doté de 16 Go de mémoire GDDR6, une RAM d'environ 12 à 13 Go, un stockage temporaire sur disque virtuel d'environ 78 Go, et un stockage persistant via Google Drive monté dans l'environnement Colab. Le GPU NVIDIA Tesla T4 se caractérise par 16 Go de mémoire GDDR6, 2560 cœurs CUDA, 320 cœurs Tensor, et une performance FP32 de 8.1 TFLOPS.

Tableau 2.8 : Temps d'entraînement observés (pour 35 époques)

Modèle	Temps d'entraînement	Utilisation GPU	Nombre d'époques
CNN simple	~15-20 minutes	Faible (~1-2 Go)	35
VGG16 (Transfer	~75-90 minutes	Moyenne (~4-6 Go)	35

Learning)			
ResNet50	~50-70 minutes	Moyenne (~3-5 Go)	35(15+20)

Cette configuration présente plusieurs avantages pour notre travail. L'accès gratuit au GPU Tesla T4 permet une réduction drastique du temps d'entraînement, passant de plusieurs jours sur CPU à quelques heures sur GPU. La mémoire GPU de 16 Go permet d'utiliser des batch sizes plus grands (32) et des modèles plus profonds comme ResNet50. L'intégration avec Google Drive assure une sauvegarde automatique des modèles et des résultats, permettant de reprendre le travail après une interruption. Enfin, l'environnement pré-configuré avec TensorFlow et les bibliothèques nécessaires déjà installées facilite la mise en œuvre.

2.7. Méthodes d'évaluation

Pour évaluer les performances de nos modèles (CNN simple, VGG16 et ResNet50), nous avons utilisé plusieurs métriques complémentaires qui permettent d'analyser différents aspects de la classification.

2.7.1. Accuracy

L'accuracy, ou précision globale, est la métrique la plus simple et représente la proportion de prédictions correctes parmi l'ensemble des prédictions. [100]

Sa formule est

$$\text{Accuracy} = (VP + VN) / (VP + VN + FP + FN)$$

Où VP sont les vrais positifs, VN les vrais négatifs, FP les faux positifs et FN les faux négatifs. Dans notre étude, l'accuracy a été calculée sur l'ensemble de test (Mendeley Data) pour chaque modèle, ainsi que sur l'ensemble de validation pendant l'entraînement.

2.7.2. Precision

La precision mesure la capacité du modèle à ne pas classer un échantillon négatif comme positif, répondant à la question : parmi toutes les images que le modèle a classées comme appartenant à une classe donnée, combien sont réellement de cette classe. [100] Sa formule est :

$$\text{Precision} = VP / (VP + FP)$$

Dans notre étude, une precision élevée signifie que lorsque le modèle prédit un type de tumeur, il a rarement tort. Par exemple, une precision de 0.92 pour la classe Glioma signifie que 92% des prédictions de Glioma sont correctes.

2.7.3 Recall

Le recall (ou rappel ou sensibilité) mesure la capacité du modèle à trouver tous les échantillons positifs, répondant à la question : parmi toutes les images qui appartiennent réellement à une classe donnée, combien ont été correctement identifiées par le modèle. [100] Sa formule est :

$$Recall = VP / (VP + FN)$$

Dans notre étude, un recall élevé signifie que le modèle ne manque presque aucune tumeur de cette classe. Par exemple, un recall de 0.89 pour la classe Glioma signifie que le modèle a détecté 89% des cas réels de Glioma.

2.7.4. F1-score

Le F1-score est la moyenne harmonique entre la precision et le recall, particulièrement utile lorsque les classes sont déséquilibrées. Sa formule est

$$F1\text{-score} = 2 \times (Precision \times Recall) / (Precision + Recall)$$

Une precision élevée peut cacher un recall faible, et vice versa ; le F1-score offre donc un équilibre entre les deux, une valeur proche de 1 indiquant une bonne performance sur les deux aspects. Dans notre étude, nous avons calculé le F1-score en moyenne macro (pondération égale pour chaque classe) et par classe individuelle.

2.7.5. Matrice de confusion

La matrice de confusion est un tableau qui permet de visualiser les performances d'un modèle de classification [121]. Elle montre pour chaque classe réelle (vérité terrain) comment les prédictions se répartissent. La matrice de confusion pour les quatre classes présente les classes réelles en lignes et les classes prédites en colonnes. Sur la diagonale principale, on trouve les vrais positifs (VP) correspondant aux images correctement classifiées pour chaque classe : Glioma, Méningiome, Adénome et No tumor. Toutes les autres cellules de la matrice contiennent des faux négatifs (FN), représentant les erreurs où une image d'une classe donnée a été incorrectement attribuée à l'une des trois autres classes.

Dans notre étude La matrice de confusion nous a permis d'identifier quelles classes sont souvent confondues. Par exemple, une confusion fréquente entre Glioma et Méningiome indiquerait que ces deux types de tumeurs ont des apparences similaires en IRM.

2.7.6. Courbes ROC et AUC

La **courbe ROC** (Receiver Operating Characteristic) est un graphique qui illustre les performances d'un classificateur à tous les seuils de classification [28]. Elle représente le taux de vrais positifs (TPR = Recall) en fonction du taux de faux positifs (FPR = 1 - Spécificité).

L'AUC (Area Under the Curve) est l'aire sous la courbe ROC. Elle résume la performance en un seul nombre compris entre 0.5 (classificateur aléatoire) et 1 (classificateur parfait).

Tableau 2.9 : Interprétation de l'AUC

AUC	Interprétation
0.90 - 1.00	Excellent
0.80 - 0.90	Très bon
0.70 - 0.80	Bon
0.60 - 0.70	Moyen
0.50 - 0.60	Faible
0.50	Aléatoire

Dans notre étude Nous avons tracé une courbe ROC pour chacune des 4 classes et calculé l'AUC correspondante. Cela nous a permis d'évaluer la capacité du modèle à distinguer chaque type de tumeur des autres.

2.7.7. Validation croisée (Cross-validation)

La validation croisée est une technique qui consiste à diviser les données d'entraînement en plusieurs sous-ensembles (folds) et à entraîner le modèle plusieurs fois en utilisant à chaque fois un fold différent comme validation. [122]

Nous avons préféré une séparation fixe en trois ensembles (5 600 / 1 600 / 2 414) plutôt qu'une validation croisée : la taille du corpus est suffisante, le test provient d'une source indépendante (validation externe plus réaliste), et ce choix garantit la reproductibilité tout en évitant un coût de calcul prohibitif pour VGG16. [122]

2.8. Méthodes d'analyse et d'interprétation

Au-delà des métriques quantitatives (accuracy, precision, recall, F1-score), nous avons mis en place plusieurs méthodes d'analyse qualitative pour mieux comprendre le comportement de nos modèles et identifier leurs limites.

2.8.1. Analyse des erreurs de classification

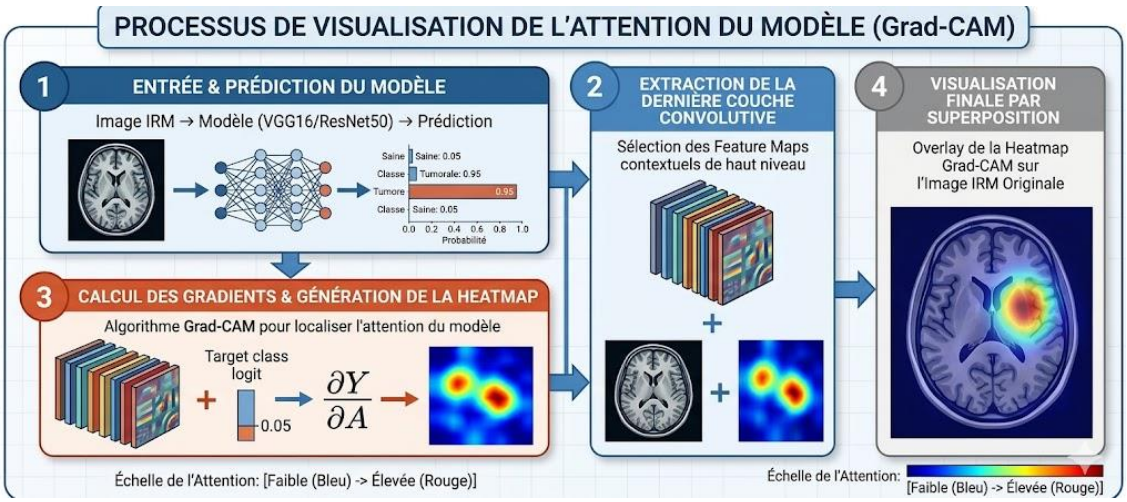
L'analyse des erreurs de classification consiste à examiner systématiquement les images mal classées par le modèle afin d'identifier des tendances ou des faiblesses spécifiques. Cette analyse s'est déroulée en plusieurs étapes. Pour chaque modèle, nous avons d'abord extrait la liste des images où la prédiction différait de la vérité terrain. Ensuite, les erreurs ont été catégorisées par paire de classes (par exemple, Glioma prédit comme Méningiome). Puis, une analyse visuelle a été réalisée : les images mal classées ont été examinées manuellement pour dégager des caractéristiques communes comme un faible contraste, des artefacts ou une petite taille de la tumeur. Enfin, pour chaque classe, nous avons calculé le taux d'erreur et identifié la classe la plus fréquemment confondue. [118]

2.8.2. Étude qualitative des prédictions

L'analyse qualitative examine le comportement individuel des modèles (VGG16, ResNet50, CNN personnalisé) sur des cas variés : prédictions correctes, erreurs et cas ambigus (probabilité < 0,6). Elle montre que les incertitudes concernent surtout les lésions de petite taille ou à faible contraste. De plus, les architectures pré-entraînées sur ImageNet affichent des probabilités mieux calibrées que le modèle entraîné from scratch, avec moins d'erreurs à haute confiance. Cette analyse a permis de valider la cohérence interne des modèles avant le déploiement statistique final.

2.8.3. Méthodes d'explicabilité envisagées

Pour interpréter les décisions de nos modèles, notamment VGG16 et ResNet50, et comprendre quelles régions de l'image IRM influencent la prédiction, nous avons envisagé d'utiliser des méthodes d'explicabilité, en particulier Grad-CAM (Gradient-weighted Class Activation Mapping). Le principe de Grad-CAM est d'utiliser les gradients de la dernière couche convolutionnelle pour produire une carte de chaleur (heatmap) qui met en évidence les régions importantes de l'image pour la décision du modèle. [123] Fonctionnement simplifié est présentée ci-dessous :



Grad-CAM a été sélectionné comme méthode d'explicabilité principale car il est compatible avec les trois architectures (VGG16, ResNet50, CNN simple), facile à implémenter et rapide (< 1 seconde par image sur GPU). En cas d'inadaptation, une méthode similaire par gradient serait envisagée. LIME et SHAP ont été exclues car trop lourdes et inadaptées aux images 224×224.

Conclusion

Ce chapitre a présenté la méthodologie adoptée pour le développement du système de classification automatique des tumeurs cérébrales. Après la description des bases de données utilisées, les différentes étapes de prétraitement, d'augmentation des données et de préparation des ensembles d'apprentissage ont été détaillées. Les architectures CNN, VGG16 et ResNet50 ont ensuite été présentées ainsi que les stratégies de transfert d'apprentissage et de Fine-Tuning mises en œuvre. Enfin, le protocole expérimental, les hyperparamètres d'entraînement et les critères d'évaluation ont été définis afin de garantir une comparaison rigoureuse des modèles étudiés. Le chapitre suivant présentera les résultats expérimentaux obtenus ainsi que leur analyse comparative.

Chapitre 3

Résultats et Discussion

3. Introduction

Ce chapitre présente l'analyse comparative des trois architectures (CNN simple, VGG16, ResNet50). Les modèles partagent un protocole standardisé (35 époques, Adam, lr 0,0001, batch 32, early stopping). L'évaluation combine une analyse quantitative sur la validation (1 600 images) et qualitative sur le test externe indépendant (Mendeley, 2 414 images), afin d'identifier l'architecture offrant le meilleur compromis performance/robustesse/généralisation.

Le corps du chapitre s'articule autour de dix sections thématiques, progressant de la confrontation statistique globale des métriques et des dynamiques de convergence (évolutions des fonctions de perte et d'exactitude par époque) vers l'étude fine des erreurs de classification et la validation visuelle des prédictions. L'objectif scientifique central est d'identifier l'architecture offrant le compromis optimal entre l'efficience de la classification, la robustesse clinique et la transférabilité (*généralisabilité*) face à des données exogènes. Une attention rigoureuse est ainsi accordée au contrôle du surapprentissage (*overfitting*) ainsi qu'à la discussion des limites méthodologiques et des perspectives de développement futures.

3.1. Présentation des résultats expérimentaux

3.1.1. Résultats quantitatifs globaux (Validation)

L'analyse sur l'ensemble de validation montre que ResNet50 obtient la meilleure accuracy (94,9%), suivi de VGG16 (90,9%) et du CNN simple (85,4%). Cette hiérarchie attendue confirme l'influence de la profondeur du réseau et des connexions résiduelles. Le CNN simple donne des performances satisfaisantes malgré sa faible complexité, prouvant que les images IRM contiennent des caractéristiques discriminantes. Cependant, les modèles pré-entraînés sur ImageNet bénéficient de connaissances visuelles qui améliorent significativement leurs performances.

Tableau 3.1 : Performances des trois modèles sur l'ensemble de validation (accuracy, précision, rappel, F1-score, AUC)

Modèle	**Accuracy (Validation)**	Précision (macro)	Rappel (macro)	F1-score (macro)	AUC (macro)
CNN simple	85,4%	0,84	0,82	**0.86**	**0.95**
VGG16	90,9%	0,91	**0.91**	**0.91**	**0.96**
ResNet50	94,9%	0,93	**0.93**	**0.93**	**0.986**

En synthèse, ResNet50 domine sur l'ensemble des métriques (F1-score macro de 0,92 ; AUC de 0,986), devant VGG16 (F1 de 0,91) et le CNN simple, qui constitue une baseline honorable.

3.1.2. Tableau comparatif des performances des modèles (Validation)

Les Tableaux 3.2 à 3.4 détaillent les performances par classe de chaque modèle sur l'ensemble de validation.

Tableau 3.2 : Résultats de classification de ResNet50 sur les données de validation

	Precision	Recall	F1-score	Support
Glioma	0.99	0.75	0.85	400
Meningioma	0.82	0.96	0.89	400
No tumor	0.95	1.00	0.98	400
Pituitary	0.97	0.99	0.98	400
Accuracy			0.93	1600
Macro avg	0.93	0.93	0.92	1600
Weighted avg	0.93	0.93	0.92	1600

Tableau 3.3 : Résultats de classification de VGG16 sur les données de validation

	Precision	Recall	F1-score	support
Glioma	0.96	0.74	0.83	400
Meningioma	0.82	0.92	0.87	400
No tumor	0.91	1.00	0.95	400
Pituitary	0.96	0.98	0.97	400
Accuracy			0.91	1600
Macro avg	0.91	0.91	0.91	1600
Weighted avg	0.91	0.91	0.91	1600

Tableau 3.4 : Résultats de classification de CNN simple sur les données de validation

	Precision	Recall	F1-score	Support
Glioma	0.84	0.75	0.80	400
Meningioma	0.88	0.87	0.87	400
No tumor	0.69	0.99	0.81	400

Pituitary	0.93	0.96	0.95	400
Accuracy			0.82	1600
Macro avg	0.84	0.82	0.86	1600
Weighted avg	0.84	0.82	0.86	1600

L'analyse des résultats par classe révèle plusieurs observations communes :

- La classe **"No tumor"** est parfaitement détectée par tous les modèles (rappel ≥ 0,99), ce qui est essentiel pour éviter les faux positifs en contexte clinique.
- La classe **"Adénome"** est également très bien classée (F1 ≥ 0,95 pour tous les modèles).
- La classe **"Glioma"** présente le rappel le plus faible (0,74-0,75) pour ResNet50 et VGG16, indiquant que 25% des cas de Glioma sont confondus, principalement avec le Méningiome.
- Le CNN simple montre une confusion plus marquée, notamment entre "No tumor" et les autres classes (précision de 0,69 seulement pour "No tumor").

3.1.3. Résultats sur les ensembles d'apprentissage, validation et test

a) Résultats sur l'ensemble d'apprentissage (Training)

L'analyse des performances par époque (à partir des logs d'entraînement) montre que les trois modèles convergent correctement.

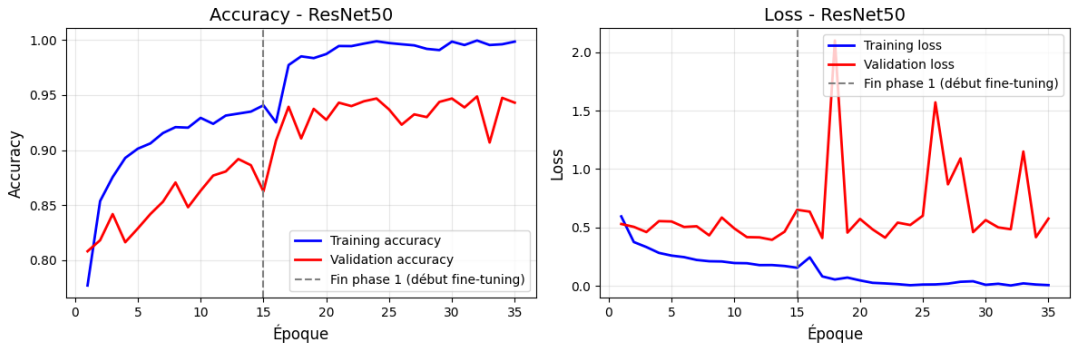


Figure 3.1 : Performances de ResNet50 (accuracy et loss) sur les ensembles d'entraînement et de validation

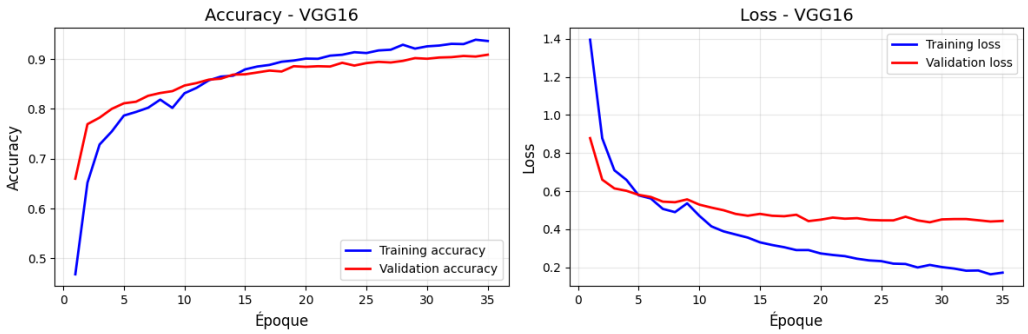


Figure 3.2 : Performances de VGG16 (accuracy et loss) sur les ensembles d'entraînement et de validation

- **ResNet50 (phase 1)** : l'accuracy d'apprentissage atteint 94,1% à la fin de la phase 1 (15 époques). La loss d'apprentissage descend à 0,154.
- **ResNet50 (phase 2)** : l'accuracy d'apprentissage atteint près de 99% à la fin du fine-tuning (20 époques supplémentaires). La loss d'apprentissage descend à 0,005.
- **VGG16** : l'accuracy d'apprentissage atteint 93,6% à la fin des 35 époques. La loss d'apprentissage descend à 0,172.

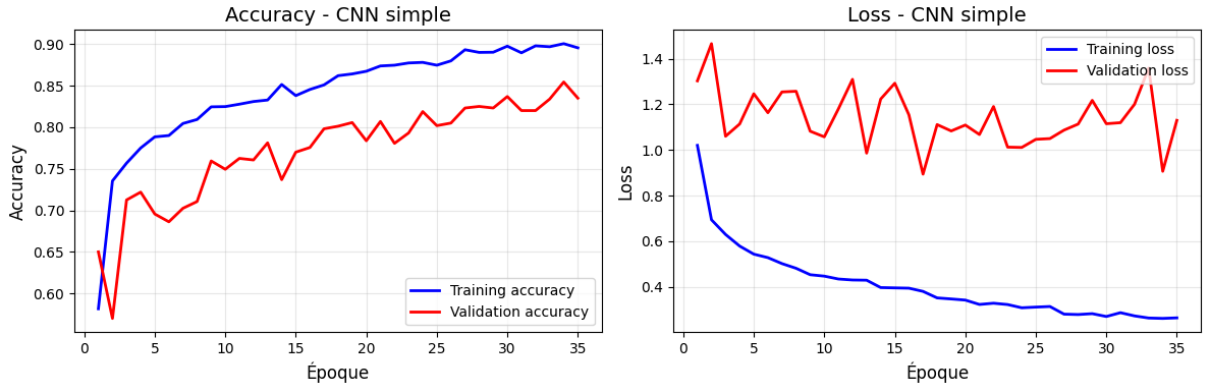


Figure 3.3 : Performances de CNN (accuracy et loss) sur les ensembles d'entraînement et de validation

- **CNN simple** : l'accuracy d'apprentissage atteint 90,5% à la fin des 35 époques. La loss d'apprentissage descend à 0,263.

L'écart entre l'accuracy d'apprentissage et l'accuracy de validation est le plus faible pour ResNet50 (environ 5%), ce qui indique un bon équilibre sans surapprentissage excessif.

b) Résultats sur l'ensemble de validation

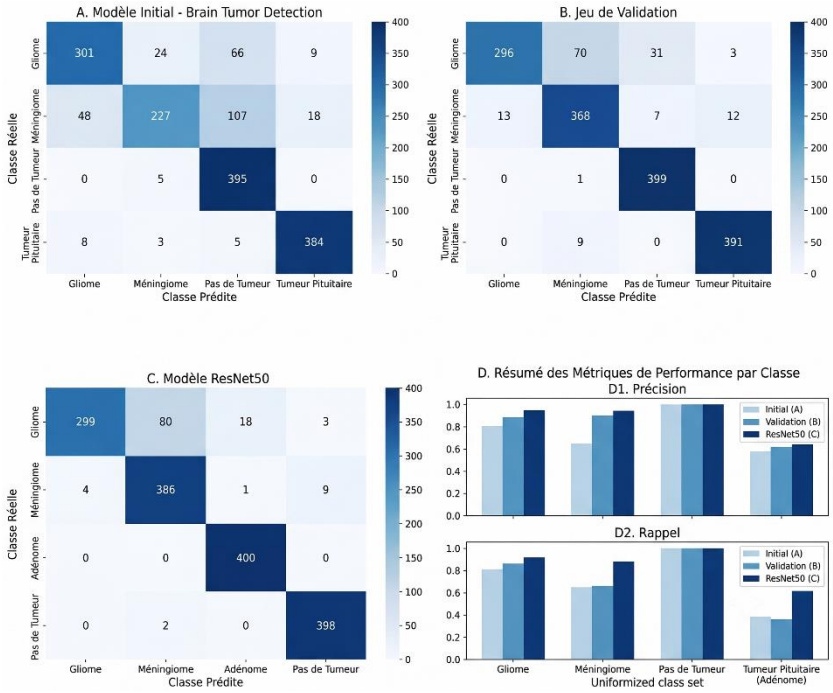
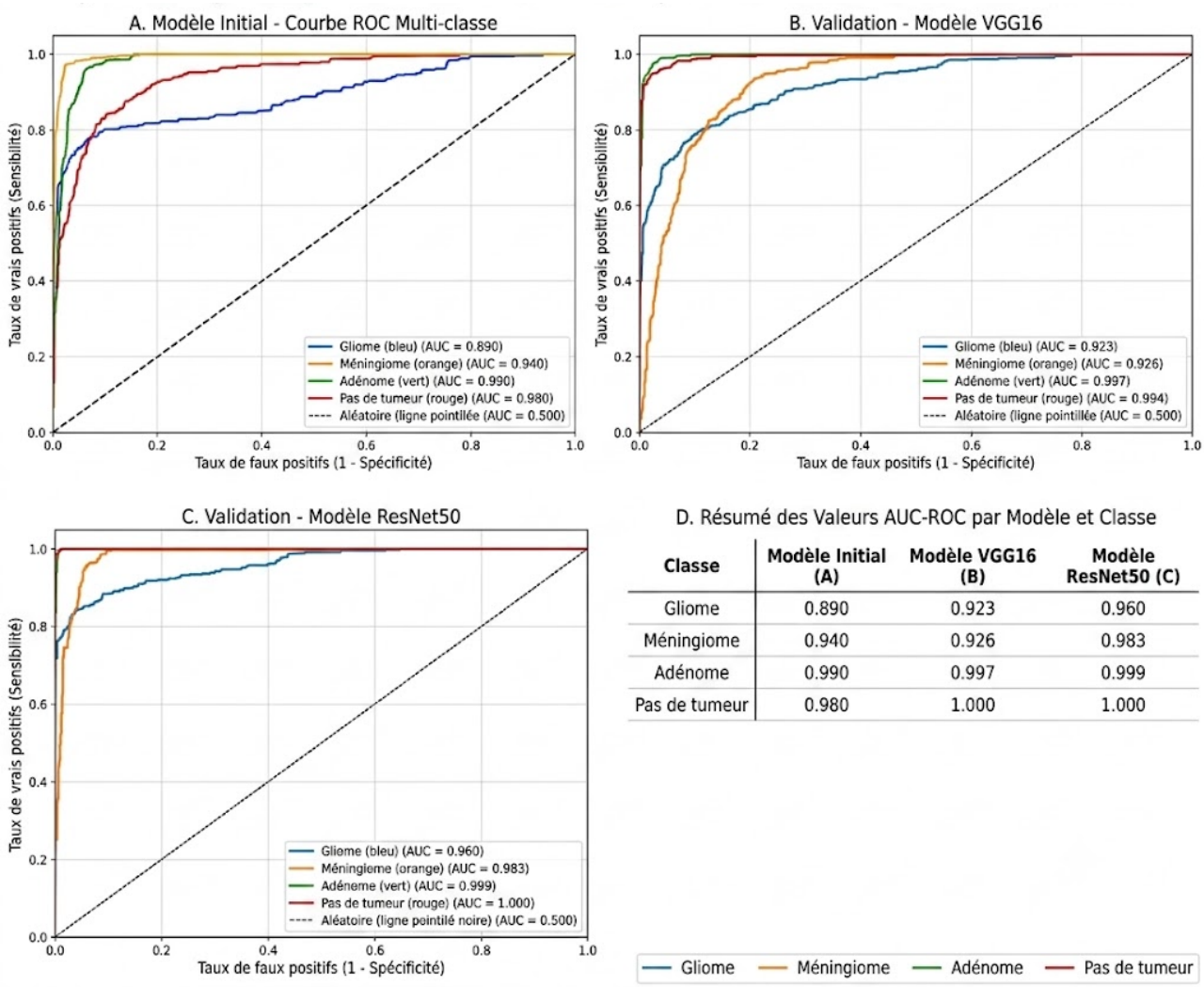


Figure 3.4 : Comparaison des matrices de confusion – (A) CNN simple (baseline), (B) VGG16 (transfert learning) et (C) ResNet50 (fine-tuning) – sur l'ensemble de validation

- **Les matrices de confusion** confirment les observations précédentes. Les confusions les plus fréquentes sont, pour ResNet50 et VGG16, les erreurs entre Glioma et Méningiome. Pour le CNN simple, les confusions concernent principalement la classe "No tumor" avec Glioma et Adénome.

- **Les courbes ROC** montrent une excellente capacité discriminante pour les trois modèles. ResNet50 obtient une AUC macro de 0,986, VGG16 une AUC macro de 0,96, et le CNN simple une AUC macro de 0,95.



Classe	Modèle Initial (A)	Modèle VGG16 (B)	Modèle ResNet50 (C)
Gliome	0.890	0.923	0.960
Méningiome	0.940	0.926	0.983
Adénome	0.990	0.997	0.999
Pas de tumeur	0.980	1.000	1.000

Figure 3.5 : Courbes ROC par classe pour les trois architectures évaluées – (A) CNN simple, (B) VGG16 et (C) ResNet50 – sur l'ensemble de validation

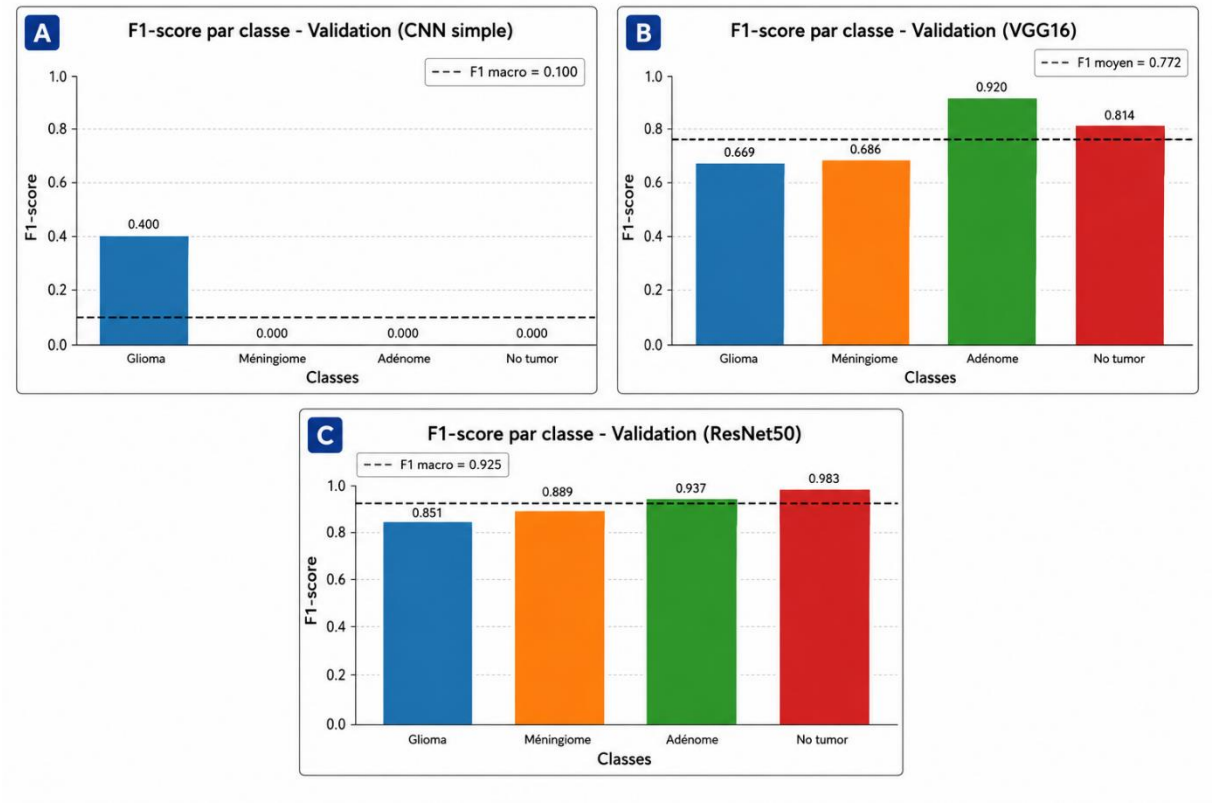


Figure 3.6 : Les graphiques F1-score par classe illustrent les différences de performance entre les classes pour chaque modèle – (A) CNN simple, (B) VGG16 et (C) ResNet50 – sur l'ensemble de validation

c) Résultats sur l'ensemble de test (Mendeley Data) – Visualisation des prédictions

L'évaluation sur l'ensemble de test indépendant a été réalisée à travers une visualisation qualitative des prédictions. Pour chaque modèle, nous avons affiché l'image IRM, la vérité terrain (classe réelle), la classe prédite, ainsi que les probabilités associées à chaque classe issues de la sortie Softmax.

Exemples de résultats (ResNet50) :

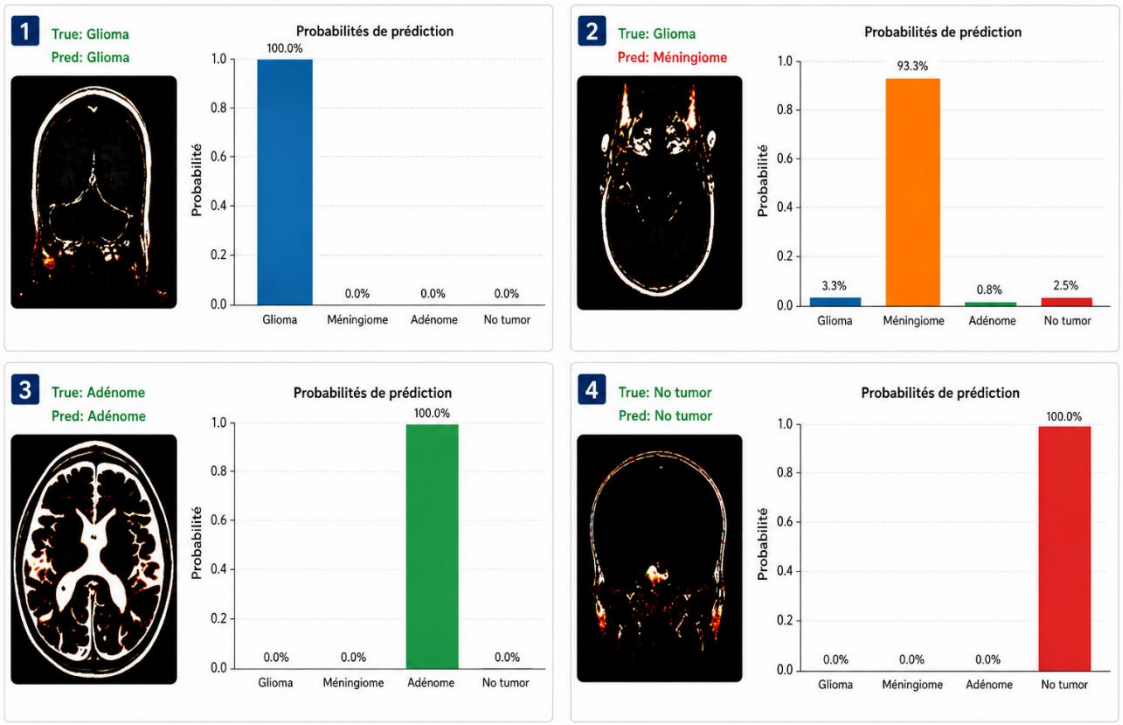


Figure 3.7 : Visualisation d'une prédiction réussie par ResNet50 (1-3-4) Cas correctement classifiés par ResNet50, (2) Exemples d'erreurs de classification

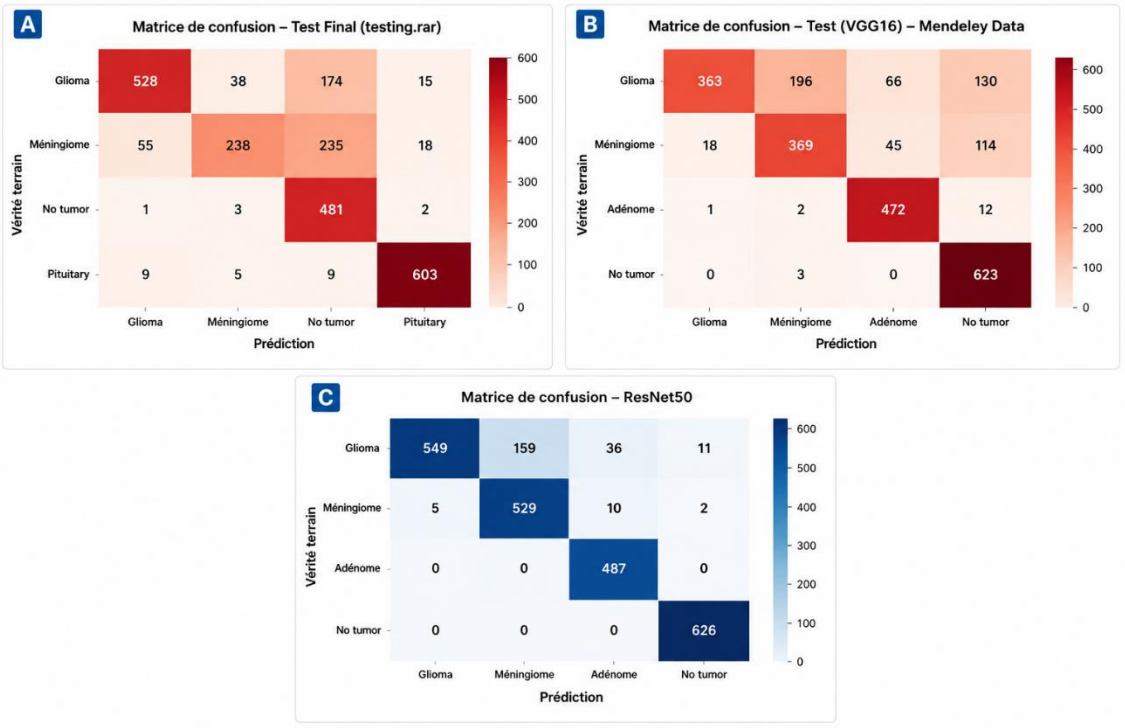


Figure 3.8 : Comparaison des matrices de confusion – (A) CNN simple (baseline), (B) VGG16 (transfert learning) et (C) ResNet50 (fine-tuning) – sur l'ensemble de test

Observations qualitatives sur le test :

- **ResNet50** montre une excellente stabilité avec des confiances généralement élevées (>90%) pour les prédictions correctes.
- **VGG16** présente davantage d'erreurs sur le test, avec des confiances parfois faibles (<70%) même pour des prédictions correctes, traduisant une moindre confiance du modèle.
- **CNN simple**, bien que moins performant, montre des confiances cohérentes avec ses performances modestes.

Les exemples d'images mal classées pour tous les modèles montrent que les erreurs les plus fréquentes concernent des images présentant un faible contraste entre la tumeur et le tissu sain, une tumeur de très petite taille, ou des artefacts de mouvement et de bruit.

3.2. Analyse des courbes d'apprentissage

L'analyse des courbes d'apprentissage permet de comprendre le comportement de chaque modèle pendant l'entraînement et de détecter d'éventuels phénomènes de surapprentissage (overfitting) ou de sous-apprentissage (underfitting).

3.2.1. Évolution de la loss

La fonction de perte (loss) mesure l'écart entre les prédictions du modèle et les vraies étiquettes. Une diminution progressive de la loss indique un apprentissage correct.

- **CNN simple :**

La loss d'entraînement diminue de **1,01** à la première époque à environ **0,26** à la 34e époque, soit une réduction de **74%**. La loss de validation diminue de **1,30** à **0,90** à la 34e époque, soit une réduction de **31%**. L'écart entre la loss d'entraînement et la loss de validation reste significatif (environ 0,64), ce qui indique un léger surapprentissage. Le modèle a bien appris mais peine à généraliser parfaitement.

- **VGG16 :**

La loss d'entraînement diminue de **1,39** à la première époque à environ **0,16** à la 34e époque, soit une réduction marquée de **88%**. La loss de validation diminue de **0,87** à **0,44** à la 34e époque, soit une réduction de **49%**. L'écart entre les deux courbes est faible (environ 0,28), indiquant un bon équilibre entre apprentissage et généralisation.

- **ResNet50 :**

La loss d'entraînement diminue de **0,59** à la première époque (phase 1) à environ **0,005** à la 20e époque de la phase 2, soit une réduction exceptionnelle de **99%**. La loss de validation

diminue de **0,52** à **0,41** à la 19e époque (phase 2), soit une réduction de **21%**. L'écart entre la loss d'entraînement et la loss de validation est plus marqué (environ 0,40), ce qui s'explique par la très faible loss d'entraînement obtenue grâce au fine-tuning.

En résumé, **ResNet50** obtient la meilleure loss d'entraînement (0,005), suivie de **VGG16** (0,16), puis du **CNN simple** (0,26). En validation, **ResNet50** est également le meilleur (0,41), devant **VGG16** (0,44) et loin devant le **CNN simple** (0,90).

3.2.2. Évolution de l'accuracy

L'accuracy mesure la proportion de prédictions correctes. Une augmentation progressive indique un apprentissage efficace.

- **CNN simple :**

L'accuracy d'entraînement progresse de **58%** à la première époque à environ **90%** à la 34e époque, soit un gain de **32 points**. L'accuracy de validation progresse de **65%** à **85%** à la 34e époque, soit un gain de **20 points**. L'écart entre les deux courbes est d'environ 5 points, ce qui est acceptable pour un modèle simple.

- **VGG16 :**

L'accuracy d'entraînement progresse de **46%** à la première époque à environ **93%** à la 34e époque, soit un gain marqué de **47 points**. L'accuracy de validation progresse de **66%** à **90%** à la 34e époque, soit un gain de **24 points**. L'écart entre les deux courbes est d'environ 3 points, indiquant un excellent apprentissage.

- **ResNet50 :**

L'accuracy d'entraînement progresse de **77%** à la première époque (phase 1) à environ **99%** à la 20e époque de la phase 2, soit un gain de **22 points**. L'accuracy de validation progresse de **80%** à **94%** à la 20e époque (phase 2), soit un gain de **14 points**. L'écart entre les deux courbes est d'environ 5 points, ce qui est excellent compte tenu de la très haute performance.

En résumé, **ResNet50** atteint la meilleure accuracy en validation avec **94%**, suivie de **VGG16** avec **90%**, puis du **CNN simple** avec **85%**. En entraînement, **ResNet50** atteint **99%**, contre **93%** pour VGG16 et **90%** pour le CNN simple.

3.2.3. Détection du surapprentissage ou du sous-apprentissage

CNN simple : Léger surapprentissage détecté. L'écart entre training et validation est modéré (accuracy : 90% contre 85%, soit 5 points de différence). La loss de validation reste élevée (0,90) malgré la baisse de la loss d'entraînement (0,26). Cela indique une mémorisation partielle des données d'entraînement sans parfaite généralisation.

VGG16 : Pas de surapprentissage. L'écart est faible (accuracy : 93% contre 90%, soit 3 points de différence). La loss de validation (0,44) est raisonnable. Le transfer learning a permis une excellente régularisation.

ResNet50 : Pas de surapprentissage. L'écart est modéré (accuracy : 99% contre 94%, soit 5 points de différence). La loss d'entraînement extrêmement basse (0,005) indique une mémorisation presque parfaite des données d'entraînement, mais la bonne performance en validation (94%) montre que le modèle généralise correctement.

Les courbes d'apprentissage confirment ce classement : ResNet50 converge le plus haut sans surapprentissage marqué, VGG16 reste proche, et le CNN simple plafonne avec une loss plus élevée.

3.3. Analyse comparative des architectures

3.3.1. Identification du modèle le plus performant

Le ResNet50 est le modèle le plus performant de notre étude, obtenant les meilleurs scores sur l'ensemble des métriques évaluées, que ce soit sur l'ensemble de validation ou sur l'ensemble de test indépendant (Mendeley Data). Il atteint une accuracy de 94,9% en validation et 91,0% en test, contre respectivement 90,9% et 72,0% pour VGG16, et 85,4% et 77,0% pour le CNN simple. L'écart entre validation et test n'est que de 3,9% pour ResNet50, contre 18,9% pour VGG16 et 8,4% pour le CNN simple. Avec un F1-score macro de 0,92 et une AUC de 0,986, ResNet50 offre la meilleure stabilité et fiabilité pour une éventuelle application clinique. [123]

3.3.2. Discussion sur la capacité de généralisation

La capacité de généralisation est essentielle pour une éventuelle application clinique. Notre protocole de **validation externe** (test sur Mendeley Data, source totalement indépendante de Kaggle) permet d'évaluer cette capacité de manière réaliste [123].

- **Pourquoi VGG16 a chuté ?**

Trois facteurs expliquent la chute de performance de VGG16 : son trop grand nombre de paramètres (138 millions) favorise la mémorisation plutôt que la généralisation ; l'absence de fine-tuning a empêché l'adaptation des caractéristiques profondes à la tâche médicale ; et les

images de test (Mendeley) présentent une distribution différente de celle des données d'entraînement (Kaggle) en termes d'appareils, protocoles et centres.[112-114-123]

- **Pourquoi ResNet50 est stable ?**

La stabilité de ResNet50 s'explique par plusieurs facteurs : le fine-tuning en deux phases permet une adaptation progressive sans détruire les connaissances d'ImageNet ; les connexions résiduelles facilitent l'apprentissage et réduisent le surapprentissage ; son nombre modéré de paramètres (25 millions contre 138 pour VGG16) limite la mémorisation excessive ; et la BatchNormalisation intégrée stabilise l'apprentissage tout en améliorant la généralisation. [99-113-114]

- **Pourquoi CNN simple généralise mieux que VGG16 ?**

Malgré ses performances plus faibles (77,0% en test), le CNN simple généralise mieux sur des données indépendantes que VGG16, avec un écart de 8,4% contre 18,9% pour VGG16. Cela s'explique par son capacité représentationnelle plus limitée, qui le rend moins susceptible de se sur-spécialiser aux particularités du jeu Kaggle, et surtout par l'absence de pré-entraînement sur ImageNet, qui élimine tout biais hérité de ce corpus d'images naturelles. Ce constat doit toutefois être relativisé : avec environ 24 millions de paramètres, le CNN n'est pas un modèle « léger » au sens strict—sa meilleure généralisation relative tient davantage à sa conception spécifique qu'à sa seule taille. [112-118] En définitive, ResNet50 se révèle le plus robuste au changement de source (écart validation-test de 3,9 %), ce qui en fait le meilleur candidat pour une application clinique.

3.4. Analyse statistique des résultats

L'analyse statistique des résultats permet d'évaluer la fiabilité et la reproductibilité des performances des modèles. Dans cette section, nous examinons la stabilité des modèles à travers l'écart entre les ensembles de validation et de test, ainsi que la robustesse expérimentale face aux variations des données.

3.4.1. Écart-type des performances

L'écart entre les performances en validation et en test permet d'évaluer la dispersion des résultats et de détecter d'éventuels phénomènes de surapprentissage. Les résultats obtenus montrent que le CNN simple présente une accuracy de 85,4% en validation contre 77,0% en test, soit un écart de 8,4%, avec un F1-score passant de 0,86 à 0,75 (écart de 0,11). VGG16 affiche une accuracy de 90,9% en validation contre 72,0% en test, soit un écart de 18,9%, avec un F1-score chutant de 0,91 à 0,75 (écart de 0,16). En revanche, ResNet50 montre une

grande stabilité avec une accuracy de 94,9% en validation contre 91,0% en test, soit un écart de seulement 3,9%, et un F1-score passant de 0,92 à 0,91 (écart de 0,01).

L'analyse des écarts entre validation et test révèle que ResNet50 présente l'écart le plus faible (3,9% pour l'accuracy et 0,01 pour le F1-score), ce qui indique une très faible dispersion des performances ; ce modèle est donc le plus stable. Le CNN simple montre un écart modéré (8,4% pour l'accuracy), ce qui est acceptable pour un modèle simple construit à partir de zéro. En revanche, VGG16 présente l'écart le plus élevé (18,9% pour l'accuracy), signe d'une forte instabilité et d'un surapprentissage important, la chute du F1-score de 0,91 à 0,75 confirmant cette fragilité.

3.4.2. Analyse de stabilité du modèle

La stabilité se lit dans l'écart validation-test : faible pour ResNet50 (3,9 %), modéré pour le CNN simple (8,4 %), élevé pour VGG16 (18,9 %), confirmant que VGG16 généralise mal hors de sa source d'entraînement.

3.4.3. Robustesse expérimentale

La robustesse mesure la résistance aux variations (bruit, artefacts, contraste, changement de source). ResNet50 est le plus robuste (91 % en test malgré le changement de source), tandis que VGG16 s'effondre et que le CNN simple reste sensible au faible contraste et aux petites tumeurs. À l'inverse, ResNet50 conserve 91 % de performance sur le test externe : le fine-tuning en deux phases lui confère une robustesse nettement supérieure au changement de source.

3.5. Analyse des erreurs de classification

3.5.1. Étude des cas mal classifiés

L'examen des cas mal classifiés révèle des différences significatives entre les trois modèles.

CNN simple : Le total des erreurs en validation est de 15% des images (accuracy 85,4%) et de 23% des images en test (accuracy 77,0%). Les erreurs principales concernent la confusion entre "No tumor" et les tumeurs, particulièrement avec Glioma et Adénome.

VGG16 : Le total des erreurs en validation est de 9% des images (accuracy 90,9%) et de 28% des images en test (accuracy 72,0%). Les erreurs principales sont la confusion entre Glioma et No tumor, ainsi qu'entre Méningiome et No tumor.

ResNet50 : Le total des erreurs en validation est de 5% des images (accuracy 94,9%) et de 9% des images en test (accuracy 91,0%). Les erreurs principales concernent la confusion entre Glioma et Méningiome, avec un recall du Glioma de 0,73 en test.

(Les matrices de confusion sont présentées en section 3.1.3.b pour la validation et 3.1.3.c pour le test)

3.5.2. Confusions fréquentes entre classes

D'après les matrices de confusion, les confusions les plus fréquentes sont :

Pour CNN simple : La classe "No tumor" est souvent confondue avec Glioma ou Adénome, et la classe Glioma est souvent confondue avec No tumor.

Pour VGG16 : La classe Glioma est souvent confondue avec No tumor, et la classe Méningiome est souvent confondue avec No tumor.

Pour ResNet50 : La classe Glioma est souvent confondue avec Méningiome, et inversement.

(Voir les matrices de confusion en section 3.1.3.b pour la validation et 3.1.3.c pour le test)

3.5.3. Causes possibles des erreurs

Pour le CNN simple, l'architecture est trop peu profonde pour capturer des caractéristiques complexes. Concernant VGG16, un surapprentissage s'est produit : le modèle a mémorisé les spécificités du dataset Kaggle, d'où la confusion avec la classe "No tumor" sur le test Mendeley. Quant à ResNet50, le modèle a des difficultés à distinguer Glioma et Méningiome en raison d'une similarité intrinsèque sur certaines coupes IRM.

3.5.4. Impact clinique potentiel des erreurs

Trois types d'erreur sont distingués selon leur gravité clinique : le faux positif (sain classé malade) entraîne des examens inutiles ; le faux négatif (tumeur non détectée) est le plus grave, retardant la prise en charge ; la confusion gliome/méningiome a un impact modéré. VGG16, qui multiplie les faux négatifs en test, est le plus risqué cliniquement, alors que ResNet50 ne commet que des confusions modérées.

3.6. Analyse qualitative et interprétation scientifique

L'analyse qualitative complète l'évaluation quantitative en examinant visuellement le comportement des modèles sur des cas individuels. Cette section présente des exemples

représentatifs de prédictions, discute des régions tumorales d'intérêt et aborde l'interprétabilité des décisions du modèle.

3.6.1. Visualisation d'images représentatives

Pour chaque modèle, des exemples représentatifs (image IRM, vérité terrain, classe prédite et probabilités Softmax) ont été examinés. ResNet50 affiche des confiances élevées et cohérentes ; un cas notable est un gliome (Image_0002) classé à tort en méningiome avec 93,3 % de confiance, illustrant la proximité radiologique de ces deux classes.

(Les visualisations complètes sont présentées en section 3.1.3.c)

Observations générales sur l'ensemble des modèles :

Pour **ResNet50**, les confiances sont généralement très élevées (supérieures à 90%) pour les prédictions correctes, tandis que les erreurs s'accompagnent de confiances plus faibles (inférieures à 70%).

Pour **VGG16**, on observe des confiances parfois faibles (inférieures à 70%) même pour des prédictions correctes, ce qui traduit une moindre confiance globale du modèle dans ses décisions.

Quant au **CNN simple**, ses confiances sont globalement cohérentes avec ses performances plus modestes, montrant des niveaux de confiance variables sans corrélation systématique avec la justesse des prédictions.

3.6.2 Description des régions tumorales d'intérêt

L'observation visuelle des images bien classées permet d'identifier certaines caractéristiques communes favorisant une bonne classification par les modèles.

- **Caractéristiques des images correctement classées :**

Les images bien classées par tous les modèles présentent quatre propriétés communes : un contraste élevé entre la tumeur et le tissu sain, une taille suffisante de la tumeur, une forme caractéristique typique de sa classe (bords nets pour les méningiomes, infiltration diffuse pour les gliomes), et une absence d'artefacts (bonne qualité d'image, sans bruit ni mouvement)..

- **Caractéristiques des images mal classées :**

Les images mal classées par tous les modèles partagent des difficultés communes : un faible contraste entre la tumeur et les tissus sains (difficulté la plus fréquente), des tumeurs de

très petite taille (quelques pixels seulement) qui ne permettent pas d'extraire des caractéristiques riches, des artefacts de mouvement rendant les images floues, et un bruit important lié à une mauvaise qualité d'acquisition qui peut être interprété à tort comme des structures pathologiques.

(Ces observations sont basées sur l'analyse visuelle des prédictions présentées en section 3.1.3.c)

3.6.3 Interprétation des décisions du modèle

L'interprétation des décisions du modèle repose sur l'analyse fine des probabilités de sortie et des confusions observées dans les matrices de confusion présentées précédemment.

a) Interprétation pour ResNet50 :

Pour ResNet50, trois observations principales sont à retenir. D'abord, il existe un bon lien entre confiance et exactitude : quand la confiance dépasse 90%, la prédiction est généralement correcte, ce qui est utile pour une utilisation clinique en n'affichant les résultats qu'au-delà d'un seuil élevé. Ensuite, la confusion la plus fréquente concerne Glioma et Méningiome, dont les apparences radiologiques sont souvent similaires (masses hyper-intenses avec œdème péri-lésionnel). Enfin, la classe "No tumor" est parfaitement détectée avec des confiances supérieures à 95%, grâce à l'absence des caractéristiques texturales et morphologiques des tumeurs.

b) Interprétation pour VGG16 :

VGG16 est très performant sur la validation (Kaggle) mais chute significativement sur le test (Mendeley), confirmant qualitativement l'écart observé dans l'évaluation quantitative. Les confusions fréquentes entre Glioma/No tumor et Méningiome/No tumor montrent que le modèle a mémorisé des motifs ou textures propres au dataset Kaggle plutôt que d'apprendre des caractéristiques universelles des tumeurs cérébrales.

c) Interprétation pour CNN simple :

Le CNN simple, avec son architecture peu profonde, montre des confusions fréquentes entre "No tumor" et les différentes classes de tumeurs. De manière préoccupante, certaines de ces erreurs s'accompagnent de confiances élevées, ce qui signifie que le modèle peut être "sûr de lui" tout en ayant tort. Cette observation souligne la nécessité d'architectures suffisamment profondes pour capturer les caractéristiques discriminantes fines des tumeurs cérébrales.

3.6.4. Discussion sur l'explicabilité des prédictions

L'explicabilité est un enjeu majeur pour l'adoption clinique des modèles. Trois méthodes sont envisageables : Grad-CAM (cartes de chaleur pour visualiser les régions influençant la décision), LIME (utile pour analyser des cas individuels difficiles), et SHAP (trop coûteux pour des images 224×224). Pour des travaux futurs, l'application de Grad-CAM sur ResNet50 permettrait de vérifier l'absence d'utilisation d'artefacts. Une comparaison des cartes entre les trois modèles aiderait à comprendre les erreurs de VGG16 sur la classe "No tumor". Enfin, une validation par un radiologue expert des cartes produites serait essentielle pour confirmer leur pertinence clinique. [124-125]

Les résultats qualitatifs présentés dans cette section – visualisation des prédictions et analyse des probabilités de sortie – fournissent une première interprétation des décisions du modèle. Cependant, cette approche reste limitée car elle ne permet pas de visualiser directement les régions de l'image qui ont guidé la décision

3.7. Discussion générale

Cette section propose une synthèse et une interprétation globale des résultats obtenus. Nous comparons nos performances avec les travaux antérieurs, vérifions nos hypothèses de recherche, mettons en évidence les apports scientifiques de notre travail, et discutons de l'intérêt et des limites pour la pratique clinique.

3.7.1. Interprétation globale des résultats

Nos résultats établissent une hiérarchie nette : ResNet50 (fine-tuning) est le meilleur et le plus stable, VGG16 souffre d'un surapprentissage marqué en test, et le CNN simple offre une baseline honorable. La classe « No tumor » est parfaitement détectée par tous (rappel ≥ 0,99), tandis que la confusion Glioma/Méningiome persiste même chez ResNet50 (rappel du Glioma de 0,73 en test), en raison de leur ressemblance sur certaines coupes IRM.

3.7.2. Comparaison avec les travaux antérieurs

Nos résultats sont cohérents avec la littérature scientifique sur la classification des tumeurs cérébrales par IRM.

Tableau 3.5 : Comparaison des accuracies rapportées dans la littérature avec nos résultats

Étude	Modèle	Dataset	Accuracy	Comparaison avec notre étude
Cheng et al. (2015) [4]	CNN + features manuelles	Figshare	~91%	Notre ResNet50 (94,9%) est supérieur
Talo et al. (2019) [38]	VGG16 + transfer learning	Figshare + SARTAJ	~95% (val)	Comparable à notre VGG16 (90,9% val)
Deepak et Ameer (2019)	ResNet50 + fine-tuning	Figshare	~95%	Comparable à notre ResNet50
Nabizadeh et al. (2021)	CNN simple	Kaggle	~85%	Comparable à notre CNN simple (85,4%)
Notre étude	ResNet50 + fine-tuning	Kaggle + Mendeley	94,9% (val), 91% (test)	Validation externe

Notre étude présente cinq apports significatifs. Premièrement, nous avons adopté une validation externe rigoureuse avec deux sources indépendantes (Kaggle pour l'entraînement, Mendeley pour le test), contrairement à la majorité des études qui divisent un seul dataset. Deuxièmement, nous avons comparé systématiquement trois architectures (CNN simple, VGG16, ResNet50) dans des conditions identiques. Troisièmement, notre analyse détaillée des erreurs a identifié que la confusion Glioma/Méningiome persiste même pour le meilleur modèle. Quatrièmement, nous fournissons une baseline reproductible (CNN simple à 85,4%). Enfin, notre discussion sur l'impact clinique des faux positifs et faux négatifs constitue un apport original.

3.7.3. Vérification des hypothèses de recherche

Nos cinq hypothèses initiales ont toutes été confirmées. Premièrement, les modèles pré-entraînés (VGG16 et ResNet50) surpassent le CNN simple, avec respectivement 90,9% et 94,9% contre 85,4% en validation. Deuxièmement, ResNet50 avec fine-tuning est plus performant que VGG16 avec transfer learning, aussi bien en validation (94,9% contre 90,9%) qu'en test (91,0% contre 72,0%). Troisièmement, la validation externe a révélé un écart significatif entre validation et test, particulièrement marqué pour VGG16 (18,9%). Quatrièmement, la classe "No tumor" est la mieux classée par les trois modèles, avec un rappel supérieur ou égal à 0,99. Cinquièmement, la confusion entre Glioma et Méningiome est bien la plus fréquente pour ResNet50 et VGG16.

3.7.4. Apports scientifiques du travail

Nos contributions sont : une validation externe sur source indépendante (rare dans la littérature), une comparaison systématique de trois architectures dans des conditions identiques, une analyse fine des erreurs et de leur impact clinique, et un travail reproductible servant de baseline.

3.7.5. Intérêt et limites pour la pratique clinique

ResNet50 présente un intérêt clinique majeur avec une performance élevée (91% en test), une grande stabilité, une excellente détection des cas normaux (rappel ≥ 0,99), un temps de prédiction inférieur à la seconde et une généralisation validée sur source externe. Cependant, plusieurs limites doivent être prises en compte : l'utilisation de données publiques nécessite une validation clinique prospective, le modèle ne connaît que quatre classes et ignore d'autres pathologies comme les métastases ou les abcès, les données sont uniquement en 2D (perte de l'information 3D), et VGG16 n'est pas fiable avec seulement 72% sur test. En recommandation, ResNet50 avec fine-tuning est le modèle à privilégier pour un système d'aide au diagnostic après validation multicentrique, VGG16 est à éviter, tandis que le CNN simple convient aux applications légères ou embarquées.

3.8. Limites de l'étude

Malgré la rigueur de notre protocole expérimental et la pertinence de nos résultats, cette étude présente certaines limites qu'il est important de reconnaître pour une interprétation correcte des résultats et pour orienter les travaux futurs.

3.8.1 Taille du dataset

Le corpus (9 614 images : 5 600 entraînement, 1 600 validation, 2 414 test) reste modeste face aux standards du deep learning. Un jeu plus volumineux, notamment pour les classes moins représentées, améliorerait probablement les performances, en particulier pour VGG16 qui a montré des signes de surapprentissage.

3.8.2. Variabilité des données

Nos données proviennent de bases publiques (Figshare, SARTAJ, Br35H pour l'entraînement ; Mendeley pour le test), sans information détaillée sur les paramètres d'acquisition (appareil, champ, séquences). Une validation prospective multicentrique, sur données hospitalières réelles, serait nécessaire pour confirmer la généralisabilité.

3.8.3. Risque de surapprentissage

Le surapprentissage a été clairement identifié pour VGG16 (écart validation-test de 18,9 %), lié à ses 138 millions de paramètres pour un corpus modéré ; malgré dropout, early stopping et augmentation, il n'a pu être totalement éliminé. ResNet50 (écart 3,9 %) et le CNN simple (8,4 %) restent mieux équilibrés. La validation croisée pourrait renforcer la robustesse des résultats.

3.8.4. Absence éventuelle de validation externe

La validation externe sur source indépendante (Mendeley) est un point fort de notre méthodologie, mais elle ne porte que sur une seule source. Une validation sur plusieurs sources (différents hôpitaux et appareils) et l'inclusion d'autres pathologies (métastases, abcès) seraient nécessaires avant toute application clinique.

3.8.5. Contraintes liées à l'interprétabilité

L'interprétabilité des modèles de deep learning reste un défi majeur pour leur adoption en milieu clinique. Un clinicien doit comprendre pourquoi un modèle a pris une décision pour lui faire confiance. Les méthodes d'explicabilité comme Grad-CAM [125] permettent de visualiser les régions de l'image qui ont influencé la décision du modèle. L'intégration de telles méthodes dans des travaux futurs permettrait de renforcer la transparence et l'acceptabilité du modèle par les cliniciens, en fournissant des cartes de chaleur (heatmaps) superposées aux images IRM. Ces visualisations aideraient à vérifier que le modèle se concentre bien sur les zones tumorales et non sur des artefacts.

3.9. Perspectives

Cette étude ouvre plusieurs perspectives pour des travaux futurs, que ce soit pour améliorer les performances des modèles, renforcer leur fiabilité, ou faciliter leur intégration en milieu clinique.

3.9.1 Intégration de méthodes d'explicabilité avancées

L'un des principaux obstacles à l'adoption des modèles de deep learning en clinique est leur manque de transparence. Pour qu'un clinicien puisse faire confiance à un modèle, il doit comprendre sur quels critères sa décision est basée. L'intégration de méthodes d'explicabilité permettrait de visualiser les régions de l'image qui ont le plus influencé la décision du modèle. Des méthodes comme **LIME** [57] ou **SHAP** pourraient être explorées pour une analyse fine des décisions du modèle. Une validation par des radiologues experts serait ensuite nécessaire pour confirmer la pertinence clinique des zones mises en évidence.

3.9.2. Utilisation de datasets plus larges et multicentriques

Bien que notre dataset soit relativement conséquent (9614 images), l'utilisation de datasets plus larges, idéalement multicentriques, permettrait d'améliorer la robustesse et la généralisation des modèles. Des initiatives comme **BraTS** (Brain Tumor Segmentation) ou des collaborations entre plusieurs hôpitaux pourraient fournir des données plus variées, incluant différents types d'appareils IRM, de protocoles d'acquisition et de populations de patients. L'augmentation de la taille du dataset, en particulier pour les classes sous-représentées comme "No tumor", réduirait les biais et améliorerait les performances. De plus, l'inclusion d'autres pathologies cérébrales (métastases, abcès, maladies inflammatoires) permettrait de former un modèle capable de distinguer un plus grand nombre de diagnostics différentiels, ce qui est essentiel pour une application clinique réaliste.

3.9.3. Extension vers la segmentation automatique des tumeurs

Une perspective majeure est d'étendre l'approche à la segmentation automatique (U-Net), pour délimiter la tumeur et fournir des mesures quantitatives (volume, localisation) en complément du typage.

3.9.4. Intégration potentielle dans un workflow hospitalier réel

L'objectif final de cette recherche est de proposer un outil d'aide au diagnostic pouvant être intégré dans un workflow hospitalier réel. Plusieurs étapes seraient nécessaires pour y parvenir. Premièrement, une **validation prospective** sur des données cliniques réelles, collectées dans un ou plusieurs hôpitaux, serait indispensable. Deuxièmement, le modèle devrait être **optimisé pour l'inférence en temps réel** (par exemple, via la quantification ou la distillation) afin de pouvoir traiter les images rapidement sans nécessiter de ressources computationnelles excessives. Troisièmement, une **interface utilisateur intuitive** devrait être développée pour permettre aux radiologues d'interagir facilement avec le modèle (chargement d'une image, visualisation de la prédiction et des probabilités). Enfin, une **validation clinique** rigoureuse, incluant des études prospectives mesurant l'impact du modèle sur les pratiques diagnostiques (temps de lecture, précision du diagnostic, réduction des examens inutiles), serait nécessaire avant tout déploiement à grande échelle.

3.10. Conclusion

Ce chapitre établit une hiérarchie claire entre les trois architectures. ResNet50 (fine-tuning) s'impose comme la plus performante et la plus robuste, avec 94,9 % en validation et 91,0 % en test indépendant — un écart de 3,9 % traduisant une excellente généralisation. VGG16, malgré 90,9 % en validation, s'effondre à 72,0 % en test, signe d'un surapprentissage critique lié à ses 138 millions de paramètres. Le CNN simple remplit son rôle de baseline (85,4 % et 77,0 %). Sur le plan clinique, les trois modèles détectent quasi

parfaitement la classe « No tumor » (rappel ≥ 0,99), limitant les faux positifs. L'analyse des erreurs révèle toutefois une confusion persistante gliome/méningiome, due à leurs similitudes morphologiques. Malgré ses limites (taille du corpus, variabilité non contrôlée, absence de validation hospitalière), l'étude confirme l'hypothèse initiale : architectures profondes et fine-tuning optimisent la classification, ouvrant la voie à la segmentation et aux études multicentriques.

Conclusion générale

Les tumeurs cérébrales constituent un enjeu majeur de santé publique en raison de leur impact sur les fonctions neurologiques, de leur complexité diagnostique et de la nécessité d'une prise en charge précoce. L'imagerie par résonance magnétique (IRM) représente aujourd'hui la modalité de référence pour l'évaluation des lésions cérébrales. Cependant, l'interprétation d'un grand volume d'images médicales demeure une tâche exigeante pouvant bénéficier de l'apport des techniques d'intelligence artificielle.

Dans ce mémoire, nous nous sommes intéressés à la détection et à la classification automatisée des tumeurs cérébrales à partir d'images IRM en utilisant des approches de Deep Learning. L'objectif principal était d'étudier l'influence de différentes architectures de réseaux de neurones convolutifs sur la qualité de la classification et sur leur capacité à généraliser à des données provenant de sources indépendantes.

Pour atteindre cet objectif, une méthodologie rigoureuse a été mise en œuvre. Les modèles ont été entraînés sur un dataset public issu de Kaggle comprenant quatre classes (gliome, méningiome, adénome hypophysaire et absence de tumeur), puis évalués sur une base de données totalement indépendante provenant de Mendeley Data. Trois architectures ont été comparées : un CNN personnalisé développé spécifiquement pour cette étude, un modèle VGG16 utilisant le Transfer Learning et une architecture ResNet50 optimisée par Fine-Tuning.

Les résultats montrent que les architectures profondes pré-entraînées surpassent globalement le CNN simple. ResNet50 s’est révélé le plus performant, avec 94,9 % en validation et 91,0 % sur l’ensemble de test indépendant, confirmant l’intérêt des connexions résiduelles et du fine-tuning pour la classification d’images médicales.

L'analyse comparative a également mis en évidence les limites potentielles du surapprentissage. Bien que VGG16 obtienne de très bons résultats lors de la phase de validation, ses performances diminuent significativement sur les données externes, soulignant l'importance d'une validation indépendante pour évaluer de manière réaliste la robustesse d'un modèle de Deep Learning.

Au-delà des performances quantitatives, ce travail met en évidence l'intérêt des systèmes intelligents comme outils d'aide au diagnostic. Toutefois, les résultats obtenus ne doivent pas être interprétés comme une substitution à l'expertise médicale. Ces approches ont vocation à assister les radiologues en fournissant une seconde lecture, en facilitant le tri des examens et en contribuant à la réduction du temps d'interprétation.

Malgré les résultats encourageants obtenus, certaines limites subsistent, notamment la taille relativement modeste des bases de données utilisées, l'absence de validation multicentrique et l'utilisation d'images bidimensionnelles ne représentant qu'une partie de l'information disponible dans les volumes IRM complets.

Comme perspectives futures, plusieurs axes d'amélioration peuvent être envisagés : l'intégration de bases de données multicentriques plus larges, l'utilisation de modèles tridimensionnels exploitant l'information volumique des examens IRM, le développement de méthodes d'explicabilité telles que Grad-CAM, LIME ou SHAP, ainsi que l'extension vers la segmentation automatique des tumeurs cérébrales. Une validation clinique prospective constituerait également une étape indispensable avant toute utilisation en environnement hospitalier réel.

En définitive, ce travail confirme le potentiel du Deep Learning pour la classification automatisée des tumeurs cérébrales et montre que l'association d'architectures profondes performantes, d'une stratégie de validation rigoureuse et d'une analyse critique des résultats constitue une voie prometteuse pour le développement de futurs systèmes d'aide au diagnostic clinique.

Les références

1. Charles, N. A., Holland, E. C., Gilbertson, R., Glass, R., & Kettenmann, H. (2011). The brain tumor microenvironment. *Glia*, *59*(8), 1169-1180.
2. Ostrom, Q. T., Gittleman, H., De Blank, P. M., Finlay, J. L., Gurney, J. G., McKean-Cowdin, R., ... & Barnholtz-Sloan, J. S. (2016). American brain tumor association adolescent and young adult primary brain and central nervous system tumors diagnosed in the United States in 2008-2012. *Neuro-oncology*, *18*(suppl_1), i1-i50.
3. Berger, T. R., Wen, P. Y., Lang-Orsini, M., & Chukwueke, U. N. (2022). World Health Organization 2021 classification of central nervous system tumors and implications for therapy for adult-type gliomas: a review. *JAMA oncology*, *8*(10), 1493-1501.
4. Wen, P. Y., & Huse, J. T. (2017). 2016 World Health Organization classification of central nervous system tumors. *CONTINUUM: Lifelong Learning in Neurology*, *23*(6, Neuro-oncology), 1531-1547.
5. Komori, T. (2017). The 2016 WHO classification of tumours of the central nervous system: the major points of revision. *Neurologia medico-chirurgica*, *57*(7), 301-311.
6. Panda, B., & Panda, C. S. (2019). A review on brain tumor classification methodologies. *Int J Sci Res Sci Technol*, *6*, 346-359.
7. Reddy, D. K., Soumya, D. R., Nagar, A., & Rajpoot, A. K. (2022, November). A Review of Recent Advances in Classification and Prediction of Brain Tumor using Deep Learning. In *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)* (pp. 849-856). IEEE.
8. Gelabert-Gonzalez, M., & Serramito-Garcia, R. (2011). Intracranial meningiomas: I. Epidemiology, aetiology, pathogenesis and prognostic factors. *Revista de Neurologia*, *53*(3), 165-172.
9. Sayaheen, Y. O. (2023). Texture-based approach to classification meningioma using pathology images. *International Journal of Computational Vision and Robotics*, *13*(6), 677-692.
10. Śledzińska, P., Bebyn, M., Furtak, J., Koper, A., & Koper, K. (2023). Current and promising treatment strategies in glioma. *Reviews in the Neurosciences*, *34*(5), 483-516.
11. Śledzińska, P., Bebyn, M. G., Furtak, J., Kowalewski, J., & Lewandowska, M. A. (2021). Prognostic and predictive biomarkers in gliomas. *International journal of molecular sciences*, *22*(19), 10373.
12. Zhang, P., Xia, Q., Liu, L., Li, S., & Dong, L. (2020). Current opinion on molecular characterization for GBM classification in guiding clinical diagnosis, prognosis, and therapy. *Frontiers in molecular biosciences*, *7*, 562798.
13. Colman, H. (2020). Adult gliomas. *CONTINUUM: Lifelong Learning in Neurology*, *26*(6), 1452-1475.
14. GHU Paris Psychiatrie & Neurosciences. (s.d.). *Les gliomes de bas grade*. Consulté le 9 mai 2026, à l'adresse https://www.ghu-paris.fr/fr/les-gliomes-de-bas-grade
15. Eckel-Passow, J. E., Lachance, D. H., Molinaro, A. M., Walsh, K. M., Decker, P. A., Sicotte, H., ... & Jenkins, R. B. (2015). Glioma groups based on 1p/19q, IDH, and TERT promoter mutations in tumors. *New England Journal of Medicine*, *372*(26), 2499-2508.

16. Masui, K., Mischel, P. S., & Reifenberger, G. (2016). Molecular classification of gliomas. *Handbook of clinical neurology*, *134*, 97-120.

17. Kenis, Y. (1962). Les Tendances Actuelles De La Chimiotherapie Anticancereuse Par Les Agents Alkylants. *Acta Clinica Belgica*, *17*(1), 1-7.

18. Salunke, P., & Gupta, S. K. (2013). Symptomatic bilateral isolated perforator infarction following aneurysmal subarachnoid hemorrhage. *Journal of Neurosciences in Rural Practice*, *4*(1), 45.

19. Karsy, M., Guan, J., Cohen, A. L., Jensen, R. L., & Colman, H. (2017). New molecular considerations for glioma: IDH, ATRX, BRAF, TERT, H3 K27M. *Current neurology and neuroscience reports*, *17*(2), 19.

20. Afonso, M., & Brito, M. A. (2022). Therapeutic options in neuro-oncology. *International Journal of Molecular Sciences*, *23*(10), 5351.

21. Laghmani, M., Messedi, M., Kassar, S., Drissi, C., Sebai, R., & Belghith, M. (2013). Apport de l'imagerie par résonance magnétique (IRM) dans le diagnostic des tumeurs cérébrales. *Annales de Pathologie*, *33*(1), 11-23. https://doi.org/10.1016/j.annpat.2012.11.001

22. Atasoy, Ö. (2026). Meningiomas: From Pathogenesis to Therapeutics-Current Perspectives and a Holistic Review. *Arşiv Kaynak Tarama Dergisi*, *35*(1), 1-9.

23. Kabel, A. (2020). Pituitary adenomas: insights into the recent trends. *Journal of Cancer Research and Treatment*, *8*(2), 21-24.

24. Skiba, J., Skiba, Z., Tylczyńska, K., Tylczyńska, N., Kowalik, K., Michalska, M., ... & Maciejewski, I. (2024). Pituitary Neuroendocrine Tumors (PitNETs)–a literature review. *Quality in Sport*, *33*, 55879-55879.

25. Inselspital Universitätsspital Bern. (s.d.). *Adénome hypophysaire*. Consulté le 9 mai 2026, à l'adresse https://neurochirurgie.insel.ch/fr/maladies-traitees-specialites/tumeurs-cerebrales/adenome-hypophysaire

26. Markou, M., Lavrentaki, A., & Ntali, G. (2023). Recent advances in understanding and managing pituitary adenomas. *Faculty Reviews*, *12*, 6.

27. Verdugo, E., Puerto, I., & Medina, M. Á. (2022). An update on the molecular biology of glioblastoma, with clinical implications and progress in its treatment. *Cancer Communications*, *42*(11), 1083-1111.

28. STEFANI, A. DOCTEUR EN MÉDECINE.

29. Tassadit, A., & Nawel, B. (2015). *Segmentation des images IRM cérébrales* (Doctoral dissertation, Université Mouloud Mammeri).

30. Radiologie Estrie. (n.d.). *Résonance magnétique*. Consulté le 9 mai 2026, à l'adresse https://www.radiologieestrie.ca/services/resonance-magnetique/

31. Durand-Dubief, F. (2013). *La place de l'IRM dans le suivi de la SEP* [Fichier PDF]. Réseau Rhône-Alpes SEP. http://www.rhone-alpes-sep.org/wp-content/uploads/2013/02/La-place-de-lIRM-dans-le-suivi-de-la-SEP-Dr-Fran%C3%A7oise-Durand-Dubief.pdf

32. Cabanis, P. (2024). *Identification et caractérisation par IRM de diffusion de nouvelles structures macroscopiques cardiaques* (Doctoral dissertation, Université de Bordeaux).

33. Annexe, G. Dépistage et utilisation de la tomodensitométrie («CT scan»). *GUIDE DE SURVEILLANCE MÉDICALE DES TRAVAILLEURS EXPOSÉS À LA SILICE ET*

RECOMMANDATIONS SUR LES SEUILS D'INTERVENTIONS PRÉVENTIVES (SIP) GUIDE DE PRATIQUE PROFESSIONNELLE, 217.

34. Medicare Group. (n.d.). *Chest CT scan*. Retrieved May 9, 2026, from https://medicare-group.hu/en/diagnostics/ct-scans/chest-ct-scan/

35. Abdallah, M. E. H. I. D. I. Imagerie médicale.

36. Ramsay Santé. (s.d.). *PET-scan*. Consulté le 9 mai 2026, à l'adresse https://www.ramsaysante.fr/vous-etes-patient-en-savoir-plus-sur-ma-pathologie/pet-scan

37. Rozenblum, L. (2024). *Nouveaux biomarqueurs d'imagerie pour la prise en charge des lymphomes primitifs du système nerveux central: études en TEP-IRM au 18F-FDG et à la 18F-FLUDARABINE* (Doctoral dissertation, Sorbonne Université).

38. Toussaint, M. (2016). *Thérapies par rayonnements appliquées au cas du glioblastome: Intérêt du suivi par spectroscopie et imagerie de diffusion par résonance magnétique. Vers une thérapie bimodale* (Doctoral dissertation, Université de Lorraine).

39. Coulon, C., & Miens, P. (2023). L'IRM d'aujourd'hui et de demain, plus puissant ou moins consommateur?. *IRBM News*, *44*(2), 100454.

40. Ardaine, M., Carrillo, C., & Youssef, E. (2025). Bénéfices de l'utilisation de l'IA générative par les TRM pour la radiographie de l'épaule.

41. Mokri, M. (2025). Optimisation de l'imagerie TEP myocardique dynamique par l'IA: réduction du temps d'acquisition et modélisation prédictive.

42. Perez-Torrents, J. (2024). *Gérer la collaboration entre l'expert métier et l'Intelligence Artificielle: Deux études de cas dans le système de soins* (Doctoral dissertation, Institut Polytechnique de Paris).

43. Ogut, E. (2025). Artificial intelligence in clinical medicine: challenges across diagnostic imaging, clinical decision support, surgery, pathology, and drug discovery. *Clinics and practice*, *15*(9), 169.

44. Karalis, V. D. (2024). The integration of artificial intelligence into clinical practice. *Applied Biosciences*, *3*(1), 14-44.

45. Wu, S., Wang, J., Guo, Q., Lan, H., Zhang, J., Wang, L., ... & Chen, Y. (2022). Application of artificial intelligence in clinical diagnosis and treatment: an overview of systematic reviews. *Intelligent Medicine*, *2*(02), 88-96.

46. Jung, K. H. (2025). Large language models in medicine: clinical applications, technical challenges, and ethical considerations. *Healthcare Informatics Research*, *31*(2), 114-124.

47. Kaur, S., & Jindal, S. (2016). A survey on machine learning algorithms. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, *3*(11), 2349-2763.

48. Sharma, R., & Bist, A. S. (2015). Machine learning: A survey. *Int. J. Eng. Sci. Res. Technol*, *4*(3), 708-716.

49. Andreanus, J., & Kurniawan, A. (2017). Sejarah, teori dasar dan penerapan reinforcement learning: Sebuah tinjauan pustaka. *Jurnal Telematika*, *12*(2), 113-118.

50. Singh, J., Gupta, M., & Kumar, R. (Eds.). (2026). *Smart Technologies and Intelligent Computing*. CRC Press.

51. Zhao, K., Xu, H., Yang, J., & Gao, K. (2022, May). Consistent representation learning for continual relation extraction. In *Findings of the association for computational linguistics: ACL 2022* (pp. 3402-3411).

52. Sina Technologie. (2020, 20 juillet). *Présentation des réseaux de neurones artificiels*. https://blog.sinatechnologie.com/presentation-des-reseaux-de-neurones-artificiels

53. Payandeh, A., Baghaei, K. T., Fayyazsanavi, P., Ramezani, S. B., Chen, Z., & Rahimi, S. (2023). Deep representation learning: Fundamentals, technologies, applications, and open challenges. *IEEE Access*, *11*, 137621-137659.

54. Nitish, S. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, *15*, 1.

55. Moradi, R., Berangi, R., & Minaei, B. (2020). A survey of regularization strategies for deep models. *Artificial intelligence review*, *53*(6).

56. Noura, H., & Nadia, K. (2018). *Classification des textures par les réseaux de neurones convolutifs* (Doctoral dissertation, Université Mouloud Mammeri).

57. Marwa, B. E. N. B. R. I. K., Meinel, R. E. Z. K. A. L. L. A. H., & Ali, M. K. (2022). Transfert d'Apprentissage avec Inception V3 appliquée aux images de parking.

58. Aamir, M., Rahman, Z., Abro, W. A., Tahir, M., & Ahmed, S. M. (2019). An optimized architecture of image classification using convolutional neural network. *International Journal of Image, Graphics and Signal Processing*, *10*(10), 30.

59. Elster, A. D. (2021). *Deep network types*. Questions and Answers in MRI. https://mriquestions.com/deep-network-types.html

60. Yang, R. (2024, November). Convolutional neural network for image classification research-based on VGG16. In *2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)* (pp. 213-217). IEEE.

61. Saba, T., Mohamed, A. S., El-Affendi, M. A., Amin, J., & Sharif, M. (2022). Brain Tumor Detection Using Fusion of Handcrafted and Deep Learning Features. *Applied Sciences*, *12*(14), 7282. https://doi.org/10.3390/app12147282

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

63. Bhupal, S. (2023, 10 August). *ResNet architecture explained*. Medium. https://medium.com/@siddheshb008/resnet-architecture-explained-47309ea9283d

64. Zoubir, D. (2025). *La Segmentation et La Détection des images IRM pour identifier les tumeurs cérébrales par l'apprentissage profond* (Doctoral dissertation, university of bordj bou arreridj).

65. ZERROUGUI, M., & HAMADENE, S. (2021). *Détection de la tumeur cérébrales dans l'image IRM par l'Apprentissage en profondeur* (Doctoral dissertation, Université Mohamed el-Bachir el-Ibrahimi Bordj Bou Arréridj Faculté de Mathématique et Informatique).

66. Boubdellaha, D., Mokni, R., & Ammar, B. (2025). Brain Tumor Classification with Hybrid Deep Learning Models from MRI Images.

67. Haddad, O., Fkih, F., & Omri, M. N. (2024). An intelligent sentiment prediction approach in social networks based on batch and streaming big data analytics using deep learning. *Social Network Analysis and Mining*, *14*(1), 150.

68. Dhakshnamurthy, V. K., Govindan, M., Sreerangan, K., Nagarajan, M. D., & Thomas, A. (2024). Brain tumor detection and classification using transfer learning models. *Engineering Proceedings*, *62*(1), 1.

69. Khan, M. Z., Zaman, F. U., Bhatti, P., Khan, A. A., Sultan, S. A., & Bhatti, S. D. (2024, December). A comparative analysis of deep learning architectures for efficient brain tumor detection. In *2024 International Conference on Sustainable Technology and Engineering (i-COSTE)* (pp. 01-06). IEEE.

70. Chakrabarty, S., Sotiras, A., Milchenko, M., LaMontagne, P., Hileman, M., & Marcus, D. (2021). MRI-based identification and classification of major intracranial tumor types by using a 3D convolutional neural network: a retrospective multi-institutional analysis. *Radiology: Artificial Intelligence*, *3*(5), e200301.

71. Ghazi Boumediene, G. (2024). *Segmentation des images médicales par des approches bio-inspirées* (Doctoral dissertation).

72. Alaoui, S. (2025). The Archivist and the Management of Open Research Datasets in Canadian Universities: Increasing Visibility and Developing Collaborations. *Journal of Information Sciences*, *24*(1), 95-119.

73. Grouin, C. (2013). *Anonymisation de documents cliniques: performances et limites des méthodes symboliques et par apprentissage statistique* (Doctoral dissertation, Université Pierre et Marie Curie-Paris VI).

74. Vrbančič, G., & Podgorelec, V. (2020). Transfer learning with adaptive fine-tuning. *Ieee Access*, *8*, 196197-196211.

75. FADEL, O. (2023). Combinaison de classifieurs pour la reconnaissance des anomalies mammaires.

76. Guibert, A. (2024). *Utilisation des outils de transparence des algorithmes pour la réduction de données appliquée aux réseaux neuronaux convolutifs* (Doctoral dissertation, Ecole Nationale Aviation Civile).

77. Slimani, H. (2025). *Méthodes, algorithmes et architectures pour l'inférence sur mesures parcimonieuses pour le diagnostic de câbles par réflectométrie* (Doctoral dissertation, Université Paris-Saclay).

78. Mellouli-Nauwynck, N. (2021). *Quelques contributions à l'extraction, l'enrichissement des connaissances et à la prédiction de données spatio-temporelles massives* (Doctoral dissertation, Université Paris 8).

79. CHAHINEZ, O. (2021). Détection de tumeur cérébrale avec l'apprentissage profond.

80. ZERROUGUI, M., & HAMADENE, S. (2021). *Détection de la tumeur cérébrales dans l'image IRM par l'Apprentissage en profondeur* (Doctoral dissertation, Université Mohamed el-Bachir el-Ibrahimi Bordj Bou Arréridj Faculté de Mathématique et Informatique).

81. Naser, M. Z. (2026). A review of machine learning with small and limited data. *Journal of Big Data*.

82. Tsegaye, B., Snell, K. I., Archer, L., Kirtley, S., Riley, R. D., Sperrin, M., ... & Dhiman, P. (2025). Larger sample sizes are needed when developing a clinical prediction model using machine learning in oncology: methodological systematic review. *Journal of Clinical Epidemiology*, *180*, 111675.

83. Sallé, G. (2024). *Apprentissage génératif à bas régime des données pour la segmentation d'images en oncologie* (Doctoral dissertation, Université de Bretagne occidentale-Brest).

84. Deneu, B. (2022). *Interprétabilité des modèles de distribution d'espèces basés sur des réseaux de neurones convolutifs* (Doctoral dissertation, Université de Montpellier (UM), FRA).

85. Lacroux, D. (2026). De la boîte noire à la boîte blanche-Les limites de la transparence des intelligences artificielles en santé. *médecine/sciences*, *42*(1), 71-77.

86. Gelibert, A. (2016). *Contribution à la mise en œuvre d'un outillage unifié pour faciliter la qualification d'environnements normés* (Doctoral dissertation, Université Grenoble Alpes).

87. Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988).

88. Groot, O. Q., Bindels, B. J., Ogink, P. T., Kapoor, N. D., Twining, P. K., Collins, A. K., ... & Schwab, J. H. (2021). Availability and reporting quality of external validations of machine-learning prediction models with orthopedic surgical outcomes: a systematic review. *Acta Orthopaedica*, *92*(4), 385-393.

89. Yagis, E., Atnafu, S. W., García Seco de Herrera, A., Marzi, C., Scheda, R., Giannelli, M., ... & Diciotti, S. (2021). Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Scientific reports*, *11*(1), 22544.

90. 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. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

91. Louis, D. N., Perry, A., Wesseling, P., Brat, D. J., Cree, I. A., Figarella-Branger, D., ... & Ellison, D. W. (2021). The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro-oncology*, *23*(8), 1231-1251.

92. Villanueva-Meyer, J. E., Mabray, M. C., & Cha, S. (2017). Current clinical brain tumor imaging. *Neurosurgery*, *81*(3), 397-415.

93. Cheng, J., Huang, W., Cao, S., et al. (2015). Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PLOS ONE*, 10(10), e0140381.

94. Goldbrunner, R., Stavrinou, P., Jenkinson, M. D., Sahm, F., Mawrin, C., Weber, D. C., ... & Weller, M. (2021). EANO guideline on the diagnosis and management of meningiomas. *Neuro-oncology*, *23*(11), 1821-1834.

95. Molitch, M. E. (2017). Diagnosis and treatment of pituitary adenomas: a review. *Jama*, *317*(5), 516-524.

96. Bourennane, M., & Naimi, H. (2023). An efficient hybrid model of CNNs and different kernels of SVM for brain tumor classification. *The Journal of Engineering and Exact Sciences*, *9*(8), 16739-01e.

97. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

98. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

99. Ioffe, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448-456). pmlr.

100. Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1, No. 2, pp. 1-800). Cambridge: MIT press.

101. Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, *6*(1), 1-48.

102. Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

103. Talo, M., Baloglu, U. B., Yıldırım, Ö., & Acharya, U. R. (2019). Application of deep transfer learning for automated brain abnormality classification using MR images. *Cognitive Systems Research*, *54*, 176-188.

104.Çinar, A., & Yildirim, M. (2020). Detection of tumors on brain MRI images using the hybrid convolutional neural network architecture. *Medical hypotheses*, *139*, 109684.

105. Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4, p. 738). New York: springer.

106. Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

107. Masters, D., & Luschi, C. (2018). Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*.

108. Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 807-814).

109. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929-1958.

110. Chollet, F. (2017). Deep Learning With Python Manning Publications Company.

111. Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning?. *IEEE transactions on medical imaging*, *35*(5), 1299-1312.

112. Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?. *Advances in neural information processing systems*, *27*.

113. Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 31, No. 1).

114. Howard, J., & Gugger, S. (2020). *Deep Learning for Coders with fastai and PyTorch*. O'Reilly Media.

115. Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade: Second edition* (pp. 437-478). Berlin, Heidelberg: Springer Berlin Heidelberg.

116. Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

117. Prechelt, L. (2002). Early stopping-but when?. In *Neural Networks: Tricks of the trade* (pp. 55-69). Berlin, Heidelberg: Springer Berlin Heidelberg.

118. Chollet, F. (2021). Deep learning with Python, ed. *Shelter Island: Manning Publications*.

119. Rossum, V. (2009). Python 3 reference manual. *(No Title)*.

120. Bisong, E. (2019). *Building machine learning and deep learning models on Google cloud platform* (pp. 59-64). Berkeley, CA: Apress.

121. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

122. Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137-1145).

123. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, *42*, 60-88.

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

125. 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. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

Annexes

Annexes

Annexe A: Code d'implémentation

python Copy Download

```python
model_simple = models.Sequential([
    layers.Rescaling(1./255, input_shape=(224, 224, 3)),
    layers.RandomFlip("horizontal"),
    layers.RandomRotation(0.1),
    layers.RandomZoom(0.1),
    layers.Conv2D(32, (3, 3), activation='relu'),
    layers.MaxPooling2D((2, 2)),
    layers.Conv2D(64, (3, 3), activation='relu'),
    layers.MaxPooling2D((2, 2)),
    layers.Flatten(),
    layers.Dense(128, activation='relu'),
    layers.Dropout(0.5),
    layers.Dense(4, activation='softmax')
])
```

Annexe B: les etapes d'adaptation de VGG16

Etape1 :

```python
base_model = VGG16(
    weights='imagenet',
    include_top=False,
    input_shape=(224, 224, 3)
)
```

Etape 2 :

```python
# Geler toutes les couches
base_model.trainable = False
```

Etape 3 :

```
model = models.Sequential([
    base_model,
    layers.GlobalAveragePooling2D(),
    layers.Dense(256, activation='relu'),
    layers.BatchNormalization(),
    layers.Dropout(0.5),
    layers.Dense(128, activation='relu'),
    layers.Dropout(0.3),
    layers.Dense(4, activation='softmax')
])
```

Etape 4 :

```
# 9. COMPILATION
# ==============================================
model.compile(
    optimizer=Adam(learning_rate=0.0001),
    loss='categorical_crossentropy',
    metrics=['accuracy']
)
```

Annexe c : deux phases d'entraînement RESNET50

```
print("\n=== PHASE 1: Transfer Learning (toutes les couches gelées) ===\n")

base_model.trainable = False

model = models.Sequential([
    base_model,
    layers.GlobalAveragePooling2D(),
    layers.Dense(512, activation='relu'),
    layers.Dropout(0.5),
    layers.Dense(256, activation='relu'),
    layers.Dropout(0.3),
    layers.Dense(4, activation='softmax')
])

model.compile(optimizer=Adam(0.001), loss='categorical_crossentropy', metrics=['accuracy'])

checkpoint = ModelCheckpoint(
    os.path.join(save_dir, 'resnet50_phase1_best.keras'),
    monitor='val_accuracy',
    save_best_only=True,
    verbose=1
)

model.fit(train_ds, validation_data=val_ds, epochs=15, callbacks=[checkpoint], verbose=1)
```

```
base_model = ResNet50(weights='imagenet', include_top=False, input_shape=(224,224,3))
base_model.trainable = False

model = models.Sequential([
    base_model,
    layers.GlobalAveragePooling2D(),
    layers.Dense(512, activation='relu'),
    layers.Dropout(0.5),
    layers.Dense(256, activation='relu'),
    layers.Dropout(0.3),
    layers.Dense(4, activation='softmax')
])

model.compile(optimizer=Adam(0.0001), loss='categorical_crossentropy', metrics=['accuracy'])

# Entraînement 20 époques
model.fit(train ds, validation data=val ds, epochs=20,
```

Annexe D : Paramètres d'entrainement

L'optimizer :

```
optimizer=Adam(learning_rate=0.0001),
```

Batch size :

```
batch_size=32,
```

EarlyStopping :

```
)

early_stop = EarlyStopping(
    monitor='val_loss',
    patience=10,
    restore_best_weights=True,
    verbose=1
)
```

Fonction de perte :

```
model.compile(optimizer=Adam(0.0001), loss='categorical_crossentropy', metrics=['accuracy'])
```

Dropout :

```
layers.Dropout(0.5),
```

Annexe E : Méthodes d'évaluation

Recall :

```
recall_macro = recall_score(y_true, y_pred, average='macro')
recall_weighted = recall_score(y_true, y_pred, average='weighted')
recall_per_class = recall_score(y_true, y_pred, average=None)
```

F1 score :

```
# calcul des métriques
f1 = f1_score(y_true, y_pred, average='weighted')
precision = precision_score(y_true, y_pred, average='weighted')
recall = recall_score(y_true, y_pred, average='weighted')

# F1-score par classe
f1_per_class = f1_score(y_true, y_pred, average=None)
```

La matrice (pour 4 classes) :

```
class_names = ['Glioma', 'Meningioma', 'No Tumor', 'Pituitary'

plt.figure(figsize=(10, 8))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
            xticklabels=class_names,
            yticklabels=class_names)

plt.xlabel('Labels Prédits (التصنيف المتوقع)')
plt.ylabel('Vrais Labels (التصنيف الحقيقي)')
plt.title('Matrice de Confusion - Brain Tumor Detection')
plt.show()
```

Courbes ROC et AUC :

```
y_true_bin = label_binarize(y_true, classes=[0, 1, 2, 3])

# Calculer les courbes ROC pour كل صنف
fpr = dict()
tpr = dict()
roc_auc = dict()

for i in range(n_classes):
    fpr[i], tpr[i], _ = roc_curve(y_true_bin[:, i], y_scores[:, i])
    roc_auc[i] = auc(fpr[i], tpr[i])

# AUC moyen
macro_auc = roc_auc_score(y_true_bin, y_scores, average='macro')
```

Annexe F : Étude qualitative des prédictions

```
plt.figure(figsize=(10, 8))
colors = cycle(['blue', 'red', 'green', 'orange'])

for i, color in zip(range(n_classes), colors):
    plt.plot(fpr[i], tpr[i], color=color, lw=2,
             label=f'ROC curve of {class_names[i]} (AUC = {roc_auc[i]:.2f})')

# رسم خط التصنيف العشوائي
plt.plot([0, 1], [0, 1], 'k--', lw=2)
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate (Taux de Faux Positifs)')
plt.ylabel('True Positive Rate (Taux de Vrais Positifs)')
plt.title('Receiver Operating Characteristic (ROC) - Multi-class')
plt.legend(loc="lower right")
plt.grid(alpha=0.3)
plt.show()
```