Normandie Université

UNIVERSITÉ DE ROUEN NORMANDIE

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité **INFORMATIQUE**

Préparée au sein de l'**Université de Rouen Normandie**

Etudes des fonctions de pertes basées sur l'entropie dans le contexte de l'apprentissage profond pour la radiomique.

Présentée et soutenue par

BROCHET THIBAUD

Thèse soutenue le 16/07/2024

devant le jury composé de :

MME RUAN SU	Professeur des Universités - URN - Université de Rouen Normandie	Directeur de thèse
M. MERIAUDEAU FABRICE	Professeur des Universités - Universite Bourgogne Europe	Président du jury
M. LAPUYADE-LAHORGUE JEROME	Maître de Conférences - URN - Université de Rouen Normandie	Membre du jury
MME GARREAU MIREILLE	Professeur des Universités - UNIVERSITE RENNES 1	Rapporteur du jury
M. ZHU YUEMIN	Directeur de Recherche - Institut National des Sciences Appliquees de Lyon	Rapporteur du jury

Thèse dirigée par **RUAN SU** (LABORATOIRE D'INFORMATIQUE DE TRAITEMENT DE L'INFORMATION ET DES SYSTEMES)

MIIS

litis

Remerciements

Je tiens à exprimer ma profonde gratitude à tous ceux qui ont contribué à la réalisation de cette thèse. Tout d'abord, un immense merci à mon directeur de thèse, le Professeur Su RUAN, dont les conseils, l'expertise et le soutien ont été inestimables tout au long de ce parcours. Sa capacité à guider mes explorations dans le vaste domaine des réseaux de neurones et des fonctions de perte entropiques a été fondamentale pour la concrétisation de ce travail. Je tiens à remercier également mon co-directeur de thèse, le Maître de Conférences Jérôme LAPUYADE-LAHORGUE, dont le soutien indéfectible et l'ardeur au travail m'ont inspiré et toujours mené sur le droit chemin. Je suis également reconnaissant envers le comité de thèse, composé de Mme GARREAU Mireille et M. ZHU Yuemin en tant que rapporteurs, ainsi que M. MERIAUDEAU Fabrice, Président du jury pour leurs retours précieux et constructifs qui ont grandement enrichi mon analyse et ma réflexion.

Mes remerciements vont également à mes collègues du laboratoire LITIS, pour l'environnement stimulant qu'ils ont su créer. Leurs questions et discussions ont souvent ouvert la voie à de nouvelles avenues de recherche et ont renforcé la rigueur scientifique de mon travail. Un remerciement spécial à AMYAR Amine et ZHOU Tongxue, avec qui j'ai partagé de nombreux défis et idées durant ces années de recherche.

Je souhaite aussi remercier l'Université de Rouen, pour avoir financé ce projet. Cette aide financière a été essentielle pour mener à bien mes recherches. Enfin, je ne saurais conclure sans adresser une pensée particulière à ma famille et mes amis, qui ont été des sources constantes de motivation et de soutien émotionnel. Leur présence et leurs encouragements ont été des piliers sur lesquels je me suis appuyé dans les moments les plus exigeants de ce parcours académique.

À tous, je vous suis profondément reconnaissant.

Table des matières

Introduction générale **7**

1 Contextes médical et scientifique **9**
1.1 Introduction . . . 9
1.2 Imagerie médicale . . . 11
1.2.1 Imagerie par résonance magnétique (IRM) . . . 11
1.2.2 Tomodensitométrie (TDM) . . . 12
1.2.3 Tomographie par émission de positons (TEP) . . . 13
1.3 Radiomique : exploration des caractéristiques quantitatives dans les images médicales . . . 14
1.3.1 Principes fondamentaux de la radiomique . . . 14
1.3.2 Caractéristiques radiomiques . . . 15
1.3.3 Applications de la radiomique . . . 18
1.3.4 Défis et perspectives de la radiomique . . . 19
1.4 Radiomique avec l'intelligence artificielle . . . 19
1.4.1 Apprentissage automatique pour l'analyse radiomique . . . 20
1.5 Conclusion . . . 24

2 Apprentissage profond et radiomique : état de l'art **27**
2.1 Introduction . . . 27
2.2 Principes fondamentaux des réseaux de neurones . . . 27
2.2.1 Généralité sur les réseaux de neurones . . . 28
2.3 Différentes architectures de réseaux de neurones profonds . . . 32
2.3.1 Couches de convolution . . . 32
2.3.2 Couches de "pooling" . . . 33
2.3.3 Couches entièrement connectées (dense) . . . 34
2.3.4 Couches de normalisation . . . 35
2.3.5 Couches de "dropout" . . . 36
2.4 Architectures d'apprentissage profond utilisées en radiomique . . . 37
2.4.1 Les Réseaux de Neurones convolutifs (CNN) . . . 37

2.4.2 Auto-encodeurs . . . 38
2.4.3 U-Net . . . 39
2.4.4 Réseaux multi-tâches . . . 40
2.4.5 Réseaux de neurones de type "Vision Transformer"(ViT) 41
2.4.6 Réseaux de neurones de type "Swin Transformers" . . . 42
2.5 Conclusion . . . 44

3 Fonctions de perte basées sur les entropies 45
3.1 Généralités sur l'entropie . . . 45
3.1.1 Introduction . . . 45
3.1.2 Entropie : de la définition de Clausius à celle de Shannon 46
3.1.3 Généralisations de l'entropie de Shannon . . . 47
3.2 Fonctions de perte basées sur les entropies : divergences et entropies-croisées . . . 51
3.2.1 Généralisations des divergences . . . 51
3.2.2 Entropies croisées . . . 51
3.2.3 Fonctions de perte basées sur les entropies : le choix des entropies croisées . . . 52
3.3 Conclusion . . . 53

4 Études de l'entropie de Havrda-Charvat dans le contexte de l'apprentissage profond pour la prédiction de l'évolution du cancer 55
4.1 Introduction . . . 55
4.2 Classification d'images vidéos . . . 55
4.2.1 Introduction . . . 55
4.2.2 Travaux connexes . . . 57
4.2.3 Architecture CNN pour la classification . . . 58
4.2.4 Résultats et analyse . . . 60
4.2.5 Conclusion . . . 62
4.3 Prédiction de la réponse au traitement . . . 64
4.3.1 Introduction . . . 64
4.3.2 Jeux de données . . . 64
4.3.3 Architecture multitâches . . . 65
4.3.4 Architecture multitâches et multidonnées pour la prédiction de la réponse au traitement . . . 69
4.4 Prédiction de la récidive . . . 73
4.4.1 Introduction . . . 73
4.4.2 Jeu de données . . . 74
4.4.3 Architecture de réseau de neurones pour la prédiction de récidive . . . 74
4.4.4 Résultats . . . 74
4.5 Discussion sur l'estimation du paramètre de l'entropie croisée d'Havrda-Charvat . . . 75
4.6 Conclusion . . . 77

5 Swin Transformers avec bloc d'attention pour la classification d'images de patients néonataux **79**
5.1 Introduction 79
5.2 Swin Transformers pour la classification d'images médicales . . . 81
5.2.1 Introduction 81
5.2.2 Méthodes 82
5.2.3 Expérimentations 85
5.2.4 Discussion 87
5.2.5 Conclusion 87

Conclusion générale **89**

Perspectives **91**

Liste des publications durant ma thèse **93**

Bibliographie **95**

Introduction générale

L'essor de l'intelligence artificielle, en particulier de l'apprentissage profond, a transformé de nombreux domaines scientifiques et industriels. Dans le secteur de la santé, l'apprentissage profond a permis de développer des modèles prédictifs de plus en plus précis pour la progression des maladies, ouvrant la voie à des diagnostics plus rapides et à des traitements personnalisés et permettant aux médecins de s'affranchir de tâches chronophages. Cependant, malgré ces avancées, certains défis subsistent, notamment en termes de gestion de petites bases de données et de la capacité à capturer la complexité des caractéristiques des maladies dans les modèles prédictifs.

La radiomique est une méthode innovante en imagerie médicale qui consiste à extraire une grande quantité de caractéristiques quantitatives des images cliniques standard, telles que les IRM, les scanners CT ou les images PET. En analysant minutieusement ces données, la radiomique permet d'obtenir des informations précieuses sur la texture, la forme, et la structure des tissus biologiques. Ces caractéristiques, souvent invisibles à l'œil nu, peuvent être utilisées pour créer des modèles prédictifs qui aident à diagnostiquer des maladies, à prédire leur évolution, et à personnaliser les traitements.

Basés sur ces principes, l'objectif principal de cette recherche est de développer des architectures de réseaux de neurones profonds des fonctions de perte qui améliorent la performance des modèles d'apprentissage profond dans la prédiction de la progression des maladies.

Dans cette thèse, nous avons proposé des architectures multitache et Transformer dans le cadre de la radiomique. Quant à la fonction de perte, l'utilisation de l'entropie de Havrda-Charvat comme base peut non seulement améliorer la précision des prédictions mais aussi offrir une meilleure gestion des cas minoritaires ou atypiques, souvent négligés par les approches classiques.

Cette thèse se concentre sur l'étude des fonctions de perte basées sur l'entropie croisée de Havrda-Charvat, cette entropie généralise l'entropie croisée de Shannon habituellement utilisée en apprentissage profond. L'entropie de Havrda-Charvat est une entropie paramétrique et l'ajustement de son para-

mètre permet d'adapter des réseaux de neurones profonds à la forme de la vraisemblance des données.

Dans les chapitres suivants, nous détaillons les architectures de réseaux utilisées, les fonctions de perte proposées, et les résultats obtenus, illustrant l'efficacité de notre approche dans divers scénarios de prédiction de maladies.

Le plan de ce manuscrit est organisé comme suit.

Le premier chapitre décrit le contexte médical et scientifique, notamment la description des modalités d'imagerie ainsi que la radiomique.

Le deuxième chapitre présente les bases des réseaux de neurones en apprentissage profond ainsi que les différentes architectures utilisées en radiomique.

Le troisième chapitre développe plus particulièrement notre méthodologie concernant les fonctions de perte basées sur les entropies croisées. Dans ce chapitre, nous détaillons également la motivation derrière la notion d'entropie et les possibilités de généralisation de l'entropie usuelle de Shannon.

Le quatrième chapitre présente le travail concernant les fonctions de perte dérivées de l'entropie généralisée d'Havrda-Charvat. Nous avons évalué notre nouvelle fonction de perte sur plusieurs bases de données avec différents types de cancer. Les résultats obtenus sont très prometteurs.

Le cinquième chapitre présente quant à lui le travail effectué pour permettre la classification des images IRM de nouveau-nés atteints d'encéphalopathie ischémique hypoxique en fonction de leur sévérité via l'utilisation d'un réseau de neurones basé sur le concept de Swin Transformer. Les améliorations obtenues par rapport aux architectures qui composent l'état de l'art sont également notables.

Au cours de cette thèse, nous avons pu contribuer aux recherches en matière d'analyse d'images médicales via des adaptations des réseaux de neurones et de nouvelles fonctions de perte pour obtenir une plus grande efficacité. Les résultats obtenus permettent de constater une amélioration des performances à travers les modifications proposées.

CHAPITRE 1

Contextes médical et scientifique

Ce premier chapitre introduit la notion de la radiomique et des méthodes d'imagerie de routine pour son étude.

1.1 Introduction

L'anticipation du traitement des cancers grâce à des approches avancées de traitement et de prédiction représente une frontière prometteuse dans la lutte contre ces maladies complexes et multifacettes. Avec l'évolution rapide des technologies biomédicales et informatiques, notamment dans les domaines de l'apprentissage profond et de la médecine personnalisée, les professionnels de la santé disposent désormais d'outils plus sophistiqués pour détecter, traiter et prédire l'évolution des cancers avec une précision sans précédent. Cette évolution marque un tournant dans la manière dont les cancers sont appréhendés, passant d'une approche réactive, centrée sur le traitement des cancers déjà développés, à une stratégie proactive et prédictive. Cette dernière vise non seulement à identifier les cancers à un stade précoce, lorsque les options de traitement sont plus nombreuses et souvent plus efficaces, mais aussi à personnaliser les interventions en fonction des caractéristiques génétiques et moléculaires spécifiques des tumeurs de chaque patient. En exploitant la puissance de l'analyse de données massives et des modèles prédictifs, les chercheurs et les cliniciens espèrent pouvoir anticiper la progression de la maladie, optimiser les schémas thérapeutiques et minimiser les effets secondaires, ouvrant ainsi la voie à une ère de traitement du cancer plus ciblée et plus efficace. Cette approche intégrée, qui combine les avancées en imagerie médicale, en génomique, en protéomique et en informatique, promet non seulement d'améliorer les résultats pour les patients atteints de cancer, mais aussi de transformer fondamentalement notre compréhension et notre gestion de cette maladie.

Les disciplines "omiques" représentent un ensemble de domaines d'étude

en biologie qui se concentrent sur l'analyse globale des différents types de molécules biologiques pour comprendre leur rôle, leur structure, leur fonction et leur interaction dans les systèmes biologiques. Chaque discipline "omique" apporte une perspective unique sur la biologie des organismes, contribuant à une compréhension holistique de la vie au niveau moléculaire.

La génomique est l'étude du génome, l'ensemble complet de l'ADN d'un organisme, incluant tous ses gènes. La génomique vise à cartographier, séquencer et analyser les gènes et leur séquence dans le but de comprendre leur structure, fonction et impact sur l'organisme. Elle joue un rôle crucial dans l'identification des variations génétiques associées à des maladies, la compréhension des bases héréditaires des traits et le développement de nouvelles approches thérapeutiques.

La transcriptomique se concentre sur l'étude du transcriptome, l'ensemble complet des ARN messagers (ARNm) produits dans une cellule. Elle examine comment les gènes sont exprimés sous différentes conditions ou dans différents types de cellules, fournissant des informations sur la régulation génique et les réponses cellulaires aux stimuli internes et externes. La transcriptomique est essentielle pour comprendre les mécanismes moléculaires des maladies et identifier de nouvelles cibles thérapeutiques.

La protéomique est l'étude du protéome, l'ensemble de toutes les protéines exprimées par le génome d'un organisme à un moment donné. Elle vise à caractériser la structure, la fonction et les interactions des protéines, offrant une perspective dynamique sur les processus biologiques. La protéomique est utilisée pour identifier les biomarqueurs de maladies, étudier les voies de signalisation cellulaire et développer des médicaments ciblés.

La métabolomique se concentre sur l'analyse globale du métabolome, l'ensemble des petits métabolites (comme les lipides, les acides aminés et les sucres) présents dans une cellule, un tissu ou un organisme. Elle permet de comprendre les réponses métaboliques aux changements environnementaux, les perturbations métaboliques associées aux maladies et les interactions entre le métabolisme et la régulation génétique.

L'épigénomique étudie les modifications épigénétiques, des changements chimiques de l'ADN et des histones qui régulent l'expression des gènes sans altérer la séquence d'ADN elle-même. Ces modifications, telles que la méthylation de l'ADN et les modifications des histones, jouent un rôle crucial dans le développement, la différenciation cellulaire et la pathogenèse des maladies. L'épigénomique offre des perspectives sur la manière dont les facteurs environnementaux peuvent influencer l'expression génique et être impliqués dans les maladies héréditaires et acquises.

Récemment la radiomique venant des images médicales est apparue comme étant une nouvelle discipline "omique". Les modalités d'imagerie médicale représentent les différentes techniques utilisées pour visualiser l'intérieur du corps humain, chacune ayant ses propres applications, avantages et limitations. Ces techniques sont cruciales pour le diagnostic, la planification du traitement et le suivi des patients dans de nombreux domaines de la médecine.

Les disciplines "omiques" fournissent une vue d'ensemble puissante et intégrée des systèmes biologiques, révolutionnant notre compréhension de la biologie moléculaire et ouvrant la voie à des avancées dans la recherche biomédicale, la médecine personnalisée et la biotechnologie. En combinant les perspectives de ces différentes disciplines, les scientifiques peuvent aborder des questions biologiques complexes avec une compréhension plus complète des mécanismes de la vie.

Cette thèse vise donc à construire des modèles prédictifs en développant des nouvelles approches basées sur l'apprentissage profond à partir des images médicales.

1.2 Imagerie médicale

L'imagerie médicale joue un rôle clé dans le diagnostic des cancers. La tomodensitométrie (TDM) et l'imagerie par résonance magnétique (IRM) fournissent des images détaillées des tissus et des structures, tandis que la tomographie par émission de positons (TEP) des informations sur le métabolisme des tissues, permettant aux médecins de localiser précisément les tumeurs, d'évaluer le grade des tumeurs. Nous nous intéressons à ces trois types d'imagerie médicales dans cette thèse.

1.2.1 Imagerie par résonance magnétique (IRM)

L'IRM est une technique d'imagerie médicale non invasive qui permet d'obtenir des images détaillées des structures internes du corps pour obtenir des images des parties non-osseuses ou des tissus mous du corps, en particulier du cerveau, des muscles, des articulations, des organes abdominaux, et également des tumeurs [36]. Elle est souvent utilisée pour la détection de certaines maladies, le diagnostic et le suivi d'un traitement.

Le fonctionnement de l'IRM repose sur le phénomène de la résonance magnétique nucléaire. Dans un champ magnétique puissant, les noyaux d'hydrogène dans les tissus corporels réagissent aux impulsions radiofréquences, absorbent de l'énergie et émettent ensuite des signaux détectables. Lorsqu'un patient est placé dans un tunnel entouré d'un champ magnétique intense (environ 4 Tesla) et constant, les noyaux d'hydrogène de son corps s'alignent avec ce champ. Des impulsions radiofréquences spécifiques sont émises dans le corps, perturbant temporairement l'alignement des noyaux d'atomes comme ceux d'hydrogène. La radiofréquence employée par le procédé dépend du type d'atomes ciblé. Après l'arrêt de l'impulsion, les noyaux d'atomes reviennent à leur alignement d'origine, émettant des signaux détectables dans le processus. Dans le cas où les atomes sont ceux d'hydrogène, les séquences **T1-pondérées** (T1) produisent des images où les substances grasses apparaissent brillantes et l'eau relativement sombre. Cette modalité est particulièrement utile pour visualiser la structure anatomique du cerveau et la moelle épinière, ainsi que pour détecter les anomalies des tissus mous, telles que des tumeurs. Les séquences

T2-pondérées (T2) rendent l'eau et les fluides dans le corps très brillants, ce qui les rend idéales pour détecter les lésions remplies de liquide, l'inflammation, les infections et l'œdème. Les tissus gras apparaissent moins brillants dans ces séquences par rapport à T1. La séquence **FLAIR** (Fluid Attenuated Inversion Recovery) est une variante de l'image T2-pondérée qui supprime le signal de l'eau libre, rendant le liquide cérébrospinal (LCS) sombre. Cette particularité permet de mieux distinguer les lésions ou les anomalies près des zones remplies de liquide, comme les ventricules cérébraux, ce qui est particulièrement utile pour diagnostiquer des maladies telles que la sclérose en plaques, les lésions ischémiques et les infections. Chacune de ces séquences d'IRM a des applications spécifiques et est choisie en fonction de l'objectif de l'examen, permettant ainsi aux médecins d'obtenir les informations les plus pertinentes pour un diagnostic précis et un plan de traitement adapté. Les signaux émis sont capturés par des antennes spéciales, convertis en données numériques, et utilisés pour reconstruire des images en trois dimensions du site examiné. L'IRM offre une résolution élevée, permettant de visualiser des détails anatomiques fins, ce qui en fait une modalité d'imagerie précise pour le diagnostic.

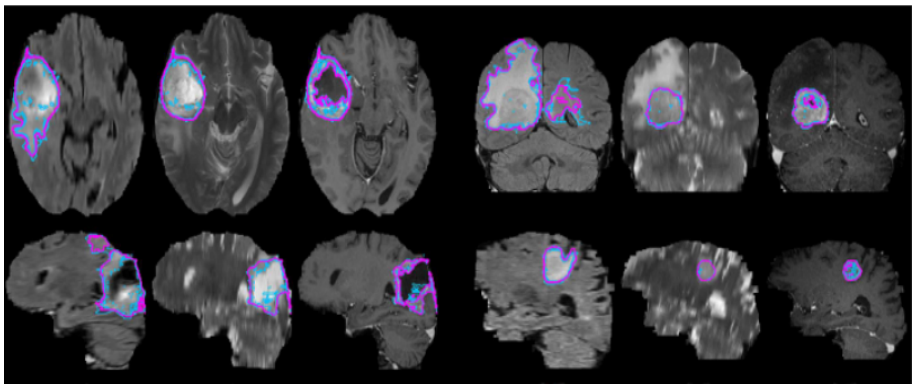


FIGURE 1.1 – Trois types d'images IRM : FLAIR, T2 et T1 montrées en différents points de vue (BratS Database http://www.braintumorsegmentation.org)

L'IRM, avec sa capacité à fournir des images détaillées et un contraste exceptionnel, demeure une modalité essentielle en radiologie diagnostique. Son utilisation continue de s'étendre, soutenant le diagnostic précoce, la planification du traitement, et la surveillance des patients dans divers domaines médicaux.

1.2.2 Tomodensitométrie (TDM)

La technique d'imagerie médicale appelée tomodensitométrie, ou scanographie, CT-scan en anglais, consiste à monitorer l'absorption de rayons X par les tissus vivants, puis via un ordinateur, de reconstruire les images bidimensionnelles ou tridimensionnelles de l'anatomie [42]. Pour obtenir ces images, on réalise un balayage de rayons X sur le patient par coupes ou de façon tomo-

graphique. Les intensités collectées sont ensuite interprétées par l'ordinateur, qui recompose des coupes en 2D puis des volumes en 3D à l'aide d'algorithme de reconstruction comme par exemple la transformée de Radon. Parfois, pour faire ressortir certains tissus, comme les vaisseaux sanguins, on a recours à un produit appelé "produit de contraste" qui absorbe les rayons X et rendent donc visibles les tissus qui concentrent ce produit. Depuis les années 1990, grâce à tomodensitomètre à multi-détecteurs et dont la table se déplace durant l'acquisition, on peut visualiser une large partie du corps humain. Bien entendu, comme pour toute radiographie, l'exposition répétée à des rayons X peut être délétère pour le corps humain. Simplement, une étude du rapport bénéfice/risque voit fortement pencher la balance en faveur de la TDM, dans le cadre d'un examen justifié. L'étude des structures osseuses permise par cet outil en fait un complément très précieux pour l'IRM et son étude des tissus mous. La figure 1.2 montre des images TDM. Bien que la tomodensitométrie ait une résolution spatiale élevée, le contraste dans les tissus mous est faible.

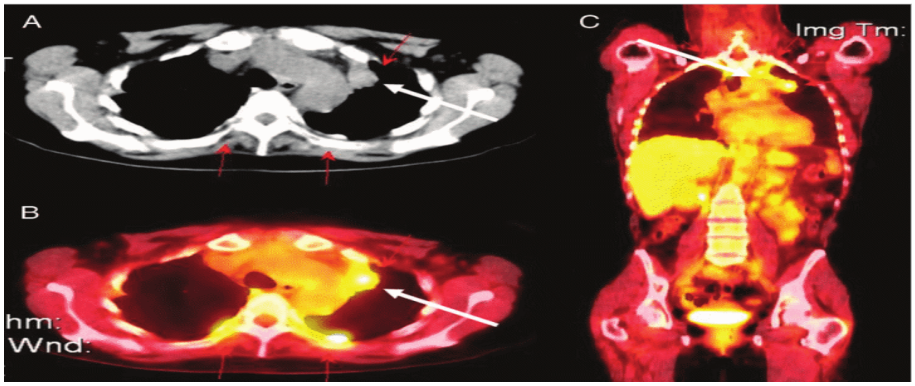


FIGURE 1.2 – Images TDM/TEP. A : image TDM, B : image TEP, C : fusion des deux images montées en direction coronale [112]

1.2.3 Tomographie par émission de positons (TEP)

L'imagerie TEP utilise l'association d'un traceur, utilise un traceur radioactif à émission β^+. Par exemple de "fluorodesoxyglucose 18" (FDG-18) qui est obtenu à partir du glucose en remplaçant les atomes d'oxygène 18 stables par des atomes de fluor 18 instables ; 18 étant le nombre total de nucléons de ces deux atomes. Le FDG-18 se comporte comme le glucose en termes d'absorption par les cellules consommatrices de glucose. Cependant, la substitution d'atome d'oxygène par atome de fluor l'empêche de sortir de la cellule et d'être métabolisé dans le cycle de Krebs après phosphorylation. Ce traceur est le plus utilisé pour détecter les cellules fortement consommatrices de glucose comme celles généralement présentes dans les tumeurs cancéreuses.
Les traceurs utilisés dans les examens TEP sont des émetteurs radioactifs β^+, ils sont instables par excès de protons. Le retour à la stabilité s'effectue par

émission de positrons, appelés aussi positons ou anti-électrons. Ce sont des particules de même masse que l'électron, de même nombre baryonique (B=0) et de charge et nombre leptonique opposés (q=+e et L=-1). Une fois l'énergie des positons déchargée par un court trajet dans la matière (0.6mm en moyenne) les positons vont s'annihiler avec des électrons, ce qui donnera deux photons gamma émis à 180 degrés l'un de l'autre. Ce sont ces photons qui seront captés par des détecteurs opposés et donneront ce qui est appelé une ligne de réponse. En utilisant ensuite les algorithmes de reconstruction adaptés, on obtient la position de fixation du FDG. Les deux photons ne sont détectables qui si leur énergie est de 511 keV et s'ils sont détectés dans un intervalle de temps très court. De plus, les positons émis ont un parcours très limité, de l'ordre du millimètre, ce qui permet d'obtenir une carte de fixation proche de la réalité. Comme la TEP est une modalité d'imagerie fonctionnelle, l'association d'une TEP à un TDM permet de superposer des modalités fonctionnelles et anatomiques, et donc de mieux repérer les organes touchés et l'étendue du cancer. La figure 1.2 montre des imafes TDM/TEP d'un patient atteint d'un cancer du poumon.

1.3 Radiomique : exploration des caractéristiques quantitatives dans les images médicales

La radiomique [53, 56, 103, 120] représente une avancée significative dans l'analyse d'images médicales en extrayant des informations quantitatives complexes à partir de données visuelles. Cette approche novatrice repose sur l'idée que les caractéristiques subtiles et non perceptibles à l'œil nu dans les images médicales peuvent fournir des informations vitales pour la prise de décision clinique.

1.3.1 Principes fondamentaux de la radiomique

La radiomique implique l'extraction d'un grand nombre de caractéristiques quantitatives à partir des images médicales, telles que la texture, la forme, l'intensité et d'autres paramètres mesurables. Les caractéristiques radiomiques peuvent servir de base au développement de biomarqueurs spécifiques à une maladie. Ces biomarqueurs peuvent être utilisés pour évaluer la progression de la maladie, la réponse au traitement, et même prédire les résultats cliniques [10, 16, 55, 74]. L'extraction des caractéristiques radiomiques, en combinant les connaissances en imagerie médicale avec les techniques d'analyse de données avancées, offre un potentiel immense pour améliorer la prise en charge des patients, faciliter la recherche médicale et conduire à des avancées significatives dans la compréhension des maladies. Ces caractéristiques peuvent être également analysées pour identifier des schémas complexes et des relations qui pourraient être liés à des aspects physiologiques, pathologiques ou à la réponse

au traitement [28, 29].
La figure 1.3 montre les étapes principales de la radiomique. Plusieurs étapes nécessitent des méthodes de traitement d'images avancées, telle que la segmentation, la calibration, le filtrage, la fusion de données ou la classification.



FIGURE 1.3 – Etapes suivies par la radiomique [104]

1.3.2 Caractéristiques radiomiques

L'extraction des données [63] en radiomique est une phase clé du processus, consistant à identifier et à quantifier les caractéristiques radiomiques pertinentes à partir des images médicales prétraitées. Ces caractéristiques, qui décrivent des aspects variés tels que la texture, la forme, l'intensité et d'autres propriétés, sont essentielles pour la caractérisation des tissus, des lésions ou des structures anatomiques. Les caractéristiques calculées préalable sont dites caractéristiques "faites-main" (Figure 1.4) pour différencier celles extraites par les méthodes de l'apprentissage profond.

1. **Texture :** L'extraction des caractéristiques de texture vise à décrire les variations locales d'intensité dans une image. Des mesures telles que la matrice de co-occurrence ainsi que d'autres paramètres statistiques capturant des motifs complexes de textures, révélant des détails souvent invisibles à l'œil nu. Nous pouvons citer la matrice de co-occurrence qui



FIGURE 1.4 – Types de caractéristiques radiomiques [104]

décrivent les relations spatiales entre les niveaux de gris, la matrice de co-répartition qui évaluent la corrélation entre différentes échelles de niveaux de gris, le contraste qui mesure la variation des niveaux de gris entre les pixels, l'homogénéité qui indique la distribution homogène des niveaux de gris, l'énergie qui représente la quantité d'énergie dans l'image [109], L'analyse de texture par ondelettes [31] qui est une méthode puissante utilisée en traitement d'images et en reconnaissance de formes pour caractériser et différencier les textures dans une image et repose sur la transformation en ondelettes, un outil mathématique qui décompose une image en plusieurs composantes à différentes échelles et orientations, permettant ainsi une analyse fine des motifs et structures présents dans l'image. Les dérivées d'ordre non entier [68], souvent appelées dérivées fractionnaires, est une méthode avancée dans le domaine du traitement d'images et de l'analyse de texture. Cette technique s'appuie sur le concept de dérivation fractionnaire, qui étend l'idée traditionnelle de dérivées entières à des ordres fractionnaires, permettant une analyse plus fine et plus nuancée des variations de texture dans une image. L'analyse fréquentielle en radiomique vise à décomposer l'image en composantes fréquentielles. Des caractéristiques, comme la fréquence spatiale dominante, peuvent révéler des informations sur la distribution des motif périodiques dans les images. Deux outils sont souvent utilisés : la transformée de Fourier permettant d'analyser les composantes fréquentielles de l'image ; le filtre Gabor permettant de détecter des structures texturées à différentes échelles [59].

2. **Forme et géométrie :** Les structures sont quantifiées à travers des caractéristiques telles que l'aire, le périmètre, la connexité qui est une caractéristique géométrique importante qui décrit la manière dont les composants d'un objet sont reliés entre eux. Dans le contexte de l'analyse d'images, la connexité peut être utilisée pour décrire la forme et la structure des objets présents dans une image. Une région connexe est une région où chaque paire de points peut être reliée par un che-

min continu entièrement contenu dans cette région. Cette propriété est souvent utilisée dans le traitement d'images pour segmenter les objets et les identifier. Par exemple, dans la segmentation d'images médicales, la connexité peut être utilisée pour séparer les différentes structures anatomiques, telles que les organes, les tissus ou les lésions. De plus, la connexité peut être utilisée comme critère de qualité pour évaluer la précision des algorithmes de segmentation d'images. En résumé, la connexité est une caractéristique géométrique fondamentale qui trouve de nombreuses applications dans le domaine du traitement d'images, en particulier dans la segmentation et l'analyse d'objets complexes. et d'autres paramètres géométriques. Ces mesures fournissent des informations sur la morphologie des régions d'intérêt. Nous pouvons citer le volume qui mesure le volume tridimensionnel de la lésion, la surface qui évalue la surface externe de la lésion, le ratio d'aspect qui établit un rapport entre la longueur et la largeur de la lésion [108]. La transformée de distance calcule la distance minimale de chaque pixel à l'arrière-plan, fournissant une carte qui peut être utilisée pour extraire des caractéristiques géométriques comme les squelettes ou les centres des objets. La transformée en squelette réduit les objets à leurs formes squelettiques, préservant la topologie et simplifiant l'analyse des structures complexes. La morphologie topologique s'intéresse aux propriétés qui sont préservées sous des transformations continues, permettant d'étudier la connexité et les caractéristiques topologiques des objets.

3. **Intensité :** Les caractéristiques liées à l'intensité comprennent des statistiques simples telles que la moyenne, la médiane, l'écart-type, mais aussi des paramètres plus avancés comme l'histogramme des intensités. Ces caractéristiques reflètent les propriétés d'intensité des pixels dans les régions délimitées, telles que la moyenne qui représente la valeur moyenne des niveaux de gris dans la région d'intérêt, l'écart-type qui mesure la dispersion des niveaux de gris, l'asymétrie qui indique l'asymétrie de la distribution des niveaux de gris [82].

4. **Caractéristiques spatiales :** Les caractéristiques spatiales s'intéressent aux relations spatiales entre les pixels. Des paramètres tels que la distance moyenne entre les pixels ou l'orientation locale peuvent fournir des informations sur la structure spatiale des tissus [46].

5. **Caractéristiques temporelles :** Dans le cas d'images séquentielles ou dynamiques, l'extraction de caractéristiques temporelles devient pertinente. Cela peut inclure des paramètres de cinétique, des courbes d'intensité temporelle et d'autres mesures liées à l'évolution temporelle des structures [46].

6. **Méthodes d'apprentissage profond :** Avec l'avènement de l'apprentissage profond, des techniques telles que les réseaux de neurones convolutifs (CNN) sont de plus en plus utilisées pour l'extraction automatique de caractéristiques. Ces approches apprennent directement à partir des

données, évitant ainsi la nécessité d'une ingénierie manuelle des caractéristiques [50].

1.3.3 Applications de la radiomique

De nombreuses applications de la radiomique ont été développées au cours des dernières décennies [1, 7, 47] :

1. **Diagnostic et classification [15, 17] :** Les difficultés inhérentes [57, 76] à la détection précoce des cancers sont multiples et complexes, reflétant la nature insidieuse de ces maladies. L'extraction des caractéristiques radiomiques joue un rôle essentiel dans le diagnostic et la classification des lésions. En analysant les motifs complexes à partir d'images médicales, la classification permet de différencier les tissus normaux des anomalies. Cette application est particulièrement cruciale dans le domaine de l'oncologie, où elle aide à caractériser les tumeurs, à évaluer leur malignité et à guider les décisions de traitement.
2. **Prédiction de la réponse au traitement [67] :** Les caractéristiques radiomiques peuvent être utilisées pour prédire la réponse d'une tumeur à des traitements spécifiques, tels que la radiothérapie ou la chimiothérapie. En analysant les propriétés texturales et fonctionnelles des images, les modèles radiomiques peuvent fournir des informations précieuses sur la probabilité de réponse d'une tumeur à un traitement particulier.
3. **Évaluation de la progression de la maladie [35, 72, 113] :** La surveillance à long terme des patients implique souvent l'évaluation de la progression de la maladie. Les caractéristiques radiomiques permettent de détecter les changements subtils dans les images au fil du temps, facilitant la détection précoce de rechutes ou de nouvelles lésions.
4. **Différenciation des types tumoraux [4, 23, 75, 87] :** Dans le cas de tumeurs complexes, telles que les tumeurs cérébrales, les caractéristiques radiomiques aident à différencier les types tumoraux. Cette distinction est cruciale pour élaborer des plans de traitement personnalisés adaptés aux caractéristiques spécifiques de chaque tumeur.
5. **Guidage de la planification du traitement [36, 41, 42, 60] :** Les caractéristiques radiomiques fournissent des informations supplémentaires qui peuvent être intégrées dans la planification du traitement. Cela inclut l'optimisation de la dose de radiation pour minimiser les dommages aux tissus sains environnants tout en maximisant l'efficacité du traitement sur la tumeur.
6. **Détection précoce de maladies [102, 38] :** Au-delà du cancer, l'extraction des caractéristiques radiomiques peut être utilisée pour la détection précoce de diverses maladies. Par exemple, dans le domaine de l'imagerie thoracique, elle peut aider à identifier des signes précurseurs de maladies respiratoires ou cardiovasculaires. L'Encéphalopathie Hypoxique-Ischémique (HIE) est une maladie neurologique grave qui

survient chez les nouveau-nés en raison d'une diminution du flux sanguin et d'un apport d'oxygène insuffisant au cerveau, généralement lors de l'accouchement. Cette situation peut entraîner des lésions cérébrales significatives et des séquelles à long terme. La détection précoce de cette maladie est essentielle pour minimiser les dommages et améliorer les résultats de traitement à long terme.

7. **Suivi de l'évolution tumorale** [71, 96, 120] **:** L'évolution temporelle des caractéristiques radiomiques permet un suivi détaillé de l'évolution tumorale. Cela peut aider les cliniciens à ajuster les protocoles de traitement en temps réel, en fonction des changements observés dans les images radiologiques.

8. **Intégration avec données cliniques et génomiques** [97] **:** L'intégration des caractéristiques radiomiques avec des données cliniques et génomiques élargit la portée de l'analyse. Cette approche multimodale offre une compréhension plus holistique des facteurs qui influent sur la maladie et permet une médecine de précision plus personnalisée.

1.3.4 Défis et perspectives de la radiomique

Un défi majeur réside dans la standardisation des protocoles radiomiques pour assurer la reproductibilité et la comparabilité des résultats entre différentes études [12]. Un autre défi réside dans la compréhension et l'interprétabilité des relations complexes identifiées par la radiomique, ce qui nécessite une collaboration étroite entre les cliniciens et les experts en analyse d'images. La détection précoce des cancers reste aussi un défi majeur en médecine, nécessitant des avancées continues dans les technologies de dépistage, une meilleure compréhension des mécanismes biologiques du cancer, ainsi qu'une approche personnalisée et sensible au dépistage, adaptée aux caractéristiques individuelles et aux risques des patients.

1.4 Radiomique avec l'intelligence artificielle

Les approches d'intelligence artificielle en médecine sont de plus en plus utilisées et sont extrêmement prometteuses de par le nombre croissant et la variété des données qu'elles permettent d'exploiter. Ainsi, l'analyse informatique des imageries médicales en particulier, la radiomique a montré de nombreux résultats très intéressants pour l'évaluation du pronostic et de la réponse chez des patients atteints de cancer [94]. L'IA peut être particulièrement efficace pour identifier des modèles non linéaires et des relations complexes entre les caractéristiques radiomiques pouvant échapper aux méthodes d'analyse traditionnelles.

1.4.1 Apprentissage automatique pour l'analyse radiomique

Les méthodes d'apprentissage automatique, en particulier les méthodes d'apprentissage profond, ont été développées pour extraire automatiquement des caractéristiques radiomiques. Elles peuvent apprendre à reconnaître les motifs significatifs et les relations entre les caractéristiques à partir de grandes quantités de données d'entraînement. Ces méthodes permettent également de construire des modèles prédictifs basés sur des caractéristiques radiomiques [63]. Ces modèles peuvent être entraînés à anticiper des résultats cliniques ou à identifier des tendances de progression de la maladie.
On distingue deux approches possibles : 1) la radiomique basée sur des caractéristiques du type "faites-main", notamment les caractéristiques de texture, qui sont préalablement calculés sur l'image. 2) la radiomique basée sur l'apprentissage profond, qui ne nécessite pas de prédéfinir des caractéristiques.
Pour une démarche classique comme montrée dans la Figure 1.3, les caractéristiques sont calculées préalablement à partir de la Figure 1.4. Comme toutes les caractéristiques ne sont pas prédictives, la sélection de celles-ci est nécessaire. Dans le cas de l'approche d'apprentissage profond, cette étape, tout comme l'extraction des caractéristiques sont intégrées dans la démarche globale de l'algorithme.

Sélection des caractéristiques

La sélection des caractéristiques en radiomique est une étape vitale visant à identifier les caractéristiques les plus informatives et discriminantes parmi la multitude de mesures extraites [109]. Cette sélection, également connue sous le terme de réduction de dimensionnalité, vise à optimiser la performance des modèles radiomiques en éliminant le bruit, en améliorant la robustesse et en accélérant les analyses. La sélection des paramètres en radiomique est un processus délicat qui nécessite une évaluation minutieuse de l'impact de chaque caractéristique sur la qualité des résultats. Elle vise à garantir que seules les caractéristiques les plus pertinentes et informatives sont utilisées, contribuant ainsi à des modèles radiomiques plus précis et généralisables. Plusieurs méthodes peuvent être employées pour cette sélection, chacune adaptée à des contextes spécifiques :

— **Analyse de variance :** L'analyse de variance permet d'évaluer la variabilité des caractéristiques radiomiques entre différentes classes ou groupes. Les caractéristiques présentant une faible variance peuvent être éliminées, car elles apportent peu d'informations utiles pour la discrimination entre ces classes [25]. L'analyse de la variance (ANOVA) est un ensemble de méthodes statistiques utilisé pour déterminer si les différences entre les moyennes de plusieurs groupes sont statistiquement significatives. Elle permet de tester l'hypothèse selon laquelle toutes les moyennes de groupe sont égales contre l'alternative que au moins une diffère. L'ANOVA à un facteur est utilisée lorsque l'intérêt est porté

sur l'effet d'un seul facteur indépendant. L'ANOVA à deux facteurs ou plus, en revanche, examine les effets de plusieurs facteurs et leurs interactions potentielles sur la variable dépendante. L'ANOVA à mesures répétées est appliquée lorsque les mêmes sujets sont testés sous plusieurs conditions ou à plusieurs reprises. L'ANCOVA, une extension de l'ANOVA, ajuste l'impact de variables continues additionnelles, appelées covariables, pour analyser plus précisément l'effet des facteurs catégoriels. Enfin, la MANOVA est une version multivariée de l'ANOVA qui analyse simultanément plusieurs variables dépendantes. Chacune de ces méthodes requiert que certaines hypothèses soient respectées, telles que la normalité des distributions, l'homogénéité des variances et l'indépendance des observations, pour garantir la fiabilité des résultats obtenus.

— **Tests statistiques :** Les tests statistiques sont des procédures essentielles en statistique pour évaluer des hypothèses sur les données. Ces tests permettent de déterminer si les observations expérimentales reflètent un phénomène réel ou sont dues au hasard. Parmi les tests les plus courants, on trouve le test t de Student pour comparer les moyennes de deux groupes, le test ANOVA pour analyser les différences entre les moyennes de trois groupes ou plus, et le test du Chi-carré pour examiner les relations entre variables catégorielles. Pour les données appariées ou mesurées à plusieurs reprises, le test t apparié est utilisé. Les tests non paramétriques comme le test de Mann-Whitney et le test de Wilcoxon sont préférés lorsque les hypothèses de normalité ne sont pas satisfaites. Le choix du test dépend de la nature des données, de l'objectif de l'étude, et des hypothèses sous-jacentes comme la distribution des données et l'indépendance des observations. Ces outils statistiques sont cruciaux pour la prise de décision basée sur des données, permettant de tirer des conclusions fiables à partir d'ensembles de données complexes [81].

— **Techniques de sélection automatique :** Les techniques de sélection automatique, souvent employées en modélisation statistique et en apprentissage automatique, visent à identifier les meilleures variables explicatives ou caractéristiques pour un modèle prédictif. Parmi ces méthodes, la sélection séquentielle avant et arrière examine systématiquement l'ajout ou la suppression de variables pour optimiser la performance du modèle. Les méthodes régularisées, telles que LASSO et Ridge [100], ajoutent une pénalité à la complexité du modèle pour encourager la simplicité et éviter le surajustement, LASSO ayant la capacité particulière de réduire certains coefficients à zéro, effectuant ainsi une sélection de variables. La méthode d'élimination récursive des caractéristiques supprime itérativement les variables les moins importantes basées sur leur contribution à la performance du modèle. L'analyse des composantes principales (PCA) [64] réduit la dimensionnalité en créant de nouvelles variables non corrélées qui capturent l'essentiel de l'information dans les données. Ces techniques aident à simplifier les modèles,

à améliorer leur interprétabilité et à réduire le risque de surajustement, en sélectionnant uniquement les variables les plus pertinentes pour la prédiction [63].

— **Arbres de décision :** Les arbres de décision peuvent être utilisés pour évaluer l'importance des caractéristiques dans la construction des arbres. Les caractéristiques les plus fréquemment utilisées pour séparer les classes ont un poids plus élevé et sont donc considérées comme plus importantes [91].

— **Algorithmes d'apprentissage automatique :** Algorithmes d'Apprentissage Automatique : Les algorithmes d'apprentissage automatique sont des méthodes informatiques conçues pour permettre aux machines d'apprendre à partir de données et de faire des prédictions ou des décisions sans être explicitement programmées pour une tâche spécifique. Ils se répartissent en trois catégories principales : l'apprentissage supervisé, où le modèle est entraîné sur un ensemble de données étiquetées pour prédire la sortie à partir des entrées ; l'apprentissage non supervisé, qui explore les données non étiquetées pour trouver des structures cachées ou des regroupements ; et l'apprentissage par renforcement, où un agent apprend à prendre des décisions en exécutant des actions dans un environnement pour maximiser une récompense. Parmi les algorithmes populaires, on trouve les arbres de décision, les forêts aléatoires, et les réseaux de neurones pour l'apprentissage supervisé ; le clustering k-means et l'analyse des composantes principales pour l'apprentissage non supervisé ; et l'algorithme Q-learning pour l'apprentissage par renforcement. Ces algorithmes trouvent des applications dans divers domaines, allant de la reconnaissance d'image et de parole à la prédiction financière et au-delà, en offrant des solutions automatisées pour analyser de grands ensembles de données et en extraire des connaissances utiles [37].

— **Méthodes d'ensemble :** Des méthodes d'ensemble, telles que la méthode de Boosting, peuvent être utilisées pour combiner les prédictions de plusieurs modèles basés sur des sous-ensembles différents de caractéristiques. Cela permet de prendre en compte la diversité des caractéristiques importantes [30].

— **Validation croisée :** La validation croisée peut être utilisée pour évaluer la performance d'un modèle sur plusieurs ensembles de données. Cela permet de vérifier que la sélection des caractéristiques généralise bien aux différentes partitions de données [82].

Classification et prédiction

La radiomique est appliquée à diverses tâches de classification, notamment la discrimination entre tissus sains et pathologiques, la classification de différentes pathologies, la prédiction de réponses au traitement et d'autres scénarios diagnostiques. La classification /prédiction en radiomique [82] constitue l'une des applications les plus importantes de cette discipline, visant à discriminer et à catégoriser automatiquement des images médicales en fonction des ca-

ractéristiques radiomiques extraites, ou à prédire la survie d'un patient ou l'effet d'un traitement sur le patient. Cette classification automatisée offre un potentiel considérable pour la prise de décision clinique en fournissant des informations quantitatives sur la nature des tissus, des lésions ou des structures anatomiques.

La classification en radiomique repose généralement sur des techniques d'apprentissage supervisé, où le modèle est entraîné sur des ensembles de données annotées. Ces ensembles contiennent des images associées à des étiquettes de classe (par exemple, sain/malade), permettant au modèle d'apprendre à associer les caractéristiques radiomiques aux catégories pertinentes. Une variété d'algorithmes de classification peut être utilisée en radiomique, tels que les machines à vecteurs de support (SVM), les arbres de décision, les réseaux de neurones et les méthodes d'ensemble comme les forêts aléatoires ou les méthodes de boosting. Le choix de l'algorithme dépend souvent de la nature des données et de la complexité de la tâche. Pour évaluer la performance des modèles de classification radiomique, la validation croisée est souvent utilisée. Cette approche divise les données en ensembles d'entraînement et de test, permettant d'évaluer la généralisation du modèle sur des données non vues.

En conclusion, la classification en radiomique représente une avancée majeure dans le domaine de l'imagerie médicale. En exploitant les caractéristiques radiomiques, elle offre une approche automatisée et quantitative pour la prise de décision clinique, contribuant ainsi à une médecine plus précise et personnalisée.

Méthode d'apprentissage profond en radiomique

La radiomique est un champ de recherche en plein essor qui vise à extraire une grande quantité de caractéristiques quantitatives à partir d'images médicales, telles que les IRM, les TDM ou les PET scans, en utilisant des algorithmes avancés d'analyse d'images. L'intégration de méthodes d'apprentissage profond (apprentissage profond) dans la radiomique ouvre de nouvelles perspectives pour améliorer la précision diagnostique, la caractérisation des tumeurs, et la prédiction de l'évolution de la maladie et de la réponse au traitement.

1. **Extraction de caractéristiques automatisée :** Contrairement aux méthodes traditionnelles de radiomique qui reposent sur l'extraction manuelle de caractéristiques prédéfinies, l'apprentissage profond permet une extraction automatique et hiérarchique de caractéristiques directement à partir des images brutes. Les réseaux de neurones convolutionnels (CNN) sont particulièrement efficaces pour identifier des motifs complexes et des caractéristiques intrinsèques des images, sans intervention humaine, réduisant ainsi la subjectivité et améliorant la reproductibilité.
2. **Amélioration de la classification et de la segmentation :** Les modèles d' apprentissage profond peuvent être entraînés pour classer les

images en fonction de la présence ou de l'absence de caractéristiques pathologiques, ou pour segmenter précisément les régions d'intérêt, comme les tumeurs ou les lésions. Cette précision dans la classification et la segmentation facilite une analyse radiomique plus détaillée et plus fiable des caractéristiques tumorales.

3. **Prédiction et Pronostic :** En intégrant les caractéristiques radiomiques extraites par les méthodes d'apprentissage profond avec des données cliniques et génomiques, il est possible de développer des modèles prédictifs pour évaluer le pronostic des patients, prédire la réponse au traitement et personnaliser les stratégies thérapeutiques. Ces modèles peuvent identifier des signatures radiomiques associées à des issues spécifiques, offrant ainsi des insights précieux pour la prise de décision clinique.
4. **Défis et Considérations :** L'intégration de l'apprentissage profond dans la radiomique nécessite de grandes bases de données d'images annotées pour l'entraînement des modèles, ce qui peut être limité par des questions de confidentialité et d'accessibilité des données. De plus, l'interprétabilité des modèles d'apprentissage profond reste un défi, car il est souvent difficile de comprendre comment les décisions diagnostiques ou pronostiques sont dérivées à partir des caractéristiques extraites.

Conclusion : L'application de méthodes d'apprentissage profond en radiomique a le potentiel de transformer significativement la manière dont les images médicales sont utilisées pour la recherche et la pratique clinique en oncologie et au-delà. En permettant une extraction automatique et précise de caractéristiques, une segmentation améliorée et des analyses prédictives robustes, l'apprentissage profond renforce le rôle de la radiomique dans la médecine personnalisée et la gestion des maladies.

1.5 Conclusion

La radiomique offre une perspective passionnante dans le domaine de l'imagerie médicale, ouvrant la voie à une compréhension plus approfondie des caractéristiques subtiles présentes dans les images. Son potentiel à contribuer au diagnostic précoce, à la prédiction de la réponse au traitement et au suivi de la progression de la maladie en fait une discipline prometteuse pour l'amélioration de la prise en charge clinique.

Compte tenu de l'état de l'art actuel en intelligence artificielle et en analyse de données, l'accent est mis sur l'apprentissage profond en raison de ses performances supérieures dans une multitude de domaines. Cette préférence pour l'apprentissage profond découle de sa capacité à apprendre automatiquement et de manière hiérarchique des caractéristiques complexes à partir de grandes quantités de données, sans nécessiter une ingénierie manuelle des caractéristiques ou une expertise de domaine approfondie. Dans cette thèse, nous développons des méthodes basées sur l'apprentissage profond pour la prédic-

tion de la réponse au traitement et la récidive, ainsi que le diagnostic à partir des données image et clinique. Nous avons donc pris le parti d'étudier des nouvelles fonctions de perte et de nouvelles architectures.

CHAPITRE 2

Apprentissage profond et radiomique : état de l'art

2.1 Introduction

Comme nous l'avons présenté au chapitre 1, l'apprentissage profond joue un rôle important en radiomique. Dans ce chapitre, nous présentons les architectures d'apprentissage profond développées dans le domaine de la radiomique.

L'apprentissage profond est une sous-catégorie de l'intelligence artificielle (IA) et de l'apprentissage automatique, qui utilise des réseaux de neurones artificiels pour modéliser et traiter des ensembles de données complexes. Cette section explore les concepts clés de l'apprentissage profond, ses applications variées et son impact croissant dans de nombreux secteurs.

L'apprentissage profond est notamment utilisé en imagerie médicale pour la détection, la segmentation et le diagnostic des maladies, plus récemment la radiomique. Les réseaux de neurones profonds sont capables d'analyser des images médicales telles que les scans IRM, PET, CT, pour identifier des anomalies, comme les tumeurs ou les pathologies [5, 43, 44, 119] et aussi pour identifier les organes sains [101].

2.2 Principes fondamentaux des réseaux de neurones

Le concept de neurone artificiel est fondamental en intelligence artificielle et en apprentissage automatique. Inspiré par le fonctionnement des neurones biologiques du cerveau humain, un neurone artificiel est une unité de calcul qui traite des données d'entrée et génère une sortie en fonction de règles apprises pendant l'entraînement [51]. Un neurone artificiel se compose de trois éléments principaux : les entrées pondérées, la fonction d'activation et la sortie comme montré dans la Figure 2.1. Les entrées pondérées représentent les

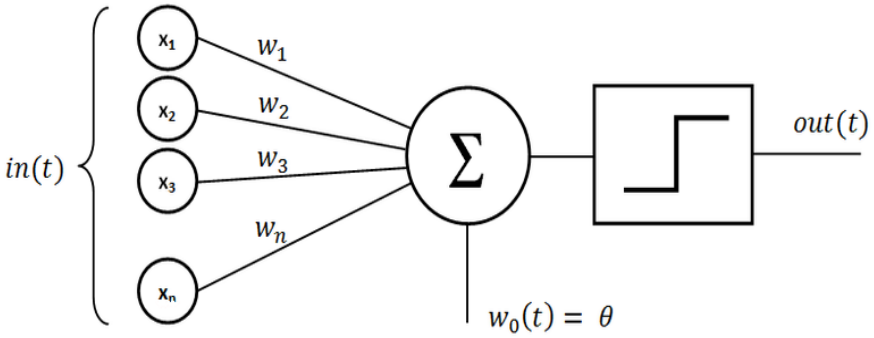


FIGURE 2.1 – Schéma du neurone artificiel (Perceptron)

données d'entrée, chacune multipliée par un poids correspondant. La somme pondérée de ces entrées est ensuite passée à travers une fonction d'activation, qui introduit de la non linéarité pour la sortie du neurone. Le fonctionnement d'un neurone artificiel est similaire à celui d'un neurone biologique. Les entrées pondérées représentent les signaux électriques reçus par les dendrites d'un neurone biologique, qui sont ensuite sommés dans le corps cellulaire. La fonction d'activation peut être vue comme le potentiel d'action déclenché dans le neurone biologique lorsque la somme des signaux dépasse un certain seuil. Enfin, la sortie du neurone artificiel correspond à l'activation ou non du neurone, simulant la propagation de l'information dans un réseau neuronal. L'apprentissage des poids d'un neurone artificiel se fait généralement par des algorithmes d'apprentissage supervisé ou non supervisé. Pendant l'entraînement, les poids sont ajustés itérativement pour minimiser une fonction de perte qui mesure l'écart entre les sorties réelles et les sorties prédites par le modèle. Ce processus permet au neurone d'apprendre à discriminer et à généraliser les données, lui permettant ainsi de mieux effectuer sa tâche.

Le concept de neurone artificiel est essentiel à la compréhension des réseaux de neurones et de l'apprentissage automatique. En simulant le comportement des neurones biologiques, les neurones artificiels permettent aux machines d'apprendre à partir de données et de réaliser une variété de tâches complexes de manière autonome.

2.2.1 Généralité sur les réseaux de neurones

Introduction

Un réseau de neurones est un ensemble de fonctions non-linéaires organisées en couche qui prend en entrée (première couche) des données à analyser et qui renvoie en sortie (dernière couche) une valeur scalaire ou vectorielle que l'on doit comparer à une valeur cible connue. Par exemple, en classification binaire, la valeur de sortie est un réel entre 0 et 1 que l'on compare avec la classe vérité terrain qui prend les valeurs 0 ou 1. En général, la valeur de sortie

du réseau en classification binaire est obtenue en utilisant une sigmoïde. Les neurones dépendent de paramètres que l'on estime au cours de l'apprentissage. Cet apprentissage est effectué de façon à minimiser la perte totale :

$$\sum_{n=1}^{N} L(y_n, \hat{y}_n), \tag{2.1}$$

où $L(y_n, \hat{y}_n)$ représente la perte individuelle pour la n-ième donnée où y_n est la vérité terrain et $\hat{y}_n$ l'estimation sortie du réseau de neurones. L est une fonction convexe et généralement quadratique appelée "fonction de perte". La procédure de minimisation se fait par une méthode itérative appelée "rétropropagation du gradient" que nous allons détailler ci-dessous. Avant toute chose, introduisons les notations mathématiques qui vont être utilisées par la suite dans la partie spécifications mathématiques d'un réseau de neurones.

Spécifications mathématiques d'un réseau de neurones

— Neurone : un neurone est une fonction f d'un espace vectoriel $\mathbb{R}^d$ vers $\mathbb{R}$ de la forme :

$$f(\vec{x}) = h\left(\sum_{k=1}^{d} w_k x_k + b\right) \tag{2.2}$$

où h est une fonction non-linéaire appelée "fonction d'activation", b est un paramètre appelé "biais" et les w_k sont des paramètres appelés "poids".

— Couche : une couche est un ensemble de neurones disposés en parallèle. Une couche à q neurones dont les entrées ont pour dimensions respectives $d_1, d_2, \ldots, d_q$ peut alors être modélisée comme une fonction de $\mathbb{R}^{d_1+\ldots+d_q}$,dans $\mathbb{R}^q$. Sans perte de généralités, quitte à imposer des poids nuls, on peut supposer que les entrées de chacun des neurones ont même dimension et que la dimension commune est égale au nombre de neurones de la couche précédente. Ceci permet alors de simplifier les calculs lors des détails de la rétropropagation.

— Réseau de neurones : nous nous donnons M le nombre de couches, $n_1, \ldots, n_M$ le nombre de neurones par couche. Les couches sont numérotées de 1 jusqu'à M où 1 est la couche la plus proche de l'entrée et M celle la plus proche de la sortie. Ainsi, la couche i est une fonction de $\mathbb{R}^{n_{i-1}}$ dans $\mathbb{R}^{n_i}$ où n_0 désigne le nombre d'entrées du réseau. Nous noterons $\vec{x}_i$ la sortie de la couche i avec pour convention $\vec{x}_0$ l'entrée du réseau et $\vec{\hat{y}} = \vec{x}_M$ sa sortie. Les paramètres de la couche i (poids et biais) sont regroupés en un paramètre multidimensionnel θ_i et l'ensemble des paramètres sera alors $\theta = \theta_{1:M}$.

— Fonction de perte : une fonction de perte est une fonction L de $\mathbb{R}^{n_M} \times \mathbb{R}^{n_M}$ dans $\mathbb{R}$, convexe et dérivable. Comme la sortie du réseau $\vec{\hat{y}}$ est une fonction de θ, la perte $L(\vec{y}, \vec{\hat{y}})$ est aussi une fonction de θ que l'on notera $\mathcal{L}$.

L'apprentissage consiste à estimer les paramètres résumés par θ de façon à ce que la sortie du réseau soit la plus proche possible au sens de L de la valeur cible vérité terrain. Pour cela nous devons minimiser la fonction $\mathcal{L}(\theta) = L(\vec{\hat{y}}(\theta), y)$. La méthode la plus élémentaire est celle de la descente du gradient :

— **Initialisation :** Se donner une valeur initiale $\theta^{(0)}$ de θ.
— **Itération :** A partir de $\theta^{(t)}$, on calcule :

$$\theta^{(t+1)} = \theta^{(t)} - \alpha \times \nabla_\theta \mathcal{L}(\theta^{(t)}), \tag{2.3}$$

où $\alpha > 0$ est un hyper-paramètre appelé "taux d'apprentissage" et $\nabla_\theta \mathcal{L}$ est le vecteur gradient.

Il existe de nombreux algorithmes plus performants et sophistiqués que la descente du gradient. Parmi eux, on peut citer le gradient conjugué [89], gradient stochastique [98] et la méthode la plus utilisée ADAM [48]. Toutes ces méthodes utilisent le gradient et certaines comme gradient conjugué et ADAM utilisent également la Hessienne de $\mathcal{L}$. Le calcul du gradient et de la Hessienne s'effectue par une technique appelée "rétropropagation". Nous détaillerons uniquement la rétropropagation dans le cas du gradient. Pour la Hessienne, le lecteur pourra se référer à [48].
La rétropropagation s'appuie sur la composition des dérivées mais on peut aussi l'expliquer par la composition des différentielles. Nous avons choisi cette dernière approche qui facilite les calculs.
Rappelons ce qu'est une différentielle de fonction. Soit f une fonction de $\mathbb{R}^n$ dans $\mathbb{R}^m$. On dit que f est différentiable en a s'il existe une fonction linaire $d_a f$ de $\mathbb{R}^n$ dans $\mathbb{R}^m$, appelée "différentielle" telle que :

$$f(a+h) = f(a) + d_a f(h) + \circ(h), \tag{2.4}$$

où $\lim_{h\to 0} \circ(h) = 0$. Dans le cas où la fonction qui à a associe $d_a f$ est continue, la dérivée partielle de f en a le long du vecteur e_i est donnée par $\dfrac{\partial f}{\partial x_i}(a) = d_a f(e_i)$, où $(e_1, \ldots, e_n)$ est la base canonique de $\mathbb{R}^n$. En calcul différentiel, il est courant de noter la base duale de la base canonique par $(dx_1, dx_2, \ldots, dx_n)$ lorsque les coordonnées d'un vecteur selon la base canonique sont notées $(x_1, \ldots, x_n)$. dx_i est l'application qui à un vecteur x associe sa i-ème coordonnée. Avec ces notations, on a alors $d_a f = \sum_{i=1}^{n} \dfrac{\partial f}{\partial x_i}(a) dx_i$. Soit maintenant g une fonction différentiable de $\mathbb{R}^m$ dans $\mathbb{R}^p$. La règle de composition des différentielles s'écrit :

$$d_a(g \circ f) = d_{f(a)} g \circ d_a f \tag{2.5}$$

De cette règle, nous pouvons déduire facilement la règle de composition des dérivées partielles :

$$\begin{aligned}
\frac{\partial(g\circ f)}{\partial x_i}(a) &= d_a(g\circ f)(e_i)\\
&= d_{f(a)}g\circ d_a f(e_i)\\
&= d_{f(a)}g\left(\frac{\partial f}{\partial x_i}(a)\right)\\
&= d_{f(a)}g\left(\sum_{j=1}^{m}\frac{\partial f_j}{\partial x_i}(a)e'_j\right)\\
&= \sum_{j=1}^{m}\frac{\partial f_j}{\partial x_i}(a)\times d_{f(a)}g(e'_j)\\
&= \sum_{j=1}^{m}\frac{\partial f_j}{\partial x_i}(a)\times\frac{\partial g}{\partial x'_j}(f(a)), \qquad (2.6)
\end{aligned}$$

où $(e'_1,\ldots,e'_m)$ est la base canonique de $\mathbb{R}^m$.
Nous pouvons remarquer, que bien que plus abstraite, la composition des différentielles est plus facile à manipuler que la composition des dérivées partielles ; il y a moins de risque de se tromper dans les indices des sommes. Ainsi, nous allons détailler la rétropropagation en utilisant la composition des différentielles. Nous retrouverons alors la rétropropagation telle qu'elle est présentée habituellement avec la composition des dérivées partielles en évaluant l'expression différentielle sur les vecteurs de la base canonique de manière similaire à (2.6).

Phase initiale de la rétropropagation :
La fonction $\mathcal{L}(\theta)=L(\vec{x}_M(\theta_M,\vec{x}_{M-1}(\theta_{1:M-1})),y)$ est égale à $L\circ\vec{x}_M\circ f_M$ avec :

$$\begin{aligned}
f_M:\theta &\to (\theta_M,\vec{x}_{M-1}(\theta_{1:M-1}))\\
\vec{x}_M:(\theta_M,\vec{x}_{M-1}) &\to \vec{x}_M(\theta_M,\vec{x}_{M-1})\\
L:\vec{x}_M &\to L(\vec{x}_M,y) \qquad (2.7)
\end{aligned}$$

La fonction f_M a pour différentielle :

$$\begin{aligned}
d_\theta f_M &= (d\theta_M, d_{\theta_{1:M-1}}\vec{x}_{M-1}(d\theta_{1:M-1}))\\
&= (d\theta_M,0)+(0,d_{\theta_{1:M-1}}\vec{x}_{M-1}(d\theta_{1:M-1})) \qquad (2.8)
\end{aligned}$$

La différentielle de $\vec{x}_M$ en $f_M(\theta)$ est :

$$\begin{aligned}
d_{(\theta_M,\vec{x}_{M-1}(\theta_{1:M-1}))}\vec{x}_M &= (\nabla_{\theta_M}\vec{x}_M(\theta_M,\vec{x}_{M-1}(\theta_{1:M-1})))^T\times d\theta_M \qquad (2.9)\\
&+ (\nabla_{\vec{x}_{M-1}}\vec{x}_M(\theta_M,\vec{x}_{M-1}(\theta_{1:M-1})))^T\times d_{\theta_{1:M-1}}\vec{x}_{M-1},
\end{aligned}$$

où A^T est la transposée de la matrice A, les applications coordonnées $d\theta_M$ et $d\vec{x}_{M-1}$ sont disposées en vecteur colonne et $\nabla_{\theta_M}\vec{x}_M$ et $\nabla_{\vec{x}_{M-1}}\vec{x}_M$ sont respectivement les matrices dont les colonnes sont les vecteurs gradients de chacune

des coordonnées de $\vec{x}_M$ par rapport respectivement à θ_M et à $\vec{x}_{M-1}$. Ainsi :

$$\begin{aligned} d_\theta \mathcal{L}(d\theta) &= (\nabla_{\vec{x}_M} L)^T \times (\nabla_{\theta_M} \vec{x}_M)^T \times d\theta_M \\ &+ (\nabla_{\vec{x}_M} L)^T \times (\nabla_{\vec{x}_{M-1}} \vec{x}_M)^T \times d_{\theta_{1:M-1}} \vec{x}_{M-1}, \end{aligned} \tag{2.10}$$

où l'opérateur $\nabla_{\vec{x}_M} L$ désigne le gradient de L par rapport à la variable $\vec{x}_M$. En remplaçant $d\theta_M$ dans le premier terme de la somme par chacun des vecteurs de la base canonique de $\mathbb{R}^{n_M}$, n_M étant le nombre de neurones de la couche M, on retrouve la contribution du gradient pour la mise à jour des paramètres (poids et biais) de la couche M (dernière couche) du réseau de neurone. Le calcul du deuxième terme est l'objet de l'itération de la rétropropagation.

Itération de la rétropropagation :

Par le même raisonnement, on montre que :

$$\begin{aligned} d_{\theta_{1:m-1}} \vec{x}_{m-1}(d\theta) &= (\nabla_{\theta_{m-1}} \vec{x}_{m-1})^T d\theta_{m-1} \\ &+ (\nabla_{\vec{x}_{m-2}} \vec{x}_{m-1})^T d\vec{x}_{m-2} \end{aligned} \tag{2.11}$$

De la même façon, en remplaçant $d\theta_{m-1}$ par les vecteurs de base de $\mathbb{R}^{n_{m-1}}$, on retrouve les composantes du gradient concernant les paramètres de la couche $m-1$.

2.3 Différentes architectures de réseaux de neurones profonds

Dans la section précédente, nous avons présenté le principe général des réseaux de neurones, nous allons maintenant nous intéresser aux différentes architectures existantes en détaillant les différentes couches.

2.3.1 Couches de convolution

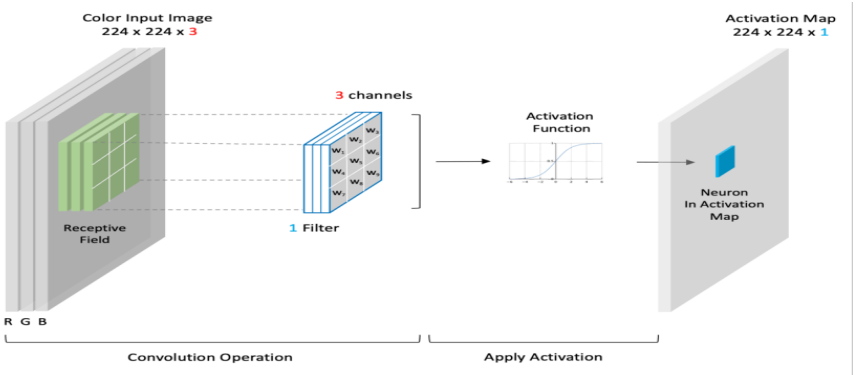


FIGURE 2.2 – Exemple de couche de convolution https://learnopencv.com

Les couches de convolution constituent la pierre angulaire des réseaux de neurones convolutifs (CNN) en raison de leur capacité à extraire automatiquement des caractéristiques pertinentes à partir des données d'entrée. La convolution est l'opération fondamentale réalisée par les couches convolutives (Figure 2.2). Elle consiste à appliquer un ensemble de filtres à l'image d'entrée pour extraire des caractéristiques. Chaque filtre parcourt l'image en effectuant des produits scalaires locaux, produisant une carte de caractéristiques qui capture les motifs présents dans l'image. Certains modèles de CNN utilisent des couches de normalisation pour régulariser et stabiliser l'apprentissage du réseau. Le padding est une technique utilisée pour contrôler la taille des cartes de caractéristiques après les opérations de convolution. Le padding consiste à ajouter des zéros autour des bords de l'image d'entrée avant d'appliquer la convolution, ce qui permet de conserver la taille spatiale des cartes de caractéristiques. Le stride détermine le décalage entre les positions successives des filtres lors de l'opération de convolution. Un stride plus grand réduit la taille de la carte de caractéristiques de sortie, tandis qu'un stride plus petit produit une carte de caractéristiques de sortie plus grande mais plus dense. En combinant ces différentes opérations, les couches convolutives sont capables d'apprendre des représentations hiérarchiques des données d'entrée, en identifiant progressivement des motifs de plus en plus abstraits et discriminants. Cette capacité permet aux CNN d'accomplir une variété de tâches complexes en imagerie médicale, telles que la détection de lésions, la segmentation d'organes et la classification de maladies.

2.3.2 Couches de "pooling"

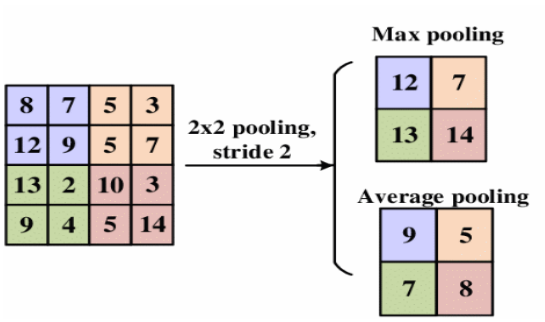


FIGURE 2.3 – Exemple de couche de "pooling"

Les couches de pooling, également connues sous le nom de sous-échantillonnage, sont une composante essentielle des réseaux de neurones convolutifs (CNN) [51]. Elles sont souvent utilisées après les couches de convolution pour réduire la dimension spatiale des cartes de caractéristiques, ce qui permet de réduire le nombre de paramètres du modèle et d'augmenter la robustesse aux variations

spatiales mineures. L'opération de pooling réduit la dimension spatiale des cartes de caractéristiques en regroupant les valeurs voisines dans des régions plus petites. Par exemple, dans une opération de max pooling, seule la valeur maximale dans chaque région de regroupement est conservée, tandis que dans une opération de average pooling, la moyenne des valeurs est calculée. Les opérations de pooling rendent le modèle plus robuste aux variations spatiales mineures en introduisant une invariance aux translations locales. Cela signifie que le modèle sera capable de reconnaître les mêmes motifs même s'ils sont légèrement décalés dans l'image. En réduisant la dimension des cartes de caractéristiques, les opérations de pooling contribuent à prévenir le surapprentissage en limitant la capacité du modèle à mémoriser les détails spécifiques aux données d'entraînement. Cela permet au modèle de généraliser plus efficacement à de nouvelles données. Les opérations de pooling permettent de sélectionner les caractéristiques les plus pertinentes de chaque région de l'image, en ne conservant que les valeurs les plus importantes. Leur utilisation contribue à rendre les modèles plus robustes, plus efficaces et moins sujets au surapprentissage.

2.3.3 Couches entièrement connectées (dense)

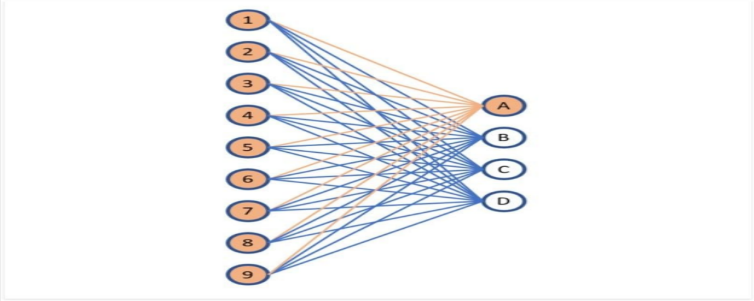


FIGURE 2.4 – Exemple de couche entièrement connectée https://builtin.com

Les couches entièrement connectées (Figure 2.4) , également connues sous le nom de couches denses, constituent une partie importante des réseaux de neurones convolutifs (CNN) [51]. Elles sont généralement situées à la fin du réseau et sont responsables de la classification ou de la régression des caractéristiques extraites par les couches convolutives précédentes. Contrairement aux couches convolutives qui utilisent des filtres locaux pour extraire des caractéristiques spatiales, les couches entièrement connectées sont connectées à toutes les entrées de la couche précédente. Chaque neurone dans une couche dense est connecté à chaque neurone de la couche précédente, formant ainsi une connexion dense. Avant d'être alimentées dans les couches denses, les caractéristiques extraites par les couches convolutives sont souvent aplaties dans un vecteur unidimensionnel à l'aide d'une opération de "flatten". Cela permet de transformer les données spatiales en un vecteur monodimensionnel, facilitant

ainsi leur traitement par les couches denses. Les couches entièrement connectées sont responsables de l'extraction de caractéristiques abstraites à partir des caractéristiques spatiales extraites par les couches convolutives. Elles combinent les informations spatiales provenant de différentes parties de l'image pour produire une représentation globale qui peut être utilisée pour la classification ou la régression. Comme pour les autres types de couches neuronales, les couches entièrement connectées utilisent souvent des fonctions d'activation non linéaires telles que ReLU, sigmoïde ou tangente hyperbolique pour introduire des non-linéarités dans le modèle et lui permettre d'apprendre des représentations plus complexes. La dernière couche dense du réseau est généralement la couche de sortie, qui produit les prédictions finales du modèle. Pour les tâches de classification, cette couche peut utiliser une fonction d'activation softmax pour produire des probabilités pour chaque classe. Pour les tâches de régression, une fonction d'activation linéaire peut être utilisée pour produire des valeurs continues à la sortie.

2.3.4 Couches de normalisation

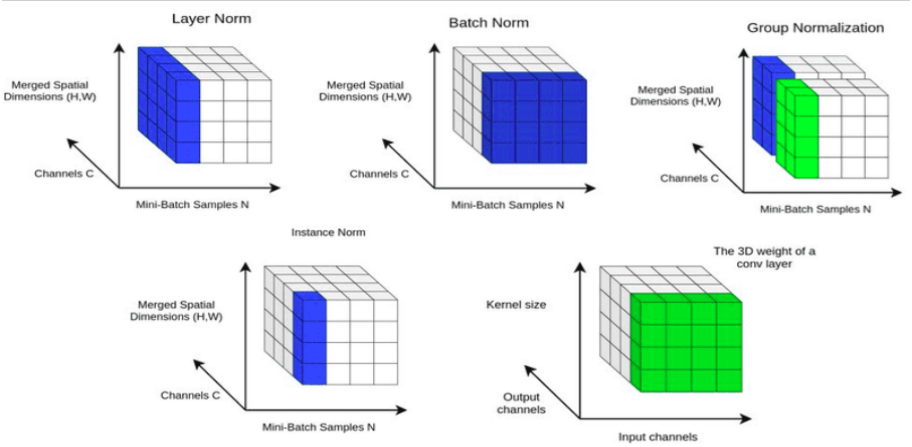


FIGURE 2.5 – Exemple de couches de normalisation

Les couches de normalisation sont des composants clés des réseaux de neurones (Figure 2.5), y compris des réseaux de neurones convolutifs (CNN), qui visent d'une part à éviter les erreurs numériques et d'autre part mettre à la même échelle des données qui ont initialement des ordres de grandeur ou unités différentes. Il existe principalement plusieurs types de normalisation utilisées dans les architectures de réseaux de neurones modernes.

1. Normalisation par lots ("batch normalisation") : La normalisation par lots normalise les activations de chaque couche pour avoir une moyenne nulle et une variance unitaire sur l'ensemble d'un mini-lot d'entraînement. Cela aide à réduire les fluctuations dans la distribution des activations pendant l'entraînement, ce qui accélère la convergence du modèle.

En outre, elle agit comme une régularisation en introduisant du bruit dans le modèle, ce qui le rend moins sensible aux variations mineures des données d'entraînement.

2. Normalisation de groupe : elle divise les activations d'une couche en groupes et normalise chaque groupe indépendamment. Contrairement à la normalisation par lots, elle ne dépend pas de la taille du mini-lot d'entraînement, ce qui la rend plus stable pour les petits mini-lots ou les cas où la taille du mini-lot varie. De plus, elle peut être plus efficace que la normalisation par lots pour les petits ensembles de données ou dans des situations où les mini-lots présentent une grande variabilité.

2.3.5 Couches de "dropout"

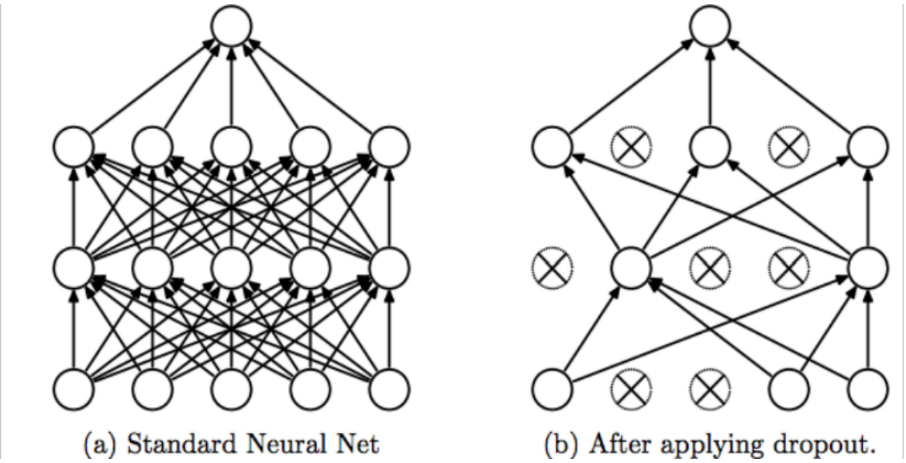


FIGURE 2.6 – Exemple de couches de "dropout" https://medium.com

Les couches de "dropout" sont une technique de régularisation largement utilisée dans les réseaux de neurones, y compris les réseaux de neurones convolutifs (CNN) [77]. Cette technique vise à réduire le surapprentissage en désactivant aléatoirement un certain pourcentage de neurones pendant la phase d'entraînement. Le surapprentissage se produit lorsque le modèle apprend à mémoriser les exemples d'entraînement au lieu d'apprendre des motifs généraux. Les couches de dropout aident à réduire ce phénomène en empêchant les neurones de coadapter trop étroitement les uns aux autres, obligeant ainsi le réseau à apprendre des représentations plus robustes et généralisables. Pendant la phase d'entraînement, chaque neurone d'une couche de dropout a une probabilité fixe, généralement définie à 0,5, d'être désactivé (mis à zéro). Cela signifie que chaque fois qu'une entrée est propagée à travers la couche de dropout, certains neurones seront aléatoirement désactivés (Figure 2.6). Ce processus simule l'effet d'apprentissage d'un sous-réseau différent à chaque itération, ce qui permet au modèle de devenir plus robuste et moins susceptible de surapprendre. Les couches de dropout sont généralement placées entre les

couches de connexion entièrement connectées du réseau. Elles peuvent être utilisées dans différentes parties du réseau, bien que leur utilisation la plus courante soit après les couches de convolution et avant les couches entièrement connectées. Il est important de noter que les couches de dropout ne sont activées que pendant la phase d'entraînement ; pendant la phase de test, toutes les activations des neurones sont conservées. Les couches de dropout offrent plusieurs avantages, notamment la réduction du surapprentissage, l'amélioration de la généralisation du modèle, et une meilleure robustesse aux variations des données d'entrée. Elles permettent également d'utiliser des architectures de réseaux plus profondes en réduisant le risque de surapprentissage.

Bien que les couches de dropout soient très efficaces pour réduire le surapprentissage, elles peuvent également réduire la capacité du modèle à apprendre des modèles complexes si le taux de dropout est trop élevé [51]. De plus, elles peuvent augmenter le temps d'entraînement du modèle en raison du besoin de plusieurs itérations pour converger.

2.4 Architectures d'apprentissage profond utilisées en radiomique

Le développement de nouvelles architectures d'apprentissage profond se déroule à grande vitesse pour différentes applications dans différents domaines. Dans cette section, nous nous intéressons aux architectures de base utilisées en radiomique. Les différentes versions de variantes sont continuellement développées pour les améliorer.

2.4.1 Les Réseaux de Neurones convolutifs (CNN)

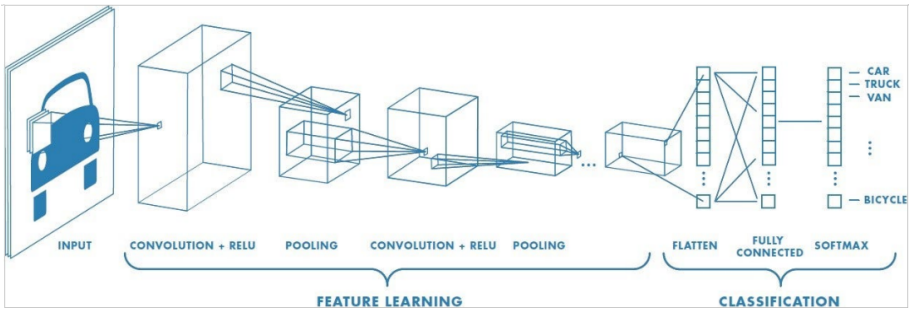


FIGURE 2.7 – Schéma de réseau de neurones convolutif https://towardsdatascience.com

Les réseaux de neurones convolutifs (CNN) [66, 77, 78, 80, 84, 90] sont une catégorie spécifique de réseaux de neurones profonds utilisés principalement dans le traitement et l'analyse d'images, notamment en radiomique [6, 85]. Proposé initialement par [58], ils ont une vaste gamme d'applications dans le

domaine de l'imagerie médicale, offrant des solutions innovantes pour diverses tâches cliniques. Son architecture de base est illustrée dans la Figure 2.7.

Les réseaux de neurones convolutifs (CNN) sont une classe de réseaux de neurones profonds, optimisés pour l'analyse d'images, qui simulent la manière dont le cortex visuel humain interprète les stimuli visuels. Les CNN se distinguent par leur architecture en couches, comprenant principalement des couches de convolution, d'activation (comme ReLU), de pooling (ou sous-échantillonnage) et des couches entièrement connectées. Les couches de convolution utilisent des filtres pour extraire des caractéristiques locales des images, comme les bords ou les textures, en appliquant des opérations de convolution à travers l'image. Les fonctions d'activation introduisent de la non-linéarité, permettant au réseau d'apprendre des motifs complexes. Les couches de pooling réduisent la dimensionnalité des données, augmentant ainsi l'efficacité computationnelle et la robustesse du modèle face aux variations mineures dans les données d'entrée. Enfin, les couches entièrement connectées intègrent les caractéristiques extraites pour effectuer des tâches telles que la classification. Au cours de l'entraînement, les CNN ajustent automatiquement les paramètres des filtres pour minimiser une fonction de perte, utilisant des techniques telles que la rétropropagation et la descente de gradient, ce qui leur permet de devenir très performants dans des tâches de vision par ordinateur comme la reconnaissance d'images et la détection d'objets.

2.4.2 Auto-encodeurs

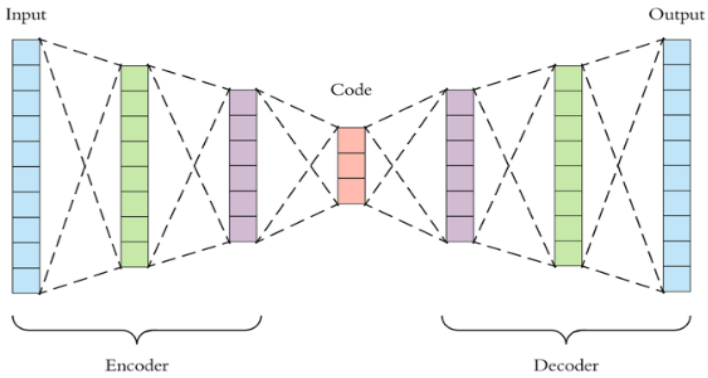


FIGURE 2.8 – Schéma de réseau de neurones auto-encodeur https://towardsdatascience.com

Les auto-encodeurs sont des modèles d'apprentissage non supervisé utilisés en deep learning pour apprendre des représentations efficaces et compactes des données d'entrée [51]. Ils se composent de deux parties principales (voir Figure 2.8) : un encodeur, qui transforme les données d'entrée en une représentation

latente de dimension réduite, et un décodeur, qui tente de reconstruire les données d'entrée à partir de cette représentation latente. Les autoencodeurs se composent généralement de trois couches principales : une couche d'entrée, une couche cachée (représentation latente) et une couche de sortie. L'encodeur est responsable de compresser les données d'entrée dans une représentation latente de dimension réduite, tandis que le décodeur reconstruit les données d'entrée à partir de cette représentation latente. Pendant l'entraînement, les autoencodeurs tentent de minimiser l'erreur de reconstruction entre les données d'entrée et les données de sortie reconstruites par le décodeur. Pour ce faire, les poids de l'encodeur et du décodeur sont ajustés itérativement à l'aide d'algorithmes d'optimisation tels que la descente de gradient. Une fois l'entraînement terminé, l'encodeur apprend à extraire des caractéristiques significatives des données d'entrée, tandis que le décodeur apprend à reconstruire fidèlement les données d'entrée à partir de ces caractéristiques.

Les autoencodeurs sont utilisés dans une variété de domaines pour des tâches telles que la réduction de dimension, la détection d'anomalies, la génération d'images, et la pré-entraînement des réseaux de neurones. En réduisant la dimension des données tout en préservant les informations importantes, les autoencodeurs permettent de mieux comprendre la structure sous-jacente des données et de les utiliser de manière plus efficace dans diverses applications. Il existe plusieurs types d'autoencodeurs, à part des autoencodeurs standard, on peut citer les autoencodeurs variationnels (VAE) [49], les autoencodeurs contractifs, les autoencodeurs sparse [21] et bien d'autres encore[18]. Chaque type d'autoencodeur présente des propriétés et des applications spécifiques, ce qui les rend adaptés à différents types de problèmes et de données.

Malgré leur efficacité dans de nombreuses applications, les autoencodeurs peuvent rencontrer certains défis, notamment la sensibilité aux données de bruit, la difficulté à apprendre des représentations significatives pour des données hautement complexes et la dépendance à l'initialisation des poids. De plus, la qualité des représentations apprises par les autoencodeurs peut dépendre fortement de la taille et de la diversité des données d'entraînement.

Les autoencodeurs sont des outils puissants en deep learning pour l'apprentissage non supervisé de représentations efficaces des données. Leur capacité à comprimer les données tout en préservant les informations importantes les rend utiles dans une variété d'applications.

2.4.3 U-Net

Conçu pour la segmentation sémantique, UNet est particulièrement efficace dans la tâche de segmentation d'images, initialement proposée par [83]. UNet se compose d'un encodeur et d'un décodeur symétriques (Figure 2.9). L'encodeur, souvent basé sur des couches convolutives, est responsable de capturer les caractéristiques hiérarchiques des images à différentes échelles spatiales. Le décodeur, quant à lui, utilise des couches de déconvolution pour reconstruire une image segmentée à partir de la représentation latente générée par l'enco-

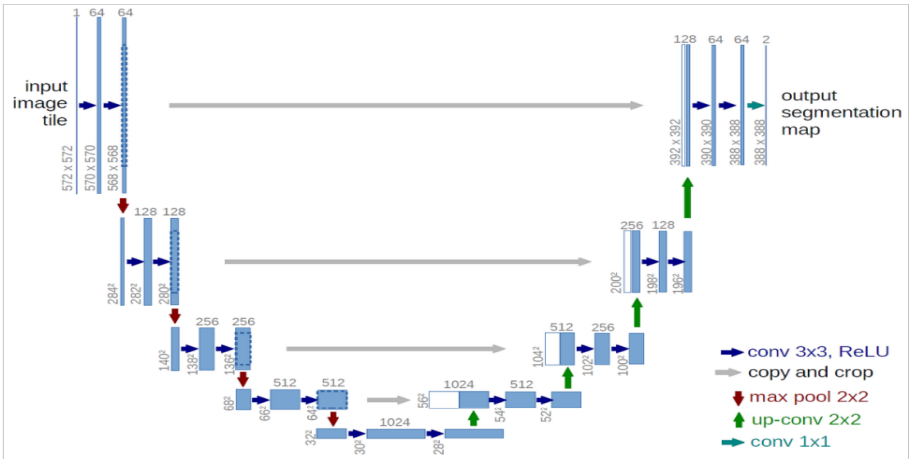


FIGURE 2.9 – Schéma de réseau de neurones U-Net https://towardsdatascience.com

deur. La caractéristique distinctive de UNet est sa structure en forme de U, où les informations sont agrégées à différentes échelles spatiales à chaque étape du décodage. Cela permet à UNet de conserver les détails spatiaux fins tout en capturant des informations contextuelles à différentes résolutions. Pour faciliter le flux d'informations à travers le réseau et atténuer le problème de la disparition du gradient, UNet utilise des connexions résiduelles qui permettent de sauter des couches dans le processus de décodage. Ces connexions aident à préserver les informations importantes et à améliorer la convergence de l'apprentissage.

UNet est largement utilisé dans divers domaines, notamment la biomédecine. Il est particulièrement efficace dans la segmentation d'images médicales pour la détection et la localisation de structures anatomiques telles que les organes, les tumeurs et les lésions [116, 117, 118]. Les principaux avantages de UNet résident dans sa capacité à capturer à la fois les détails locaux et les informations contextuelles à différentes échelles spatiales, ce qui en fait un choix idéal pour la segmentation d'images médicales et d'autres applications nécessitant une analyse fine des structures spatiales.

2.4.4 Réseaux multi-tâches

L'architecture multitâches permet d'effectuer simultanément plusieurs tâches, telles que la segmentation, la classification et la reconstruction dans une seule architecture unifiée, comme illustrée dans la Figure 2.10. Cette approche offre une solution efficace pour des problèmes où il n'a pas assez de données pour chaque tâche, car plusieurs tâches peuvent partager des caractéristiques communes. L'UNet standard qui comprend un encodeur et un décodeur symétriques est très souvent utilisée pour l'architecture multitâches. L'encodeur est responsable de capturer les caractéristiques hiérarchiques de l'image tandis que le décodeur reconstruit l'image à partir de la représentation latente générée

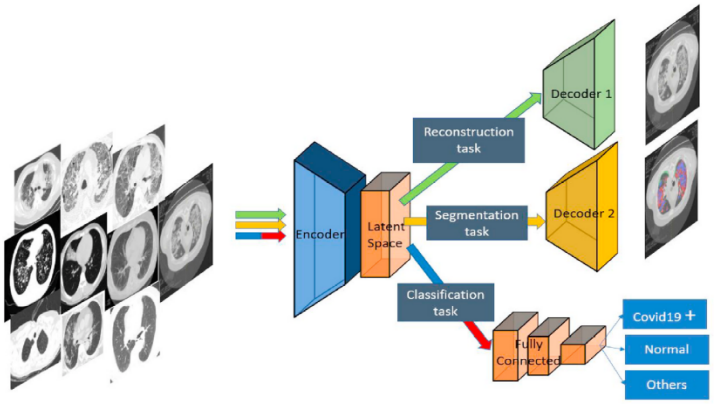


FIGURE 2.10 – Exemple de réseau multi-tâches (images extraite de [5])

par l'encodeur. En plus des couches de décodage utilisées pour la reconstruction, l'architecture multitâches UNet comprend également des branches (décodeurs) supplémentaires pour la segmentation et la classification. Ces branches prennent la représentation latente générée par l'encodeur et l'utilisent pour effectuer des tâches de segmentation et de classification. La tâche de reconstruction de l'image d'entrée est une tâche associée aux tâches principales, telles que la segmentation et la classification. Son but est d'aider à extraire des caractéristiques. Comme dans l'UNet standard, des connexions résiduelles peuvent être utilisées pour faciliter le flux d'informations à travers le réseau et améliorer la convergence de l'apprentissage. Ces connexions aident à préserver les informations importantes et à atténuer les problèmes de disparition du gradient lors de l'entraînement.

L'architecture multitâche UNet est particulièrement utile dans les applications de classification, car la tâche de localiser précisément les pathologies dans une image permet de concentrer les caractéristiques sur ces régions afin de mieux classer les images. Cette approche présente l'avantage de combiner des tâches importantes dans une seule architecture, ce qui peut simplifier le processus de modélisation et permettre un apprentissage plus efficace en utilisant des informations contextuelles pour améliorer la classification. L'architecture multitâches offre une solution efficace pour combiner des différentes tâches dans une seule architecture, notamment pour la radiomique [5, 14].

2.4.5 Réseaux de neurones de type "Vision Transformer"(ViT)

Les "Vision Transformer" sont une classe d'architectures de réseaux neuronaux qui ont révolutionné le domaine de la vision par ordinateur en appliquant les principes fondamentaux des transformers, initialement développés pour le traitement de texte, à la tâche de traitement d'images [24]. Contrairement

aux architectures conventionnelles basées sur des réseaux de neurones convolutifs (CNN), les "Vision Transformer" exploitent des mécanismes d'attention pour capturer les relations spatiales entre les différentes parties de l'image sans utiliser de convolutions. Son architecture initiale est illustrée dans la Figure 2.11. Le fonctionnement des "Vision Transformer" commence par la décomposition de l'image en une grille de patches. Contrairement aux CNN qui utilisent des filtres de convolution pour extraire des caractéristiques locales, les "Vision Transformer" découpent l'image en une séquence de patches de taille fixe, qui sont ensuite aplatis et traités comme des séquences de données unidimensionnelles. Chaque patch est associé à une position unique dans l'image, et ces informations de position sont encodées en ajoutant des vecteurs de position à chaque patch. Cela permet au modèle de conserver des informations spatiales sur la disposition des patches dans l'image, ce qui est essentiel pour capturer les relations spatiales entre les différentes parties de l'image. Une fois que les patches et leurs positions ont été encodés, l'image est ensuite alimentée dans une série de couches transformer. Chaque couche transformer est composée de deux sous-modules principaux : le mécanisme d'attention multi-têtes et les réseaux de neurones entièrement connectés (feedforward). Le mécanisme d'attention permet au modèle de capturer les relations à longue distance entre les différentes parties de l'image en calculant des poids d'attention pour chaque paire de patches. Ces poids d'attention sont ensuite utilisés pour agréger l'information provenant de tous les patches de l'image, en mettant l'accent sur les patches les plus pertinents pour la tâche à accomplir. Après le mécanisme d'attention, les représentations des patches sont alimentées dans des réseaux de neurones feedforward, qui sont chargés de transformer les caractéristiques extraites par l'attention en représentations plus riches et complexes. Ces réseaux de neurones feedforward sont généralement composés de couches entièrement connectées et de fonctions d'activation non linéaires. Enfin, la sortie de la dernière couche transformer est alimentée dans une couche de sortie, qui est responsable de produire les prédictions finales du modèle. Pour les tâches de classification, cette couche de sortie peut utiliser une fonction d'activation softmax pour

2.4.6 Réseaux de neurones de type "Swin Transformers"

Les "Swin Transformers" [62] représentent une évolution des "Vision Transformer" , conçus spécifiquement pour améliorer l'efficacité et la performance dans les tâches par ordinateur (Figure 2.12). Ils sont une variante innovante des architectures transformer qui visent à améliorer la capacité des modèles à capturer des informations à différentes échelles spatiales de manière efficace. Leur structure innovante ouvre de nouvelles perspectives pour le traitement d'images à grande échelle et continue de susciter un vif intérêt dans la communauté de recherche en vision par ordinateur. Les "Swin Transformers" se reposent sur le mécanisme d'attention à fenêtre coulissante des "Swin Transformers", qui constitue l'un de leurs aspects les plus distinctifs et innovants.

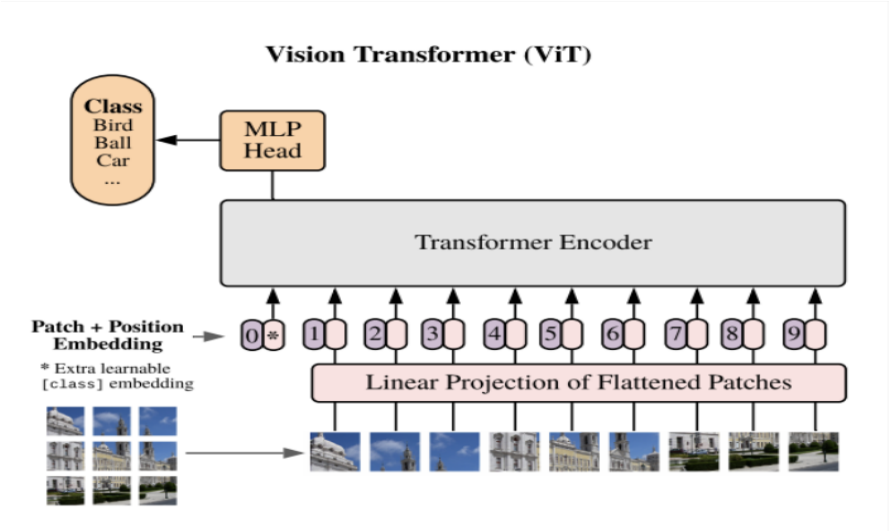


FIGURE 2.11 – Schéma d'un réseau de neurone "transformer" (image extraite de [24])

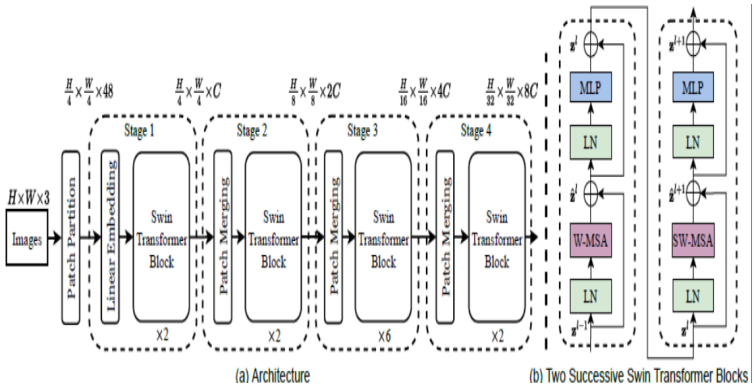


FIGURE 2.12 – Schéma d'un réseau de neurone "swin transformer" https://arshren.medium.com

Contrairement aux mécanismes d'attention traditionnels qui opèrent sur l'ensemble de l'image, le mécanisme d'attention à fenêtre coulissante divise l'image en patches et permet au modèle de se concentrer sur des sous-régions spécifiques de manière sélective. Ce mécanisme fonctionne en plusieurs étapes :

1. **Division de l'image en patches :** Tout d'abord, l'image en entrée est divisée en une grille de patches de taille fixe. Chaque patch représente une sous-région de l'image et est traité comme une unité distincte par le modèle.
2. **Génération des clés, valeurs et requêtes :** Pour chaque patch, des vecteurs de clés, de valeurs et de requêtes sont générés. Ces vecteurs sont utilisés pour calculer l'importance de chaque patch par rapport aux autres patches dans l'image.

3. **Calcul de l'attention :** En utilisant les vecteurs de clés, de valeurs et de requêtes, le modèle calcule l'importance de chaque patch par rapport à tous les autres patches. Cette opération est effectuée en calculant un score d'attention pour chaque paire de patches, qui est ensuite utilisé pour pondérer les valeurs correspondantes.
4. **Agrégation des informations :** Une fois que les scores d'attention ont été calculés, les informations des patches sont agrégées en fonction de ces scores. Les patches les plus importants reçoivent une pondération plus élevée lors de l'agrégation, ce qui permet au modèle de se concentrer sur les parties les plus pertinentes de l'image.
5. **Propagation de l'attention :** Les informations agrégées sont ensuite propagées à travers les différentes couches du modèle, permettant au modèle d'apprendre des relations complexes entre les patches à différentes échelles spatiales.

En combinant ce mécanisme d'attention à fenêtre coulissante avec la structure à étages alternés des "Swin Transformers", le modèle est capable de capturer efficacement des informations à différentes résolutions spatiales, ce qui lui permet de traiter des images de haute résolution avec une grande efficacité. Ce mécanisme est aussi efficace dans le domaine de la radiomique [13].

2.5 Conclusion

Les architectures de réseaux de neurones profonds et les fonctions mathématiques correspondantes fournissent d'excellents outils pour servir l'objectif que nous nous sommes fixés au début de cette thèse. Nous avons pu décrire et comprendre les différents composants disponibles à assembler afin de construire les réseaux les plus appropriés, ainsi que l'état de l'art dans le domaine et les architectures usuelles, des CNN aux "Swin Transformers". Nous avons proposé, lors de nos travaux, de focaliser nos efforts sur les architectures les plus utilisées pour l'analyse d'images, et les mettre à l'épreuve sur le sujet spécifique de l'analyse des images médicales via la radiomique. Ces architectures sont le réseau de convolution usuel, le UNet et le swin transformer. Nous avons pris le parti d'adapter ces architectures qui font partie de l'état de l'art à nos besoins, via des réglages spécifiques, des branches supplémentaires pour accomplir plusieurs tâches simultanément et des modifications pour accepter les données tridimensionnelles et les données au format tableur. Ces réseaux spécifiques ont le potentiel de devenir de nouvelles références dans l'étude d'images médicales. Un autre aspect reste à étudier afin de circonscrire totalement les tenants et les aboutissants de l'apprentissage profond, les fonctions de pertes utilisées à la sortie du réseau que l'on souhaite entraîner qui permettent de s'assurer que le réseau construit puisse apprendre à partir des données fournies en entrée. Dans le chapitre suivant, nous allons détailler les fonctions de pertes basées sur les entropies qui ont été étudiées au cours de nos travaux.

CHAPITRE 3

Fonctions de perte basées sur les entropies

3.1 Généralités sur l'entropie

3.1.1 Introduction

L'entropie [22] est un concept fondamental qui concerne plusieurs disciplines, de la thermodynamique à la théorie de l'information, reflétant des notions de désordre, d'incertitude, de rareté des évènements aléatoires ainsi que la non-réversibilité d'un phénomène physique. En thermodynamique, l'entropie peut être vue sous deux angles. Pour Clausius, celle-ci est liée à la non réversibilité d'un phénomène physique dans un système isolé comme par exemple le transfert de la chaleur d'un corps chaud vers un corps froid [19]. Le physicien thermodynamicien Boltzmann interprête quant à lui l'entropie comme une mesure du désordre [11]. La non-réversibilité se traduisant alors par l'augmentation du désordre au cours du temps. C'est également Boltzmann qui a proposé sa célèbre loi de probabilité qui est d'ailleurs la loi de probabilité de répartition des niveaux d'énergie ayant l'entropie au sens de Boltzmann la plus grande pour une température donnée ; la température étant la variance des états d'énergie [11]. En théorie de l'information, développée par Claude Shannon dans les années 1940, l'entropie quantifie l'incertitude ou l'imprévisibilité inhérente à l'information contenue dans un message [88]. Elle mesure la quantité d'information "perdue" lors de la transmission d'un message ou la quantité d'information nécessaire pour décrire complètement l'état d'un système [40]. Dans ce contexte, une entropie élevée signifie une plus grande incertitude et une plus grande richesse d'information potentielle, car plus un événement est rare et plus il est informatif. Ceci rejoint la notion de désordre proposé un peu moins d'un siècle avant par Boltzmann. Un haut degré de désordre se traduit par un grand nombre d'évènements rares, tandis qu'une entropie faible indique un degré de prédictibilité plus élevé. Dans ses premières recherches, Shannon a

intitulé sa fonction “quantité d’information” ; la forte similitude avec la formule de Boltzmann a amené Shannon à l’appeler “entropie de Shannon”. L’entropie trouve des applications dans divers domaines tels que l’analyse statistique d’images [34], les méthodes de clustering [2, 93], la théorie du codage [105], les statistiques théoriques [45, 70], l’apprentissage profond [9, 32, 79, 106] ainsi qu’en physique statistique [61, 73].

3.1.2 Entropie : de la définition de Clausius à celle de Shannon

La notion du concept d’entropie est née des travaux de Clausius sur l’irréversibilité des processus physiques dans un système isolé comme par exemple le transfert de chaleur. Clausius a alors imaginé une quantité abstraite appelée “entropie” telle que sa variation soit liée au transfert de chaleur via sa célèbre formule :

$$\delta Q = TdS, \tag{3.1}$$

où δQ désigne le transfert de chaleur, T la température du corps et dS la variation d’entropie. La notion de chaleur étant, à son époque, difficilement définissable, il a fallu attendre les travaux de Boltzmann pour avoir une définition rigoureuse de l’entropie et ainsi définir le transfert de chaleur comme la variation d’énergie interne due à la variation d’entropie. Pour se faire, Boltzmann considère la matière macroscopique comme étant constituée de micro-éléments (micro-état) tels que l’énergie du corps suit une répartition aléatoire au sein des micro-états. Le formalisme de Clausius dit que la variation d’énergie moyenne est donné par :

$$dU = \underbrace{TdS}_{\text{chaleur}} + \underbrace{\mathbb{E}\left[\frac{\partial \phi(\alpha, X)}{\partial \alpha}\right] d\alpha}_{\text{travail mécanique}}, \tag{3.2}$$

où ϕ est la fonction d’énergie microscopique dépendant d’un paramètre α telle que $\phi(\alpha, x)$ est l’énergie du micro-état x. Les travaux de Boltzmann ont alors consisté à chercher la loi de probabilité de la variable aléatoire X ainsi que la définition mathématique de l’entropie de façon à ce que, d’une part, la loi de probabilité maximise l’entropie à énergie donnée car celle-ci n’augmente plus à l’équilibre et d’autre part, on puisse retrouver l’équation (3.2). L’entropie de Boltzmann est alors donnée par :

$$S = -\int \log(p(x))p(x)dx$$

et la loi dite de Maxwell-Boltzmann est celle qui maximise l’entropie à énergie moyenne U fixée et est donnée par :

$$p(x|\alpha, U) = \frac{1}{Z(\alpha, U)} \exp\left(-\frac{\phi(\alpha, x)}{T(\alpha, U)}\right),$$

où $T(\alpha, U)$ est la température.
Quant à Shannon, ce dernier a introduit en premier lieu la notion de quantité

d'information en la définissant comme le rapport entre la taille en bits du signal compressé sur la taille en bits du signal original après compression maximale. Celui-ci trouva alors la formule suivante :

$$H = -\left[p_0 \log_2(p_0) + p_1 \log_2(p_1)\right], \tag{3.3}$$

formule à rapprocher de celle de Boltzmann. Cette quantité fut alors nommée "entropie de Shannon". Dans les faits, les entropies de Boltzmann et de Shannon correspondent à la même formule et le choix du nom est surtout motivé selon le domaine d'application : entropie de Boltzmann en thermodynamique et physique statistique, entropie de Shannon en théorie de l'information et en statistiques.
Ainsi, l'entropie de Shannon d'une variable aléatoire X à valeurs dans un espace mesurable E est définie par :

$$\begin{aligned} H(X) &= -\sum_{x\in E} p(x)\log(p(x)) \text{ si } E \text{ est dénombrable et discret} \\ &= -\int_E p(x)\log(p(x))dx \text{ si } E \simeq \mathbb{R}^n \text{ muni de la tribu borélienne} \end{aligned} \tag{3.4}$$

Par la suite, nous utiliserons, sauf dans des cas particuliers, essentiellement la formule dans le cas discret étant donné que les variables aléatoires étudiées dans cette thèse de doctorat sont à valeurs discrètes.

3.1.3 Généralisations de l'entropie de Shannon

Depuis la découverte de l'entropie de Shannon, de nombreuses recherches aussi bien en sciences des données qu'en théorie de l'information ont conduit à des généralisations de celle-ci. M. Basseville présente dans un rapport les différentes façons de généraliser l'entropie de Shannon ainsi que les divergences associées [8]. Parmi ces généralisations, on peut utiliser une axiomatique décrivant l'entropie mais on peut également utiliser une équation fonctionnelle. Nous avons choisi cette dernière façon de faire qui s'avère beaucoup plus souple. Pour généraliser l'entropie de Shannon, remarquons dans un premier temps que celle-ci s'écrit :

$$H(X) = -\sum_{x\in E} h(p(x)) \tag{3.5}$$

avec :

$$h(p) = p \times \log(p) \tag{3.6}$$

Cette fonction h est convexe et s'annule en 0. Ainsi, une première façon de généraliser est de remplacer h par une fonction convexe s'annulant en 0. Nous donnons ainsi quelques exemples :

— **Entropie de Havrda-Charvat :** Celle-ci est définie pour $\alpha \in]0, +\infty[\backslash\{1\}$ par :

$$h_\alpha(p) = \frac{p^\alpha - p}{\alpha - 1}$$

et donne donc :

$$H_\alpha(X) = \frac{1}{\alpha - 1} \times \left(1 - \sum_{x \in E} (p(x))^\alpha\right) \tag{3.7}$$

On remarque que pour x fixé, $\lim_{\alpha \to 1} h_\alpha(p(x)) = p(x) \times \log(p(x))$, quitte à restreindre α à l'intervalle $\left[\frac{1}{2}, 2\right]$, on peut majorer $|h_\alpha(p(x))|$ par une fonction sommable qui ne dépend pas de α. Ainsi, on peut appliquer le théorème de convergence dominée qui nous donne que $\lim_{\alpha \to 1} H_\alpha(X) = H_1(X)$ où $H_1(X)$ désigne l'entropie de Shannon. L'entropie de Havrda-Charvat s'interprète facilement pour les valeurs entières de α supérieures ou égales à 2. En effet, si $\alpha = n \geq 2$, $(p(x))^n$ est la probabilité de la réalisation $\underbrace{(x, \ldots, x)}_{\text{n fois}}$ pour le vecteur aléatoire $(X_1, \ldots, X_n)$ où les X_k sont k copies indépendantes de la variable aléatoire X de loi p. Ainsi, $\sum_{x \in E} (p(x))^n$ est la probabilité que n tirages indépendants de la variable aléatoire X donnent des ex-aequo. C'est une façon plus fine de mesurer le désordre qu'avec l'entropie de Shannon. En effet, on peut interpréter l'entropie d'Havrda-Charvat d'ordre $n \geq 2$ par le fait que plus celle-ci est grande et moins il y a de chance d'avoir n valeurs égales si je tire n réalisations indépendantes de la variable aléatoire X [20]. Lorsque $\alpha = 2$, nous retrouvons l'indice de diversité de Gini utilisé par les économistes et sociologues pour mesurer le taux d'inégalité dans une population, par exemple inégalité salariale. Plus l'indice de Gini est élevé, et plus la probabilité que deux personnes prises au hasard ont le même salaire est faible, ce qui traduit une forte inégalité. Lorsque α tend vers 0, on peut montrer que l'entropie de Havrda-Charvat tend nombre d'évènements de probabilité non nulle auquel on soustrait 1. On en déduit que plus α est grand et plus on sera sensible à la différence de désordre. La Figure 3.1 donne la comparaison entre l'entropie de Shannon et l'entropie de Havrda-Charvat pour différentes valeurs de α pour une variable aléatoire à valeurs dans $\{0, 1\}$. Comme nous pouvons le constater, pour les valeurs extrêmes de α ($\alpha = 0$ et $\alpha = \infty$), l'entropie de Havrda-Charvat tend vers une fonction constante valant respectivement 1 pour $\alpha = 0$ et 0 pour $\alpha = \infty$. Ceci peut justifier le choix de α dans la fonction de perte définie à partir de l'entropie croisée d'Havrda-Charvat (voir sous-section suivante). Plus α est petit, plus on pourra discriminer une loi de Dirac d'une loi non Dirac ; ce qui justifie le choix de petites valeurs dans le cas de classification dure (cad. non floue) ; tout au moins tant qu'il n'y a pas de problèmes numériques dûs aux trop petites valeurs de α.

La Figure 3.2 donne la même comparaison pour une loi normale centrée en fonction de l'écart-type de la loi normale. Les entropies, dans ce cas, sont obtenues en remplaçant la somme par une intégrale. L'entropie de Shannon d'une loi normale centrée et d'écart-type $\sigma > 0$ est donnée

par :

$$H_1 = \log(\sqrt{2\pi e}\sigma), \tag{3.8}$$

tandis que l'entropie d'Havrda-Charvat de paramètre $\alpha > 0$ est donnée par :

$$H_\alpha = \frac{1}{\alpha - 1} \times \left(1 - \frac{1}{(2\pi)^{\frac{\alpha-1}{2}}\sqrt{\alpha}\sigma^{\alpha-1}}\right) \tag{3.9}$$

— **Entropie d'Hellinger :** Celle-ci est donnée par :

$$H_{\frac{1}{2}}(X) = \sum_{x \in E} \sqrt{p(x)} - 1 \tag{3.10}$$

et correspond à la moitié de l'entropie d'Havrda-Charvat d'ordre $\frac{1}{2}$. Celle-ci a la même interprétation que l'entropie de Shannon comme mesure de désordre, sauf qu'au lieu de prendre l'opposé de la log-vraisemblance, on fait intervenir la racine carrée de la vraisemblance. Ainsi, contrairement à l'entropie de Shannon, si X et Y sont deux variables aléatoires indépendantes, $H_{\frac{1}{2}}((X,Y)) \neq H_{\frac{1}{2}}(X) + H_{\frac{1}{2}}(Y)$. Celle-ci n'est par contre qu'un cas particulier de l'entropie d'Havrda-Charvat.

— **Entropie quadratique :** Celle-ci est définie par :

$$Q(X) = -\sum_{x \in E} q(x,y)p(x)p(y), \tag{3.11}$$

où q est un produit scalaire, c'est-à-dire une forme bilinéaire, symétrique et définie positive. Dans le cas où $E = \{x_1, \ldots, x_k\}$, on peut définir

$$q(x,y) = \sum_{i=1}^{k} x_i y_i$$

. Il est aussi possible de paramétrer q en utilisant une matrice symétrique définie positive.

Une autre façon de généraliser l'entropie de Shannon, est de partir de la formule suivante :

$$H(X) = -\log \phi^{-1}\left(\sum_{x \in E} p(x)\phi(p(x))\right) \tag{3.12}$$

avec $\phi(p) = \log(p)$. On peut alors généraliser l'entropie de Shannon en remplaçant ϕ par une fonction concave qui s'annule en 1. En exemple :

— **Entropie de Rényi :** Elle est définie pour $\alpha \in]0, +\infty[\backslash 1$ avec :

$$\phi_\alpha(p) = \frac{p^{\alpha-1} - 1}{\alpha - 1} \tag{3.13}$$

et on trouve :

$$\tilde{H}_\alpha(X) = \frac{1}{1-\alpha} \log\left(\sum_{x \in E} (p(x))^\alpha\right) \tag{3.14}$$

Comme on a fait pour l'entropie de Havrda-Charvat, en utilisant le théorème de convergence dominée, celle-ci se prolonge en $\alpha = 1$ et on

retrouve l'entropie de Shannon. Celle-ci est liée à l'entropie de Havrda-Charvat par la relation :

$$\tilde{H}_\alpha(X) = \frac{1}{1-\alpha} \log\left(1 - (\alpha - 1) H_\alpha(X)\right) \tag{3.15}$$

et donc hérite de la même interprétation lorsque α est un entier supérieur ou égal à 2.
Lorsque α tend vers l'infini, l'entropie de Rényi tend vers $-\log \max p(x)$. Ainsi, plus l'entropie est grande et plus la loi de la variable aléatoire tend vers une loi uniforme.

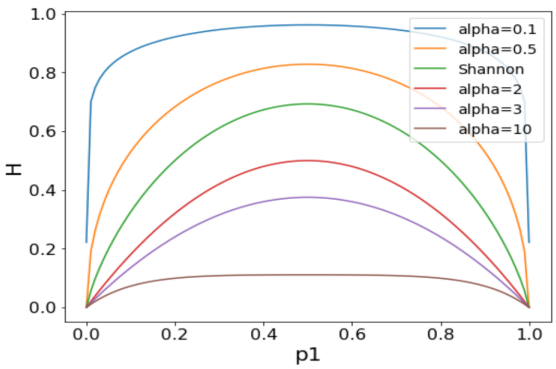


FIGURE 3.1 – Différentes entropies de Havrda-Charvat d'une variable aléatoire à valeurs dans $\{0, 1\}$ en fonction de la probabilité p_1 de la réalisation 1.

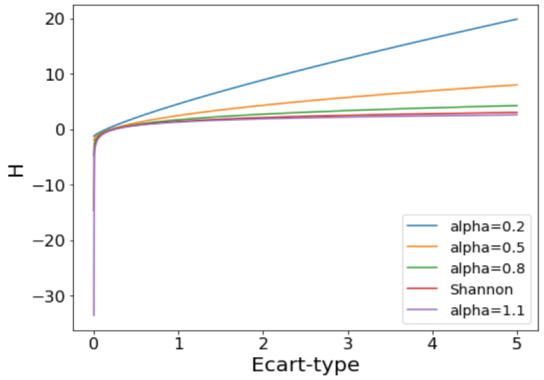


FIGURE 3.2 – Différentes entropies de Havrda-Charvat pour une loi normale centrée en fonction de son écart-type.

3.2 Fonctions de perte basées sur les entropies : divergences et entropies-croisées

Les fonctions de perte utilisées généralement en classification comparent deux probabilités : une probabilité dite de "Dirac" q représentant la vérité terrain ; par exemple si $q = \delta_j$ cela signifie que la donnée est dans la classe j et une probabilité p représentant la sortie du réseau de neurones. Il y a deux grandes façons d'établir une "métrique" entre probabilité, toutes deux issues des entropies utilisées : la divergence et l'entropie croisée.

3.2.1 Généralisations des divergences

Soient p et q deux lois de probabilité sur un espace discret E et soit H une fonction d'entropie choisie. La divergence issue de H de q conditionnellement à p est obtenue en remplaçant dans l'opposé de l'entropie $-H(r)$, r par $\frac{q}{p}$ et en pondérant la somme par p. Dans le cas où l'entropie H est de la forme (3.5), la divergence est donnée par la formule :

$$K(q||p) = \sum_{x \in E} h\left(\frac{q(x)}{p(x)}\right) p(x) \tag{3.16}$$

Ainsi, dans le cas où H désigne l'entropie de Shannon, on retrouve la divergence de Kullback-Leibler donnée par :

$$K_1(q||p) = \sum_{x \in E} \log\left(\frac{q(x)}{p(x)}\right) q(x) \tag{3.17}$$

Tandis que la divergence issue de l'entropie d'Havrda-Charvat est donnée par :

$$K_\alpha(q||p) = \frac{1}{1-\alpha} \times \left[1 - \sum_{x \in E} q(x)^\alpha p(x)^{1-\alpha}\right] \tag{3.18}$$

Dans le cas où l'entropie H est de la forme (3.12), la divergence est donnée par :

$$\tilde{K}(q||p) = -\log \phi^{-1}\left(\sum_{x \in E} \phi\left(\frac{q(x)}{p(x)}\right) q(x)\right) \tag{3.19}$$

On en déduit l'expression de la divergence pour le cas de l'entropie de Rényi :

$$\tilde{K}_\alpha(q||p) = \frac{1}{\alpha - 1} \log\left(\sum_{x \in E} q(x)^\alpha p(x)^{1-\alpha}\right) \tag{3.20}$$

3.2.2 Entropies croisées

Les entropies croisées d'une loi de probabilité q par rapport à une loi de probabilité p s'obtiennent de manière similaire à partir d'une entropie. Partant

d'une entropie $H(r)$, l'entropie croisée $H(q||p)$ est obtenue en remplaçant r par p et en pondérant la somme par $\frac{q}{p}$. Ainsi, si H est de la forme (3.5) :

$$H(q||p) = -\sum_{x \in E} h(p(x)) \frac{q(x)}{p(x)} \tag{3.21}$$

Dans le cas où H est l'entropie de Shannon, l'entropie croisée est alors :

$$H_1(q||p) = -\sum_{x \in E} \log(p(x)) q(x) \tag{3.22}$$

et dans le cas où H est l'entropie d'Havrda-Charvat, on obtient :

$$H_\alpha(q||p) = \frac{1}{\alpha - 1} \times \left(1 - \sum_{x \in E} p(x)^{\alpha - 1} q(x)\right) \tag{3.23}$$

Dans le cas où H est de la forme (3.12), l'entropie croisée est de la forme :

$$\tilde{H}(q||p) = -\log \phi^{-1} \left(\sum_{x \in E} q(x) \phi(p(x))\right) \tag{3.24}$$

Ainsi, pour Rényi, on trouve :

$$\tilde{H}_\alpha(q||p) = \frac{1}{1 - \alpha} \log \left(\sum_{x \in E} p(x)^{\alpha - 1} q(x)\right) \tag{3.25}$$

3.2.3 Fonctions de perte basées sur les entropies : le choix des entropies croisées

Dans nos expériences, nous avons choisi les fonctions de perte basée sur les entropies croisées pour les tâches de classification. En effet, en apprentissage profond, c'est l'entropie croisée de Shannon qui est utilisée le plus couramment. Ceci s'explique par le fait suivant. L'entropie croisée de Shannon se décompose de la manière suivante :

$$\begin{aligned} H_1(q||p) &= -\sum_{x \in E} \log(p(x)) q(x) \\ &= \sum_{x \in E} \log\left(\frac{q(x)}{p(x)}\right) q(x) \\ &- \sum_{x \in E} \log(q(x)) q(x) \\ &= K_1(q||p) + H_1(q) \end{aligned} \tag{3.26}$$

Ainsi, minimiser l'entropie croisée de Shannon constitue un compromis entre minimiser la divergence de Kullback-Leibler et minimiser l'entropie de Shannon de la vérité terrain. Dans les faits, la vérité terrain est représentée par une loi de Dirac, ainsi l'entropie croisée de Shannon coïncide avec la divergence de Kullback-Leibler.

Dans le cas général des entropies de Havrda-Charvat et de Rényi, bien que la divergence et l'entropie croisée ne coïncident pas, on peut montrer facilement que dans le cas où la vérité terrain suit une loi de Dirac, alors les minimisations de la divergence et de l'entropie croisée sont équivalentes.
La fonction de perte que nous allons définir est utilisée dans le cadre de la classification de données médicales (images et données cliniques) de patients traités pour cancer en deux classes comme la récidive (réponse au traitement) versus non récidive (non réponse au traitement). Soit N le nombre de données, soit q_n la vérité terrain donnée par $q_n = 1$ si récidive et 0 sinon et p_n la sortie du réseau de neurone obtenue à l'aide d'une fonction sigmoïde après la dernière couche du réseau. La fonction de perte est alors donnée par :

$$L((q_n)_n, (p_n)_n) = \frac{1}{N} \sum_{n=1}^{N} H(q_n||p_n), \tag{3.27}$$

où H est une des fonctions d'entropie croisée définies précédemment.

3.3 Conclusion

Ce chapitre a exploré en profondeur les diverses fonctions de perte dérivées des entropies de Shannon, Havrda-Charvat, et Rényi, mettant en lumière leur applicabilité et efficacité dans le domaine de l'apprentissage automatique, en particulier dans le traitement des réseaux de neurones profonds pour des tâches de classification. L'entropie de Shannon, bien que largement utilisée en raison de sa simplicité et de son efficacité éprouvée, présente certaines limitations, notamment dans le cas de jeux de données déséquilibrés ou lorsqu'une discrimination plus fine entre les classes est requise. L'entropie de Havrda-Charvat, avec son hyperparamètre ajustable α, offre une flexibilité supplémentaire, permettant une adaptation plus précise au comportement spécifique des données. Les résultats obtenus montrent que pour certaines configurations de données, l'ajustement de α peut nettement améliorer les performances par rapport à l'entropie de Shannon, comme démontré dans les études de cas sur les données de cancer ORL ainsi que du poumon. D'autre part, l'entropie de Rényi généralise davantage le concept en introduisant un autre paramètre, transformant la fonction de perte en un outil encore plus adaptable, capable de moduler la sensibilité du modèle aux erreurs de différentes classes. Cependant, cette généralisation vient avec une complexité accrue dans le calcul et l'optimisation des paramètres, ce qui nécessite des recherches supplémentaires pour exploiter pleinement son potentiel. En conclusion, alors que l'entropie de Shannon reste un pilier dans la conception des fonctions de perte, les entropies de Havrda-Charvat et de Rényi émergent comme des alternatives puissantes offrant des capacités d'adaptation et de performance potentiellement supérieures, selon le contexte et la nature des données traitées. Les travaux futurs devront se concentrer sur l'exploration de ces fonctions de perte dans des contextes plus variés et sur le développement de méthodologies d'optimisation qui maximisent

leur efficacité tout en minimisant la complexité computationnelle. Lors de la suite de cette thèse, nous allons présenter les expérimentations basées sur les notions présentées dans les chapitres précédents.

CHAPITRE 4

Études de l'entropie de Havrda-Charvat dans le contexte de l'apprentissage profond pour la prédiction de l'évolution du cancer

4.1 Introduction

Dans le domaine de l'apprentissage profond, le choix de la fonction de perte revêt une importance cruciale pour la formation efficace des modèles. Traditionnellement, des fonctions de perte telles que l'entropie croisée binaire ou catégorielle basée sur l'entropie de Shannon ont été largement utilisées dans les tâches de classification. Nous avons détaillé dans le chapitre précédent comment l'entropie de Shannon peut être généralisée notamment grâce à l'entropie de Havrda-Charvat et quelles sont les nouvelles fonctions de perte associées. Dans ce chapitre, nous utiliserons ces nouvelles fonctions de perte dans les architectures de réseaux neuronaux pour différentes tâches d'apprentissage profond. Nous étudierons comment les nouvelles fonctions de perte basées sur l'entropie de Havrda-Charvat améliorent les résultats.

Les données utilisées sont issues de plusieurs bases de données médicales concernant des cas de cancers pulmonaires et ORL (oto-rhino-laryngologiques) dans le but d'améliorer la personnalisation des traitements.

4.2 Classification d'images vidéos

4.2.1 Introduction

L'endomicroscopie pulmonaire utilise la microscopie à fluorescence confocale par fibres, qui peut fournir des informations diagnostiques sur la fibrose et l'inflammation des espaces aériens distaux associés aux maladies pulmonaires. Il s'agit d'une nouvelle technologie d'imagerie en temps réel qui fournit des

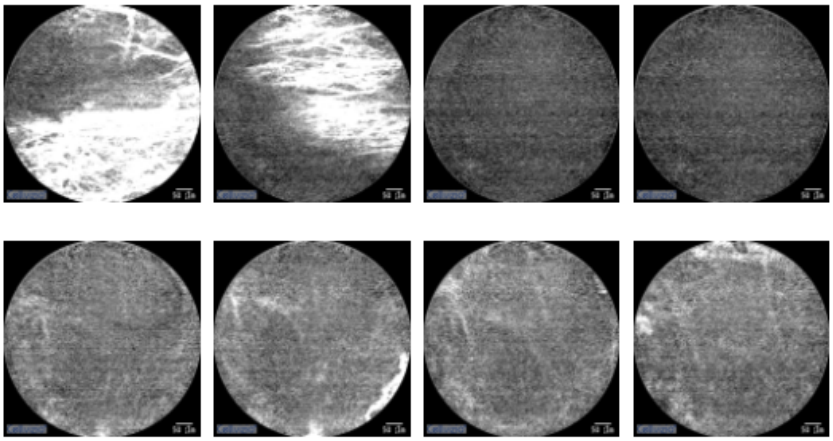


FIGURE 4.1 – Images pulmonaires informatives (1ère ligne) et non-informatives (2ème ligne).

images des alvéoles pulmonaires à un niveau microscopique. Cependant, lors de son utilisation clinique, une séquence d'images de MEEP (Micro Endoscopie Electronique Pulmonaire) peut comporter une proportion de plus de 25 % d'images non informatives à cause du bruit et des artéfacs de mouvement. Pour une future analyse des données, ces images non informatives doivent d'abord être supprimées de l'ensemble de données. Lors de l'examen clinique, la détection des images non informatives est effectuée manuellement. Cette opération manuelle est chronophage et laborieuse. Par conséquent, une détection automatique est nécessaire pour accélérer l'analyse des données et raccourcir le temps de diagnostic. Notre travail vise à développer une méthode de détection automatique [C1] pour supprimer les images non informatives d'une séquence vidéo. La figure 4.1 montre quatre images informatives et quatre images non informatives. Nous pouvons observer que les textures dans ces deux types d'images sont différentes.

Afin d'effectuer la classification, nous proposons d'utiliser un réseau de neurones convolutif (CNN) [R1]. La figure 4.2 présente l'architecture générale des CNN. De nombreuses méthodes de classification basées sur les CNN ont été développées dans les domaines de l'imagerie médicale. La définition de la fonction de perte pour l'entraînement du réseau est importante pour obtenir de bonnes performances en termes d'exactitude. La plupart des études utilisant l'entropie en apprentissage profond se sont concentrées sur les formes les plus courantes telles que la métrique quadratique ou l'entropie croisée de Shannon. La métrique quadratique est généralement utilisée lorsque la sortie désirée est déterministe ; par exemple, dans la segmentation d'images. Les fonctions de perte d'entropie croisée sont généralement utilisées pour la prise de décision probabiliste ; par exemple, dans la classification. L'entropie de Shannon est l'entropie la plus connue et l'entropie croisée dérivée peut être facilement interprétée comme une "distance" entre les probabilités. De plus, l'entropie de Shannon est plus pertinente lorsque les données suivent une distribution de Gibbs (c'est-à-dire

l'exponentielle de l'opposé d'une fonction d'énergie convexe). Contrairement à l'entropie de Shannon, l'entropie de Havrda-Charvat ne nécessite pas de distribution spécifique des données pour conserver sa spécificité et sa précision. Dans ce chapitre, nous proposons de mettre en œuvre un CNN avec une fonction de perte dérivée de l'entropie de Havrda-Charvat pour classer les images étudiées dans leurs catégories appropriées (informatives/non informatives) et comparer les résultats avec une entropie croisée basée sur Shannon afin de déterminer quelles améliorations peuvent être obtenues. Afin de valider nos résultats, nous avons la précision qui est définie par :

$$\text{Précision} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

où TP, TN, FP et FN désignent respectivement le taux de vrais positifs, de vrais négatifs, de faux positifs et de faux négatifs.

4.2.2 Travaux connexes

Il existe de nombreuses méthodes de classification des données. Certaines méthodes, telles que les Machines à Vecteurs de Support (SVM) [26], sont géométriques. Les SVM sont inspirées du théorème de Hahn-Banach et sont une méthode d'apprentissage supervisé consistant à séparer deux sous-ensembles convexes par un hyperplan. Les SVM ont été généralisées dans le cas non linéaire par l'introduction d'un noyau métrique, et propose d'estimer le noyau via une inférence bayésienne. La méthode de l'arbre de décision [92] construit des classificateurs par arbre à la fois en régression et en classification. Un arbre est construit en divisant progressivement une population en deux sous-populations afin d'optimiser l'homogénéité des populations selon leur étiquette. Les arbres de décision ont été généralisés en forêts aléatoires, qui regroupent un grand nombre d'arbres de décision indépendants. Ceux-ci sont construits à partir de la même base d'apprentissage en utilisant différents processus aléatoires. Le fait de combiner plusieurs arbres de décision permet de réduire l'influence des données bruitées pendant la phase d'apprentissage. Les forêts aléatoires ont été utilisées avec succès en classification ; par exemple en radiomique. D'autres méthodes de classification sont statistiques. Parmi les méthodes statistiques, la régression logistique [39] consiste à apprendre les paramètres de la fonction de classification logistique à partir de données annotées. Dans la maximisation de la vraisemblance [69], un modèle probabiliste représentant la probabilité de chaque classe en fonction de l'observation est donné ; la décision consiste à choisir la classe la plus probable. Une telle maximisation de vraisemblance est également utilisée dans les champs aléatoires conditionnels (CRF) [95] plus récents pour la segmentation d'images et pour la classification. Un autre type de méthodes statistiques est les méthodes bayésiennes dans lesquelles nous avons une connaissance a priori de l'appartenance à une classe. Les méthodes bayésiennes ont été largement utilisées en segmentation d'images et en classification de données. Plus récemment, les méthodes d'apprentissage en profondeur sont

de plus en plus utilisées. Parmi les applications de l'apprentissage en profondeur en classification de données, on peut citer la reconnaissance de texte, la reconnaissance faciale ou la détection de spams. La méthode d'apprentissage en profondeur la plus classique pour la classification supervisée est basée sur les réseaux neuronaux convolutifs (CNN). Il existe différentes architectures basées sur CNN. Parmi elles, l'architecture LeNet est le premier CNN réussi utilisé pour classifier des chiffres ; AlexNet a été le premier CNN appliqué à la vision par ordinateur et a été soumis au défi ImageNet ILSVRC en 2012 ; ZFNet est une amélioration d'AlexNet proposée dans [114]. La détection d'images non informative a été explorée en 2019 [110] et 2022 [111], c'est à dire de façon contemporaine par rapport à cette étude. Ces deux études ont utilisé l'apprentissage profond pour classer automatiquement les images informatives dans des flux issus respectivements de coloscopies et de laryngoscopies. Les réseaux de neurones ont été utilisés soit isolément, soit couplés avec un algorithme d'extraction des caractéristiques principales pour réaliser la tâche dédiée de classification. Les résultats étaient très prometteurs, avec un F1-score de 0.775 et une précision de 92% en moyenne respectivement.

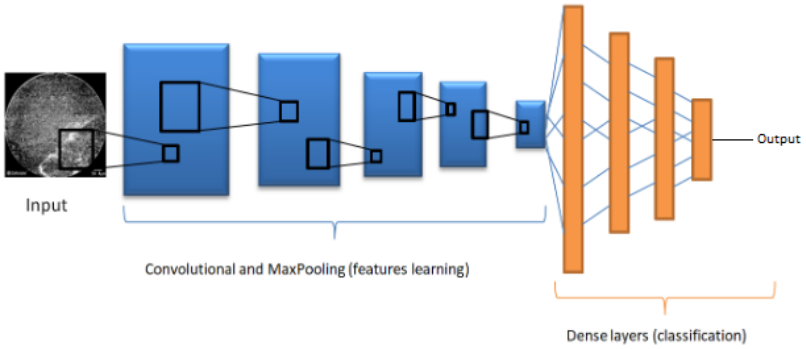


FIGURE 4.2 – Architecture CNN pour classification d'images pulmonaires.

4.2.3 Architecture CNN pour la classification

Au regard de l'implémentation, nous avons proposé d'adopter une architecture CNN simple. Les deux principaux types de couches dans un CNN sont les suivants :

1. Les premières couches sont les couches de convolution. Leur fonction est de réduire l'entrée à ses caractéristiques les plus saillantes, afin de conserver les informations les plus importantes tout en réduisant les dimensions de l'image d'entrée. Ces couches sont souvent associées à une fonction d'activation, utilisée pour déterminer la valeur de la sortie, et éventuellement à une opération de MaxPooling, qui vise à réduire encore davantage les dimensions et à extraire les caractéristiques.

2. Les secondes couches sont les couches entièrement connectées, dont le rôle est d'établir une règle de décision qui conduit, dans la dernière couche, à la classification de l'image d'entrée.

Nous avons implémenté les couches suivantes : cinq couches convolutionnelles de taille décroissante (128, 64, 32, 16 et 8 neurones) et ayant un filtre de convolution de dimensions 3×3. La fonction d'activation utilisée est l'Unité Linéaire Rectifiée (ReLU), puis quatre couches denses de taille décroissante (128, 64, 32 et 16 neurones). Le nombre total de paramètres entraînables dans cette architecture était d'environ 6.6 millions. Le partage de validation de ce réseau était fixé à 70/30, c'est-à-dire que 70 % servaient de set d'entraînement, et les 30 % restants étaient utilisés comme set de validation afin que le CNN puisse s'améliorer. La taille des lots, le nombre de données passées simultanément à travers le CNN, était de 64. Nous avons également ajouté des couches de Dropout après les trois premières couches denses. Ces couches obligent le réseau à abandonner certains liens, réduisant ainsi le phénomène de sur-apprentissage.

D'un point de vue technique, il existe plusieurs façons de généraliser l'entropie de Shannon classique et les différentes métriques et divergences comme résumé dans d'autres études. Il existe deux principales façons de généraliser l'entropie de Shannon : la première consiste à remplacer la fonction intégrée dans l'expression de l'entropie de Shannon et la seconde selon un axiome. Comme souligné dans le chapitre précédent, les entropies de Havrda-Charvat et de Shannon sont les seules à satisfaire la propriété de récursivité forte. Même si la généralisation de l'entropie de Shannon comme Havrda-Charvat n'est pas récente ; la plupart de leurs applications sont la classification ou la théorie du codage. Dans [52], une propriété de récursivité affaiblie est étudiée et la généralisation de la Havrda-Charvat satisfaisant cette propriété est proposée. Cette propriété facilite son utilisation dans le cas de la classification multi-étiquettes, comme par exemple pour l'analyse de l'expression des gènes. Dans [54], les résultats de la classification d'images de mammographie sont comparés en utilisant différentes entropies paramétrées : Rényi, Havrda-Charvat, entropie de Tsallis et entropie de Kapur. Il apparaît que l'entropie de Tsallis donne les meilleurs résultats dans le cadre de leur étude. Cependant, l'entropie de Tsallis a deux paramètres alors que l'entropie de Havrda-Charvat n'en a qu'un seul et donne des résultats légèrement moins performants. L'entropie de Havrda-Charvat a également été utilisée pour le regroupement dans en remplaçant l'entropie de Shannon par l'entropie de Havrda-Charvat dans la divergence de Jensen-Shannon. Dans [65], une équation fonctionnelle à partir de laquelle les entropies de Havrda-Charvat, de Shannon et de Tsallis sont proposées. De plus, la stabilité au sens de Hyers-Ulam de cette équation fonctionnelle est étudiée.

Dans cette section, nous présentons le réseau de neurones que nous utilisons pour la classification des images optiques. Nous utilisons une méthode supervisée pour la classification des images optiques des poumons en deux classes : classe informative et classe non informative. La méthode que nous utilisons est un réseau de neurones convolutif (CNN) dont l'architecture est représentée

à la figure 4.4. Les fonctions de perte que nous utilisons pour l'apprentissage supervisé sont des fonctions basées sur l'entropie. Ces fonctions de perte comparent deux lois de probabilité. La première est la sortie du CNN et représente la probabilité estimée de l'appartenance de l'image d'entrée à la classe informative. Cette probabilité estimée est calculée par l'étape de sigmoïde du CNN. La deuxième probabilité est une probabilité de Dirac dont la valeur est 1 si l'image est informative et 0 sinon. Une fonction de perte basée sur l'entropie est construite en choisissant une fonction d'entropie et une divergence pour comparer les entropies des deux distributions. La première entropie est l'entropie de Shannon classique et la deuxième est l'entropie de Havrda-Charvat qui généralise l'entropie de Shannon.

Dans cette étude, nous nous concentrons sur la classification des images optiques en deux classes : images informatives et non informatives. L'ensemble des états possibles est $\Omega = \{0, 1\}$ où 1 correspond à l'événement "image informative" et 0 à "image non informative". En classification supervisée, nous entraînons le réseau à partir d'une base de données annotée. Soit N le nombre d'images annotées. À la n-ième image, nous associons la probabilité de Dirac suivante : $q_n(1) = \{1$ si l'image est informative, 0 sinon$\}$. Soit p_n la sortie du CNN pour la n-ième image. La fonction de perte à minimiser est :

$$L_\alpha(q, p) = \frac{1}{N} \sum_{n=1}^{N} H_\alpha(q_n || p_n) \tag{4.2}$$

où H_α est l'entropie croisée de Havrda-Charvat. Pour rappel, les détails de la formule de Havrda-Charvat sont explicités dans le chapitre 3.

4.2.4 Résultats et analyse

Données

Dans cette étude, l'ensemble de données se compose de différents types de données. Les données ont été obtenues sous forme de séquences vidéo et converties en images individuelles pour l'analyse. Les maladies possiblement observables dans ces vidéos sont : l'asbestose, la fibrose pulmonaire idiopathique, la pneumopathie d'hypersensibilité, la sclérodermie et également des personnes en bonne santé comme groupe de contrôle. Nous avons utilisé 3895 images dont 2313 étaient des images informatives et 1582 étaient des images non informatives. Ces images ont été obtenues à partir de séquences d'images fournies par le CHU (Centre Hospitalier Universitaire) de Rouen, en Normandie. La taille des images varie entre 512×512 et 500×500 pixels, en fonction de la séquence. Nous avons proposé de normaliser toutes les images dans une résolution de 128×128 pixels, afin d'entraîner le CNN avec une résolution raisonnable en considérant un équilibre entre la taille de l'image et le temps de traitement.

Conditions d'apprentissage

Nous avons utilisé des ensembles de données équilibrés, avec environ la moitié des images non informatives et la moitié des images informatives. De cette manière, nous avons évité le biais de disponibilité qui aurait orienté le modèle vers le type de données le plus disponible. Un tel biais peut potentiellement fausser les données où rien de tel ne peut être expliqué par la distribution des données. Nous avons utilisé le langage Python avec la bibliothèque Keras qui est une puissante bibliothèque Python facile à utiliser pour développer et évaluer des modèles d'apprentissage profond. Les critères d'évaluation étaient ceux mis en œuvre avec le CNN, c'est-à-dire le nombre d'identifications correctes sur la longueur totale de l'ensemble de données.

Expérimentation

Lors de nos expérimentations, nous avons employé la fonction de perte issue de l'entropie de Shannon comme baseline de performances pour comparaison avec celle dérivée de l'entropie d'Havrda-Charvat. Dans notre algorithme de travail, lorsque le paramètre α était égal à 1, nous utilisions la fonction de perte usuelle pour la classification.

Une étude comparative entre différentes entropies est réalisée en fonction du nombre d'époques pour définir leur précision (voir Tableau 4.1). N_e représente le nombre d'époques défini dans l'apprentissage du réseau, et α est le paramètre de l'entropie croisée d'Havrda-Charvat (Eq. 3.23). Le tableau suivant a été réalisé pour un ensemble de données de 2947 images (1625 informatives / 1322 non informatives). Dans cette étude, nous avons effectué une comparaison entre différentes valeurs de l'entropie de Havrda-Charvat pour diverses époques afin de déterminer leur précision définie en (4.1) (voir Tableau 4.1).

N_e	α=**1.0**	α=**1.1**	α=**1.3**	α=**1.5**	α=**2.0**
20	0,59	0,65	0,62	0,63	0,59
30	0,59	0,68	0,73	0,59	0,59
40	0,70	**0,79**	0,77	0,76	0,59
50	0,64	0,68	0,69	0,70	0,59

TABLE 4.1 – Comparaison quantitative de la précision avec différentes valeurs de α pour N_e époques pour des images pulmonaires.

Nous avons constaté que le nombre d'époques a une influence modérée sur les résultats finaux, à l'exception des cas où le nombre d'époques est délibérément trop faible pour atteindre une convergence correcte. Pendant les tests, nous avons remarqué que les résultats présentaient une plus grande disparité lorsque le nombre d'époques était faible. Cela s'explique par le fait que le réseau sélectionne un ensemble aléatoire d'images d'entrée à partir de l'ensemble de données pour effectuer à la fois l'entraînement et les tests. En augmentant le nombre d'époques, nous nous assurons que le réseau atteint la convergence et diminue ainsi le facteur aléatoire dans l'analyse du sous-ensemble.

Les valeurs de sensibilité et de spécificité nous permettent de conclure que la valeur de $\alpha = 2.0$ doit être écartée. Ceci peut s'expliquer par le fait que plus α est grand et plus l'entropie croisée à des difficultés à discriminer une variable aléatoire binomiale d'une variable aléatoire de Dirac (déterministe) comme illustré dans la Figure 3.1. Ces résultats concordent avec nos observations antérieures. Les valeurs restantes de α sont satisfaisantes en termes de précision du modèle.

Les meilleurs résultats pour les éléments supervisés ont été obtenus avec $\alpha = 1.1$ et 40 époques. Ces résultats démontrent qu'un choix judicieux du paramètre α dans l'entropie croisée de Havrda-Charvat peut donner des améliorations en comparaison avec l'entropie croisée de Shannon.

4.2.5 Conclusion

Dans ce travail nous avons conçu un classificateur CNN avec une nouvelle fonction de perte basée sur l'entropie de Havrda-Charvat. La plupart des classificateurs CNN utilisent l'entropie de Shannon, tandis que l'entropie de Havrda-Charvat est une entropie de Shannon généralisée. Par conséquent, elle peut la surpasser si la nature des données ne peut pas satisfaire certaines conditions. Notre application vise à classer les images d'endomicroscopie pulmonaire dans lesquelles les images informatives et non informatives ne sont pas faciles à distinguer. Le classificateur proposé peut atteindre une précision de 79%, meilleure que celle obtenue par l'entropie de Shannon. Ce travail a été publié dans [14]. Nous discutons à présent des avantages et limitations.

A propos de la fonction de perte

Nous remarquons, grâce au tableau 4.1, que pour $\alpha = 1.0$ (Shannon), $\alpha = 1.1$ et $\alpha = 1.3$, les erreurs d'entraînement et de validation diminuent au même rythme pour $N_e = 40$. En ce qui concerne l'entropie de Shannon, où les deux erreurs sont plus élevées que pour les deux modèles basés sur Havrda-Charvat, nous pouvons en déduire que, bien qu'elles diminuent également légèrement, l'entropie de Shannon est moins pertinente et moins fiable. Cependant, pour un nombre d'époques N_e supérieur à 40, le phénomène de surajustement commence à apparaître. Nous avons testé les modèles pour $N_e = 100$ époques et le phénomène de surajustement apparaît et augmente au fur et à mesure des époques. Pour la taille de l'ensemble de données étudié, 40 époques peuvent être considérées comme suffisantes, mais à mesure que la taille de l'ensemble de données augmente, un plus grand nombre d'époques sera disponible sans déclencher de surajustement.

En résumé, pour des études ultérieures, la valeur $\alpha = 1.1$ semble être la plus prometteuse et une amélioration par rapport à l'entropie de Shannon communément utilisée. En ce qui concerne l'autre valeur, nous pouvons remarquer qu'après 40 à 50 époques, la perte de validation tend à augmenter, tandis que la perte d'entraînement atteint son minimum. Si cette valeur devait être utilisée, il faudrait prendre en compte que, pour un ensemble de données de taille

comparable, le nombre d'époques devrait être maintenu en dessous de 50, ou adapté pour refléter ce niveau d'ajustement, car l'augmentation des époques au-delà de ce point diminuerait bientôt l'adaptabilité du réseau en raison du surajustement. Nous remarquons que l'erreur de validation tend à augmenter loin de celle d'entraînement lorsque N_e devient supérieur à 30. Ce phénomène tend à s'accentuer à mesure que les époques progressent. Encore une fois, nous pouvons remarquer que les époques ne doivent pas être augmentées au-delà de 40, car les deux pertes semblent s'éloigner de plus en plus l'une de l'autre à un rythme rapide. De plus, nous remarquons que le phénomène devient plus important lors de l'utilisation de l'entropie de Shannon que de l'équation de Havrda-Charvat, ce qui constitue un autre élément en faveur de la nouvelle fonction de perte par rapport à l'entropie de Shannon plus couramment utilisée. Cela commence déjà à $N_e = 40$ époques pour la perte basée sur Shannon, comme le montre la table 4.1 , d'où la moindre fiabilité des résultats obtenus grâce à elle, y compris la sensibilité et la spécificité.

Parmi les images correctement classées, nous trouvons celles avec des structures évidentes et clairement identifiées que le réseau identifie comme informatives. À l'inverse, les images clairement non informatives sont reconnues par l'algorithme et considérées comme du bruit comme présenté dans la Figure 4.1. Plusieurs images non informatives correctement classées Malgré les résultats encourageants obtenus par le réseau, certaines limitations existent.

Limitations

Lors des expériences, nous avons remarqué que plusieurs images étaient classées dans la mauvaise catégorie. Par exemple, les images suivantes peuvent être classées comme informatives alors qu'elles étaient non informatives. Pour expliquer cette erreur, nous pouvons dire que les données sont très bruitées et que toute structure dans l'image peut passer pour une information et ainsi rendre l'image informative, selon le réseau de neurones. Les structures sont difficiles à prendre en compte, car la carte des caractéristiques à la fin des couches de convolution est aplatie pour répondre aux exigences des couches entièrement connectées en matière d'entrées. Lorsque les véritables structures deviennent plus floues ou sont mélangées avec beaucoup de bruit, cela peut entraîner la classification de l'image comme non informative alors qu'une structure est présente :

Ainsi, les images bruitées présentant toujours des éléments informatifs sont susceptibles d'être considérées comme ne contenant que du bruit et donc comme non informatives, les données utiles étant perdues parmi les valeurs de bruit. Une solution possible serait de prétraiter les données avec des filtres de lissage afin de réduire l'influence du bruit sur le signal.

Rareté des données et surapprentissage

Comme présenté dans la partie précédente, il y a un phénomène de surapprentissage plus nous augmentons le nombre d'époques. Ce phénomène pro-

vient du petit ensemble de données que nous utilisons. En général, plus il y a d'époques, mieux c'est, mais avec des bases de données plus petites, le réseau a tendance à trop bien apprendre son ensemble d'entraînement et devient incapable de s'adapter à d'autres données. C'est pourquoi, pour des recherches ultérieures, nous aurons besoin de plus de données labellisées. Une autre option concernant le surapprentissage est de réduire la complexité du réseau en réduisant le nombre de neurones dans les couches existantes ou en supprimant certaines couches sans altérer la structure globale du réseau. En ce qui concerne la rareté des données labellisées, une solution peut être d'augmenter le jeu de données par des algorithmes de type génératif.

Hyperparamètre α

Dans cette étude, nous avons proposé de fixer le paramètre α à des valeurs prédéterminées et avons obtenu les résultats correspondants. Nous avons remarqué que, selon les données disponibles, la meilleure valeur de l'hyperparamètre avait tendance à varier. C'est pourquoi nous pouvons supposer qu'il peut exister des valeurs de α produisant de meilleurs résultats dans l'intervalle considéré. Afin d'améliorer encore nos résultats, l'algorithme doit être modifié pour tester automatiquement les valeurs de α et fournir la valeur qui a donné les meilleurs résultats.

4.3 Prédiction de la réponse au traitement

4.3.1 Introduction

Dans cette section nous proposons un réseau de neurones multitâches et multi-données, avec l'utilisation de l'entropie croisée d'Havrda-Charvat comme fonction de perte [R2]. L'utilisation d'un réseau multi-tâches consiste à entraîner un réseau de neurones profonds pour effectuer simultanément plusieurs tâches avec plusieurs types de données, telles que la prédiction et la classification, en exploitant la capacité de l'entropie croisée d'Havrda-Charvat à gérer efficacement les incertitudes.

4.3.2 Jeux de données

Les jeux de données sont composés de 580 patients, parmi lesquels 434 souffrent de cancer ORL et 146 de cancer du poumon. Comme nous pouvons le constater, les deux jeux de données sont de petite taille. Nous avons choisi de mener des expériences sur les deux sous-ensembles (tête-cou et poumon) ensemble et séparément. Comme nous le verrons, la valeur optimale de α dépend du type de données utilisées. Les résultats utilisant le jeu de données total de 580 sont médiocres. Nous montrons donc les résultats pour les jeux de données séparés. Les images CT utilisées dans le réseau de neurones ont été redimensionnées avec une résolution d'image de $128 \times 128 \times 64$ voxels. Les

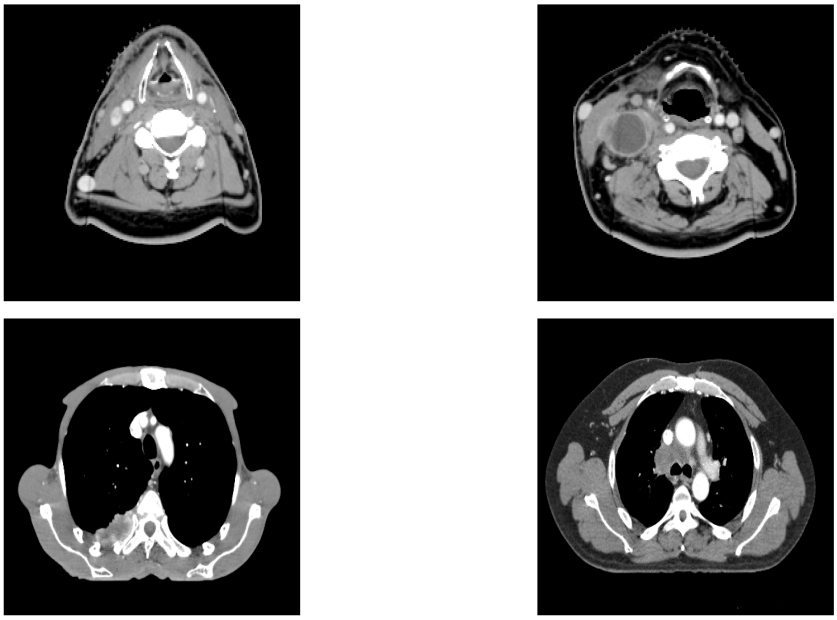


FIGURE 4.3 – Images CT de deux patients : ORL (première ligne) et pulmonaires (deuxième ligne).

informations sur les patients utilisées comme données d'entrée dans le réseau de neurones sont de deux types, à savoir quantitatives et qualitatives, comme le montrent les tableaux 4.2 et 4.3.

Donnée Clinique	**Modalité**
Hémoglobine	g/dL
Lymphocytes	Giga/L
Leucocytes	Giga/L
Thrombocytes	Giga/L
Albumine	g/L
Durée du traitement	Jours
Dose totale d'irradiation	Gy
Nombre de fractions	/
Dose moyenne par fraction	Gy
Poids au début et à la fin du traitement	kg

TABLE 4.2 – Données cliniques quantitatives traitées par le réseau.

4.3.3 Architecture multitâches

L'utilisation d'architectures U-Net dans le domaine de la reconstruction d'images médicales représente une avancée significative en imagerie biomédicale. L'architecture U-Net, reconnue pour son efficacité en matière de segmentation d'images, est dans notre cas, un cadre multitâches pour traiter simulta-

Donnée Clinique	**Modalité**
Genre	M/F
Tabacologie	Fumeur, non-fumeur, ancien fumeur
Utilisation de chimiothérapie d'induction	Oui/Non
Utilisation de chimiothérapie concomitante	Oui/Non
TNM	Tumeur, Ganglion (Node), Métastase

TABLE 4.3 – Données cliniques qualitatives traitées par le réseau.

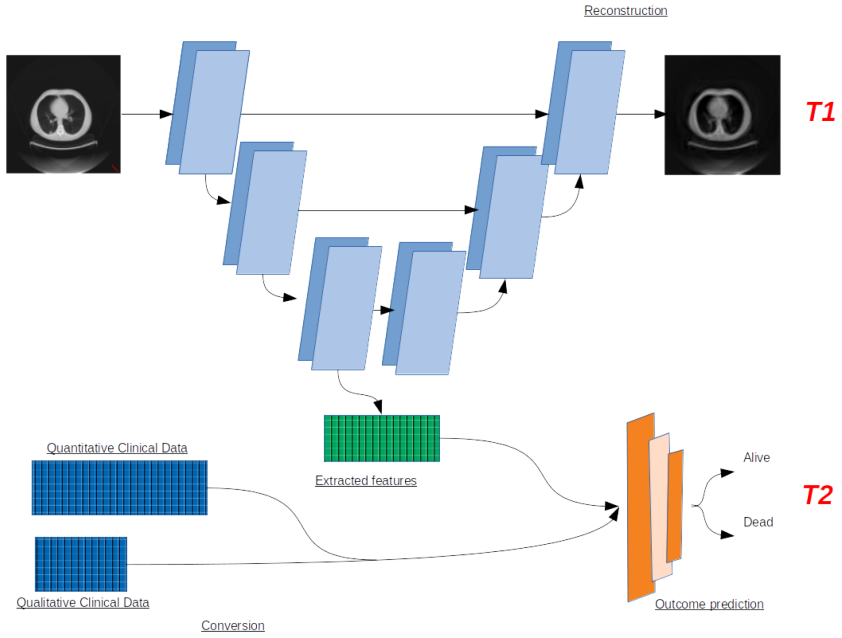


FIGURE 4.4 – Architecture multitâches. T1 : reconstruction de l'image, T2 : prédiction.

nément la tâche complexe de reconstruction d'images de haute qualité à partir de données incomplètes ou bruitées, tout en identifiant les caractéristiques d'image critiques qui peuvent indiquer un risque de récidive de maladie. Cette approche intégrée permet non seulement une amélioration significative de la clarté et de la précision des images médicales, comme les IRM ou les tomodensitogrammes, mais ouvre également la voie à des analyses prédictives basées sur des caractéristiques d'image subtilement dérivées, qui pourraient échapper à l'analyse humaine. En exploitant des réseaux de neurones profonds pour apprendre des représentations riches et des caractéristiques discriminantes à partir d'ensembles de données médicales, ils peuvent prédire la probabilité de récidive de maladies telles que le cancer, en fournissant des informations précieuses pour la prise de décision clinique. Cette synergie entre reconstruction d'image et analyse prédictive dans une architecture unifiée permet non seulement d'améliorer l'efficacité diagnostique, mais aussi de personnaliser les traitements en fonction des risques identifiés, contribuant ainsi à une médecine plus précise et orientée vers le patient.

L'architecture multitâches que nous proposons comprenant deux tâches : reconstruction et prédiction ; la fonction de perte totale est la somme de deux contributions. La première contribution, pour la reconstruction, est la fonction de perte quadratique donnée par :

$$L_{\text{rec}}(y, \hat{y}) = \frac{1}{N} \sum_{n=1}^{N} \|y_n - \hat{y}_n\|^2, \tag{4.3}$$

où y_n est l'image vérité-terrain pour le n-ième patient, $\hat{y}_n$ est la sortie de l'étage de reconstruction et la norme euclidienne $\|.\|^2$ est donnée par :

$$\|y_n - \hat{y}_n\|^2 = \sum_{k=1}^{P} (y_{n,k} - \hat{y}_{n,k})^2, \tag{4.4}$$

où P est le nombre de pixels et $y_{n,k}, \hat{y}_{n,k}$ les valeurs respectives d'intensité pour le k-ième pixel.
La deuxième contribution, pour la prédiction est la fonction de perte basée sur l'entropie croisée de Havrda-Charvat $L_\alpha(q, p)$ définie par (4.2). Au final, la fonction de perte totale est :

$$L((q, y), (p, \hat{y})) = L_{\text{rec}}(y, \hat{y}) + L_\alpha(q, p) \tag{4.5}$$

Étant donné que l'étude [C2] a été menée sur de petits jeux de données (434 et 146 patients), une stratégie de validation des résultats était nécessaire. Nous avons proposé d'utiliser la validation croisée k-fold pertinente pour la validation de petites données. Dans notre travail, nous avons utilisé une validation croisée à cinq plis. La procédure se déroule comme suit :

1. Mélanger le jeu de données.
2. Le diviser aléatoirement en 5 sous-ensembles.
3. Pour chaque sous-ensemble :
   — Prendre le sous-ensemble comme jeu de données de test.
   — Prendre les autres ensembles comme données d'entraînement.
   — Ajuster le modèle aux données d'entraînement et l'évaluer sur le jeu de données de test.
   — Conserver le score d'évaluation.
4. Résumer la compétence du modèle à partir des échantillons de scores d'évaluation du modèle.

De plus, la précision a été proposée comme métrique d'évaluation. Dans ce cas, elle consistait à comparer les valeurs de la vérité terrain et de la prédiction et à sommer toutes ces occurrences sur la taille du jeu de données.

Expérimentations

Résultats de prédiction de la réponse au traitement Les tableaux suivants ont été obtenus suite aux expérimentations sur la prédiction de la

réponse au traitement. Les chiffres 1 à 5 correspondent aux 5 volets de la validation croisée "5-folds". Les résultats présentés dans les Tableaux 4.4 et 4.5 illustrent les performances de la fonction de perte dérivée de l'entropie de Havrda-Charvat sur deux jeux de données distincts, à savoir ceux des cancers ORL ainsi que du poumon. Pour le jeu de données sur le cancer ORL, une amélioration notable de la précision est observée lorsque la valeur de α dépasse 1, avec un pic de performance pour $\alpha = 1.5$, où la précision moyenne atteint 0.80, significativement supérieure à celle obtenue avec l'entropie de Shannon. En revanche, pour le cancer du poumon, les valeurs optimales de α sont légèrement plus élevées, démontrant les meilleures précisions autour de $\alpha = 2.1$ à $\alpha = 2.3$, avec des précisions moyennes excédant 0.79, ce qui suggère une adaptation plus spécifique du paramètre selon le type de données traitées. Ces résultats confirment que la fonction de perte de Havrda-Charvat, ajustée via α, peut être particulièrement efficace pour certains types de données médicales, tout en indiquant l'importance d'un réglage soigneux de l'hyperparamètre pour maximiser la performance du modèle.

α	1	2	3	4	5	Moyenne	Écart-type	p-valeur
0.1	0.68	0.53	0.60	0.58	0.63	0.60	0.06	0.01
0.3	0.60	0.70	0.70	0.70	0.43	0.63	0.12	0.28
0.5	0.58	0.58	0.60	0.70	0.73	0.64	0.07	0.27
0.7	0.85	0.70	0.60	0.70	0.65	0.70	0.09	0.25
0.9	0.58	0.60	0.60	0.60	0.68	0.61	0.04	0.07
1.0	0.75	0.65	0.68	0.58	0.70	0.67	0.06	N/A (Shannon)
1.1	0.68	0.75	0.75	0.73	0.75	0.73	0.03	0.09
1.3	0.63	0.73	0.68	0.75	0.70	0.70	0.05	0.32
1.5	0.70	0.78	0.85	0.80	0.88	0.80	0.07	0.03
1.7	0.80	0.73	0.63	0.75	0.78	0.74	0.07	0.07
1.9	0.75	0.73	0.73	0.75	0.83	0.76	0.05	0.02
2.1	0.68	0.63	0.60	0.62	0.575	0.63	0.04	0.12
2.3	0.73	0.73	0.73	0.73	0.70	0.72	0.01	0.09
2.5	0.75	0.58	0.68	0.68	0.60	0.66	0.07	0.34
2.7	0.68	0.53	0.45	0.63	0.60	0.58	0.09	0.04
2.9	0.73	0.73	0.73	0.75	0.73	0.73	0.01	0.07
3.1	0.70	0.55	0.65	0.57	0.63	0.62	0.06	0.02
3.3	0.65	0.65	0.65	0.58	0.55	0.62	0.05	0.07
3.5	0.73	0.75	0.73	0.75	0.70	0.73	0.02	0.08
3.7	0.73	0.70	0.60	0.60	0.58	0.64	0.07	0.20
3.9	0.68	0.70	0.55	0.58	0.53	0.61	0.08	0.09

TABLE 4.4 – Précision obtenue par la fonction de perte dérivée de l'entropie de Havrda-Charvat en fonction de α pour le jeu de données sur le cancer ORL (les p-valeurs inférieures à 0.05 et les précisions supérieures à celles de Shannon sont mises en évidence en bleu).

α	1	2	3	4	5	Moyenne	Écart-type	p-valeur
0.1	0.58	0.58	0.47	0.58	0.63	0.57	0.06	0.23
0.3	0.58	0.58	0.58	0.58	0.68	0.60	0.04	0.13
0.5	0.63	0.58	0.52	0.58	0.52	0.56	0.03	0.26
0.7	0.58	0.58	0.63	0.63	0.58	0.60	0.03	0.13
0.9	0.63	0.53	0.68	0.47	0.53	0.57	0.08	0.21
1.0	0.73	0.47	0.47	0.47	0.47	0.52	0.12	N/A (Shannon)
1.1	0.63	0.63	0.52	0.52	0.52	0.56	0.06	0.18
1.3	0.68	0.53	0.53	0.47	0.53	0.55	0.08	0.15
1.5	0.58	0.53	0.53	0.47	0.53	0.53	0.04	0.44
1.7	0.73	0.47	0.63	0.57	0.42	0.56	0.12	0.37
1.9	0.69	0.63	0.58	0.53	0.68	0.62	0.07	0.04
2.1	0.79	0.84	0.73	0.79	0.79	0.79	0.04	0.004
2.3	0.84	0.78	0.84	0.73	0.84	0.81	0.05	0.002
2.5	0.79	0.84	0.74	0.68	0.63	0.74	0.08	0.007
2.7	0.79	0.74	0.69	0.74	0.74	0.74	0.04	0.003
2.9	0.79	0.79	0.79	0.79	0.73	0.78	0.03	0.004
3.1	0.74	0.74	0.78	0.78	0.68	0.74	0.04	0.008
3.3	0.79	0.79	0.74	0.74	0.74	0.76	0.03	0.003
3.5	0.74	0.73	0.68	0.33	0.58	0.61	0.17	0.12
3.7	0.68	0.63	0.47	0.63	0.73	0.63	0.09	0.07
3.9	0.78	0.53	0.47	0.47	0.63	0.58	0.13	0.07

TABLE 4.5 – Précision obtenue par la fonction de perte dérivée de l'entropie de Havrda-Charvat en fonction de α pour le jeu de données sur le cancer du poumon (les p-valeurs inférieures à 0.05 et les précisions supérieures à celles de Shannon sont mises en évidence en bleu).

De plus, la reconstruction des images d'entrée a été suivie avec attention pour s'assurer que les caractéristiques extraites par les couches de convolution sont pertinentes pour représenter et donc reconstruire correctement l'image. Les résultats de la reconstruction sont présentés dans la Figure 4.5

4.3.4 Architecture multitâches et multidonnées pour la prédiction de la réponse au traitement

En considérant que les données sont peu nombreuses, nous proposons de mélanger les deux bases de données. L'objectif est d'augmenter le jeu de données, puisque les images possèdent les mêmes caractéristiques. Dans cette expérimentation, nous avons conservé les mêmes données et proposé une architecture différente pour classifier les images d'entrée entre cancer ORL et du poumon en plus de prédire la réponse au traitement. L'architecture proposée est présentée dans la Figure 4.6. On considère que les trois tâches peuvent partager les caractéristiques communes.

Trois tâches principales sont exécutées conjointement par le réseau. T1 est similaire à celle du réseau précédent, et utilise la même fonction de perte

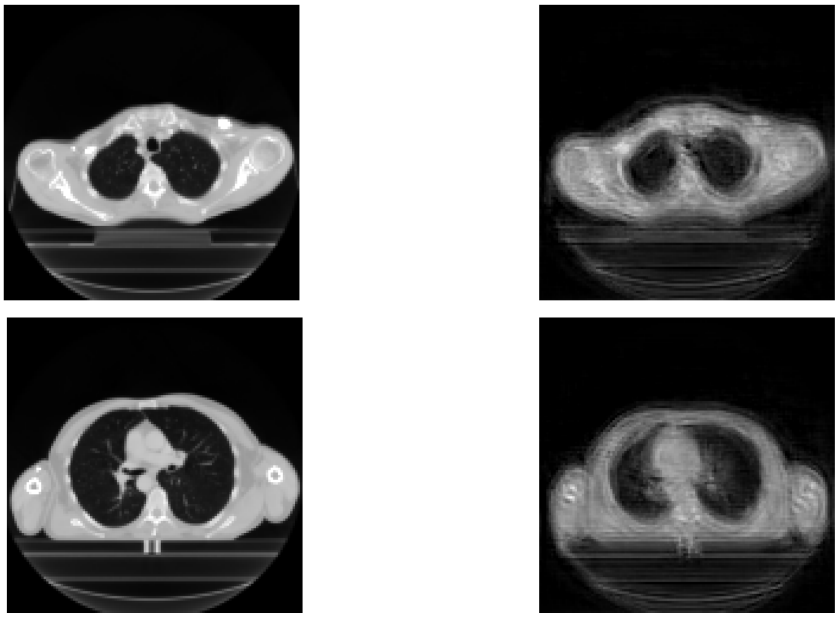


FIGURE 4.5 – Images d'entrée (gauche) vs. images reconstruites (droite).

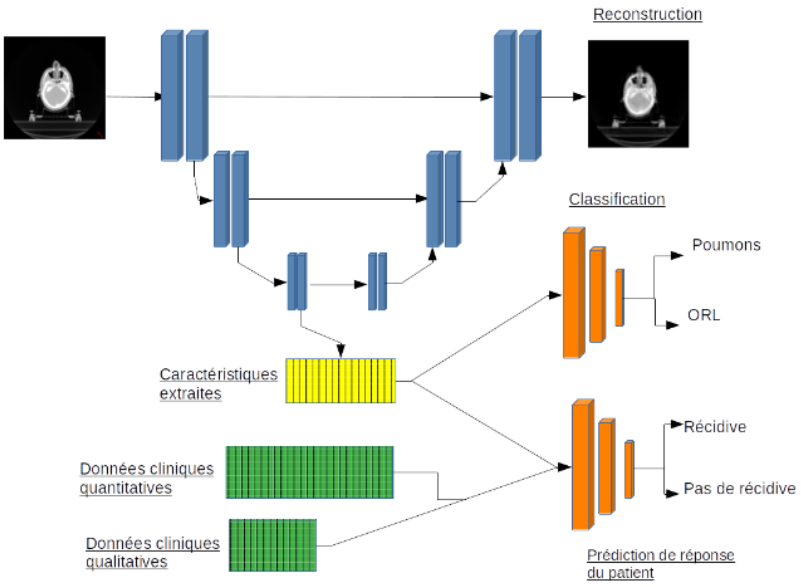


FIGURE 4.6 – Architecture multitâches et multidonnées proposée. T1 : reconstruction d l'image, T2 : classification de patients, T3 : prédiction de la réponse au traitement.

décrite dans la formule 4.3. T2 est la tâche de classification et la fonction de perte associée est donnée par 4.2 pour $\alpha = 1$, ce qui correspond à l'entropie croisée de Shannon. La dernière tâche, T3 concerne la prédiction de la réponse au traitement. Comme toute tâche de classification, nous utilisons également l'entropie croisée de Havrda-Charvat. La fonction de perte totale est alors donnée par :

$$L_{\text{total}}((y, \hat{y}), (q_2, p_2), (q_3, p_3)) = L_{\text{rec}}(y, \hat{y}) + L_1(q_2, p_2) + L_\alpha(q_3, p_3), \qquad (4.6)$$

où y, q_2, q_3 sont les vérités-terrain pour respectivement la reconstruction, la classification ORL/poumon et la prédiction de la réponse au traitement et $\hat{y}$, p_2 et p_3 les estimations par le réseau correspondantes. La branche de prédiction est l'objet d'intérêt pour cette partie, les tâches de reconstruction et de classification étant utilisées pour aider la prédiction dans l'étape d'extraction des caractéristiques.

Concernant le temps d'exécution et son changement en raison de l'augmentation ou de la diminution de la complexité du réseau, nous avons proposé de verrouiller notre réseau à un certain niveau de complexité et de trouver un temps d'apprentissage acceptable de 5 à 7 heures pour un nombre d'époques de 100. Cela a été considéré comme un bon compromis entre obtenir un nombre suffisant d'époques pour atteindre des résultats significatifs et ne pas en obtenir tant que cela provoque un surajustement. Nous avons décidé d'utiliser des couches de convolution dans notre structure de base U-Net en raison de la puissance de calcul limitée des machines disponibles et donc de la nécessité de limiter le nombre de paramètres dans le réseau. Permettre quelques heures de calculs a permis de lancer les expériences la nuit afin que les résultats puissent être disponibles et prêts pour l'analyse le lendemain.

Expérimentations

Méthode d'évaluation Étant donné que l'étude est menée sur de petites bases de données, une stratégie de validation des résultats est nécessaire. Nous proposons d'utiliser la validation croisée k-fold. La validation croisée est une procédure de rééchantillonnage utilisée pour évaluer des modèles sur des petites bases de données. Dans notre travail, nous utilisons une validation croisée quintuple. De plus, comme mesure d'évaluation, nous utilisons le taux de bonne prédiction. Nous calculons également la sensibilité et la spécificité de la prédiction.

Résultats de prédiction de la réponse au traitement Les résultats obtenus lors des tests sont affichés dans cette section. Les images reconstruites sont principalement utilisées afin de vérifier la pertinence des caractéristiques extraites pour les autres tâches. Par conséquent, la performance de la reconstruction n'est pas très importante ici, car l'objectif est la prédiction de la réponse au traitement. Concernant l'entropie croisée de Havrda-Charvat, il a été proposé d'étudier les variations de précision lorsque son hyperparamètre α

varie de 0.1 à 1.3. La p-value affichée mesure si les résultats acquis par Havrda-Charvat sont statistiquement différents de ceux obtenus en utilisant celle de Shannon avec une validation croisée quintuple. Deux conditions doivent être satisfaites pour que l'on accepte que l'entropie de Havrda-Charvat donne de meilleurs résultats que l'entropie de Shannon. Un seuil de 0.05 est pris pour la p-value, et la moyenne des résultats doit être cinq fois supérieure à celle obtenue avec Shannon. Les résultats sont décrits dans le tableau 4.6.

α	Précision	Écart-type	p-value
0.1	0.62	0.08	0.36
0.2	0.69	0.03	0.01
0.3	0.62	0.03	0.24
0.4	0.63	0.09	0.41
0.5	0.66	0.01	0.02
0.6	0.65	0.06	0.28
0.7	0.67	0.04	0.02
0.8	0.63	0.04	0.31
0.9	0.60	0.04	0.04
1.0	0.64	0.02	N/A (Shannon's entropy)
1.1	0.68	0.05	0.11
1.2	0.63	0.03	0.46
1.3	0.66	0.06	0.23

TABLE 4.6 – Résultats obtenus par la fonction de perte issue de l'entropie croisée de Havrda-Charvat en fonction de α (p-values inférieures à 0.05 sont soulignées en bleu).

Quand on regarde les résultats obtenus via Havrda-Charvat, on peut confirmer que pour la plupart des valeurs de α, le taux d'erreur de prédiction n'est pas supérieur à celui obtenu par la fonction de perte de Shannon. On peut également observer que la fonction de perte issue de Havrda-Charvat peut donner de meilleurs résultats que celle de Shannon dans plusieurs cas. Cependant, il est difficile de savoir a priori quelle valeur de α est bonne pour une application donnée. Son choix reste un défi, mais ici nous pouvons démontrer que l'entropie de Shannon n'est pas toujours le meilleur choix. Pour démontrer l'avantage de notre réseau multitâches, nous proposons d'évaluer l'importance de chaque tâche du réseau avec Havrda-Charvat en utilisant le α le plus favorable. Les métriques utilisées sont la précision (4.1), la sensitivité définie par :

$$\text{Sensitivité} = \frac{TP}{TP + FN} \tag{4.7}$$

et la spécificité définie par :

$$\text{Spécificité} = \frac{TN}{TN + FP} \tag{4.8}$$

où TP désigne les vrais positifs, TN les vrais négatifs, FP les faux positifs et FN les faux négatifs.

Les résultats sont affichés dans le tableau 4.7.

Tâches	Précision	Sensitivité	Specificité
T2,T3	0.64	0.37	0.79
T1,T3	0.65	0.18	0.90
T1,T2,T3	0.69	0.52	0.77

TABLE 4.7 – Résultats obtenus pour différentes combinaisons de tâches avec Havrda-Charvat en prenant $\alpha = 0.2$.

4.4 Prédiction de la récidive

4.4.1 Introduction

Les expériences menées avec les données CT ont permis d'obtenir des résultats prometteurs. Pour étudier la performance de notre architecture de prédiction de récidive, nous avons utilisé les deux modalités disponibles dans notre base de données, à savoir PET et CT [C3,C4]. Dans l'évaluation des performances de notre réseau de neurones, nous avons opté pour remplacer la métrique de précision par le F1-score. Cette décision repose sur le fait que la précision seule peut être trompeuse, surtout dans les cas de déséquilibre des classes où les prédictions correctes d'une classe majoritaire pourraient dominer le résultat. Le F1-score, en revanche, est une mesure harmonique de la précision et du rappel, offrant un équilibre entre la capacité du modèle à identifier correctement les cas positifs (rappel) et la proportion de ces identifications qui sont effectivement correctes (précision). Ce choix est particulièrement pertinent pour les applications où il est crucial de maintenir un équilibre entre la performance pour les différentes classes, comme dans notre étude. En somme, le F1-score fournit une vision plus complète et équilibrée de la performance du modèle, surtout dans des contextes où chaque classe est d'importance critique. La formule du F1-score est la suivante :

$$F1\text{-score} = 2 \cdot \frac{\text{Précision} \cdot \text{Rappel}}{\text{Précision} + \text{Rappel}} \tag{4.9}$$

où la précision est donnée par la formule (4.1) et le rappel par la formule (4.10).

$$\text{Rappel} = \frac{TP}{TP + FN} \tag{4.10}$$

Le rappel mesure la proportion de cas positifs réels qui ont été correctement identifiés par le modèle et la précision indique la proportion de prédictions positives qui sont réellement correctes, ce qui donne la formule suivante pour le F1-score :

$$F1\text{-score} = \frac{TP}{TP + \frac{1}{2}(FN + FP)} \tag{4.11}$$

4.4.2 Jeu de données

Le jeu de données initial provient du défi MICCAI 2022 (https://hecktor.grand-challenge.org/) et est nommé Hecktor. Le jeu de données sélectionné est composé de 495 patients, tous souffrant d'un cancer ORL. La valeur optimale de α dépend du type de données utilisé. Les images CT et les images PET utilisées dans le réseau de neurones ont été modifiées pour avoir la même résolution d'image de $128 \times 128 \times 64$ voxels. Nos expériences consistaient à établir l'efficacité de Havrda–Charvat sur notre jeu de données, et à tester l'influence du paramètre α. De plus, la précision a été proposée comme métrique d'évaluation. Elle consistait, dans ce cas, à comparer les valeurs de la vérité terrain et de la prédiction et à additionner toutes ces occurrences sur la taille du jeu de données.

4.4.3 Architecture de réseau de neurones pour la prédiction de récidive

L'architecture proposée utilisée pour la prédiction de récidive est un réseau de neurones multitâche basé sur U-Net, avec une branche capable d'effectuer des tâches de prédiction et une autre pour reconstruire l'image d'entrée pour extraire des caractéristiques aidant à la prédiction [C4]. L'architecture est présentée dans la Figure 4.7.
La structure de base U-Net est composée de trois couches convolutionnelles sautées par des concaténations pour ajouter les caractéristiques extraites des convolutions descendantes aux ascendantes. Au sein des couches de convolution, nous utilisons des fonctions d'activation ReLU. Au niveau de la couche convolutive la plus basse, les caractéristiques extraites sont envoyées en entrée à une branche de couches entièrement connectées en charge de prendre une décision pour déterminer si le patient est à risque de récidive.

4.4.4 Résultats

Les résultats obtenus lors des tests sont présentés dans cette section. Concernant Havrda–Charvat, nous avons étudié son hyperparamètre α variant de 0,1 à 4,0. Lorsque $\alpha = 1$, cette entropie correspond à l'entropie de Shannon. Les résultats sont décrits dans le tableau suivant. Les résultats obtenus pour le jeu de données sur les cancers ORL sont décrits dans le tableau 4.8. La recherche en grille est une méthode courante pour effectuer l'optimisation des hyper-paramètres. En utilisant cette méthode, nous avons obtenu un ensemble d'hyperparamètres optimaux. Les résultats obtenus via la formule de Havrda–Charvat confirment que, pour des valeurs de l'hyperparamètre α inférieures à 1, la précision finale n'est pas supérieure à la précision obtenue par la fonction de perte de Shannon. Il est également observable que la fonction de perte dérivée de l'équation de Havrda–Charvat peut fournir de meilleurs résultats que celle de Shannon pour des valeurs de α supérieures à 1. Jusqu'à 1, toutes les valeurs du F1-score sont égales à 0, puis montent jusqu'à 0,96 par la

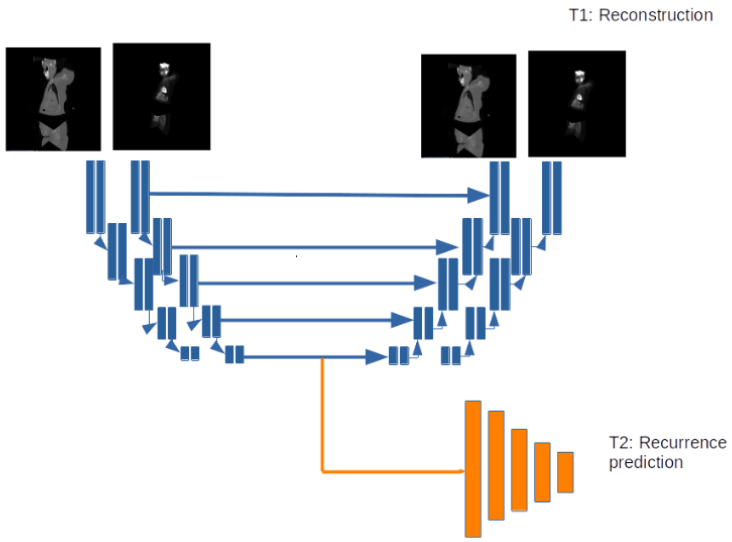


FIGURE 4.7 – Architecture du réseau de neurones pour la prédiction de récidive.

suite. Les valeurs les plus prometteuses ont été obtenues pour α supérieur à 1. À cet égard, nous pouvons dire que les résultats obtenus par Havrda–Charvat peuvent être prometteurs. La métrique utilisée pour les expériences était le F1-score, en raison du déséquilibre dans le jeu de données. Dans ce jeu de données, 80% des personnes n'ont pas connu de récidive. Ainsi, si le réseau prédisait que personne n'avait connu de récidive, il n'aurait rien appris mais aurait tout de même été précis à 80%, ce qui ne serait pas représentatif de la quantité d'apprentissage réalisée par le réseau. Nous avons également expérimenté en concaténant les deux modalités avant de les fournir à un U-Net classique avec une branche de prédiction. Les caractéristiques extraites sont donc un mélange de PET/CT. Nos expériences ont montré que, quelle que soit la valeur définie pour α, le réseau n'a jamais appris à distinguer les personnes connaissant une récidive de celles qui n'en connaissent pas.

4.5 Discussion sur l'estimation du paramètre de l'entropie croisée d'Havrda-Charvat

Il a été déterminé que la fonction de perte de Havrda–Charvat peut se montrer plus performante que la fonction de perte de Shannon. Cela ouvre la porte à des applications plus larges de Havrda–Charvat dans les travaux futurs. On peut dire que la fonction de perte de Havrda–Charvat, selon la valeur de son hyperparamètre α, peut s'adapter à un plus large éventail de données d'entrée et potentiellement produire de meilleurs résultats. En revanche, il faut noter que Havrda–Charvat peut donner de mauvaises performances, surtout si l'hyperparamètre α est inférieur à 1 ou dépasse 2. Dans le cas où l'hyperparamètre

α	F1-score	Précision
0.8	0	0.80
1.0	0	0.80
1.4	0.92	0.97
1.7	0.92	0.97
2.0	0.93	0.97
3.0	0.95	0.98
3.5	0.95	0.98
3.9	0.95	0.98

TABLE 4.8 – F1-score et précision obtenus par la fonction de perte dérivée de l'entropie de Havrda–Charvat en fonction de α pour le jeu de données sur le cancer ORL. $\alpha = 1$ correspondant à l'entropie croisée basée sur Shannon.

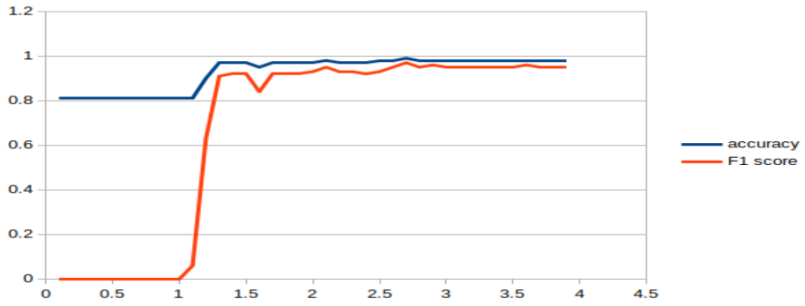


FIGURE 4.8 – Représentation graphique des résultats en termes de précision (en bleu) et de F1-score (en rouge). La valeur de α est représentée en abscisse et la performance en ordonnée.

est inférieur à 1, ceci s'explique par des problèmes numériques (apparition de valeurs trop grandes et donc de "Not A Number") et dans le cas où l'hyper-paramètre dépasse 2, cela s'explique par la difficulté de discriminer les lois de probabilité. Ainsi, nous avons proposé un algorithme d'estimation du paramètre α. La Figure 4.9 donne la description schématique de l'algorithme d'estimation de α. Pour effectuer l'estimation de α, nous avons modifié le calcul du gradient en utilisant la dérivée de (4.2) par rapport à α. Pour cela, nous calculons à chaque époque la quantité :

$$L_{d,\alpha}(q,p) = \frac{1}{N}\sum_{k=1}^{N}\frac{\partial H_\alpha}{\partial \alpha}(q_n||p_n), \tag{4.12}$$

où :

$$\begin{aligned}\frac{\partial H_\alpha}{\partial \alpha}(q||p) &= \frac{1}{(\alpha-1)^2}\times\left(\sum_{x\in E} p(x)^{\alpha-1}q(x)\right)\\ &+ \frac{1}{1-\alpha}\sum_{x\in E}\log(p(x))p(x)^{\alpha-1}q(x)\end{aligned} \tag{4.13}$$

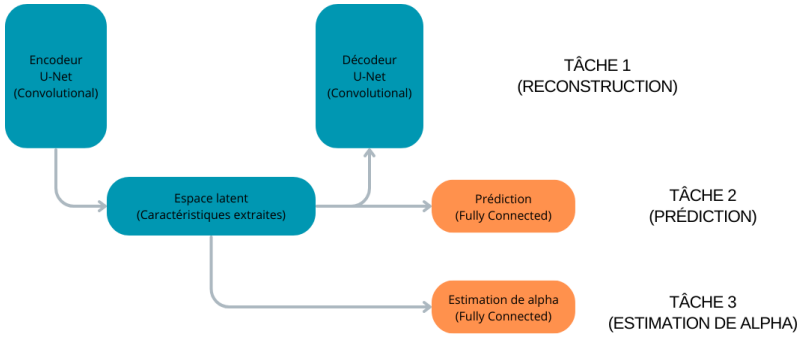


FIGURE 4.9 – Architecture automatique d'estimation d'α

Puis nous remettons à jour la valeur de α en ajoutant la quantité (4.12) à la valeur précédente. L'algorithme se résume ainsi :

1. **Initialisation :** On choisit une valeur initiale α_0 pour α ($\alpha_0 = 1.5$).
2. **Itération :** A l'époque t, soit α_t la valeur courante de α :
   (a) On effectue la reconstruction/prédiction comme dans le cas du réseau multitâche exposé auparavant.
   (b) On calcule $L_{d,\alpha_t}(q,p)$ avec la valeur courante p obtenue en sortie du réseau.
   (c) On pose $\alpha_{t+1} = \alpha_t - \lambda \times L_{d,\alpha_t}(q,p)$, où λ est un paramètre similaire au taux d'apprentissage. Nous avons pris $\lambda = 0.1$.

Malgré nos espoirs, cet algorithme s'est avéré peu concluant. En effet, de nombreuses fois, α devenait de plus en plus grand et des problèmes numériques se produisaient. Ceci peut s'expliquer par le fait que α est un paramètre d'un estimateur et n'est pas relatif aux données et que notre méthode est peut-être un peu trop naïve pour estimer le paramètre d'un estimateur. La sélection d'estimateur est encore un problème délicat de nos jours. Une idée serait de trouver une fonction de perte entre estimateurs ou au moins un moyen de prévoir le comportement de l'estimateur par avance ; ce qui est délicat car cela dépend grandement de la nature des données.

4.6 Conclusion

Dans cette partie, nous avons établi que, pour nos données et dans plusieurs cas, l'entropie croisée de Havrda-Charvat est plus performante que la fonction de perte de Shannon. La meilleure performance moyenne de Havrda-Charvat est supérieure à celle de Shannon. Dans les applications médicales, même une amélioration de 1 ou 2% est intéressante. Cela fait de la formule de Havrda-Charvat le meilleur candidat pour des recherches plus approfondies sur l'ensemble de données. Dans ces recherches, nous pourrions adapter l'entropie croisée binaire de Havrda-Charvat pour qu'elle devienne catégorique, permettant ainsi de réaliser des prédictions multi-classes, notamment en estimant la

durée entre la fin du traitement du cancer et la récidive. Un autre axe de développement pourrait être de trouver un moyen d'automatiser la détermination de la valeur appropriée de α pour une application donnée. Les études en deep learning sur l'utilisation de l'entropie d'Havrda-Charvat comme fonction de perte par rapport à l'entropie de Shannon ont démontré des résultats prometteurs. Bien que l'entropie de Shannon soit largement utilisée dans les réseaux de neurones, l'entropie d'Havrda-Charvat offre une alternative intéressante, en particulier dans les cas où les données ne répondent pas aux conditions spécifiques requises par l'entropie de Shannon.

Les recherches ont montré que l'entropie d'Havrda-Charvat peut être plus performante dans des scénarios où les données sont bruitées, complexes ou ne répondent pas aux distributions de probabilité classiques. Elle offre une généralisation de l'entropie de Shannon, ce qui lui permet de mieux capturer la structure des données dans certains contextes. Les expériences menées ont montré que l'utilisation de l'entropie d'Havrda-Charvat comme fonction de perte peut conduire à des améliorations significatives en termes de précision et de performances par rapport à l'entropie de Shannon. En outre, l'ajustement du paramètre α dans l'entropie d'Havrda-Charvat offre une flexibilité supplémentaire pour adapter la fonction de perte aux caractéristiques spécifiques des données. Cependant, des recherches supplémentaires sont nécessaires pour explorer pleinement les avantages et les limitations de l'utilisation de l'entropie d'Havrda-Charvat dans les applications de deep learning. Il est également important de développer des méthodes pour estimer automatiquement les valeurs optimales du paramètre α en fonction des caractéristiques des données. Les études en deep learning sur l'entropie d'Havrda-Charvat offrent une perspective prometteuse pour améliorer la modélisation des données complexes et bruitées, ouvrant la voie à de nouvelles avancées dans le domaine de l'apprentissage automatique.

CHAPITRE 5

Swin Transformers avec bloc d'attention pour la classification d'images de patients néonataux

5.1 Introduction

Toujours dans le contexte radiomique, nous nous concentrons cette fois-ci sur les architectures de réseaux de neurones. Plus précisement, nous étudions une nouvelle architecture pour un nouveau problème de classification d'images. L'encéphalopathie hypoxique-ischémique (HIE) [3] chez les nouveau-nés représente une cause majeure de morbidité et de mortalité à l'échelle mondiale. Cette condition, résultant d'une privation d'oxygène et d'un flux sanguin insuffisant au cerveau au moment de la naissance, peut conduire à des conséquences neuro-développementales à long terme et nécessite une évaluation rapide pour un traitement approprié. Dans ce contexte, la classification précise de la gravité des atteintes est cruciale pour orienter les interventions médicales et améliorer les pronostics. Cependant, l'interprétation des images médicales pour diagnostiquer et classer les sévérités de HIE reste un défi, notamment en raison de la subtilité des marqueurs de lésions cérébrales sur les images de diagnostic par résonance magnétique (IRM). Les récents progrès en intelligence artificielle et, plus particulièrement, dans les domaines de l'apprentissage profond, offrent des perspectives prometteuses pour améliorer la précision diagnostique dans les environnements médicaux. Les Swin Transformers, une nouvelle classe de modèles de vision par ordinateur basés sur des mécanismes d'attention, se distinguent particulièrement par leur efficacité à capter des contextes hiérarchiques complexes dans les images, ce qui les rend particulièrement adaptés à l'analyse d'images médicales. Ce chapitre se concentre sur l'exploration et l'adaptation des architectures de Swin Transformers pour développer un système automatisé capable de classifier les images IRM de patients néonataux avec HIE selon la gravité des atteintes. Nous débutons par une revue des fondements théo-

riques des Swin Transformers [62], mettant en lumière le fonctionnement et les avantages des blocs d'attention qui sont au cœur de ces modèles. Ensuite, nous présentons une série d'expérimentations visant à évaluer et optimiser ces architectures pour notre cas d'usage spécifique. L'objectif est de déterminer dans quelle mesure ces modèles peuvent surpasser les approches conventionnelles de traitement d'images et contribuer à une prise de décision clinique plus éclairée et plus rapide. À travers l'adaptation de ces technologies avancées, ce travail aspire non seulement à faire progresser le champ de la neuro-radiologie pédiatrique mais également à poser les bases pour de futures recherches appliquées dans l'amélioration des diagnostics médicaux grâce à l'intelligence artificielle. Ajoutée à la malléabilité du cerveau des nouveau-nés, qui de fait crée une grande disparité de forme et d'organisation d'un patient à un autre, la classification des différents grades de sévérité de la HIE pose un autre problème. Pour un oeil non-expert, il est difficile, voire impossible de reconnaître et donc de classifier les images qui témoignent de grades 1, 2 ou 3. La figure 5.1 présente des exemples d'images IRM prises chez des patients souffrant de HIE à des grades différents.

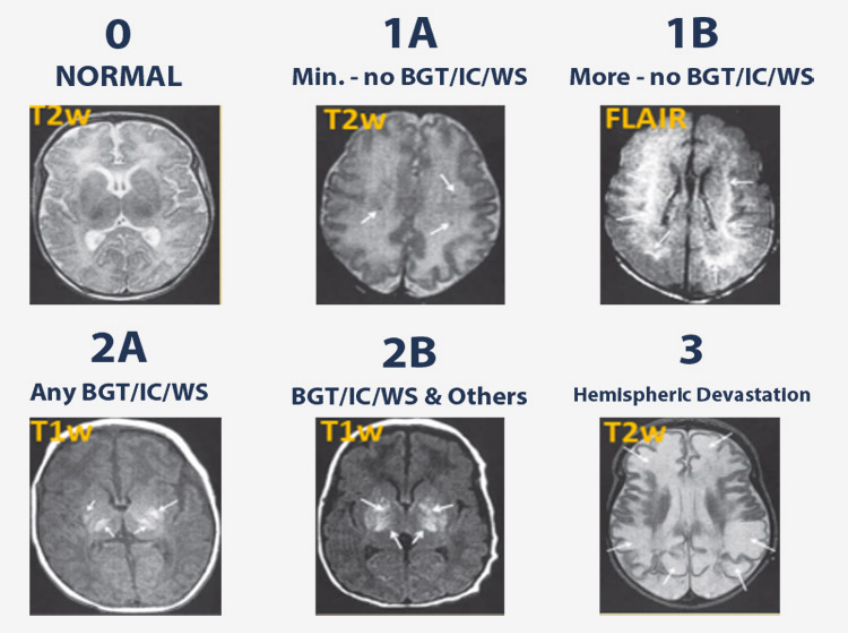


FIGURE 5.1 – Images IRM de patients avec différents grades de HIE. (https ://www.nationalbirthinjurylaw.com)

5.2 Swin Transformers pour la classification d'images médicales

5.2.1 Introduction

Les Transformers [99], initialement conçus pour des tâches de traitement du langage naturel, ont été adaptés avec succès au domaine de la vision par ordinateur [24], démontrant une capacité remarquable à modéliser des interactions complexes entre les caractéristiques visuelles à différentes échelles. Le modèle Swin Transformer, une variante des architectures Transformer traditionnelles, introduit une approche novatrice basée sur des fenêtres décalées, permettant une efficacité computationnelle accrue et une meilleure intégration des contextes locaux et globaux. Cette spécificité rend les Swin Transformers particulièrement aptes à gérer les images médicales, qui requièrent souvent une analyse détaillée des structures fines et une compréhension globale des contextes anatomiques pour un diagnostic précis. Dans le domaine médical, la classification précise des images est cruciale, car elle influe directement sur les décisions thérapeutiques et les pronostics des patients. Les images médicales, comme les IRM, les scans CT et les radiographies, contiennent des informations subtiles qui sont souvent difficiles à interpréter, même pour des spécialistes expérimentés. L'utilisation de Swin Transformers dans ce contexte promet de surmonter certaines limites des méthodes traditionnelles de vision par ordinateur, telles que la convolution, qui peut être moins efficace pour capter les interactions non-contigües entre les caractéristiques de l'image. Ce chapitre présente d'abord une revue des principes de base des Transformers et de leur adaptation aux Swin Transformers. Nous explorons ensuite leur application dans l'analyse d'images médicales, en soulignant les défis spécifiques de ce domaine, tels que la haute dimensionnalité des données, la nécessité d'une interprétation précise et la gestion des variations inter-patient. L'efficacité des Swin Transformers est illustrée à travers des études de cas sur la classification d'images de pathologies spécifiques, mettant en lumière leur potentiel à améliorer significativement les capacités de diagnostic. En intégrant les Swin Transformers dans le flux de travail diagnostique, nous ouvrons la voie à des systèmes de santé plus intelligents et plus réactifs, capables de fournir des évaluations précises et rapides, essentielles pour le traitement efficace des patients. L'Imagerie par Résonance Magnétique (IRM) joue un rôle crucial dans le diagnostic clinique et la classification de la HIE, car la détection précoce et le traitement approprié par hypothermie modérée augmenteraient les chances de meilleur développement cérébral du nouveau-né. Dans ce chapitre, pour aborder cette problématique, nous proposons un nouveau réseau basé sur le Transformer Swin pour la classification automatique de la gravité de la HIE en utilisant des données IRM. Plus précisément, nous intégrons les modules d'attention spatiale et de canal dans le réseau pour extraire des caractéristiques informatives et complémentaires des images IRM multimodales d'entrée. Des expériences sur un ensemble de données incluant des images IRM T1w et T2w

de 109 nouveau-nés souffrant de trois stades de HIE ont montré la performance prometteuse du réseau proposé, comparée aux méthodes de pointe.

5.2.2 Méthodes

Les Transformers, tels que le Swin Transformer, ont démontré leur efficacité dans divers domaines de la vision par ordinateur, y compris la classification, la segmentation et la détection d'objets, et ont été adaptés avec succès à l'analyse d'images médicales. Initialement développés pour le traitement du langage naturel, les Transformers, notamment le Transformer Visuel (ViT) introduit en 2021, ont rivalisé et même surpassé les réseaux de neurones convolutionnels (CNN) en termes d'efficacité et de précision, offrant une meilleure gestion des ressources de calcul. Ils traitent les images d'entrée en les décomposant en fragments (tokens), permettant une intégration plus efficace des informations. Plus récemment, l'attention s'est portée sur les architectures multimodales, notamment le Swin Transformer multimodal, qui gère efficacement différentes modalités d'entrée grâce à l'auto-attention, facilitant l'intégration des données et la préservation des relations spatiales entre les modalités. Cette approche améliore la performance dans des applications telles que l'imagerie médicale, où nous avons intégré des modules d'attention spatiale et de canal pour mieux exploiter les séquences IRM T1w et T2w, créant une représentation hiérarchique des caractéristiques plus robuste et complète.

Méthode et architecture

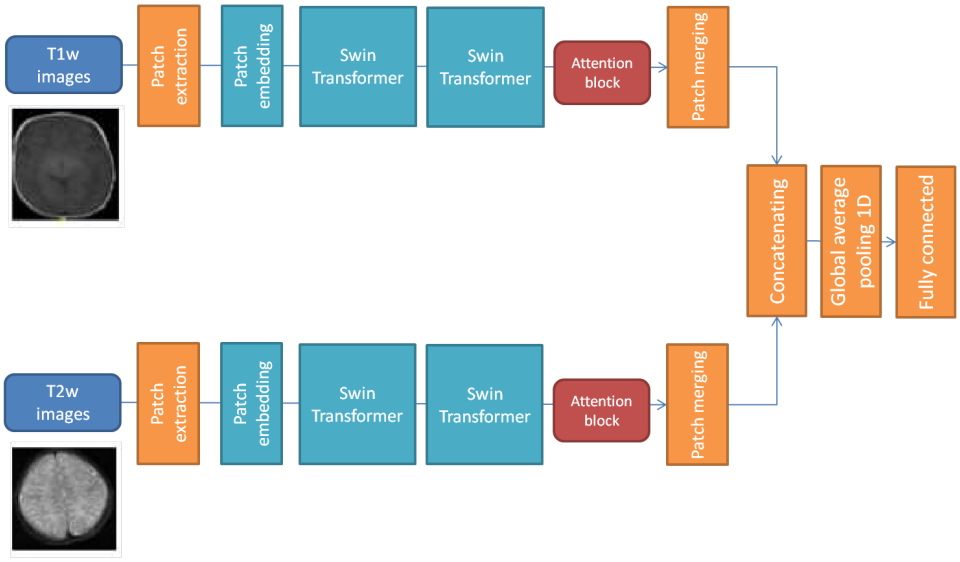


FIGURE 5.2 – Architecture du réseau de neurones ASwin pour la classification multiclasse HIE.

Notre architecture [C5], nommée ASwin (Attention Swin), est représentée dans la figure 5.2 et comprend deux branches pour les coupes 2D extraites

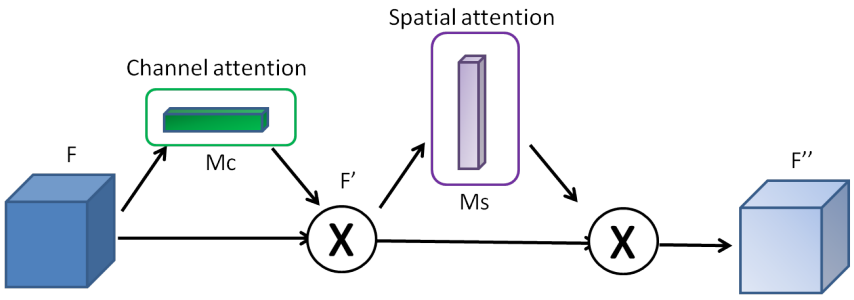


FIGURE 5.3 – Architecture du module d'attention.

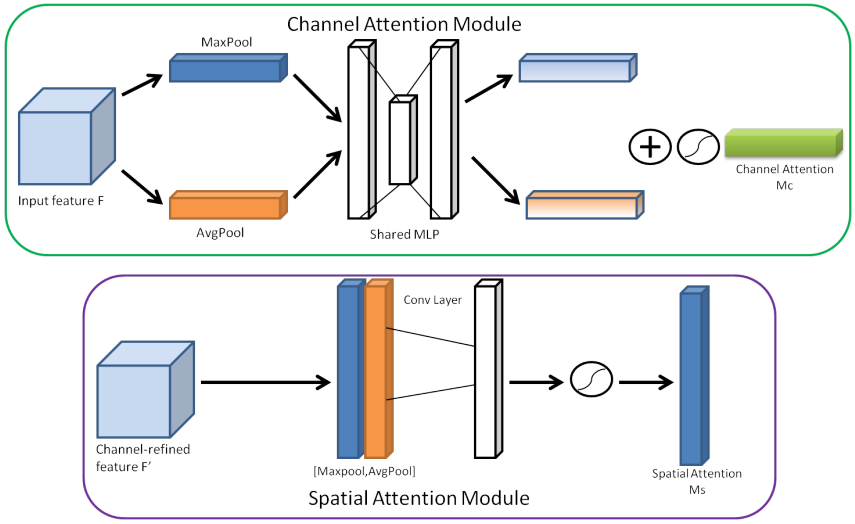


FIGURE 5.4 – Architecture des sous-modules.

respectivement des images T1w et T2w. Similaire au Transformer Swin, notre structure limite le calcul de l'attention aux fenêtres locales, tout en permettant une connexion entre les fenêtres. En détail, chaque branche du réseau comprend un bloc d'extraction de patchs, un bloc d'encapsulation de patchs, deux blocs de Transformer Swin, un bloc d'attention et un bloc de fusion de patchs. Ensuite, les cartes de caractéristiques générées par chaque branche sont concaténées et transformées en un ensemble de caractéristiques 1D par un bloc de pooling moyen global. Finalement, l'ensemble 1D de caractéristiques est transmis à une couche dense pour prédire le grade final. La taille des patchs dans les modules d'extraction et de fusion de patchs est de 2 × 2 pixels, et la dimension d'encapsulation de chaque bloc de Transformer Swin est de 32. Comme illustré dans les Figures 5.3 et Figure 5.4, nous avons appliqué le module de bloc convolutif d'attention (CBAM) [107] pour aider le réseau entraîné à se concentrer davantage sur les cartes de caractéristiques et les sous-régions les plus pertinentes des images IRM cérébrales infantiles avec une résolution insuffisante, un bruit important et un faible contraste des tissus. La méthode de fusion des patchs produit un nombre défini de patchs lorsqu'elle est appliquée aux "tokens" d'entrée. Des patchs similaires sont acheminés de manière similaire, ce qui signifie qu'ils finissent par contribuer au même sous-ensemble de patchs de sortie. Cela peut aider à réduire le temps perdu sur des arrière-plans redondants. Concernant le processus d'attention, il peut être résumé ainsi :

$$F' = M_c(F) \odot F, \tag{5.1}$$

$$F'' = M_s(F') \odot F', \tag{5.2}$$

où $\odot$ représente la multiplication élément par élément, F est l'entrée des éléments, F' est le résultat du sous-module d'attention de canal multiplié par l'entrée, et F'' est la sortie finale du module d'attention. M_c est le processus du sous-module de canal et M_s est le processus du sous-module spatial. Les deux opérateurs M_c et M_s sont détaillés respectivement dans les équations 5.4 et 5.6.

$$Mc(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))), \tag{5.3}$$

$$= \sigma(W_1(W_0(F_{\text{avg}})) + W_1(W_0(F_{\max}))), \tag{5.4}$$

où σ symbolise la fonction sigmoïde, $W_0 \in \mathbb{R}^{C/r \times C}$, et $W_1 \in \mathbb{R}^{C \times C/r}$ où r est un facteur par lequel la dimensionnalité des canaux ou des caractéristiques est réduite. Notez que les poids MLP (Perceptron Multicouches), W_0 et W_1, sont partagés pour les deux entrées et la fonction d'activation ReLU est suivie par W_0.

$$Ms(F) = \sigma(f_{7\times7}([AvgPool(F); MaxPool(F)])), \tag{5.5}$$

$$= \sigma(f_{7\times7}([F_{\text{avg}}; F_{\max}])), \tag{5.6}$$

où σ symbolise la fonction sigmoïde et $f_{7\times7}$ représente une opération de convolution avec une taille de filtre de 7×7.

Étant donné une image d'entrée, les deux modules d'attention, attention par canal et attention spatiale, calculent l'attention de façon complémentaire, se concentrant respectivement sur le *quoi* et le *où*. À cet égard, ces deux modules peuvent être disposés de manière parallèle ou séquentielle. Nous avons opté pour la disposition séquentielle commençant par le canal plus susceptible de générer de meilleurs résultats.

Les modules d'attention de canal exploitent les relations inter-canaux, qui se concentrent sur les parties significatives étant donné l'entrée. Tandis que les cartes d'attention spatiales se concentrent sur les coordonnées des parties informatives. La fonction de perte utilisée dans ce travail est l'entropie croisée catégorique. La formule de l'entropie croisée catégorique est définie ci-après dans l'équation 5.7.

$$H(y, \hat{y}) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{5.7}$$

Dans cette formule :
— $H(y, \hat{y})$ représente l'entropie croisée entre la distribution réelle y et la distribution prédite $\hat{y}$.
— C est le nombre de classes.
— y_i est un indicateur binaire (0 ou 1) si la classe i est la vraie classe pour l'observation donnée.
— $\hat{y}_i$ est la probabilité prédite que l'observation appartienne à la classe i.
Cela donne la fonction de perte suivante, à savoir l'équation 5.8

$$L = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{C} y_{n,i} \log(\hat{y}_{n,i}), \tag{5.8}$$

où C est le nombre de classes et N le nombre de patients.

5.2.3 Expérimentations

Matériel et Logiciel

Toutes les expériences ont été réalisées sur un ordinateur équipé d'un processeur Intel Core i9-12900F de 12e génération, de 32 Go de RAM et d'une carte graphique NVIDIA RTX 3080 Ti. Le système d'exploitation utilisé est Windows. Le langage de programmation est Python 3.9.13 en utilisant la librairie TensorFlow 2.

Jeu de Données et Méthodes Concurrentes

Dans nos expériences, nous avons utilisé 109 patients néonataux possédant à la fois des IRM cérébrales cliniques T1w et T2w. Ces patients comprennent 19 sujets légers, 58 sujets modérés et 32 sujets sévères. Après correction des inhomogénéités d'intensité des images, dix coupes centrées sur le cerveau ont

été extraites de chaque modalité pour tenir compte de l'épaisseur importante des coupes. Nous avons comparé notre architecture en utilisant une validation croisée 10 fois avec différentes méthodes de pointe, y compris le Transformer Visuel ViT[62] et quelques réseaux neuronaux convolutionnels largement adoptés, y compris MobileNet[86], ShuffleNet[115], DenseNet-121[27] et ResNet-34[33]. L'entraînement a duré 50 époques, après quoi les données de test ont été fournies aux modèles obtenus pour les tests. Les métriques de classification, y compris la précision, la sensibilité, la spécificité et le F1-score, ont été utilisées pour l'évaluation des modèles. Le taux d'apprentissage sélectionné était de 0,001 et l'optimiseur était Adam [48].

Réseau	**Métrique**	**Précision**	**Spécificité**	**Sensibilité**
Aswin (Proposé)	Moyenne	**0.956**	**0.982**	**0.918**
	Variance	0.0029	0.0004	0.0163
Swin Transformer	Moyenne	0.924	0.914	0.957
	Variance	0.0039	0.0021	0.0040
ShuffleNet	Moyenne	0.912	0.868	0.968
	Variance	0.0116	0.0283	0.0013
ViT	Moyenne	0.882	0.927	0.892
	Variance	0.0101	0.0034	0.0065
DenseNet-121	Moyenne	0.872	0.911	0.810
	Variance	0.0222	0.0088	0.0363
MobileNet	Moyenne	0.858	0.858	0.894
	Variance	0.0143	0.0104	0.0052
ResNet-34	Moyenne	0.762	0.885	0.793
	Variance	0.0090	0.0013	0.0086

TABLE 5.1 – Résultats des différentes méthodes.

Résultats

Les principaux résultats obtenus au cours de l'expérience sont résumés dans le Tableau 5.1. Comme nous pouvons le constater, le réseau multimodal proposé surpasse toutes les autres architectures concurrentes sur tous les critères. De plus, pour évaluer davantage la pertinence des images multimodales, nous comparons les résultats de notre réseau avec seulement des images T1w et seulement des images T2w dans le Tableau 5.2, ce qui valide le choix d'utiliser les images T1w et T2w conjointement. Nous avons également testé les performances de notre ASwin avec différents nombres de coupes dans le Tableau 5.3, ce qui a démontré qu'il est plus approprié d'extraire 10 coupes pour la classification de la HIE, car cela pourrait fournir suffisamment d'informations sans induire de contenu non pertinent pouvant influencer négativement la performance.

Modalité	Métrique	Précision	Spécificité	Sensibilité
T1w	Moyenne	0.881	0.927	0.909
	Variance	0.0101	0.0076	0.0105
T2w	Moyenne	0.828	0.852	0.827
	Variance	0.0339	0.0328	0.0461
T1w+T2w	Moyenne	**0.956**	**0.982**	**0.918**
	Variance	0.0029	0.0004	0.0163

TABLE 5.2 – Comparaison de ASwin avec T1w seules vs T2w seules vs Les Deux

Nombre de coupes	Métrique	Précision	Spécificité	Sensibilité
5	Moyenne	0.94	0.971	0.902
	Variance	0.0064	0.0013	0.0214
10	Moyenne	**0.956**	**0.982**	**0.918**
	Variance	0.0029	0.0004	0.0163
15	Moyenne	0.93	0.957	0.939
	Variance	0.0052	0.0018	0.0038

TABLE 5.3 – Comparaison de ASwin en utilisant différents nombres de coupes

5.2.4 Discussion

Nous avons montré que notre réseau atteint une grande précision sur le jeu de données que nous avons étudié. Les deux modalités utilisées semblent contenir suffisamment d'informations pour que le réseau fonctionne correctement et atteigne des résultats encourageants. Cependant, il existe certaines limitations à notre étude. Premièrement, nous n'avons pas utilisé l'augmentation de données, qui pourrait être une bonne solution au problème de notre jeu de données relativement petit. Nous devrons déterminer si l'utilisation de données augmentées est bénéfique ou préjudiciable à notre architecture. Ensuite, nous avons uniquement utilisé des images T1w et T2w, car ce sont les modalités les plus courantes et les plus largement utilisées. Cependant, d'autres types d'images IRM sont disponibles, par exemple DWI (imagerie pondérée en diffusion), qui pourraient fournir des informations complémentaires pour améliorer les résultats, toutefois, nous devrons résoudre le problème de l'absence de modalité chez de nombreux sujets. Cette limitation pourrait affecter la généralisabilité et l'efficacité de notre approche sur une population plus large et plus diverse. À l'avenir, l'intégration de modalités supplémentaires, telles que les images DWI, pourrait non seulement améliorer la robustesse de nos prédictions mais aussi nous rapprocher d'un score quasi-parfait.

5.2.5 Conclusion

Dans cette partie, nous avons présenté un réseau de Transformer Swin avec attention multimodale (ASwin) capable de tirer parti conjointement des images T1w et T2w pour atteindre une forte capacité de prédiction (jusqu'à

99% de précision) sur un jeu de données HIE, ce qui en fait un solide candidat pour de futurs tests et expérimentations sur divers ensembles de données dans le domaine médical. C'est également le premier travail utilisant des données IRM pour la classification de la gravité de la HIE basée sur l'apprentissage profond. Un travail futur intéressant serait d'intégrer des données cliniques et/ou d'autres modalités d'image dans notre architecture, car cela pourrait apporter plus d'informations utiles au réseau et ainsi aider à prendre de meilleures décisions.

Conlusion générale

Cette thèse a exploré les fonctions de perte basées sur l'entropie de Havrda-Charvat dans le cadre de l'apprentissage profond, avec un accent particulier sur la prédiction de l'évolution des maladies. Les résultats obtenus démontrent que ces fonctions de perte offrent des avantages significatifs en termes de robustesse et de sensibilité par rapport aux fonctions de perte conventionnelles, en particulier dans les cas de données déséquilibrées et de problématiques complexes de classification de stades de maladies. Au cours de cette recherche, plusieurs architectures de réseaux de neurones ont été expérimentées telles que l'architecture multitâches, et il a été constaté que l'intégration de l'entropie de Havrda-Charvat améliore la capacité des modèles à généraliser à partir de données limitées ou bruitées. Cela a été particulièrement pertinent pour les prédictions de réponse au traitement, où la précision peut significativement influencer les décisions thérapeutiques. Les implications de cette étude sont vastes, ouvrant de nouvelles voies pour l'utilisation des mesures d'entropie dans l'amélioration des algorithmes d'apprentissage profond. En outre, l'adaptabilité de l'entropie de Havrda-Charvat à différents types de données et scénarios de prédiction en fait un outil précieux pour la recherche médicale future. Les travaux futurs pourraient explorer l'intégration de ces fonctions de perte dans d'autres domaines de l'intelligence artificielle, tels que le traitement du langage naturel et la vision par ordinateur, ainsi que leur application à d'autres maladies chroniques ou conditions de santé complexes. En conclusion, cette thèse pose une pierre angulaire pour des recherches approfondies sur les fonctions de perte basées sur l'entropie et leur potentiel pour révolutionner les domaines de l'apprentissage automatique et de la santé numérique. De plus, cette thèse a exploré les profondeurs et les potentialités de l'apprentissage profond dans divers domaines d'application, mettant en lumière son impact transformateur. Un exemple particulièrement éloquent de cette transformation se trouve dans le domaine de la radiomique. Grâce à l'application des techniques de l'apprentissage profond, l'analyse radiomique a franchi des seuils significatifs en permettant une extraction automatique des caractéristiques radiomiques, qui

étaient auparavant extraites de façon manuelle (par exemple, calculer des indices de texture). Cette avancée ouvre la voie à des diagnostics plus précis et à des traitements personnalisés, révolutionnant ainsi le domaine de la médecine diagnostique. Ainsi, l'apprentissage profond, en synergie avec la radiomique, promet de pousser les frontières de la médecine prédictive et personnalisée, illustrant parfaitement l'impact et la portée de cette recherche.

Grâce aux architectures et fonctions de perte proposées, des résultats bien supérieurs aux réseaux considérés comme baseline ont été obtenus. Au regard de la prédiction de la réponse au traitement, nous sommes passés de 73% pour l'entropie de Shannon à 79% de précision pour l'entropie de Havrda-Charvat, ce qui représente des possibilités d'amélioration conséquentes dans le domaine de la santé et de l'analyse d'images médicales, où chaque pourcent gagné signifie davantage de maladies correctement diagnostiquées et de patients correctement traités. De plus, le tri des images préalable à la prédiction est très fiable, avec une précision de classification proche de 99%. Lors des tests qui concernent le F1-score, nous observons des valeurs de l'hyperparamètre α pour lesquelles les résultats sont bien supérieurs à ceux de l'entropie de Shannon.

Au regard de la classification des sévérités des HIE, l'architecture proposée basée sur les Transformers Swin permet d'augmenter la précision de classification jusqu'à 99%, ce qui nous permet d'espérer des améliorations considérables dans le diagnostic de cette maladie et le traitement approprié.

Nos contributions ont été résumées comme suit :

— De nouvelles architectures adaptées aux données médicales. Les réseaux de neurones proposés sont de type multitâches pour la prédiction de la réponse au traitement et basés sur les Swin Transformers avec des blocs d'attention convolutive pour la prédiction de la sévérité des HIE.

— Une nouvelle fonction de perte dérivée de l'entropie de Havrda-Charvat. Cette généralisation de l'entropie de Shannon permet une plus grande adaptabilité de la fonction de perte aux données médicales d'entrée et donc de maximiser l'efficacité du réseau de neurones employé.

— Dans les différentes expérimentations proposées, nous avons constaté des résultats prometteurs dans toutes les prédictions effectuées, binaires ou multiclasses. Cela constitue un axe de recherche pertinent pour des projets de recherche supplémentaires.

Perspectives

Cette thèse a exploré les potentiels des fonctions de perte basées sur des entropies généralisées dans les réseaux de neurones, ouvrant plusieurs voies prometteuses pour des recherches futures. Premièrement, l'approfondissement de l'étude sur l'impact des différentes formes d'entropie, telles que celles de Rényi ou de Tsallis, pourrait conduire à une meilleure compréhension des dynamiques de convergence des modèles en apprentissage profond. La diversité des comportements de ces entropies sous différentes conditions et architectures suggère une potentialité d'adaptation et d'optimisation ciblée selon les spécificités des données traitées.

Deuxièmement, l'intégration de ces entropies dans des architectures de réseaux neuronaux plus complexes, comme les réseaux de neurones récurrents, pourrait améliorer la gestion des séquences temporelles et la génération de données, ouvrant ainsi de nouvelles applications en traitement du signal et en synthèse d'image.

Troisièmement, la conceptualisation de nouvelles méthodes d'optimisation qui tirent parti des propriétés spécifiques des entropies généralisées pourrait mener à des algorithmes d'apprentissage plus rapides et plus robustes. Enfin, l'exploration de l'application de ces concepts à des problèmes de grande dimension, tels que la reconnaissance de motifs dans des ensembles de données massifs ou l'analyse de réseaux complexes, pourrait transformer substantiellement les approches actuelles.

Ces perspectives, en exploitant pleinement le potentiel des fonctions de perte basées sur les entropies généralisées, pourraient non seulement affiner la précision des prédictions des modèles de réseaux de neurones, mais également élargir considérablement le champ d'application de l'apprentissage automatique, comme d'affiner la prédiction de survie des patients atteints de cancers.

Il devient envisageable, au regard des expériences menées au cours de cette thèse, que les réseaux de neurones, et plus précisément ceux dévolus à l'analyse d'images médicales, continuent de révéler leur potentiel et les avancées qu'ils permettraient.

De plus, concernant la fonction de perte, nous pouvons étayer notre hypothèse initiale concernant la plus grande efficacité et adaptabilité de l'entropie d'Havrda-Charvat. Il serait profitable de poursuivre les analyses sur cette fonction de perte afin de prendre la mesure des progrès qu'elle permet et produire un algorithme apte à gérer automatiquement l'adaptation de son hyperparamètre α, par exemple via un réseau de neurones attitré ou un pipeline qui réinitialise les réseaux de neurones en fonction des nouvelles valeurs d'α.

Liste des publications durant ma thèse

Articles de revues

[R1] **Thibaud Brochet**, Jérôme Lapuyade-Lahorgue, Stéphane Bougleux, Matthieu Salaün, Su Ruan, "Deep Learning Using Havrda-Charvat Entropy for Classification of Pulmonary Optical Endomicroscopy," *IRBM*, vol 42, 6, 2021. *10.1016/j.irbm.2021.06.006*.
[R2] **Thibaud Brochet**, Jérôme Lapuyade-Lahorgue, Alexandre Huat, Sébastien Thureau, David Pasquier, Isabelle Gardin, Romain Modzelewski, David Gibon, Juliette Thariat, Vincent Grégoire, Pierre Vera, Su Ruan, "A Quantitative Comparison between Shannon and Tsallis–Havrda–Charvat Entropies Applied to Cancer Outcome Prediction," *Entropy*, Application of Entropy to Computer Vision and Medical Imaging, vol 24, 4, 2022.

Conférences

[C1] **Thibaud Brochet**, Jérôme Lapuyade-Lahorgue ; Sébastien Bougleux ; Mathieu Salaün ; Su Ruan, "Deep learning based automatic detection of uninformative images in pulmonary optical endomicroscopy," *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Paris, France, 2020.
[C2] **Thibaud Brochet**, Jérôme Lapuyade-Lahorgue ; Alexandre Huat ; Sébastien Thureau ; David Pasquier ; Isabelle Gardin ; Romain Modzelewski ; David Gibon ; Juliette Thariat ; Vincent Grégoire ; Pierre Vera ; Su Ruan , "Deep Learning Based Radiomics to Predict Treatment Response Using Multi-datasets," *2022 Medical Imaging and Computer-Aided Diagnosis (MICAD)*, Leicester, United Kingdom, 2022.
[C3] **Thibaud Brochet**,Jérôme Lapuyade-Lahorgue ; Alexandre Huat, Sébastien Thureau, David Pasquier, Isabelle Gardin, Romain Modzelewski, David Gibon, Juliette Thariat, Vincent Grégoire, Pierre Vera, Su Ruan, "Apprentis-

sage profond multitâche pour la prédiction de la récidive du cancer utilisant l'entropie d'Havrda-Charvat," *XXVIIIème Colloque Francophone de Traitement du Signal et des Images, GRETSI 2022*, Nancy, France, 2022.

[C4] **Thibaud Brochet**,Jérôme Lapuyade-Lahorgue ; Hua Li ; Pierre Vera ; Pierre Decazes ; Su Ruan, "Prediction of Head-Neck Cancer Recurrence from Pet/CT Images with Havrda-Charvat Entropy," *2023 Twelfth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Paris, France, 2023.

[C5] **Thibaud Brochet**,Kangfu Han, Jiale Cheng, Fenqiang Zhao, Jerome Lapuyade-Lahorgue, Su Ruan, Yi-Fang Tu, Sheng-Che Hung, and Gang Li, "Neonatal Hypoxic Ischemic Encephalopathy severity grading using multimodal Swin Transformer," *2024 21st IEEE International Symposium on Biomedical Imaging (ISBI*, Athens, Greece, 2024.

Bibliographie

[1] A. Ahmed, P. Gibbs, M. Pickles, and L. Turnbull. Texture analysis in assessment and prediction of chemotherapy response in breast cancer. *Journal Magn. Reson. Imaging*, 33(1) :89–101, 2013.

[2] E. Aldana-Bobadilla and A. Kuri-Morales. A Clustering Method Based on the Maximum Entropy Principle. *Entropy*, 17 :151–180, 2015.

[3] K.A. Allen and D.H. Brandon. Hypoxic ischemic encephalopathy : Pathophysiology and experimental treatments. *Newborn Infant Nurs Rev*, 2011.

[4] B. A. Altazi, G. G. Zhang, D. C. Fernandez, and et al. Reproducibility of (18 F)-FDG PET radiomic features for different cervical tumor segmentation methods, gray-level discretization, and reconstruction algorithms. *Journ. Appl. Clin. Med. Phys.*, 18(6) :32–48, 2017.

[5] A. Amyar, R. Modzelewski, P. Vera, V. Morard, and S. Ruan. Multi-task multi-scale learning for outcome prediction in 3D PET images. *Computers in Biology and Medicine*, 151, 2022.

[6] A. Amyar, S. Ruan, I. Gardin, and et al. 3D RPET-NET : Development of a 3D PET Imaging Convolutional Neural Network for Radiomics Analysis and Outcome Prediction. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 3(2) :225–231, 2019.

[7] A. Ashraf, B. Gaonkar, Mies C., and et al. Breast DCE-MRI kinetic heterogeneity tumor markers : preliminary associations with neoadjuvant chemotherapy response. *Transl. Oncol.*, 8(3) :154–162, 2015.

[8] M. Basseville. Information : entropies, divergences et moyennes. Technical report, INRIA, 1996.

[9] H. Bhatia, W. Paul, F. Alajaji, and et al. Least k-th-Order and Rényi Generative Adversarial Networks. *Neural Computation*, 33(9) :2473–2510, 2021.

[10] N. Bhoostan, M. L. Giger, S. A. Jansen, and et al. Cancerous breast lesions on dynamic contrast-enhanced MR images : computerized characterization for image-based prognostic markers. *Radiology*, 254(3) :680–690, 2010.

[11] L. Boltzmann. Über die beziehung dem zweiten Haubtsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respektive den Sätzen über das Wärmegleichgewicht. *Wiener Berichte*, 76 :373–435, 1877.

[12] N. Braman, M. Etesami, P. Prasanna, and et al. Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neadjuvant chemotherapy based on breast DCE-MRI. *Breast Cancer Research*, 19(57), 2017.

[13] T. Brochet, K. Han, J. Cheng, and et al. Neonatal Hypoxic Ischemic Encephalopathy Severity Grading Using Multimodal Swin Transformer. Athen, Greece, 2024. IEEE-ISBI.

[14] T. Brochet, J. Lapuyade-Lahorgue, P. Vera, and Ruan S. Deep Learning Based Radiomics To Predict Treatment Response Using Multi-Datasets. MICAD 2022, 2022.

[15] L. Cai, X. Wang, Y. Wang, and et al. Robust phase-based texture descriptor for classification of breast ultrasound images. *Biomed. Eng. Online.*, pages 14–26, 2015.

[16] Y.-C. Chang, C.-S. Huang, Y.-J. Liu, and et al. Angiogenic response of locally advanced breast cancer to neoadjuvant chemotherapy evaluated with parametric histogram from dynamic contrast-enhanced MRI. *Phys. Med. Biol.*, 49(16) :3593–3602, 2004.

[17] C.-H. Chen, C.-K. Chang, C.-Y. Tu, and et al. Radiomic features analysis in computed tomography images of lung nodule classification. *PLoS ONE*, 3(2), 2018.

[18] S. Chen and W. Guo. Auto-Encoders in Deep Learning—A Review with New Perspectives. *Mathematics*, 11(8), 2023.

[19] R. Clausius. *Théorie mécanique de la chaleur*. Eugène Lacroix, 1868.

[20] G. E. Crooks. On Measures of Entropy and Information. Technical report, 2018.

[21] H. Cunningham, A. Ewart, L. Riggs, R. Huben, and L. Sharkey. Sparse Autoencoders Find Highly Interpretable Features in Language Models, 2023.

[22] C. Das and K. Bhattacharyya. Information entropy and bound quantum states. *Phys. Rev.*, 79(1), 2009.

[23] M.-C. Desseroit, F. Tixier, W. A. Weber, and et al. Reliability of PET/CT shape and heterogeneity features in functional and morphologic components of non-small cell lung cancer tumors : a repeatability analysis in a prospective multicenter cohort. *Journ. Nucl. Med.*, 58(3) :406–411, 2017.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, and et al. An Image is Worth 16x16 Words : Transformers for Image Recognition at Scale. ICLR 2021, 2021.

[25] D. Emsley. Variance analysis and performance : two empirical studies. *Accounting, Organizations and Society.*, 25(1) :1–12, 2000.

[26] T. Evgeniou and M. Pontil. Support vector machines : Theory and applications. volume 2049, pages 249–257, 09 2001.

[27] D. Ezzat, A.E. Hassanien, and H.A. Ella. An optimized deep learning architecture for the diagnosis of covid-19 disease based on gravitational search optimization. *Appl Soft Comput*, 2021.

[28] M. Fan, H. Cheng, P. Zhang, and et al. DCE-MRI texture analysis with tumor subregion partitioning for predicting Ki-67 status of estrogen receptor-positive breast cancers. *J. Magn. Reson. Imaging*, 48(1) :237–247, 2018.

[29] M. Fan, H. Li, S. Wang, and et al. Radiomic analysis reveals DCE-MRI features for prediction of molecular subtypes of breast cancer. *PloS One*, 12(2), 2017.

[30] A. Ferreira and M. Figueiredo. Boosting algorithms : A review of methods, theory, and applications. *Ensemble Machine Learning : Methods and Applications*, 3 :35–85, 01 2012.

[31] J. P. Fortin, D. Parker, B. Tunc, and et al. Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, 161(149-170), 2017.

[32] M. Gabrié, A. Manoel, C. Luneau, and et al. Entropy and mutual information in models of deep neural networks. NeurIPS, 2018.

[33] M. Gao, J. Chen, H. Mu, and D. Qi. A transfer residual neural network based on resnet-34 for detection of wood knot defects. *Forests*, 12 :212, 02 2021.

[34] B. Gaonkar, J. Beckett, M. Attiah, and et al. Eigenrank by committee : Von-Neumann entropy based data subset selection and failure prediction for deep learning based medical image segmentation. *Medical Image Analysis*, 67, 2021.

[35] L. Gjesteby, B. D. Man, Y. Jin, and et al. Metal artifact reduction in ct : where are we after four decades ? *IEEE Access*, 4 :2169–3536, 2016.

[36] J. Goya-Outi, F. Orlhac, Calmon, and et al. Computation of reliable textural indices from multimodal brain MRI : suggestions based on a study of patients with diffuse intrinsic pontine glioma. *Phys. Med. Biol.*, 63(10), 2018.

[37] L. Guo, R. Sharma, L. Yin, and et al. Automated competitor analysis using big data analytics : Evidence from the fitness mobile app business. *Business Process Management Journal*, 23, 06 2017.

[38] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.*, 3(6) :610–621, 1973.

[39] JK. Harris. Primer on binary logistic regression. *Fam Med Community Health*, 2021.

[40] R. V. L. Hartley. Transmission of information. *Bell Syst. Techn. Journal*, 7(3) :535–563, 1928.

[41] M. Hatt, M. Majdoub, M. Vallières, and et al. (18F)-FDG PET uptake characterization through texture analysis : investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. *Journ. Nucl. Med.*, 56(1) :38–44, 2015.

[42] M. Hatt, F. Tixier, L. Pierce, and et al. Characterization of PET/CT images using texture analysis : the past, the present, ... any future ? *European Journal Nucl. Med. Mol. Imaging*, 44(1), 2017.

[43] H. Hu, L. Shen, Q. Guan, and et al. Deep Co-supervision and Attention Fusion Strategy for Automatic COVID-19 Lung Infection Segmentation on CT Images. *Pattern Recognition*, 124, 2022.

[44] L. Huang, S. Ruan, and T. Denœux. Semi-supervised multiple evidence fusion for brain tumor segmentation. *Neurocomputing*, 535(28) :40–52, 2023.

[45] O. Johnson and C. Vignat. Some results concerning maximum Rényi entropy distributions. *Annales de l'institut Henri Poincaré*, 43(339-351), 2007.

[46] W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostat*, 8(1) :118–127, 2007.

[47] J.-H. Kim, E. S. Ko, Y. Lim, and et al. Cancer heterogeneity : MR imaging texture analysis and survival outcomes. *Radiology*, 282(3) :665–675, 2017.

[48] D. P. Kingma and J. Ba. ADAM : A METHOD FOR STOCHASTIC OPTIMIZATION. Amsterdam, Netherlands. ICLR 2015.

[49] Diederik P. Kingma and Max Welling. An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*, 12(4) :307–392, 2019.

[50] S. et al. Kiritchenko. ExaCT : automatic extraction of clinical trial characteristics from journal publications. *BMC Med Inform Decis Mak.*, 10(56), 2010.

[51] L. R. Koetzier, D. Mastrodicasa, T. P. Szczykutowicz, and et al. Deep Learning Image Reconstruction for CT : Technical Principles and Clinical Prospects. *Radiology*, 306(3), 2023.

[52] S. Kumar and G. Ram. A Generalization of the Havrda-Charvat and Tsallis Entropy and Its Axiomatic Characterization. *Abstract and Applied Analysis*, 2014, 2014.

[53] V. Kumar, Y. Gu, S. Basu, and et al. Radiomics : the process and the challenges. *Magn. Reson. Imaging*, 30(9) :1234–1248, 2012.

[54] T. Kuttiyannan and R. Roselin. *Mammogram Mining Using Genetic Ant-Miner*, pages 775–792. 01 2013.

[55] P. Lambin, R. Leijenaar, T. M. Deist, and et al. Radiomics : the bridge between medical imaging and personalized medicine. *Nat. Rev. Oncol.*, 14(12) :749–762, 2017.

[56] P. Lambin, E. Rios-Velazquez, R. Leijenaar, and et al. Radiomics : extracting more information from medical images using advanced feature analysis. *Eur. Journ. Cancer*, 48(4) :441–446, 2012.

[57] S. Lambrechts, D. Smeets, M. Moisse, and et al. Genetic heterogeneity after first-line chemotherapy in high-grade serous ovarian cancer. *Eur. Journal Cancer*, 53 :51–64, 2016.

[58] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11) :2278–2324, 1998.

[59] L. Lefebvre and P. Poulin. Analysis and Synthesis of Structural Textures. pages 77–86, 01 2000.

[60] R. T. H. Leijenaar, G. Nalbantov, S. Carvalho, and et al. The effect of SUV discretization in quantitative FDG-PET Radiomics : the need for standardized methodology in tumor texture analysis. *Sci. Rep.*, 5, 2015.

[61] A. Lesne. Shannon entropy : a rigorous notion at the crossroads between probability, information theory, dynamical systems and statistical physics. *Cambridge University Press*, 2014.

[62] Z. Liu, Y. Lin, Y. Cao, and et al. Swin Transformer : Hierarchical Vision Transformer using Shifted Windows, 2021.

[63] F. Lucia, D. Visvikis, M. Vallières, and et al. External validation of a combined PET and MRI radiomics model for prediction of recurrence in cervical cancer patients treated with chemoradiotherapy. *Eur. Journal Nucl. Med. Mol. Imaging*, 2018.

[64] A. Mackiewicz and W. Ratajczak. Principal components analysis (PCA). *Computers and Geosciences*, 19(3) :303–342, 1993.

[65] G. Maksa. The stability of the entropy of degree alpha. *Journal of Mathematical Analysis and Applications*, 346(1) :17–21, 2008.

[66] S. Mannor, D. Peleg, and R. Rubinstein. The cross entropy method for classification. Bonn, Germany, 2005. Proceedings of the 22nd International Conference on Machine Learning.

[67] M. E. Mayerhoefer and et al. Introduction to radiomics. *J Nucl Med.*, 61(4) :488–495, 2020.

[68] K. S. Miller. Derivatives of Noninteger Order. *Mathematics Magazine.*, 68(3), 1995.

[69] K. Miura. An introduction to maximum likelihood estimation and information geometry. *Interdisciplinary Information Sciences (IIS)*, 17, 11 2011.

[70] T. Mora and A. M. M. Walczak. Rényi entropy, abundance distribution, and the equivalence of ensembles. *Physical Review*, 93, 2016.

[71] C. Nioche, F. Orlhac, S. Boughdad, and et al. LIFEx : a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. *Cancer Res.*, 78(16) :4786–4789, 2018.

[72] C. A. Owens, C. B. Peterson, C. Tang, and et al. Lung tumor segmentation methods : Impact on the uncertainty of radiomics features for non-small cell lung cancer. *PLoS One*, 13(10), 2018.

[73] J. A. Pachter, Y.-J. Yang, and K. A. Dill. Entropy, irreversibility and inference at the foundations of statistical physics. *Nature*, 2024.

[74] C. Parmar, P. Grossmann, J. Bussink, and et al. Machine learning methodsf or quantitative radiomic biomarkers. *Sci. Rep.*, 5, 2015.

[75] C. Parmar, E. Rios Velazquez, R. Leijenaar, and et al. Robust Radiomics feature quantification using semiautomatic volumetric segmentation. *PloS One*, 9(7), 2014.

[76] A.-M. Patch, E. L. Christie, D. Etemadmoghadam, and et al. Whole-genome characterization of chemoresistant ovarian cancer. *Nature*, 521(7553) :489–494, 2015.

[77] Z. Qu, J. Mei, L. Liu, and D. Y. Zhou. Crack Detection of Concrete Pavement With Cross-Entropy Loss Function and Improved VGG16 Network Model. Melbourne, Australia, 2020. Proceedings of the 30th International Telecommunication Networks and Applications Conference (ITNAC).

[78] V. Rajinikanth, K. P. Thanaraj, S. C. Satapathy, and et al. Shannon's Entropy and Watershed Algorithm Based Technique to Inspect Ischemic Stroke Wound. *Smart Intelligent Computing and Applications*, 2 :23–31, 2018.

[79] Z. Ramezani and A. Pourdarvish. Transfer learning using Tsallis entropy : An application to Gravity Spy. *Physica A*, 561, 2021.

[80] D. Ramos, J. Franco-Pedroso, A. Lozano-Diez, and J. Gonzalez-Rodriguez. Deconstructing Cross-Entropy for Probabilistic Binary Classifiers. *Entropy*, 20, 2018.

[81] P. Ranganathan. An Introduction to Statistics : Choosing the Correct Statistical Test. *Indian J Crit Care Med.*, 25(2) :184–S186, 2021.

[82] S. Reuzé, F. Orlhac, C. Chargari, and et al. Prediction of cervical cancer recurrence using textural features extracted from (18F)-FDG PET images acquired with different scanners. *Oncotarget*, 8(26), 2017.

[83] O. Ronneberger, P. Fisher, and T. Brox. U-Net : Convolutional Networks for Biomedical Image Segmentation. *arXiv :1505.04597*, 2015.

[84] U. Ruby and V. Yendapalli. Binary cross entropy with deep learning technique for Image classification. *Int. J. Adv. Trends Comput. Sci. Eng.*, 9 :5393–5397, 2020.

[85] M. Saad, S. He, W. Thorstad, and et al. Learning-based Cancer Treatment Outcome Prognosis using Multimodal Biomarkers. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 2021.

[86] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen. Mobilenetv2 : Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[87] M. Shafiq-Ul-Hassan, G. G. Zhang, K. Latifi, and et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med. Phys.*, 44(3) :1050–1062, 2017.

[88] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Techn. Journal*, 27(3) :379–423, 1948.

[89] J. R. Shewchuk. An Introduction to the Conjugate Gradient Method Without the Agonizing Pain. 1994.

[90] L. Silva, J. Marques de Sa, and L. A. Alexandre. Neural network classification using Shannon's entropy. Bruges, Belgium. Proceedings of the ESANN 2005 Proceedings—13th European Symposium on Artificial Neural Networks.

[91] Y.-Y. Song. Decision tree methods : applications for classification and prediction. *Shanghai Arch Psychiatry.*, 27(2), 2015.

[92] YY. Song and Y Lu. Decision tree methods : applications for classification and prediction. *Shanghai Arch Psychiatry*, 27(2) :(130)–135, 2015.

[93] J. Sublime, B. Matei, G. Cabanes, and et al. Entropy based probabilistic collaborative clustering. *Pattern recognition*, 72 :144–157, 2017.

[94] R. Sun and et al. Intelligence artificielle en radiothérapie : radiomique, pathomique, et prédiction de la survie et de la réponse aux traitements. *Cancer/Radiothérapie*, 25(6-7) :630–637, 2021.

[95] C. Sutton and A. McCallum. An introduction to conditional random fields, 2010.

[96] I. Thomassin-Naggara, E. Aubert, A. Rockall, and et al. Adnexal masses : development and preliminary validation of an MR imaging scoring system. *Radiology*, 267(2) :432–443, 2013.

[97] I. Thomassin-Naggara, I. Trop, J. Chopier, and et al. Nonmasslike enhancement at breast MR imaging : the added value of mammography and us for lesion categorization. *Radiology*, 261(1) :69–79, 2011.

[98] Y. Tian, Y. Zhang, and H. Zhang. Recent Advances in Stochastic Gradient Descent in Deep Learning. *Mathematics*, 11(3), 2023.

[99] L. Tianyang, W. Yuxin, L. Xiangyang, and Q. Xipeng. A survey of transformers. *AI Open*, 3 :111–132, 2022.

[100] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Mathematics Magazine.*, 58(1), 1996.

[101] R. Trullo, C. Petitjean, C. Dubray, and S. Ruan. Multiorgan segmentation using distance-aware adversarial networks. *SPIE, Journal of Medical Imaging*, 6(1), 2019.

[102] J. J. M. Van Griethuysen, A. Fedorov, C. Parmar, and et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.*, 77(21), 2017.

[103] P. Van Helden. Data-driven hypotheses. *EMBO Rep.*, 14(2) :104, 2013.

[104] S. Vande Perre, L. Duron, A. Milon, and et al. Radiomique : mode d'emploi. Méthodologie et exemples d'application en imagerie de la femme. *Imagerie de la Femme*, 29(1) :25–33, 2019.

[105] J.-J. Weng, F. Alajaji, and T. Linder. An Information Bottleneck Problem with Rényi's Entropy. IEEE International Symposium on Information Theory, 2021.

[106] K. K. Wickstrø m, S. Lø kse, M. C. Kampffmeyer, and et al. Analysis of Deep Convolutional Neural Networks Using Tensor Kernels and Matrix-Based Entropy. *Entropy*, 25(6), 2023.

[107] S. Woo, J. Park, J. Lee, and I.S. Kweon. Cbam : Convolutional block attention module. In *Computer Vision – ECCV 2018*, pages 3–19. Springer International Publishing, 2018.

[108] H. Wu, R. Wang, S. Li, and et al. Aspect ratio-dependent dual-regulation of the tumor immune microenvironment against osteosarcoma by hydroxyapatite nanoparticles. *Acta Biomater.*, 2023.

[109] J. Yan, J. L. Chu-Shern, H. Y. Loi, and et al. Impact of Image reconstruction settings on texture features in (18F)-FDG PET. *Journal Nucl. Med.*, 56(11) :1667–73, 2015.

[110] H. Yao, R. Stidham, and R. et al. Soroushmehr. Automated detection of non-informative frames for colonoscopy through a combination of deep learning and feature extraction. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2402–2406, 2019.

[111] P. Yao, D. Witte, and H. et al. Gimonet. Automatic classification of informative laryngoscopic images using deep learning. *Laryngoscope Investigative Otolaryngology*, 7(2) :460–466, 2022.

[112] L. R. Young, D. N. Franz, P. Nagarkatte, and et al. Utility of fdg-18-pet in Sporadic and Tuberous Sclerosis Associated Lymphangioleiomyomatosis. *Chest*, 136(3) :926–933, 2009.

[113] P. A. Yushkevich, J. Piven, H. C. Hazlett, and et al. User-guided 3D active contour segmentation of anatomical structures : Significantly improved efficiency and reliability. *NeuroImage*, 31(3) :1116–1128, 2006.

[114] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing.

[115] X. Zhang, X. Zhou, M. Lin, and J. Sun. Shufflenet : An extremely efficient convolutional neural network for mobile devices. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6848–6856, 2018.

[116] T. Zhou, S. Canu, and S. Ruan. Automatic COVID-19 CT Segmentation Using U-Net Integrated Spatial and Channel Attention Mechanism. *Int. Journal of Imaging Systems and Technology*, 31(1) :16–27, 2021.

[117] T. Zhou, S. Canu, P. Vera, and S. Ruan. Latent Correlation Representation Learning for Brain Tumor Segmentation with Missing MRI Modalities. *IEEE Trans.on Image Processing*, 30 :4263–4274, 2021.

[118] T. Zhou, A. Noeuveglise, R. Modzelewski, and et al. Prediction of Brain Tumor Recurrence Location Based on Multi-modal Fusion and Nonlinear Correlation Learning. *Computerized Medical Imaging and Graphics*, 106, 2023.

[119] T. Zhou, S. Ruan, P. Vera, and S. Canu. A Tri-attention Fusion Guided Multi-modal Segmentation Network Pattern Recognition. *Pattern Recognition*, 124, 2022.

[120] A. Zwanenburg, S. Leger, M. Vallières, and S. Löck. Initiative for the IBS. image biomarker standardisation initiative. 2016.