Titre: Title:	Apprentissage multi-tâches pour l'évaluation du positionnement des seins en mammographie
Auteur: Author:	Valérie Gauvin
Date:	2026
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Gauvin, V. (2026). Apprentissage multi-tâches pour l'évaluation du positionnement des seins en mammographie [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/76415/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/76415/
Directeurs de recherche: Advisors:	Lama Séoud, & François Guibault
Programme: Program:	Génie logiciel

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Apprentissage multi-tâches pour l'évaluation du positionnement des seins en mammographie

VALÉRIE GAUVIN
Département de génie informatique et génie logiciel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie informatique

Mai 2026

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Apprentissage multi-tâches pour l'évaluation du positionnement des seins en mammographie

présenté par **Valérie GAUVIN**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Farida CHERIET, présidente
Lama SÉOUD, membre et directrice de recherche
François GUIBAULT, membre et codirecteur de recherche
Luc DUONG, membre

DÉDICACE

À Lauriane et Victor,
les deux tiers qui me complètent

REMERCIEMENTS

Je tiens d'abord à remercier ma directrice de recherche, Lama Séoud, pour son soutien pendant ces deux années de maîtrise. Elle a toujours été disponible et à l'écoute, et a su m'encourager tout au long de ce projet. Merci d'avoir rendu mon expérience à la maîtrise aussi agréable et enrichissante ! Un grand merci également à François Guibault, mon codirecteur de recherche, pour son encadrement et ses bons commentaires tout au long du projet.

J'aimerais remercier l'équipe d'IMD Research, particulièrement Nathaniel Lasry, pour avoir rendu ce projet possible. Je souhaite également remercier les trois technologues qui ont participé au projet : Karine Morency, Isabelle Jean et Ginette Goudreau. Le projet n'aurait pas été possible sans votre précieuse collaboration. Merci également à Philippe Debanne pour son travail sur le projet et les articles qui en ont découlé.

Je tiens à remercier chaleureusement mes deux acolytes du projet Mammo, Lauriane et Victor. Quelle chance d'avoir pu faire ma maîtrise aux côtés de mes proches amis ! Merci pour votre support et votre écoute dans les bons comme dans les moins bons moments. Ces deux dernières années n'auraient pas été les mêmes sans vous. Je souhaite aussi remercier mes camarades du VisionIC pour ces deux années de recherche partagées avec vous. Merci pour les conseils, les rires et les débats sémantiques au dîner. Je suis extrêmement chanceuse de vous avoir rencontrés ; les moments passés avec vous sont toujours un plaisir.

Finalement, je souhaite remercier ma famille et mes proches qui m'ont toujours encouragée et soutenue dans mon parcours. Votre présence à mes côtés est précieuse et me motive tous les jours.

Ce projet a été financé au travers d'une double subvention CRSNG Alliance et Medteq+.

RÉSUMÉ

Le cancer du sein est un enjeu de santé important au Canada et dans le monde, et son diagnostic précoce est primordial afin de maximiser l'efficacité des traitements. La modalité de dépistage du cancer du sein la plus répandue à ce jour est la mammographie ; cette technique d'imagerie par rayons X permet de visualiser les différents tissus et structures du sein, y compris les masses cancéreuses. L'acquisition des images radiographiques nécessite une forte compression des seins, rendant cet examen inconfortable pour les patient.e.s et très dépendant de leurs morphologies variées. Le positionnement mammaire est donc un grand enjeu en mammographie, un positionnement inadéquat pouvant causer l'exclusion de tissu mammaire et mener à des rappels de patient.e.s ou à des faux négatifs au cancer du sein.

Le travail présenté dans ce mémoire s'inscrit dans un projet plus large impliquant d'autres étudiants, et visant au développement d'un système automatique d'évaluation du positionnement mammaire. Plus particulièrement, ce mémoire se penche sur l'étude des liens intrinsèques existants entre les critères de positionnement mammaire établis, et comment ces liens peuvent être exploités dans le cadre de l'évaluation automatique de ces critères. À cet effet, un jeu de données adapté à l'entraînement de modèles d'apprentissage profond a été créé, puis des modèles de référence ont été implémentés afin d'établir des performances de base pour l'évaluation de différents critères de positionnement. Des modèles d'apprentissage multi-tâches ont ensuite été élaborés dans le but de mettre à profit les interdépendances entre certains critères. Les expérimentations effectuées ont révélé que la combinaison de certains critères permettait une amélioration des performances par rapport aux modèles de référence. Ce travail est le premier à aborder l'interdépendance des critères de positionnement mammaire dans le cadre de leur évaluation par des modèles d'apprentissage profond, et ouvre de nouvelles perspectives dans ce domaine de recherche.

Un objectif complémentaire de ce travail est de déterminer s'il est possible de vérifier le positionnement mammaire sur des images acquises à plus faible doses de rayons X, comme par exemple les images *pre-shot* de certains systèmes d'acquisition. Pour ce faire, et en l'absence de telles images très basse-dose, une méthode de simulation de réduction de dose a été implémentée et appliquée aux images de notre jeu de données. L'inférence des modèles d'apprentissage profond sur ces images a validé cette hypothèse, ouvrant la voie à la possibilité d'une irradiation moins grande des patient.e.s avant l'obtention d'un positionnement mammaire adéquat.

ABSTRACT

Breast cancer is a major health issue in Canada and around the world, and early diagnosis is essential to maximize the effectiveness of treatments. The most widely used breast cancer screening method nowadays is mammography; this X-ray imaging technique allows for the visualization of the various tissues and structures of the breast, including cancerous masses. Acquiring radiographic images requires significant compression of the breasts, making this exam uncomfortable for patients and highly dependent on their morphology. Breast positioning is therefore a major challenge in mammography, as inadequate positioning can result in the exclusion of breast tissue and lead to patient recalls or false-negative breast cancer cases.

The work presented in this thesis is part of a larger project involving other students, aimed at developing an automated system for evaluating breast positioning. More specifically, this thesis examines the intrinsic relationships between established breast positioning criteria and how these relationships can be leveraged in the context of their automated evaluation. To this end, a dataset suitable for training deep learning models was created, and baseline models were implemented to establish basic performance. Multi-task learning models were then developed to leverage the interdependencies between certain criteria. The experiments conducted revealed that combining certain criteria led to improved performance compared to the baseline models. This work is the first to address the relationships between breast positioning criteria in the context of their evaluation by deep learning models, and opens up new avenues in this field of research.

An additional objective of this study is to determine whether it is possible to verify breast positioning using images acquired at lower X-ray doses, such as pre-shot images available in some mammographic systems. To this end, and without such images at-hand, a dose-reduction simulation method was implemented and applied to the images in our dataset. Inference of deep learning models on these images validated this hypothesis, paving the way for the possibility of reducing patient radiation exposure before achieving adequate breast positioning.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

ACR	American college of radiology
AUC	Aire sous la courbe (*Area Under the Curve*)
BCE	Entropie croisée binaire (*Binary cross-entropy*)
BI-RADS	*Breast Imagning Reporting and Data System*
CAR	Association canadienne des radiologistes
CBIS-DDSM	*Curated Brast Imaging Subset of* DDSM
CC	Cranio-caudal
CNN	Réseau de neurones convolutif (*Convolutional Neural Network*)
CrossAtt	*Cross-Attention module*
CTAB	*Cross-task attention bottleneck*
CTAE	*Cross-task attention encoder*
CVAT	*Computer Vision Annotations Tool*
DDSM	*Digital Database for Screening Mammography*
EUREF	*European Reference Organization for Quality Assured Breast Screening and Diagnostic Services*
FC	Couche entièrement connectée (*Fully connected layer*)
FP	Faux positifs
FN	Faux négatifs
FPN	*Feature Pyramid Network*
HPS	*Hard Parameter Sharing*
IA	Intelligence artificielle
IMA	Angle inframmammaire
K	*Key*
lr	*Learning rate*
MIAS	*Mammography Image Analysis Society*
ML	Multi-label
MLO	Médio-latéral oblique
MTAN	*Multi-task Attention Network*
MultiHead	*Multi-head Attention module*
NHSBSP	*National Health Service Breast Screening Program*
OTIMROEPMQ	Ordre des technologues en imagerie médicale, en radio-oncologie et en électrophysiologie médicale du Québe
PGMI	*Perfect, Good, Moderate, Inadequate*

PL	Ligne du muscle pectoral
PNL	Ligne postérieure du mamelon
PQDCS	Programme québécois de dépistage du cancer du sein
PR	Précision-rappel (*Precision-recall*)
Q	*Query*
ResNet	*Residual network*
RPN	*Region Proposal Network*
ROC	Fonction d'efficacité du récepteur (*Receiving Operator Characteristic*)
ROI	Région d'intérêt
SPS	*Soft Parameter Sharing*
TP	Vrais positifs
TN	Vrais négatifs
V	*Value*
ViT	Modèle transformeur visuel (*Vision transformer*)
WCE	BCE pondéré (*Weighted binary cross-entropy*)
YOLO	*You Only Look Once*

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

Le cancer du sein figure parmi les formes de cancer les plus répandues chez les femmes au Canada. En 2025, 31 900 femmes ont reçu un diagnostic de cancer du sein, représentant 26% de tous les nouveaux cas de cancer dans cette population [9]. Dans le monde, on estimait à 670 000 le nombre de décès par cancer du sein en 2022 [10]. Le cancer du sein est une pathologie qui se développe dans les cellules du sein ; il est caractérisé par la formation d'une tumeur cancéreuse, soit un groupe de cellules ayant une prolifération anormale et pouvant envahir les tissus voisins, entraînant leur destruction [11].

Le sein est un organe composé d'une glande mammaire ainsi que du tissu de soutien. Le tissu de soutien est constitué majoritairement de graisse, de fibres et de vaisseaux, alors que la glande mammaire comprend des lobules et des canaux ayant pour fonction de produire le lait et de l'acheminer vers le mamelon [1]. Les cancers apparaissent le plus fréquemment dans le tissu glandulaire, plus précisément dans les cellules adjacentes aux canaux et dans les cellules des lobules [11]. L'anatomie du sein est représentée à la Figure 1.1.

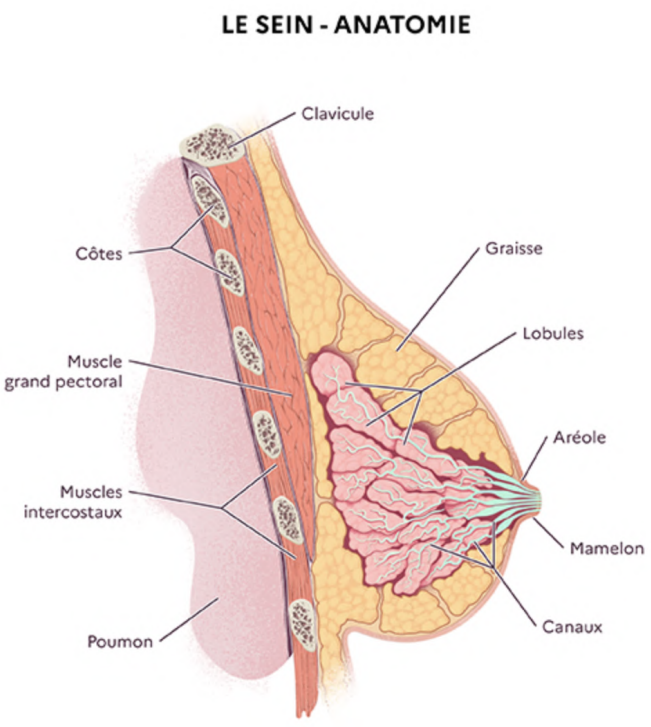
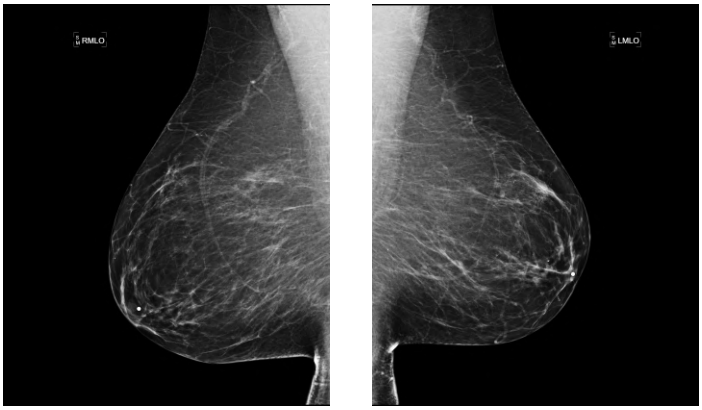
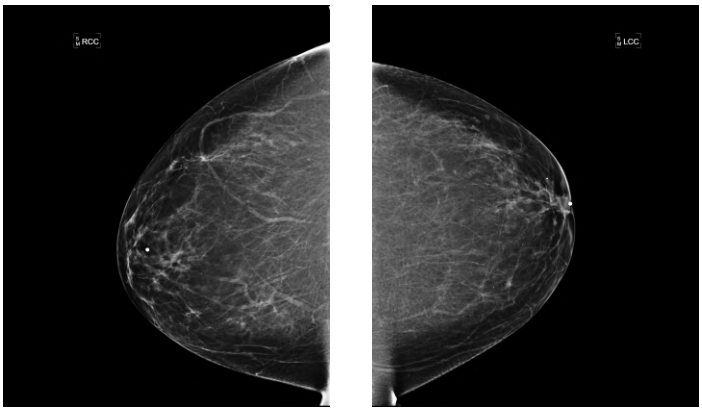


FIGURE 1.1 Anatomie du sein [1]

Le diagnostic précoce du cancer du sein est primordial afin de maximiser les chances de réussite des traitements. Plusieurs modalités d'imagerie sont disponibles à cette fin : les ultrasons, l'IRM et la mammographie. Cette dernière est la plus répandue pour le dépistage du cancer du sein. Il s'agit d'une technique d'imagerie par rayons X à faible dose permettant de visualiser les différentes structures du sein, dont les masses cancéreuses, les lésions et les calcifications [12]. La mammographie consiste en l'acquisition de deux images radiographiques par sein ; une image pour la vue médio-latérale oblique (MLO) et une image pour la vue cranio-caudale (CC). Ces deux vues sont nécessaires afin de couvrir l'entièreté du tissu mammaire d'un sein. La MLO offre une visualisation du quadrant supérieur externe du sein et permet de voir la majeure partie du tissu mammaire. Elle est prise à un angle d'environ 45 degrés. La CC est une vue prise du dessus du sein et permet la visualisation du quadrant interne du sein. Cette vue permet de voir des lésions qui ne seraient pas visibles en MLO. La Figure 1.2 présente les quatre images standards d'un examen.



(a) MLO droite (b) MLO gauche



(c) CC droite (d) CC gauche

FIGURE 1.2 Quatre vues standards composant un examen mammographique

Lors de l'examen, le sein est placé entre deux plateaux de compression par un technologue en mammographie. Pour l'acquisition d'une image, la compression dure entre 10 et 15 secondes, et l'examen complet se déroule sur environ 20 minutes. La compression du sein est nécessaire afin de réduire le chevauchement des tissus, de réduire la dose de radiation, de limiter le mouvement durant l'examen et d'uniformiser la forme des seins [13]. Certains fabricants d'appareils de mammographie offrent aussi l'option d'acquérir des images dites *pre-shot*, soit des clichés obtenus à plus faible dose de rayons X dans le but de régler automatiquement les paramètres de l'imageur comme la dose d'irradiation optimale. Le rôle du technologue est d'effectuer l'acquisition des clichés mammaires, de guider les patientes tout au long du processus et de s'assurer de la qualité des images acquises. L'interprétation des images est ensuite effectuée par un radiologiste, soit un médecin spécialisé dans les techniques d'imagerie. À cause de la forte compression des seins, il s'agit d'un examen souvent inconfortable pour les patientes. De plus, la mammographie est sensible aux différentes morphologies des patient.e.s. Il s'agit de l'un des facteurs les plus importants à prendre en compte lors de la réalisation d'un examen : les technologues tentent d'obtenir la meilleure qualité d'image possible tout en respectant les limites morphologiques et le seuil de tolérance à la douleur des patient.e.s. Le dépistage étant le meilleur moyen de prévention, le Gouvernement du Québec a mis en place en 1998 le Programme québécois de dépistage du cancer du sein (PQDCS) [14]. Ainsi, les femmes de 50 à 74 ans sont invitées à passer une mammographie tous les deux ans.

Un des plus grands enjeux en mammographie est le positionnement mammaire [15–18]. En effet, il est nécessaire que la totalité du tissu mammaire soit visible sur les images acquises afin de garantir un bon diagnostic. Un positionnement inadéquat peut entraîner des faux négatifs au cancer du sein ou des reconduites d'examens, causant du stress pour les patient.e.s et une perte de ressources financières et temporelles pour le système de santé. Dans un audit de l'Ordre des technologues en imagerie médicale, en radio-oncologie et en électrophysiologie médicale du Québec (OTIMROEPMQ) réalisé sur des mammographies acquises entre 2004 et 2005 au Québec dans le cadre du PQDCS, il a été révélé que 49.7% des mammographies ne respectaient pas les standards de positionnement de l'Association Canadienne des Radiologistes (CAR) [18]. En 2015, l'*American College of Radiology* (ACR) rapportait que 92% des échecs d'examens étaient dûs à un positionnement mammaire inadéquat [17]. Une autre étude impliquant deux radiologistes évaluant la qualité des mammographies a révélé que le problème le plus fréquemment signalé comme étant à l'origine d'une mauvaise qualité d'image était le positionnement mammaire [15]. Ces constats mettent en évidence les lacunes persistantes concernant le positionnement mammaire, ainsi que le besoin de travaux plus approfondis sur cette problématique.

Le travail présenté dans ce mémoire s'inscrit dans un projet plus large visant à établir des mo-

dèles d'apprentissage profond pour l'évaluation automatique du positionnement mammaire en temps réel dans le contexte clinique québécois. Ce projet regroupe une équipe de trois étudiants à la maîtrise et d'un stagiaire post-doctoral. La recherche présentée dans ce mémoire aborde particulièrement la mise en place de modèles d'apprentissage profond multi-tâches afin d'exploiter l'interdépendance entre certains critères de positionnement. Ce travail comprend aussi une étude de l'impact d'une réduction de dose de rayons X lors de la mammographie sur les modèles d'évaluation.

La structure du mémoire est la suivante : le chapitre 2 présente les travaux existants de la littérature et les connaissances en lien avec le projet, puis les questions de recherche et les objectifs du projet sont énoncés dans le chapitre 3. Le chapitre 4 décrit la méthodologie employée pour réaliser les différentes parties du projet, incluant la création du jeu de données, l'analyse de la variabilité inter-technologues, la définition des modèles *baselines* ainsi que la conception des modèles d'apprentissage multi-tâches. Les résultats obtenus sont présentés et discutés dans le chapitre 5, puis le mémoire est conclu dans le chapitre 6, qui présente un résumé des travaux, les limitations ainsi que les recommandations pour les travaux futurs.

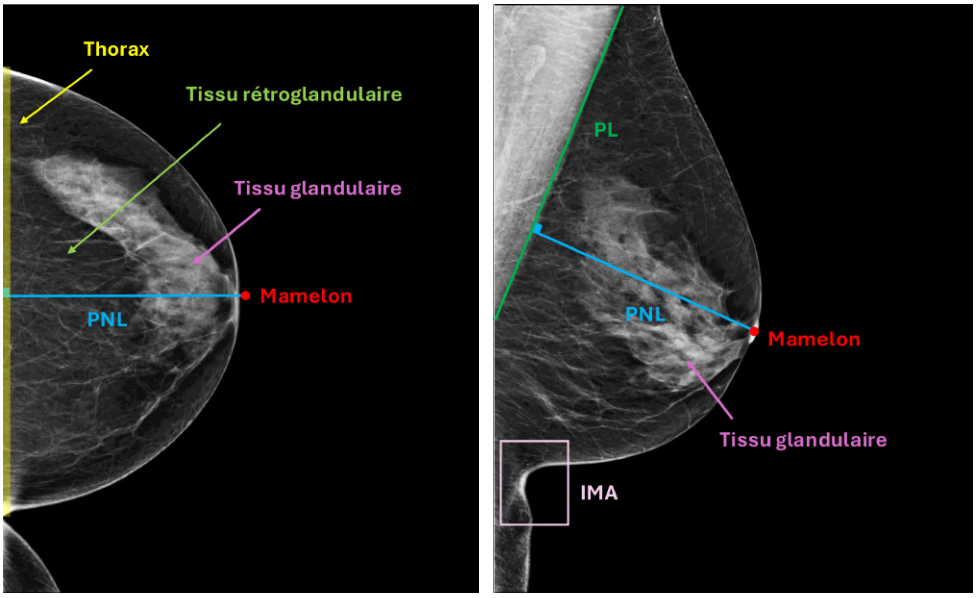
CHAPITRE 2 REVUE DE LITTÉRATURE ET DES CONNAISSANCES

Ce chapitre résume l'état des connaissances en lien avec le projet de recherche. Dans un premier temps, les différents standards de positionnement existants seront énoncés, avec une attention particulière portée sur les critères de positionnement utilisés au Québec. Ensuite, l'interdépendance entre les critères sera abordée et discutée. En second lieu, une revue des travaux traitant des méthodes de vision par ordinateur appliquées aux mammographies sera exposée. Les méthodes portant plus précisément sur l'évaluation automatique des critères de positionnement mammaire seront ensuite présentées, incluant des approches par traitement d'images classiques et des approches par apprentissage profond. Par la suite, les méthodes d'apprentissage multi-tâches seront abordées, en ciblant les travaux appliqués au domaine médical ainsi qu'aux mammographies. Finalement, des méthodes de simulation de réduction de dose en radiographie seront présentées.

2.1 Critères de positionnement

Des critères de positionnement ont été établis par plusieurs instances dans le monde dans le but de standardiser les exigences en matière de qualité du positionnement mammaire, ainsi que de mieux encadrer la pratique en clinique. Afin de comprendre ces critères de positionnement, il est primordial de connaître les différents repères anatomiques et les structures importantes du sein, telles que visibles dans les mammographies. La Figure 2.1 présente ces repères anatomiques et ces structures dans les deux vues acquises lors d'un examen. Cette figure a été élaborée par les membres de l'équipe du projet dans le cadre de l'écriture d'un article portant sur la définition des critères de positionnement d'un point de vue de la vision par ordinateur [2]. Elle apparaît sous forme traduite dans ce mémoire.

Dans les deux vues, le mamelon est identifiable comme une protubérance sur le contour du sein ou une plaque brillante. Dans certains cas, un marqueur radio-opaque est posé sur le mamelon en amont de l'examen ; on le reconnaît alors par un disque rond, homogène et brillant. Ensuite, la glande mammaire est constituée de tissu glandulaire et est caractérisée par une structure de couleur claire dans le sein, dont les fibres convergent vers le mamelon. Dans les seins très denses, cette glande est clairement visible, alors que dans les seins moins denses, elle est moins apparente. Dans les cas où la glande mammaire est clairement visible, il est possible de distinguer le tissu rétroglandulaire, c'est-à-dire le tissu graisseux se trouvant derrière la glande et de couleur sombre. Dans la vue MLO, la ligne du muscle pectoral (PL) est définie comme la ligne partant de l'intersection entre le muscle et le côté proximal de



(a) (b)

FIGURE 2.1 Repères anatomiques et structures du sein dans les vues (a) CC et (b) MLO [2]

l'image, et suivant le muscle dans la direction des fibres. À partir de la PL et du mamelon, il est possible de tracer la ligne postérieure du mamelon (PNL). En MLO, celle-ci est définie comme la ligne reliant le mamelon à la PL à un angle de 90 degrés. Dans la vue CC, la PNL est perpendiculaire au thorax, soit le bord proximal de l'image. Finalement, dans la vue MLO, l'angle inframammaire (IMA) est identifiable dans le bas du sein. Il s'agit de l'endroit où le sein et le corps se rejoignent, formant un angle plus ou moins grand dépendamment de la morphologie du sein et du positionnement lors de l'examen.

Plusieurs organisations ont élaboré des listes de critères pour évaluer la qualité du positionnement mammaire. L'ACR propose des critères bilatéraux séparés en deux catégories, majeurs et mineurs [19]. Les critères majeurs incluent que, pour les deux vues, aucune portion du sein ne doit être coupée, aucune autre partie du corps ne doit cacher de tissus mammaires, que le mamelon doit être vu de profil et que les tissus profonds doivent être visibles. Particulièrement pour la vue CC, la longueur de la PNL ne doit pas être à plus d'un centimètre de la longueur de la PNL en MLO. Quant aux critères mineurs, ceux-ci comprennent l'orientation adéquate du mamelon, l'absence de plis de peau dans l'IMA et dans le reste du sein, le muscle pectoral allant jusqu'au niveau du mamelon et toujours croissant, ainsi que l'alignement du sein au centre du détecteur et de l'image.

En Europe, les lignes directrices utilisées sont tirées de l'*European Reference Organization for Quality Assured Breast Screening and Diagnostic Services* (EUREF) [20]. Les lignes direc-

trices présentées dans l'*European guidelines for quality assurance in breast cancer screening and diagnosis* incluent pour la vue CC une bonne visualisation des bords médial et latéral du sein, un mamelon de profil et des images symétriques. Pour la vue MLO, les critères de positionnement incluent la bonne visualisation du tissu mammaire, le muscle pectoral au niveau du mamelon, des images symétriques, un mamelon vu de profil et un IMA clairement démontré. Particulièrement au Royaume-Uni, le *National Health Service Breast Screening Program* (NHSBSP) publie des directives pour les technologues concernant le positionnement mammaire [21]. Parmi les recommandations, on retrouve la vérification des plis de peau et des artéfacts dans l'image pour s'assurer que ceux-ci ne nuisent pas au diagnostic, le mamelon vu de profil, la symétrie, le muscle pectoral au niveau du mamelon et à un angle adéquat, et l'IMA clairement démontré.

Au Canada, la référence en positionnement mammaire est définie par la CAR, et les critères utilisés sont semblables à ceux définis par l'ACR [18]. Au Québec, Rouette et al. [16] ont élaboré une liste de critères de positionnement inspirée de celle de la CAR, et adaptée au contexte clinique québécois. Il s'agit de la liste de critères utilisée en clinique et enseignée par les technologues au Québec. Le tableau 2.1 présente ces critères de positionnement avec une courte description ainsi qu'une définition des causes d'échec des critères.

Tableau 2.1 Liste des critères de positionnement

Vue	Critère (version courte)	Description	Évaluation
CC et MLO	Image évaluable	Évaluation de la qualité de l'image pour le diagnostic, y compris l'exposition appropriée, le contraste et l'absence de flou de mouvement.	**Respect du critère :** Qualité de l'image adéquate. **Échec du critère :** Image de qualité insuffisante.
	Plis de peau dans le sein (Plis de peau)	Détection de plis de peau proéminents susceptibles de comprimer les tissus ou d'entraver la visibilité des masses et donc la détection des lésions.	**Respect du critère :** Pas de plis de peau ou plis n'obstruant pas l'image. **Échec du critère :** Présence de plis de peau obstruant l'image (pli(s) opaque(s) de plus de 1 cm de long, masquant du tissu mammaire).
	Artéfacts et superpositions (Artéfacts)	Identification de tout artefact (par exemple, traces d'antisudorifiques, cheveux ou artefacts liés à l'équipement d'imagerie) susceptible de masquer le tissu mammaire ou d'imiter une pathologie.	**Respect du critère :** Pas d'artéfacts ou artéfacts n'obstruant pas l'image. **Échec du critère :** Artéfact(s) masquant du tissu mammaire.
	Mamelon vu de profil	Vérification de la visibilité du mamelon de profil sans superposition du tissu mammaire.	**Respect du critère :** Le mamelon est vu de profil sans superposition du tissu mammaire ou est identifié par un marqueur. **Échec du critère :** Présence de superposition avec le tissu mammaire et pas de marqueur.
	Partie du sein coupée (Sein coupé)	Évaluation visant à déterminer si une partie du tissu mammaire a été exclue de l'image, en particulier au niveau des bords médial, latéral et postérieur.	**Respect du critère :** Pas de partie du sein coupée, ou seulement la peau. **Échec du critère :** Une partie du tissu mammaire est coupée.
	Bonne visualisation des tissus profonds (Tissus profonds)	Évaluation de l'inclusion complète des tissus rétroglandulaires et glandulaires afin de garantir une évaluation mammaire complète.	**Respect du critère :** Inclusion de tous les tissus profonds. Tissu rétroglandulaire visible pour les seins denses. **Échec du critère :** L'entièreté du tissu mammaire n'est pas visible.

Suite à la page suivante

Vue	Critère (version courte)	Description	Évaluation
CC	La PNL est perpendiculaire au bord de l'image (Orientation du mamelon)	Vérification de l'orientation adéquate du mamelon, favorisant l'inclusion de tous les tissus mammaires dans l'image.	**Respect du critère :** Le mamelon est orienté correctement, ou le mamelon est désorienté mais les parties internes et externes du sein sont incluses. **Échec du critère :** Le mamelon est désorienté avec parties internes ou externes du sein non incluses.
	La mesure de la ligne en CC (mamelon-corps) est égale à celle prise en MLO (Profondeur CC vs MLO)	Comparaison de la longueur de la PNL entre les vues CC et MLO.	**Respect du critère :** La PNL en CC n'est pas plus courte de plus de 1 cm que la PNL en MLO. **Échec du critère :** La PNL en CC est plus courte de plus de 1 cm que la PNL en MLO.
MLO	Quantité adéquate de muscle pectoral (Quantité de muscle)	Évaluation de l'extension suffisante du muscle pectoral le long du bord proximal de l'image, indiquant une profondeur adéquate du tissu inclus.	**Respect du critère :** L'intersection entre la PL et la PNL se trouve à l'intérieur de l'image. **Échec du critère :** L'intersection entre la PL et la PNL se trouve à l'extérieur de l'image.
	Vue de la largeur maximale du muscle pectoral (Largeur du muscle)	Évaluation de l'épaisseur suffisante du muscle pectoral.	**Respect du critère :** La PL forme un angle de 10 degrés ou plus avec la verticale et le muscle est croissant sur toute sa longueur. **Échec du critère :** La PL forme un angle de moins de 10 degrés avec la verticale ou le muscle n'est pas croissant sur toute sa longueur.
	IMA bien ouvert et montré (IMA ouvert)	Vérification de l'ouverture suffisante de l'IMA sans plis de peau nuisible.	**Respect du critère :** L'IMA est bien ouvert et montré, sans obstruction du tissu mammaire. **Échec du critère :** L'IMA n'est pas assez ouvert ou du tissu mammaire est masqué.

2.2 Interdépendance des critères

Les critères de positionnement établis ont tous pour but ultime d'assurer une visualisation optimale de l'ensemble du tissu mammaire, permettant la visibilité de toutes les lésions possibles. Ainsi, certains critères sont intrinsèquement liés par leurs définitions et par leurs implications, et sont parfois basés sur les mêmes éléments de l'image.

D'abord, le critère du Sein coupé dans la vue MLO est lié au critère de l'IMA ouvert. En effet, si l'IMA n'est pas visible sur l'image ou s'il n'est pas assez ouvert, le risque qu'une partie du tissu mammaire soit coupée augmente. De manière similaire, le critère des Tissus profonds dans la vue CC est lié au critère de l'Orientation du mamelon : si le mamelon est désorienté, cela implique un débalancement des quadrants proximal et distal du sein. De manière générale, cela implique aussi l'exclusion d'une partie des tissus profonds. Dans la vue MLO, si la largeur ou la quantité de muscle pectoral n'est pas adéquate, le risque d'un manque de visualisation des tissus profonds augmente aussi. Finalement, les critères du Sein coupé et des Tissus profonds dans chacune des vues sont intimement liés, visant tous deux à déterminer directement si une partie du tissu mammaire est exclue de l'image.

La Figure 2.2 illustre les interdépendances entre les critères de positionnement et inclut également l'identification préalable des repères anatomiques nécessaires à l'évaluation de certains d'entre eux. Elle permet de décrire le partage d'informations entre certains critères de positionnement. Cette figure a été élaborée par les membres de l'équipe de recherche dans le cadre de l'écriture d'un article (soumis pour publication) détaillant les critères de positionnement et leur évaluation [2]. Elle est présentée en version traduite dans ce mémoire.

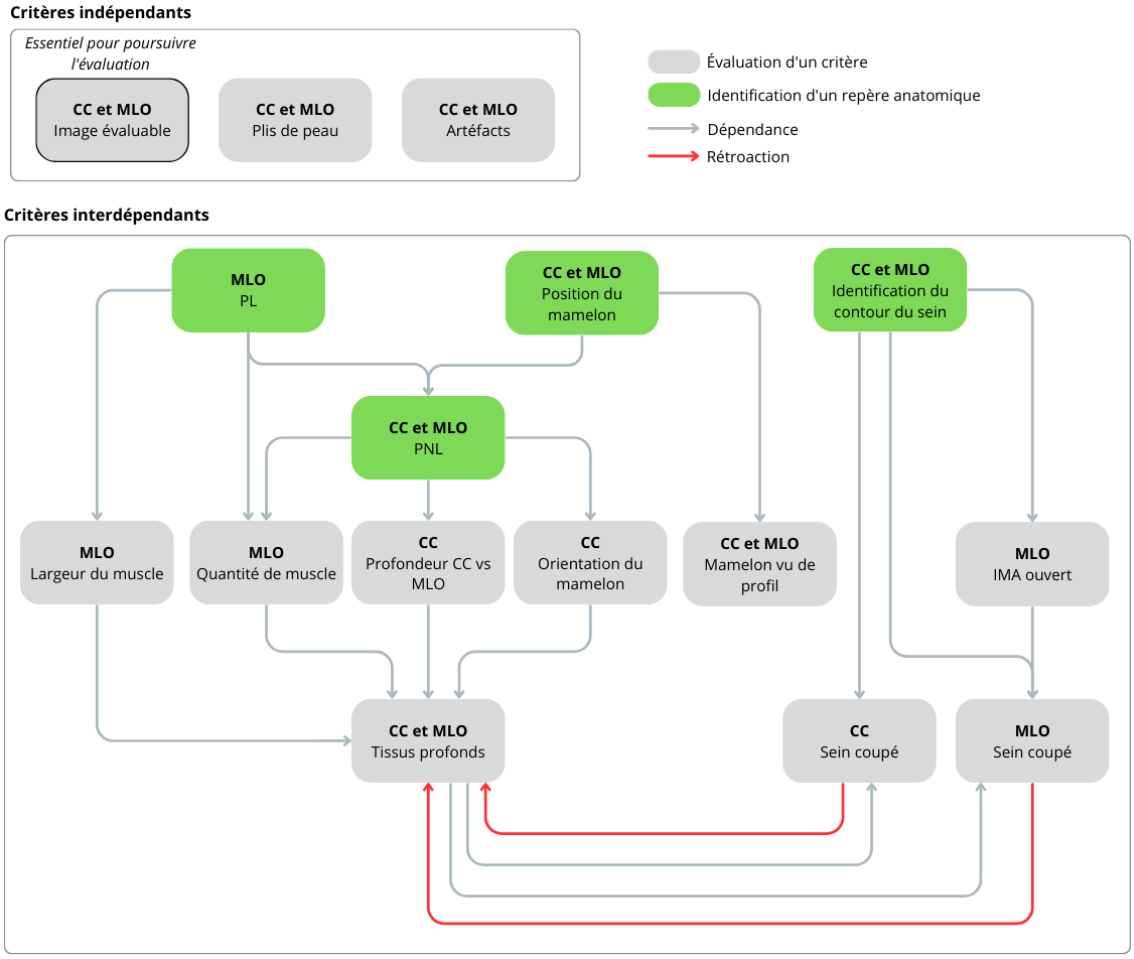


FIGURE 2.2 Graphe de l'interdépendance des critères de positionnement [2]

2.3 Vision par ordinateur en mammographie

2.3.1 Jeux de données en mammographie

Afin de comprendre les enjeux de l'apprentissage profond en mammographie, il est important de connaître l'état des données disponibles pour l'entraînement de modèles. En effet, le développement de modèles d'apprentissage profond requiert l'accès à des données pour leur entraînement, leur validation et leur test. L'accès aux données est souvent un enjeu dans le domaine médical puisque les données sont souvent sensibles, elles nécessitent généralement une anonymisation, et le consentement des patients est habituellement requis. De plus, l'annotation des données requiert la participation d'experts, ce qui peut rendre le processus long et onéreux.

Il existe plusieurs bases de données publiques de mammographies qui sont fréquemment utili-

sées en recherche. Parmi celles-ci, on retrouve le *Digital Database for Screening Mammography* (DDSM) et sa version réduite, le *Curated Breast Imaging Subset of* DDSM (CBIS-DDSM). DDSM contient environ 10 000 images annotées selon les masses et les calcifications présentes, alors que CBIS-DDSM contient un sous-ensemble 6 671 images de DDSM [22, 23]. Ce sous-ensemble de données contient les images standardisées et décompressées ainsi que des annotations mises à jour par des technologues expérimentés sur l'état des masses (bénignes ou malignes) et des boîtes englobantes des régions d'intérêt. Un autre jeu de données fréquemment utilisé est celui du *Mammographic Image Analysis Society* (MIAS) [24]. Il s'agit d'un petit jeu de données comprenant 322 images annotées selon les régions anormales du sein. Similairement, la base de données INbreast comprend 410 images avec des annotations concernant tous les types de lésions présentes [25]. Une autre base de données importante est VinDr-Mammo, qui figure parmi les plus grandes bases de données en mammographie. Celle-ci contient 20 000 images et inclut plusieurs annotations, dont les régions normales et anormales ainsi que la densité mammaire [26]. La disponibilité de ces jeux de données a permis à de nombreuses études d'aborder la détection automatique du cancer du sein. Toutefois, parmi tous les jeux de données publics de mammographies, aucun ne contient d'annotations relatives au positionnement mammaire.

2.3.2 Détection du cancer du sein

La vaste majorité des travaux existants en mammographie portent sur la détection du cancer du sein, soit par segmentation des lésions, soit par classification des images. Dans les années 2010, plusieurs travaux ont porté sur la détection du cancer par segmentation des masses [27, 28]. D'autres chercheurs ont opté pour des méthodes par extraction de vecteurs de caractéristiques des images, suivie d'une classification [29–32].

L'apparition des travaux utilisant des réseaux de neurones convolutifs (*Convolutional Neural Networks*) (CNNs) s'est faite progressivement, et il s'agit encore aujourd'hui de la méthode la plus répandue. Selon une étude de 2023, environ 57 % des travaux portant sur les mammographies et l'apprentissage profond étaient consacrés à l'identification et à la classification du cancer du sein [12]. En 2015, Ertosun et al. [33] développent une méthode de recherche et de localisation des masses dans les mammographies en utilisant des CNNs. Pour la classification des images (avec ou sans masse), les auteurs testent trois CNNs établis, soient AlexNet, VGG16 et GoogleNet. Un second CNN sert de moteur de localisation (*localization engine*) afin de localiser les masses dans l'image avec une approche basée sur les probabilités régionales (*regional probabilistic approach*). Par la suite, de nombreux travaux se sont intéressés à l'utilisation de CNNs pour la classification de régions d'intérêt (ROIs) en mammographie

(normales, bénignes, malignes) [34–37]. D'autres utilisent une approche combinant l'image globale avec des régions locales afin de classer les images [38, 39].

Suzuki et al. [40] et Guan et al. [41] utilisent aussi des CNNs pour la classification des mammographies, mais en introduisant le transfert de connaissances. Leurs modèles sont donc d'abord entraînés pour un autre type de classification avant d'appliquer un réglage fin (*fine-tuning*) sur la tâche principale. En effet, l'entraînement de modèles d'apprentissage profond peut être difficile lorsque le nombre d'images disponibles pour l'entraînement est limité, comme c'est souvent le cas dans le domaine médical. Le transfert de connaissances offre alors une option intéressante, puisqu'il permet de pré-entraîner le réseau de neurones sur une grande quantité d'images. Puisque les premières couches des CNN apprennent des caractéristiques plus générales, le pré-entraînement sur un grand jeu de données d'images naturelles aide le réseau à comprendre des motifs de base qui s'appliquent à tous types de tâches et d'images. Suzuki et al. [40] utilisent un CNN personnalisé de huit couches pré-entraîné sur le jeu de données ImageNet, et effectuent la classification de ROIs de mammographies. Guan et al. [41] utilisent plutôt un VGG-16 avec un pré-entraînement sur ImageNet, et effectuent aussi la classification de ROIs. Ces deux travaux rapportent d'excellentes performances pour la classification des ROIs de mammographies et démontrent le potentiel du transfert de connaissances pour l'entraînement de modèles pour la détection du cancer du sein.

Avec l'apparition de modèles de détection automatique d'objets, plusieurs travaux se sont penchés sur leur utilisation pour la détection de lésions. C'est le cas d'Alantari et al. [42] et de Hamed et al. [43] qui utilisent le modèle YOLO (*You Only Look Once*) pour d'abord détecter les masses ou les lésions, puis un CNN pour la classification des ROIs. Finalement, certains travaux présentent des méthodes originales afin d'améliorer les performances des CNNs simples pour la classification des mammographies. Ces méthodes incluent la combinaison des deux vues d'un examen, l'exploitation des différences entre les mammographies séquentielles, la combinaison de différents ensembles de caractéristiques extraites des images et l'utilisation de modèles transformeurs visuels (ViTs) [44–47].

2.3.3 Classification de la densité

La densité mammaire correspond à la proportion de tissu dense (tissu glandulaire et tissu fibreux) par rapport au tissu non dense (tissu graisseux) dans le sein. Il s'agit d'une caractéristique importante du sein, car un sein plus dense est plus à risque de développer un cancer [48]. Pour catégoriser la densité mammaire, la grande majorité des cliniques et des centres de recherche a recours à la classification du *Breast Imaging Reporting and Data System* (BI-RADS). Ce système permet de classer la densité mammaire en quatre catégories,

allant d'un sein surtout composé de tissu graisseux à un sein surtout composé de tissu dense.

La classification de la densité mammaire fait l'objet de plusieurs travaux dans la littérature. Plusieurs approches reposent sur la segmentation de la glande mammaire pour ensuite effectuer la classification [49, 50]. D'autres travaux portent sur la classification de la densité mammaire directement par des modèles d'apprentissage profond, soit par des CNN personnalisés ou des modèles établis comme des ResNets [51–53].

2.3.4 Identification des repères anatomiques

Une autre grande proportion de la recherche en mammographie est dédiée à l'identification des repères anatomiques du sein, plus particulièrement du mamelon et du muscle pectoral. Les travaux de Casti et al. [54] portent sur la détection du mamelon à l'aide de méthodes classiques de traitement d'images incluant une analyse du champ vectoriel de gradients et l'application de conditions locales basées sur la forme. Moran et al. [55] et Zeng et al. [56] proposent des méthodes pour la détection du mamelon et la segmentation du muscle pectoral en utilisant les informations de texture et de forme du sein. De nombreux travaux se consacrent à la segmentation du muscle pectoral par différentes méthodes basées sur la morphologie, les contraintes géométriques, la topographie ou l'intensité des pixels [57–60].

2.4 Évaluation automatique du positionnement mammaire

Contrairement aux tâches de détection du cancer, de classification de la densité et d'identification des repères anatomiques, l'évaluation automatique du positionnement mammaire fait l'objet de moins de travaux dans la littérature. D'abord, des approches par traitement d'images classique seront abordées, puis les méthodes par apprentissage profond seront présentées. Il est à noter que les critères de positionnement traités dans les différents travaux de la littérature proviennent de sources variées et ne correspondent pas forcément à ceux utilisés dans le contexte québécois.

2.4.1 Méthodes par traitement d'images classique

Les méthodes d'évaluation du positionnement par traitement d'images classique se décomposent en deux étapes : d'abord, l'identification de repères anatomiques dans le sein, puis l'évaluation de certains critères en fonction de ces repères.

En 2004, Zhou et al. [61] développent un algorithme pour identifier la position du mamelon et pour déterminer l'état de celui-ci : visible ou invisible. L'algorithme comprend plusieurs

étapes : d'abord, une détection du contour du sein est effectuée à l'aide d'une approche par suivi des contours basée sur les gradients (*gradient-based boundary tracking algorithm*). Une analyse des niveaux de gris près du contour et une analyse de convergence géométrique sont aussi utilisées pour limiter la recherche du mamelon dans une ROI. La détection du mamelon est ensuite basée sur les informations de niveaux de gris ainsi que sur les informations géométriques de la ROI. La classification du mamelon comme étant visible ou invisible s'apparente à l'évaluation du critère Mamelon vu de profil ; dans les deux cas, il s'agit de déterminer si la visualisation du mamelon est affectée par une superposition des tissus mammaires.

Ensuite, Bulow et al. [62] présentent des méthodes pour détecter automatiquement le muscle pectoral, le mamelon et l'IMA. Une analyse de la mammographie par lignes de l'image et à l'aide de dérivées est effectuée afin d'identifier les points potentiels du contour du muscle pectoral. Par la suite, l'identification du mamelon et de l'IMA se fait par l'analyse de la courbure du contour du sein. Une fois les repères anatomiques identifiés, des critères de positionnement s'apparentant à Mamelon vu de profil, IMA ouvert, Largeur du muscle, Quantité de muscle et Profondeur CC vs MLO sont évalués à l'aide de contraintes géométriques.

Zeng et al. [63] s'intéressent à la détection et à la reconnaissance de l'IMA. D'abord, la détection de la région de l'IMA est effectuée par une analyse de la courbure du sein. Ensuite, des caractéristiques sont extraites de cette région, puis un classificateur multi-décisionel est utilisé pour la reconnaissance de l'IMA. L'algorithme permet donc de déterminer si l'IMA est présent ou non dans l'image.

Les méthodes par traitement d'images classique sont rapides et souvent simples à implémenter, mais leur utilisation se limite à l'étude des critères directement évaluables à partir des repères anatomiques. De plus, les méthodes de détection du muscle pectoral, du mamelon et de l'IMA employées sont très sensibles à la qualité des images et aux morphologies des patient.e.s.

2.4.2 Méthodes par apprentissage profond

Dans les dernières années, de plus en plus de travaux introduisent l'apprentissage profond pour pallier le problème du positionnement mammaire.

Certains travaux emploient encore une méthode en deux temps : d'abord la détection des repères anatomiques, puis l'évaluation des critères. C'est le cas des travaux de Gupta et al. [64], qui ciblent l'évaluation de deux critères de positionnement : Profondeur CC vs MLO et Quantité adéquate de muscle pectoral sur l'image. L'évaluation des critères repose sur la détection du muscle pectoral et du mamelon, puis sur les relations géométriques entre

ces repères anatomiques. Une transformée de Hough circulaire est utilisée afin d'identifier le marqueur posé sur le mamelon. Pour la détection du muscle pectoral, les auteurs utilisent un réseau de neurones ; la base de leur réseau est Inception V3, avec la dernière couche remplacée par quatre sorties correspondant aux deux paires de coordonnées *(x, y)* qui définissent la ligne du muscle pectoral. 442 images ont été utilisées pour l'entraînement de ce modèle, provenant d'un jeu de données privé. L'évaluation des critères de positionnement se fait ensuite de manière géométrique, en traçant la PNL en MLO et en CC : les auteurs déterminent si l'intersection entre la PL et la PNL se trouve dans l'image et si la longueur de la PNL en CC est plus courte de plus de 1 cm que celle en MLO.

Similairement, les travaux de Tanyel et al. [65] portent sur l'évaluation du critère Quantité adéquate de muscle pectoral. Leur méthode comprend d'abord la détection de la PL et de la position du mamelon à l'aide d'un modèle d'apprentissage profond, puis la classification des images selon le critère de positionnement. Les données utilisées dans le cadre de ces travaux correspondent à un sous-ensemble de 5000 images du jeu de données VinDr-Mammo. Ce sous-ensemble de données a été annoté avec la ligne du muscle pectoral et le mamelon par des technologues experts en mammographie. Un seul modèle est employé pour effectuer conjointement la détection de la PL et du mamelon ; il s'agit d'un CoordAtt U-Net, soit un U-Net classique incluant un module d'attention. Ensuite, l'étape de classification est effectuée simplement à l'aide de contraintes géométriques sur les repères anatomiques : si l'intersection entre la PL et la PNL se trouve dans l'image, le critère est respecté. Il s'agit d'une méthode intéressante, car la détection de la PL et du mamelon en amont de l'évaluation des critères donne un sens clinique et de l'explicabilité au modèle de bout en bout. Les auteurs ont aussi rendu publiques les annotations de PL et du mamelon effectuées sur le sous-ensemble de VinDr-Mammo : il s'agit jusqu'à présent du seul jeu de données public incluant des annotations relatives aux repères anatomiques du sein.

D'autres études se penchent plutôt sur l'utilisation de modèles d'apprentissage profond pour évaluer directement de bout en bout certains critères de positionnement, sans passer d'abord par la détection des repères anatomiques. Ces méthodes permettent l'évaluation de tous les critères et ne requièrent que des annotations sous forme de classification au niveau de l'image pour l'entraînement.

Dans leurs travaux effectués au début des années 2020, Zeng et al. [66–68] abordent la reconnaissance des plis de peau, l'évaluation du critère Mamelon vu de profil et la classification des mammographies selon la visibilité du tissu rétroglandulaire. Pour détecter les plis de peau dans des ROIs de mammographies, les auteurs entraînent trois CNNs différents : Densenet-121, VGG-16 et Resnet-101 [66]. Pour ce faire, 46 200 ROIs d'images CC et 28 066

ROIs d'images MLO sont utilisées pour l'entraînement, la validation et le test des modèles. Le Densenet-121 est le modèle qui obtient les meilleures performances de classification en termes d'aire sous la courbe de la fonction d'efficacité du récepteur (AUC ROC), de spécificité, de sensibilité et d'exactitude. Une méthodologie similaire est employée pour évaluer le critère Mamelon vu de profil [67]. Trois catégories sont utilisées pour la classification du critère : mamelon vu de profil, mamelon pas vu de profil et mamelon absent. Parmi les modèles Squeezenet, Alexnet, Densenet-121, Resnet-18, Resnet-101, VGG-16 et VGG-19, les auteurs rapportent que le Densenet-121 est le meilleur choix avec une exactitude de 0,982. Le dernier article de Zeng et al. se penche sur la classification des images selon la qualité de la visualisation des tissus profonds [68]. Plus particulièrement, ils s'intéressent à la présence ou l'absence du tissu rétroglandulaire dans les images. Des régions candidates sont extraites du côté proximal des images, puis sont classées à l'aide de CNNs. 1 014 images de mammographies ont été utilisées, desquelles ont été extraites 6 702 régions négatives et 5 274 régions positives. Les CNNs testés sont Densenet-121, Resnet-101, VGG-19, Alexnet et Squeezenet. Encore une fois, le Densenet-121 se démarque des autres modèles avec une exactitude de 0,893. Les résultats de ces travaux montrent le potentiel de l'évaluation automatique du positionnement en mammographie. Par contre, tous les modèles présentés prennent en entrée des ROIs d'images et non des mammographies complètes, ce qui limite leur utilisation dans un contexte clinique.

Ensuite, Watanabe et al. [69] se sont penchés particulièrement sur les régions de l'IMA et du mamelon. Pour réaliser cette étude, 1 631 images tirées du jeu de données CBIS-DDSM ont été annotées par deux technologues experts qui devaient se consulter pour parvenir à un consensus. L'évaluation des critères de positionnement associés à l'IMA et au mamelon a été définie selon une échelle *Poor, Average* et *Excellent*. La méthode proposée s'effectue en deux temps : d'abord, la détection des régions d'intérêt est effectuée à l'aide de méthodes de traitement d'images classiques, puis la classification selon les trois catégories est réalisée à l'aide de CNNs. Pour la détection de l'IMA, une approche classique par analyse des pixels de régions d'intérêt le long du côté proximal de l'image est employée. La région du mamelon est quant à elle déterminée en se basant sur l'identification du point maximal du contour du sein. Par la suite, ces régions d'intérêt sont données en entrée à des modèles de classification : VGG-16, Inception V3, Xception, Inception-Resnet-V2 et EfficientNet-B0. Les auteurs rapportent que VGG-16 est le modèle qui performe le mieux en termes de F1-score sur les deux tâches de classification.

Dans des travaux de 2023, Hejduk et al. [70] utilisent des modèles d'apprentissage profond variés pour la classification des images selon le positionnement ainsi que pour la détection de repères anatomiques. Les critères étudiés concernent la visualisation du parenchyme en CC

et en MLO, la présence de l'IMA en MLO, la présence du muscle pectoral en CC ainsi que le mamelon vu de profil en CC et en MLO. Les images utilisées proviennent d'une collecte de données privée, totalisant 11 733 images pour l'entraînement des modèles. Les modèles développés sont des CNNs comprenant 13 couches convolutives, et des valeurs d'exactitude allant de 87 à 97 % ont été obtenues pour l'évaluation des critères.

Finalement, le travail de Brahim et al. [5] représente l'étude la plus complète à ce jour sur l'évaluation de différents critères de positionnement mammaire. Les auteurs utilisent deux CNNs personnalisés qui contiennent cinq couches convolutives pour classer les images selon certains critères de positionnement. Les critères abordés dans cet ouvrage ne correspondent pas directement à ceux des références reconnues, mais s'y apparentent fortement. En MLO, quatre critères de positionnement sont inclus : Quantité de muscle, Largeur du muscle, Mamelon vu de profil et Recouvrement de tout le tissu mammaire pertinent. Ce dernier s'apparente à un mélange entre Sein coupé et Tissus profonds, d'après sa définition. En CC, les critères traités sont Mamelon vu de profil et Recouvrement de tout le tissu mammaire pertinent. Le jeu de données utilisé comprend 1 556 images provenant de centres de dépistage en Allemagne. Pour la plupart des critères, on se trouve en situation de déséquilibre des classes, c'est-à-dire que le nombre d'images de la classe positive n'est pas égal au nombre d'images de la classe négative. Ici, il s'agit de la quantité d'images pour lesquelles les critères échouent qui est limitante. Les auteurs utilisent donc des augmentations pour générer des images inadéquates à partir d'images adéquates et ainsi rééquilibrer artificiellement leur jeu de données. Le test de leurs modèles se fait aussi sur un ensemble de données équilibré, et les valeurs d'exactitude et de F1-score obtenues sont au-dessus de 94% pour tous les critères.

Ces travaux montrent le potentiel de l'apprentissage profond pour l'évaluation automatique du positionnement mammaire. Par contre, les critères employés varient d'une étude à l'autre et manquent de définitions claires. D'ailleurs, aucun ouvrage ne considère les critères exacts utilisés dans le contexte clinique québécois et définis par Rouette et al. [16]. D'autre part, les méthodes mentionnées proposent un modèle par critère ; l'évaluation complète de la qualité du positionnement mammaire requiert donc l'entraînement de multiples modèles.

Outils commercialisés

Certains outils d'évaluation automatique sont déjà commercialisés dans différentes régions du monde, notamment aux États-Unis, en Suisse et au Canada [71–73]. Or, les informations concernant le développement de ces méthodes et leur validation ne sont pas disponibles dans des publications scientifiques, ce qui rend difficile leur évaluation.

2.5 Apprentissage multi-tâches

L'apprentissage multi-tâches est une technique d'apprentissage profond qui consiste à entraîner un réseau de neurones à effectuer plusieurs tâches simultanément. Il a été montré que les modèles basés sur ce type d'apprentissage peuvent avoir une meilleure généralisation et permettent d'obtenir de meilleures performances pour les tâches sur lesquelles ils sont entraînés [3]. Les tâches qui se prêtent bien à ce type de modèles sont celles qui possèdent des corrélations, des dépendances ou des liens inhérents les unes avec les autres.

Ruder [3] présente une vue d'ensemble de l'apprentissage multi-tâches dans les réseaux de neurones profonds. Deux grandes méthodes d'apprentissage multi-tâches sont adressées dans cet ouvrage : le *Hard parameter sharing* (HPS) et le *Soft Parameter Sharing* (SPS). Dans un modèle HPS, toutes les tâches partagent un nombre de couches cachées du réseau, alors que d'autres couches sont spécifiques à chaque tâche. Au contraire, dans un modèle SPS, chaque tâche possède ses propres couches cachées avec ses propres paramètres, mais le modèle est encouragé à établir des paramètres similaires entre les tâches par le biais d'une régularisation. Par exemple, la distance l_2 est souvent utilisée comme régularisation dans ces types de modèles. Des schémas de ces deux types d'architectures sont présentés à la Figure 2.3.

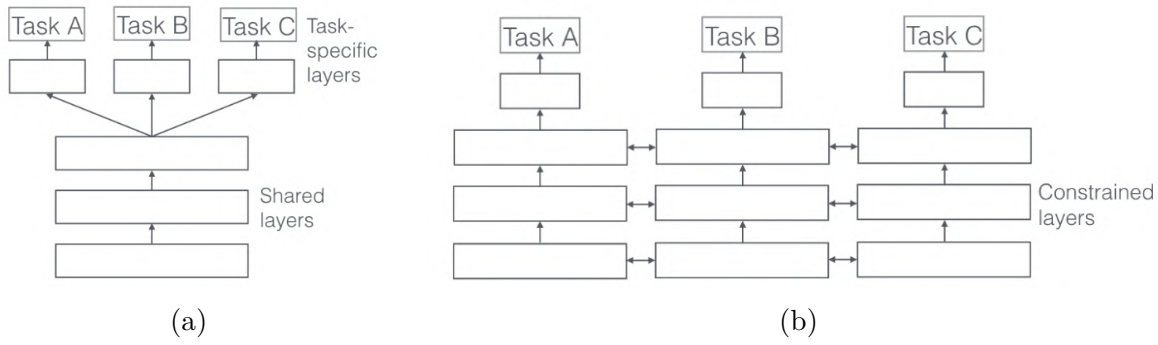


FIGURE 2.3 Deux grandes méthodes d'apprentissage multi-tâches. (a) HPS et (b) SPS [3]

À partir de ces architectures de base, une multitude de modèles peut être construite, et ce pour des tâches de domaines variés. Par exemple, Liu et al. [74] proposent le *Multi-Task Attention Network* (MTAN). Il s'agit d'un modèle multi-tâches qui comprend un réseau partagé entre toutes les tâches, et k réseaux d'attention spécifiques à chacune d'entre elles. Chaque réseau d'attention applique un masque d'attention à des parties du réseau partagé afin de favoriser l'apprentissage des caractéristiques spécifiques à chaque tâche. Deux versions de ce modèle existent : une permettant d'effectuer des tâches de segmentation, comprenant un encodeur et des décodeurs, et une autre permettant d'effectuer de la classification d'images, qui comprend

un encodeur et des têtes de classification. Dans leur étude, les auteurs testent ce modèle sur plusieurs combinaisons de tâches, dont la segmentation de plusieurs classes sémantiques sur des images de rues et la classification d'images selon 10 tâches dans des domaines variés.

2.5.1 Applications au domaine médical

Zhao et al. [75] propose une revue des travaux portant sur l'apprentissage multi-tâches dans le domaine médical. Parmi les régions anatomiques les plus étudiées à l'aide de ce type d'apprentissage, on retrouve le cerveau, les yeux, le thorax et le cœur. En imagerie cérébrale, la majorité des travaux portent sur la segmentation de diverses régions d'intérêt ou la segmentation d'une région combinée avec une tâche de régression, de classification ou de reconstruction. De nombreux travaux ont aussi été effectués sur des images de rétine, entre autres grâce à leur facilité d'acquisition. Les tâches les plus couramment effectuées en apprentissage multi-tâches sont la classification de pathologies, ainsi que la segmentation de vaisseaux et de lésions. L'architecture parallèle, correspondant à une méthode HPS, est la plus utilisée dans les travaux du domaine médical.

Les travaux de Kim et al. [76] portent sur le développement d'un modèle multi-tâches spécialement conçu pour les images du domaine médical. La méthode développée supporte aussi bien des tâches au niveau des pixels qu'au niveau de l'image et inclut deux modules d'attention afin de favoriser les interactions inter-tâches : un *cross-task attention encoder* (CTAE) et un *cross-task attention bottleneck* (CTAB). L'encodeur partagé par les deux tâches est un ResNet-50. Le modèle proposé est versatile et s'applique à une variété de modalités d'imagerie et de combinaisons de tâches. Dans cette étude, quatre combinaisons de tâches sont testées. La première est la combinaison d'une tâche de segmentation et d'une tâche de planification de dose sur des jeux de données de prostate et d'imagerie cérébrale. Ensuite, une tâche de segmentation et une tâche de diagnostic de lésions sont effectuées sur des images dermatologiques. Finalement, deux tâches au niveau de l'image sont combinées, soient le diagnostic et l'évaluation du niveau de sévérité de la COVID-19 dans des images tomodensitométriques de thorax.

À ce jour, seulement quelques travaux appliquent des méthodes multi-tâches dans le domaine de l'imagerie mammaire, dont certains en imagerie ultrasons et d'autres en mammographie. En 2021, Zhang et al. [77] développent un modèle multi-tâches pour la segmentation et la classification du cancer du sein dans les images d'ultrasons. L'architecture utilisée est celle d'un encodeur-décodeur avec une branche de classification ajoutée. Un DenseNet-121 est utilisé comme squelette (*backbone*) de l'encodeur. Le modèle inclut aussi un *Attention-Gated Unit* pour amplifier l'influence des régions cancéreuses et supprimer l'information non

pertinente. Les résultats de classification obtenus atteignent 94 % d'exactitude, dépassant les performances d'un modèle entraîné en tâche unique. De manière similaire, Chowdary et al. [78] présentent un modèle multi-tâches pour la segmentation et la classification des tumeurs dans les images d'ultrasons. Le modèle proposé est un U-Net résiduel pour la segmentation, auquel est ajoutée une branche pour la classification de la tumeur (bénigne/maligne). Selon leurs expérimentations, cette méthode multi-tâches permet d'obtenir de meilleurs résultats de classification qu'avec des méthodes à tâches uniques.

2.5.2 Applications aux mammographies

La vaste majorité des travaux utilisant l'apprentissage multi-tâches en mammographie s'intéresse aux tâches de détection, de segmentation et de classification des masses ou lésions dans les images [79–83]. Sainz de Cea et al. [79] entraînent un modèle pour la détection des masses et la classification des images (normale/anormale) en utilisant un encodeur partagé, puis une branche de classification ainsi qu'une branche de détection comprenant un *Feature Pyramid Network* (FPN) et un réseau pour la détection de boîtes englobantes. En 2020, Shen et al. [80] introduisent un modèle qui détecte les régions suspectes dans les mammographies, effectue la segmentation des lésions possibles et classe ces lésions. Leur méthode se déroule en deux phases : d'abord, la détection des régions possibles se fait à l'aide d'un encodeur et d'un décodeur, et retourne en sortie une carte de chaleur (*heatmap*) correspondant aux différentes régions. Ensuite, la classification des régions est effectuée en parallèle à la segmentation des masses à l'aide d'un encodeur partagé et de couches spécifiques à chacune des tâches. Similairement, Gao et al. [81] utilisent une architecture HPS comprenant un encodeur commun, un *Region Proposal Network* (RPN) et trois branches spécifiques aux tâches de détection, de classification et de segmentation des masses.

Ensuite, Hou et al. [82] proposent aussi un modèle multi-tâches pour la détection, la segmentation et la classification des masses dans les mammographies. L'architecture de leur modèle comprend des couches partagées entre toutes les tâches ainsi que des branches spécifiques aux tâches, suivant la forme d'un modèle HPS. Plus précisément, les couches partagées comprennent un CNN pour extraire les caractéristiques des images, un FPN et un RPN. Un module d'attention est aussi intégré dans les couches convolutives afin de permettre au modèle de se concentrer sur les canaux de caractéristiques les plus riches en information. Le réseau se sépare ensuite en deux branches : une pour la détection et la localisation des masses, et une autre pour la segmentation. De manière similaire, Wang et al. [83] introduisent un modèle multi-tâches pour la segmentation et la classification du cancer du sein, mais qui tire profit d'une stratégie de guidage entre les tâches. Le modèle proposé est composé de couches parta-

gées, puis de couches distinctes pour la classification et la segmentation. La particularité de ce modèle est le guidage sémantique et de localisation que la branche de classification donne à la branche de segmentation.

Dans les travaux de Nebbia et al. [84], la tâche principale de classification des mammographies (normales/anormales) est aidée par l'introduction de deux tâches auxiliaires : la classification BI-RADS de la densité mammaire et la classification BI-RADS du niveaux de sévérité du cancer du sein. Le modèle proposé est simple : il s'agit d'une architecture HPS avec un encodeur partagé par toutes les tâches, puis des têtes de classification différentes. Plusieurs encodeurs ont été testés dans cette étude, et les meilleurs résultats sont obtenus avec un DenseNet-121 pré-entraîné sur ImageNet. Zhou et al. [85] utilisent aussi une tâche auxiliaire en complément de la tâche de diagnostic du cancer du sein : le changement de la densité mammaire dans le temps. Comme mentionné plus tôt, la densité est un important facteur de risque de développement d'un cancer du sein. Le modèle développé par les auteurs prend en entrée deux mammographies séquentielles, puis un ViT est utilisé pour extraire les caractéristiques de ces images. Ensuite, un réseau siamois est utilisé pour comparer les vecteurs de caractéristiques des deux images et effectuer la prédiction de changement de densité, et une autre branche contenant un module d'attention inter-vues est élaborée pour la classification de la mammographie.

Finalement, dans un article de 2024, Nguyen et al. [86] développent un modèle multi-tâches simple pour la classification BI-RADS de la densité mammaire et pour la classification de la sévérité du cancer. L'architecture employée est basée sur le HPS, avec un encodeur commun aux deux tâches, suivi de deux couches connectées spécifiques à chacune des tâches. Plusieurs versions du ResNet sont testées dans ce travail et le ResNetXt-101 est celui qui obtient les meilleures performances avec des valeurs d'exactitude de 95 à 98 % pour les deux tâches de classification.

Cette revue permet de constater que la majorité des modèles multi-tâches sont conçus pour effectuer une tâche au niveau des pixels, souvent la segmentation, en combinaison avec une tâche au niveau de l'image, par exemple, la classification binaire de l'image. Ils suivent pour la plupart la même architecture de base, soit une structure HPS avec un encodeur commun à toutes les tâches, puis des couches spécifiques à chacune. De plus, la quasi-totalité des travaux effectués en mammographie cible comme tâche principale la classification des images par rapport au diagnostic du cancer du sein. Peu de travaux s'intéressent à l'évaluation du positionnement mammaire, et aucun n'a exploité l'interdépendance entre les critères de positionnement dans le cadre de leur évaluation automatique.

2.6 Simulation de réduction de dose

Alors que l'acquisition d'images radiographiques de grande qualité est nécessaire pour garantir une analyse et un diagnostic adéquats lors du dépistage du cancer du sein, l'exposition à de fortes doses de rayons X n'est cependant pas souhaitable pour les patients. Comme mentionné dans le chapitre 1, certains fabricants d'appareils de radiographie offrent la possibilité d'acquérir des images *pre-shot* à dose de rayons X réduite. L'évaluation du positionnement mammaire sur ces images, avant l'irradiation plus importante des patientes, pourrait permettre une réduction de la dose totale de rayons X reçue lors d'un examen nécessitant la reprise d'images. Or, les images *pre-shot* sont extrêmement difficiles à obtenir puisqu'elles appartiennent aux fabricants des appareils. Plusieurs travaux en vision par ordinateur se penchent donc sur des approches de simulation de réduction de dose et sur l'évaluation de l'impact d'une telle réduction sur les performances de modèles d'intelligence artificielle (IA).

De nombreux auteurs modélisent le bruit associé à une réduction de dose comme un modèle de Poisson ou de mélange Poisson-Gaussien [87–90]. Le bruit gaussien est celui associé aux composantes électroniques, tandis que le bruit de Poisson est plutôt lié à la nature des photons. Lors d'une simulation de réduction de dose, un bruit de Poisson signal-dépendant est ajouté à l'image. Ainsi, le bruit est de plus grande amplitude aux endroits où le signal est le plus fort et de plus faible amplitude où le signal est plus faible. Ceci reproduit le comportement du bruit dans une radiographie à faible dose. Dans la plupart des travaux, des fantômes d'imagerie sont aussi utilisés afin de caractériser précisément le bruit.

Dans un article de 2016, Borges et al. [4] présentent une méthode pour simuler une réduction de dose en mammographie à l'aide de la transformation d'Anscombe. La transformation d'Anscombe est une méthode stabilisatrice de la variance, qui permet de transformer des données suivant une loi de Poisson en données suivant une loi approximativement gaussienne. La méthode proposée par les auteurs consiste à appliquer un bruit gaussien indépendant du signal à une image dans le domaine d'Anscombe, puis à appliquer la transformation d'Anscombe inverse à l'image pour la ramener dans le domaine d'origine. Ainsi, le bruit indépendant du signal ajouté à l'image dans le domaine d'Anscombe est transformé en bruit signal-dépendant, permettant une simulation d'une réduction de dose. Afin de réaliser cette méthode, une image à dose élevée et une image à dose réduite sont nécessaires pour estimer le bruit qui doit être simulé. La figure suivante présente le processus d'addition de bruit gaussien dans le domaine d'Anscombe.

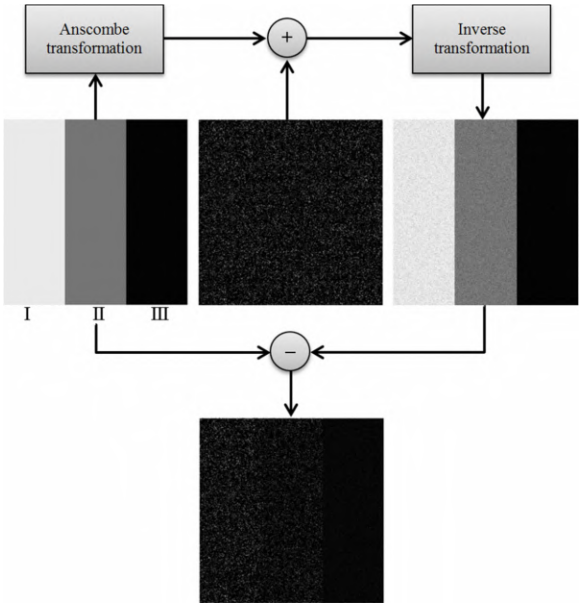


FIGURE 2.4 Processus d'addition de bruit gaussien dans le domaine d'Anscombe [4]

CHAPITRE 3 QUESTIONS ET OBJECTIFS DE RECHERCHE

Les chapitres précédents permettent de dégager les constats suivants :

1. La mammographie est la modalité d'imagerie la plus utilisée pour le dépistage du cancer du sein, et le positionnement mammaire fait partie des enjeux les plus importants lors de cet examen.
2. Aucune base de données en mammographie ne contient d'annotations relatives au positionnement mammaire, et les travaux existants utilisent, pour la plupart, des jeux de données privés.
3. Aucun travail n'aborde les critères de positionnement utilisés dans le contexte clinique québécois.
4. L'interdépendance entre les critères de positionnement n'a encore jamais été exploitée dans les méthodes d'évaluation automatique du positionnement mammaire existantes.
5. L'acquisition d'images à dose réduite est bénéfique pour les patientes, leur permettant une exposition réduite aux rayons X.

On note que, puisque les standards de positionnement diffèrent à travers les régions du monde, il est primordial que les outils d'évaluation automatiques développés soient basés sur les critères de positionnement employés en clinique. Bien que quelques solutions commerciales soient déjà accessibles, il n'existe à ce jour aucun outil d'aide à la décision concernant le positionnement mammaire développé spécifiquement pour le contexte clinique québécois. De plus, aucune documentation ni validation n'est présentée avec les outils disponibles actuellement, ce qui contribue à la réticence de certains professionnels de la santé envers les solutions d'IA. À la lumière des résultats de l'audit de 2017 sur le positionnement mammaire au Québec, nous croyons qu'il est important de se pencher sur le développement d'outils d'évaluation automatiques conçus à l'aide de données provenant du Québec et utilisant les standards de positionnement appliqués par les technologues d'ici en clinique. De plus, une validation rigoureuse doit être mise en place afin de situer les performances des modèles développés dans le contexte de la variabilité existante entre les technologues. Cette problématique permet de faire ressortir l'objectif principal de ce projet, qui est le suivant :

Objectif général : Développer un système d'évaluation automatique du positionnement mammaire pour le contexte clinique québécois.

Par ailleurs, l'interdépendance existante entre certains critères de positionnement constitue une source d'informations qui, à notre connaissance, n'a jamais été exploitée par des modèles

d'apprentissage profond. Il s'agit d'une avenue qui mérite d'être explorée et qui amène la question de recherche suivante :

Q1 : Est-il possible de tirer profit des relations entre les critères à l'aide d'une approche par apprentissage multi-tâches afin d'améliorer l'évaluation automatique du positionnement mammaire ?

Pour répondre à cette question de recherche, les sous-objectifs suivants ont été élaborés :

SO1 : Établir un jeu de données contenant des annotations relatives au positionnement mammaire et pouvant être utilisé dans le cadre de l'entraînement de modèles d'apprentissage profond.

SO2 : Effectuer une analyse de la variabilité inter-technologue quant à l'évaluation des critères de positionnement.

SO3 : Définir, entraîner et valider des modèles d'apprentissage profond pour l'évaluation séparée des critères de positionnement.

SO4 : Définir, entraîner et valider des modèles d'apprentissage multi-tâches pour l'évaluation des critères de positionnement mammaire, dans le but de tirer profit des interdépendances entre ceux-ci.

Finalement, l'utilisation des images *pre-shot* de mammographie pour l'évaluation du positionnement mammaire permettrait de réduire la dose de rayons X donnée aux patientes dans les cas où une mammographie doit être reprise. Il est donc pertinent de se pencher sur l'impact d'une réduction de dose sur les performances des modèles proposés. Un travail complémentaire au projet consiste donc à répondre à la question de recherche suivante :

Q2 : Est-ce possible de vérifier le positionnement mammaire sur des images acquises à plus faible dose de rayons X, avant l'irradiation complète des patientes ?

Le sous-objectif fixé afin de répondre à cette question est le suivant :

SO5 : Évaluer les performances des modèles développés en SO3 et SO4 sur des images simulées en basse dose.

La méthodologie adoptée pour réaliser chacun des objectifs est décrite dans le chapitre suivant, et les résultats obtenus sont ensuite présentés et discutés dans le chapitre 5.

CHAPITRE 4 MÉTHODOLOGIE

Ce chapitre détaille les choix méthodologiques effectués dans le but de rencontrer les objectifs de recherche. Il comprend l'élaboration de la base de données, incluant le processus d'annotation des images et l'analyse de la variabilité inter-technologues, le choix des critères à l'étude, le pré-traitement des images et l'élaboration des modèles baselines, le développement des modèles multi-tâches, le choix des métriques d'évaluation, le protocole d'entraînement et la méthode de simulation de réduction de dose.

4.1 Création du jeu de données Mammo-Pos

La création du jeu de données utilisé dans ce travail a été effectuée en collaboration avec les deux autres étudiants à la maîtrise participant au projet. Un ensemble d'images a été mis à notre disposition par notre partenaire industriel, IMD Research. Cet ensemble comprend des images de mammographies acquises dans diverses cliniques du Québec en 2017 dans le cadre d'un audit de performance initié par l'OTIMROEPMQ. Tous les examens comprennent les quatre images standards d'une mammographie (deux vues par sein), sauf exceptions. Les données ne comprennent pas d'images de patients masculins, de patientes ayant des implants mammaires ni de patientes ayant subi une double mastectomie.

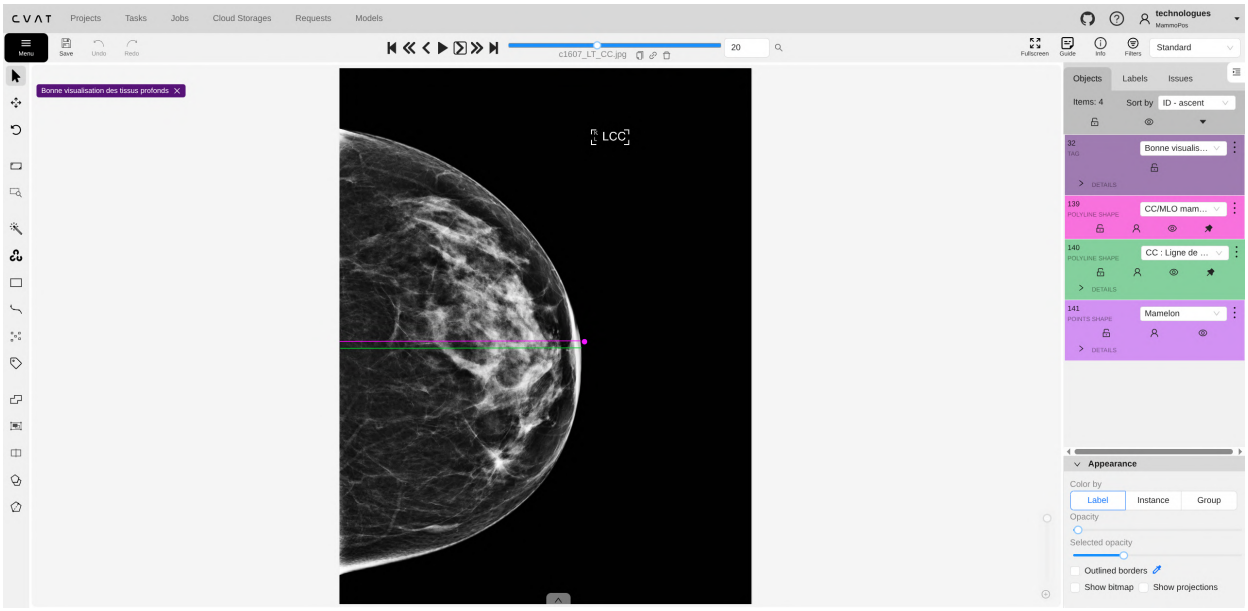
Dans le cadre de l'audit, les images ont été annotées selon les critères de positionnement proposés par Rouette et al. [16]. Toutefois, les technologues chargés de l'évaluation disposaient d'examens antérieurs des patientes et d'informations sur les antécédents chirurgicaux et sur la morphologie des patientes. Le jeu de données mis à notre disposition ne contient pas ces informations, il contient uniquement les images de l'examen. Dans un contexte d'entraînement supervisé de modèles d'apprentissage profond, seules les données images dont nous disposons seront fournies en entrée aux modèles ; c'est pourquoi une phase d'annotations supplémentaire a été mise en place pour refléter l'appréciation par les technologues du positionnement mammaire en présence uniquement des images qui composent l'examen. Le jeu de données utilisé dans le cadre de ce mémoire est celui issu de cette seconde phase d'annotation, appelé Mammo-Pos.

Processus d'annotation

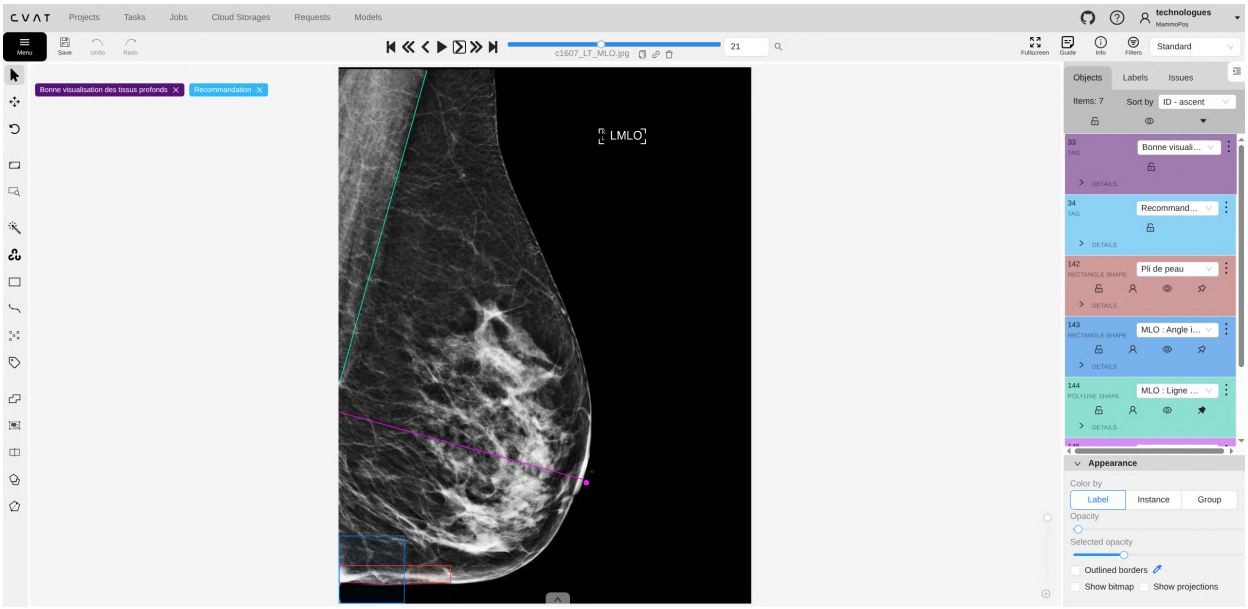
Trois technologues expertes faisant partie de l'OTIMROEPMQ ont été recrutées pour effectuer l'annotation des images. Celles-ci ont toutes plus de 10 ans d'expérience et enseignent

ou donnent des formations en mammographie régulièrement. Un ensemble de 1 215 examens a été annoté par les technologues, totalisant 4821 images, dont 2387 pour la vue CC et 2386 pour la vue MLO. La majorité des images a été annotée par une seule technologue à la fois (ensemble A), et un sous-ensemble de 250 examens a été annoté communément par les trois technologues (ensemble B).

Les annotations ont été effectuées à l'aide de la plateforme en ligne *Computer Vision Annotations Tool* (CVAT). Il s'agit d'un outil d'annotation d'images incluant plusieurs types d'annotations, dont des lignes, des points, des boîtes englobantes, des cases à cocher et des menus déroulants. Avant le début de la phase d'annotation, les technologues ont participé à une demi-journée de formation sur l'utilisation de la plateforme. Par la suite, une phase de calibration a été effectuée sur 10 examens qui n'ont pas été inclus dans le jeu de données final. Il a été demandé aux technologues d'effectuer les annotations en faisant abstraction de toute information morphologique concernant les patientes et en évaluant les images indépendamment les unes des autres. La Figure 4.1 présente des exemples d'images annotées dans les vues CC et MLO sur la plateforme CVAT.



(a)



(b)

FIGURE 4.1 Images lors du processus d'annotation sur la plateforme CVAT. (a) Vue CC avec exemples d'annotations et (b) vue MLO avec exemples d'annotations

Les annotations effectuées par les technologues comprennent l'identification des repères anatomiques du sein (PL, PNL, mamelon et IMA), ainsi que l'évaluation des critères de positionnement. Les tableaux 4.1 et 4.2 présentent les structures annotées et les critères évalués, ainsi que les outils utilisés pour y parvenir. Ceux-ci sont tirés d'un article réalisé par l'équipe

du projet [91] portant sur l'analyse de la variabilité inter-technologues. Les tableaux sont traduits en français pour ce mémoire.

Tableau 4.1 Structures identifiées

Vue	**Structures**	**Type d'annotation**
CC et MLO	Plis de peau	Boîtes englobantes
	Artéfacts	Boîtes englobantes
	Position du mamelon	Point
MLO	PL	Ligne
	PNL	Ligne
	IMA	Boîtes englobantes
CC	PNL	Ligne
	Ligne de profondeur maximale du sein	Ligne

Tableau 4.2 Évaluation des critères

Vue	**Critère**	**Type d'annotation**
CC et MLO	Image évaluable	Étiquette (Vrai/Faux) Zone de texte pour commentaires additionnels
	Plis de peau	Cases à cocher
	Artéfacts	Cases à cocher
	Mamelon vu de profil	Case à cocher
	Sein coupé	Cases à cocher
	Tissus profonds	Étiquette (Vrai/Faux) Cases à cocher
MLO	Quantité de muscle	Case à cocher
	Largeur du muscle	Case à cocher
	IMA ouvert	Menu déroulant
CC	Orientation du mamelon	Case à cocher
	Profondeur CC vs MLO	Lignes (calcul à posteriori)

L'évaluation de certains critères de positionnement a été effectuée directement par les technologues, à savoir les critères de l'Image évaluable, des Plis de peau, des Artéfacts, de l'IMA ouvert, du Mamelon vu de profil et de l'Orientation du mamelon. Le critère des Tissus profonds a été évalué à l'aide de plusieurs raisons possibles à cocher, incluant l'exclusion de tissu glandulaire, l'obstruction du tissu par un pli de peau ou un artéfact, la visualisation inadéquate d'une structure importante du sein, l'absence de l'IMA dans l'image (pour les images MLO) ou l'échec des critères de Largeur du muscle ou de Quantité de muscle (pour les images MLO). Un sous-ensemble de ces raisons a été utilisé pour l'évaluation du critère

du Sein coupé. Le seul critère qui n'a pas été évalué directement par les technologues est celui de la Profondeur CC vs MLO, faute d'un outil de mesure sur la plateforme CVAT. L'évaluation de ce critère a été réalisée par l'équipe à la suite des annotations, en mesurant et en comparant les longueurs des PNLs en CC et en MLO.

Statistiques du jeu de données

Une fois la phase d'annotation terminée, les statistiques concernant la distribution des classes pour les différents critères ont été obtenues. Les Figures 4.2 et 4.3 présentent ces statistiques pour chacun des critères sur les ensembles A et B respectivement. Pour le sous-ensemble B, la distribution des classes présentée est basée sur le vote majoritaire des technologues. Les figures ont été élaborées par un membre de l'équipe dans le cadre de la rédaction de l'article [91] et sont présentées en version traduite dans ce mémoire.

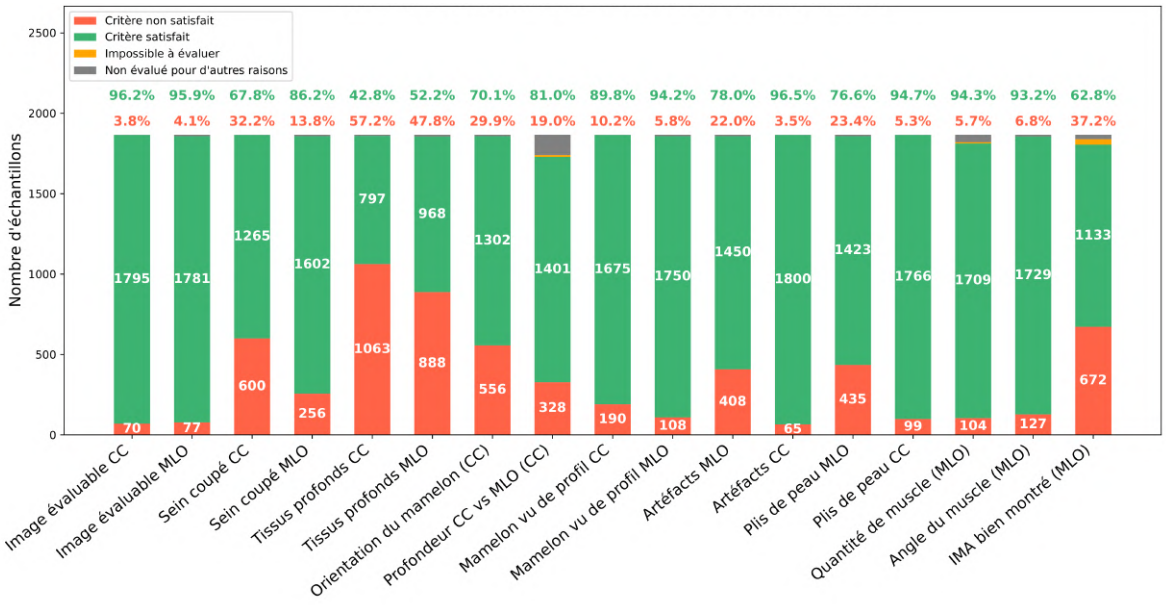


FIGURE 4.2 Distribution des classes pour l'ensemble A

D'abord, on constate que la distribution des classes varie d'un critère à l'autre. Pour le critère des Tissus profonds, les classes sont relativement bien équilibrées, tandis que pour tous les autres critères de positionnement, la classe des critères satisfaits prédomine. Ce déséquilibre des classes est plus ou moins prononcé sur l'ensemble A, allant de 37,2% d'images avec un positionnement inadéquat pour le critère de l'IMA ouvert à 3,5% pour le critère des Artéfacts dans la vue CC. Par ailleurs, on observe aussi que la distribution des classes semble similaire dans l'ensemble B, avec des proportions d'images avec un positionnement adéquat et inadéquat dans le même ordre de grandeur que l'ensemble A.

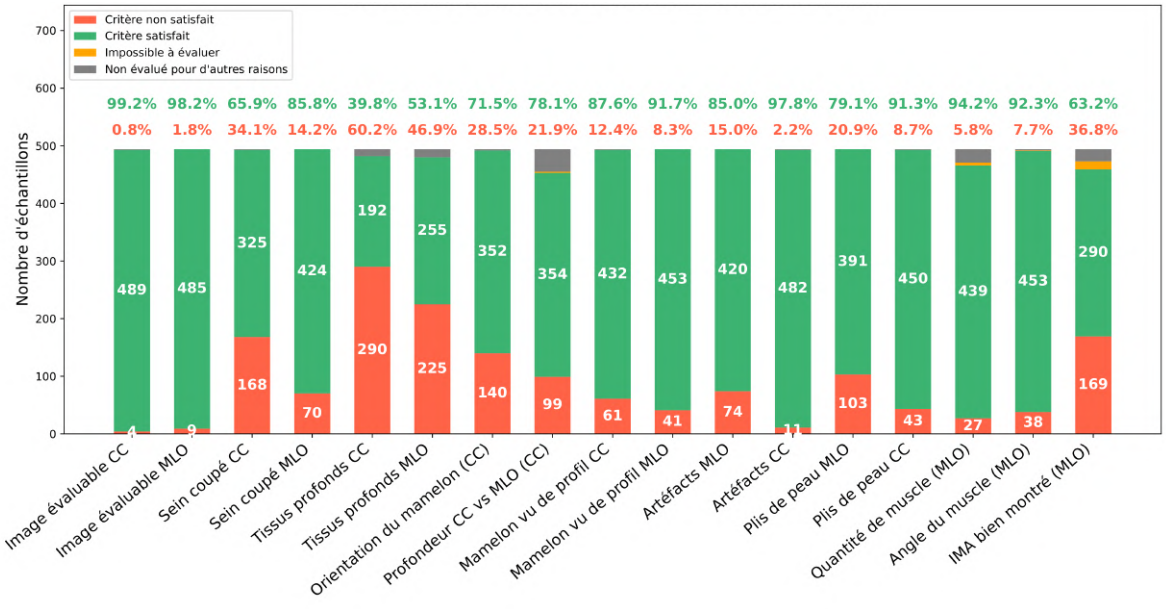


FIGURE 4.3 Distribution des classes pour l'ensemble B

Il est à noter que les trois technologues participant au projet n'ont pas annoté le même nombre d'images de l'ensemble A, certaines ayant plus de temps à accorder au projet et d'autres moins. Les annotations des images de l'ensemble A sont réparties ainsi : la technologue 1 a annoté 33% des images, la technologue 2 a annoté 45% des images et la technologue 3 a annoté 22% des images.

Analyse de la variabilité inter-technologues

La création de l'ensemble B, contenant des examens annotés communément par les trois technologues, sans consultation entre elles, a permis à l'équipe d'analyser la variabilité entre les annotations des technologues. Dans ce mémoire sera présentée l'analyse de la variabilité dans l'évaluation des critères de positionnement. L'analyse de la variabilité pour la détection des repères anatomiques du sein a été effectuée par un autre membre de l'équipe.

Deux métriques ont été utilisées pour analyser la variabilité dans l'évaluation des critères. La première est le taux d'accord entre les technologues. Cette métrique représente la proportion d'images du jeu de données pour lesquelles les trois technologues étaient unanimes dans leur évaluation. Le taux d'accord est calculé à l'aide de la formule suivante :

$$Taux\,d'accord = \frac{N_{unanimes}}{N_{total}} \tag{4.1}$$

où $N_{unanimes}$ correspond au nombre d'images évaluées de façon unanime par les technologues, et N_{total} correspond au nombre total d'images. Cette métrique permet de mettre en évidence quelle proportion des images a suscité un désaccord entre les technologues.

La seconde métrique utilisée est le kappa de Cohen. Il s'agit d'une métrique couramment utilisée lors d'analyses de variabilité inter-opérateurs ou intra-opérateurs [7]. Elle varie entre -1 et 1, 1 indiquant un accord parfait entre les opérateurs, 0 indiquant un accord pouvant être attendu par hasard, et -1 indiquant aucun accord. Alors que le taux d'accord est sensible aux déséquilibres de classes, le kappa est une métrique plus robuste puisqu'il tient compte de l'accord attendu par hasard, offrant ainsi une estimation plus fidèle de la concordance réelle entre les évaluateurs. Pour l'évaluation de la variabilité entre les trois technologues, le kappa de Fleiss a été utilisé. Il s'agit d'une version du kappa de Cohen adaptée pour plus de deux opérateurs. La formule utilisée pour calculer le kappa de Fleiss est la suivante :

$$kappa\ de\ Fleiss = \frac{p_0 - p_e}{1 - p_e} \tag{4.2}$$

avec p_0 l'accord observé, et p_e l'accord attendu par hasard.

4.2 Choix des critères à l'étude

Un sous-ensemble de critères a été choisi pour être étudié dans ce mémoire. Plusieurs raisons ont mené à ces choix, notamment l'interdépendance entre les critères, la présence de corrélations dans les cas de non-respect, les discussions avec les technologues et la distribution des classes dans notre jeu de données.

D'abord, le graphe d'interdépendance des critères présenté dans le chapitre 2 permet de dégager les liens inhérents entre certains critères de positionnement. Les critères du Sein coupé et des Tissus profonds sont intrinsèquement liés, puisqu'ils visent tous deux à déterminer si une partie du tissu mammaire est exclue de l'image. Le non respect de l'un de ces critères entraîne une augmentation du risque que l'autre critère ne soit pas respecté. Ces deux critères, en CC, sont aussi reliés au critère de l'Orientation du mamelon : dans une image où le mamelon est désorienté, il est probable qu'une partie du tissu mammaire soit manquante ou non visible. Le graphe permet aussi de constater la relation entre le critère de l'IMA ouvert et du Sein coupé en MLO. Effectivement, si l'IMA n'est pas visible, cela entraîne automatiquement le non respect du critère du Sein coupé. Dans le cas où l'IMA est visible, mais pas assez ouvert, l'évaluation du critère du Sein coupé n'est pas nécessairement affectée.

Afin d'étudier la corrélation entre les annotations, les matrices de cooccurrences des non-

respects des critères ont été générées pour les critères des deux vues respectives. Ces matrices permettent de visualiser la fréquence à laquelle deux critères ne sont pas satisfaits pour une même image, fournissant des informations sur de possibles liens entre ceux-ci. Les matrices sont présentées aux Figures 4.4 et 4.5.

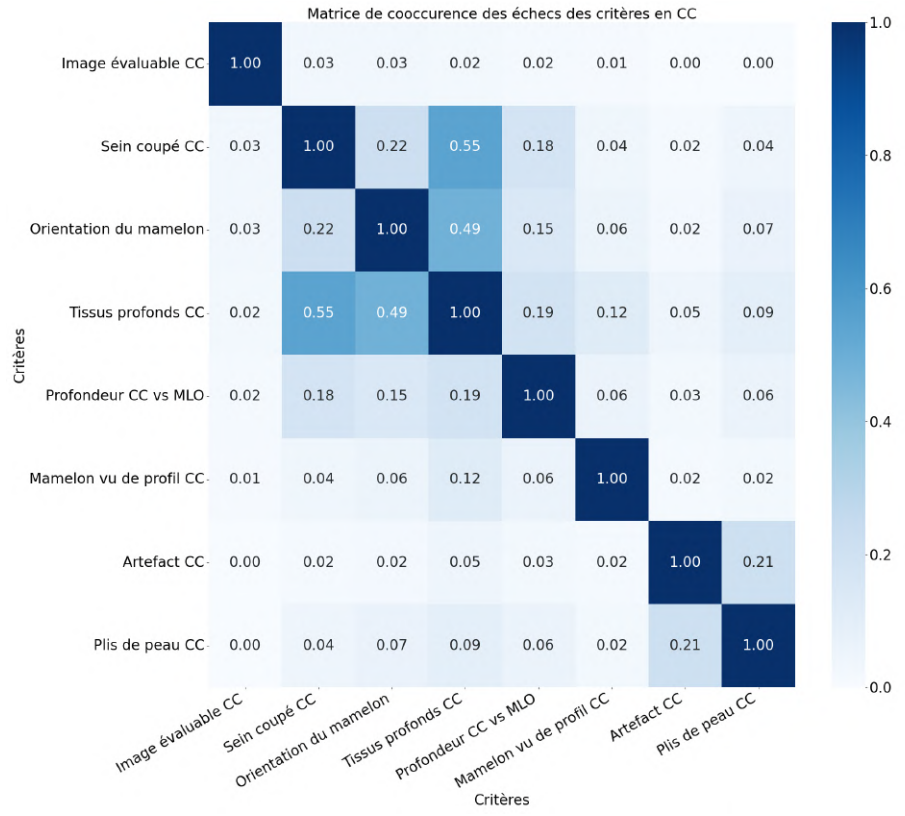


FIGURE 4.4 Matrice de cooccurrences des échecs entre les critères pour la vue CC

Ces matrices permettent de confirmer certains liens établis entre les critères de positionnement. En effet, dans la vue CC, on observe les plus fortes cooccurrences entre les critères du Sein coupé et des Tissus profonds, ainsi qu'entre les critères des Tissus profonds et de l'Orientation du mamelon. L'occurrence entre Orientation du mamelon et Sein coupé est plus faible, mais demeure plus élevée que pour toutes les autres combinaisons de critères. Ceci permet de renforcer l'intuition selon laquelle l'apprentissage multi-tâches pourrait être bénéfique pour l'évaluation de ces critères. Dans la vue MLO, on remarque la plus grande cooccurrence entre les critères de l'IMA ouvert et des Tissus profonds. Cette corrélation n'est cependant pas illustrée par le graphe des interdépendances et n'a pas été mentionnée lors de nos échanges avec les technologues. D'autres combinaisons de critères présentent des cooccurrences élevées pour cette vue, incluant les critères des Artéfacts et des Plis de peau entre eux et avec le critère de l'IMA ouvert. Cependant, aucun lien n'existe entre ces critères par

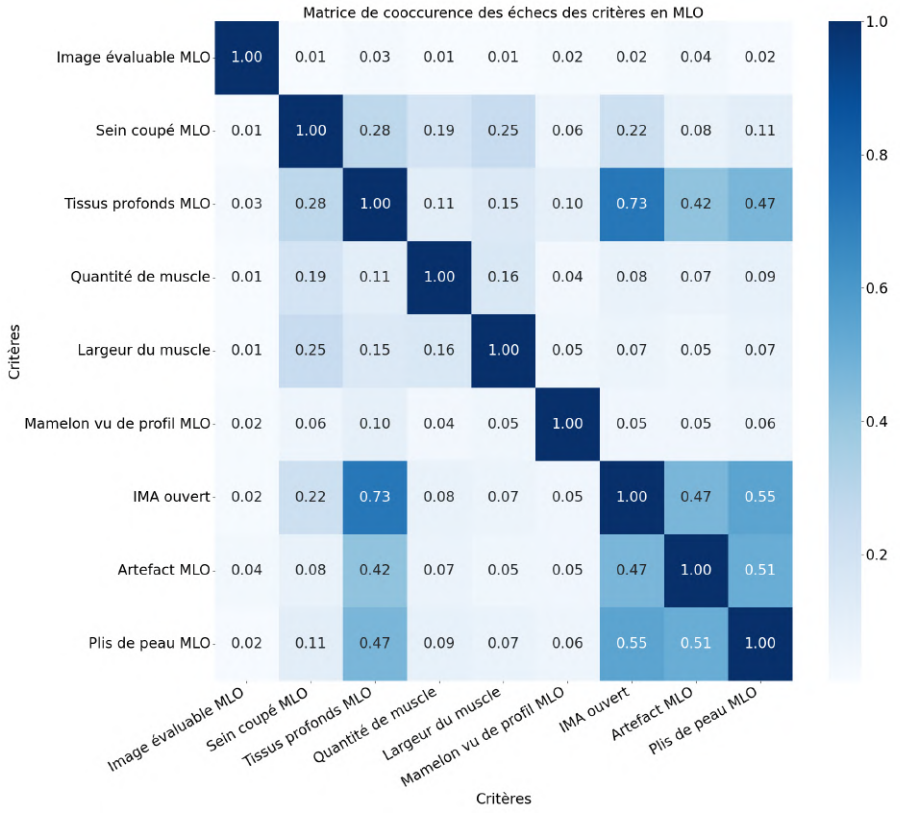


FIGURE 4.5 Matrice de cooccurrences des échecs entre les critères pour la vue MLO

leur définition. Les critères des Artéfacts et des Plis de peau n'ayant pas d'interdépendance avec d'autres critères, tant selon le graphe du chapitre 2 que selon les discussions que l'équipe a eues avec les technologues, nous avons choisi de ne pas les inclure dans ce mémoire. La détection d'artéfacts et de plis de peau fera l'objet d'un travail de recherche séparé, par un autre étudiant.

Lors de nos discussions avec les technologues participant à la phase d'annotation, il nous a été mentionné que le non-respect du critère du Mamelon vu de profil n'entraînait pas la reprise de l'examen. Il s'agit d'un critère que les technologues essaient de satisfaire lors de l'acquisition des images, mais qu'il n'est pas primordial de respecter jusqu'à reprendre les clichés mammaires. Pour cette raison, ce critère ne sera pas considéré dans ce travail.

Finalement, les critères de la Quantité de muscle et de la Largeur du muscle, en MLO, concernent tous deux le muscle pectoral. L'approche par apprentissage multi-tâches pourrait donc être pertinente pour l'évaluation de ces critères, puisque celle-ci se base sur le même élément de l'image. Or, puisque l'occurrence des images inadéquates pour ces critères est d'au plus seulement 7%, nous avons jugé que l'évaluation de ces critères par une approche

d'apprentissage supervisé n'était pas optimale. Dans le cadre du projet global, l'évaluation de ces critères est effectuée par une étudiante de l'équipe et par une approche basée sur la détection des repères anatomiques.

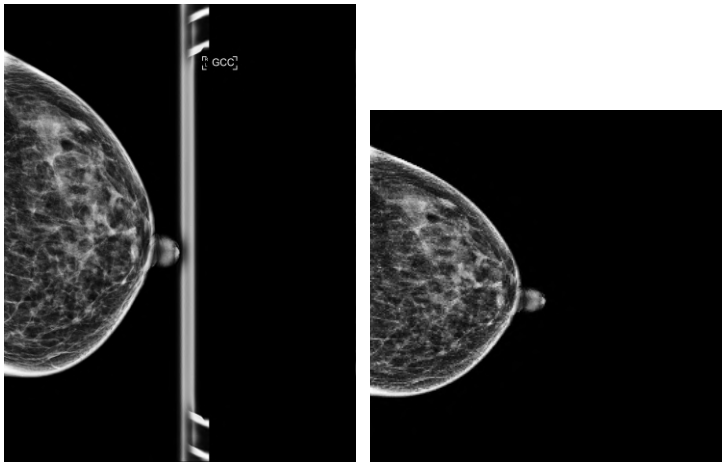
Le tableau ci-dessous présente les critères retenus dans le cadre de ce mémoire.

Tableau 4.3 Critères considérés dans ce mémoire

Vue	**Critères**
CC et MLO	Sein coupé
	Tissus profonds
CC	Orientation du mamelon
MLO	IMA ouvert

4.3 Pré-traitement des images

Le pré-traitement des données est une étape essentielle afin d'homogénéiser les images qui serviront à l'entraînement des modèles d'apprentissage profond. Chaque image du jeu de données contient une étiquette indiquant la vue (CC ou MLO) et le côté (droit ou gauche) de l'image. Ces étiquettes ont été supprimées à l'aide d'un algorithme basé sur les seuils d'intensité des pixels et sur les contours présents dans l'image. Certaines formes claires apparaissant sur les images et dues au plateau de compression ont aussi été supprimées à l'aide de cette méthode. Les détails du fonctionnement et de l'implémentation de l'algorithme de pré-traitement sont présentés à l'annexe A. Il est à noter que plusieurs images ont dû être traitées individuellement à cause de l'hétérogénéité des intensités des pixels d'arrière-plan des images et des différentes formes d'artéfacts. De plus, la taille originale des images n'étant pas uniforme et trop grande pour l'entraînement de modèles dans des délais raisonnables, un redimensionnement des images à une taille de 256 x 256 pixels a été effectué. La Figure 4.6 présente un exemple d'image avant et après le pré-traitement.



(a) (b)

FIGURE 4.6 Effet du pré-traitement des images. (a) Image originale avec étiquette et forme claire due au plateau de compression et (b) image après le pré-traitement

4.4 Élaboration des modèles *baselines*

Une étape importante dans la réalisation du projet est l'élaboration des modèles *baselines*, c'est-à-dire les modèles servant de référence pour la suite des expérimentations. Dans le cadre du projet, les modèles *baselines* sont des modèles d'évaluation des critères en tâche unique qui serviront de comparaison pour l'approche par apprentissage multi-tâches.

Un premier modèle *baseline* correspond à un de ceux développés par Brahim et al. [5]. Comme mentionné dans le chapitre 2, l'équipe de Brahim et al. a élaboré deux CNNs simples pour l'évaluation de plusieurs critères de positionnement ; il s'agit du travail de la littérature qui se rapproche le plus de notre projet. Le modèle choisi comme *baseline* est celui utilisé pour l'évaluation des critères du Mamelon vu de profil (CC et MLO) et du Recouvrement de tout le tissu mammaire pertinent (CC et MLO) tels que définis par les auteurs. Bien que ces critères soient un peu différents de ceux que nous utilisons, leur définition se rapproche des critères que nous considérons, entre autres Sein coupé et Tissus profonds. L'architecture du modèle est présentée à la Figure 4.7.

Le modèle prend en entrée une image de 256 x 256 pixels et retourne en sortie les probabilités d'appartenir à l'une ou l'autre des deux classes (positionnement adéquat/positionnement inadéquat). Il s'agit d'un CNN composé de cinq couches convolutives entre lesquelles sont placées des couches de *Max Pooling*. Celles-ci réduisent la dimensionalité des cartes de caractéristiques (*feature maps*), ce qui permet d'améliorer la robustesse du modèle [92]. Finalement, deux couches entièrement connectées (*fully connected layers*) (FCs) sont utilisées pour la

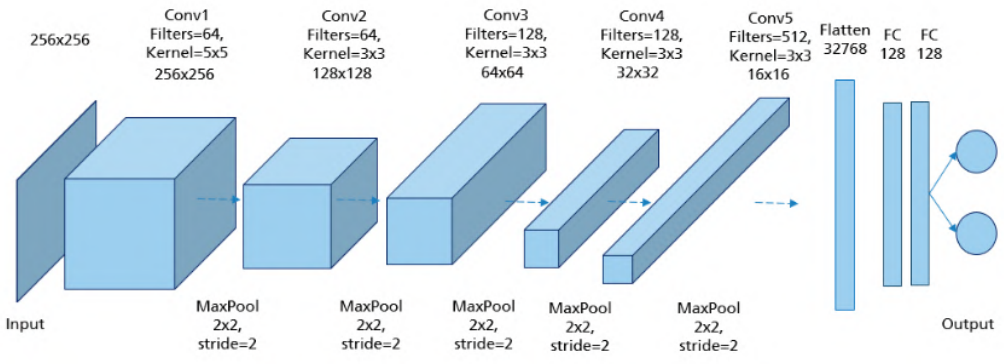


FIGURE 4.7 Architecture du modèle de Brahim et al. [5]

classification des images. Puisque les données utilisées par Brahim et al. sont privées, nous avons recréé l'architecture du modèle et avons entraîné le modèle sur les images de notre jeu de données selon le protocole d'entraînement décrit dans la section 4.7.

Afin d'élaborer une deuxième *baseline*, nous avons considéré une architecture DenseNet-121. Il s'agit d'un modèle ayant démontré d'excellentes performances sur des tâches de classification d'images dans des domaines variés ; cependant, ce modèle n'a jamais été utilisé dans le cadre de l'évaluation automatique du positionnement mammaire. Le choix du DenseNet-121 découle également d'expérimentations effectuées en amont, présentées à l'annexe B.

Ce modèle a été développé par Huang et al. [93] en 2017. Un DenseNet est composé de plusieurs blocs denses ainsi que de couches de transition. Chaque bloc dense est composé d'une normalisation de batch (*batch normalization*), d'une activation ReLu et de couches convolutives. La particularité de ce bloc est qu'il reçoit en entrée la concaténation de toutes les sorties précédentes, donc les cartes de caractéristiques obtenues en sortie de toutes les couches antérieures du réseau. Quant aux couches de transition entre les blocs denses, elles permettent de réduire la taille des cartes de caractéristiques et la résolution spatiale. Un DenseNet-121 contient quatre blocs denses. L'architecture de base d'un DenseNet est illustrée à la Figure 4.8.

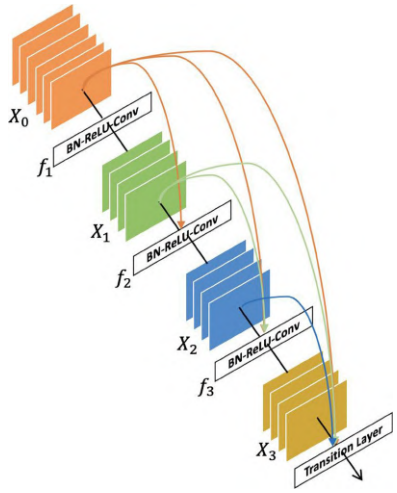


FIGURE 4.8 Architecture de base d'un DenseNet [6]

4.5 Élaboration des modèles multi-tâches

Afin de répondre à la première question de recherche (Q1), plusieurs architectures multi-tâches ont été élaborées. Nous proposons trois modèles principaux, auxquels s'ajoutent des variantes.

Modèle MT HPS

Le premier modèle développé est basé sur une architecture HPS. Dans ces types de modèles, les tâches partagent un certain nombre de couches cachées du réseau, tandis que d'autres couches sont spécifiques à chaque tâche. Puisqu'aucun travail ne s'est encore penché sur l'apprentissage multi-tâches pour l'évaluation des critères de positionnement mammaire, nous avons décidé d'élaborer un modèle HPS très simple, comprenant un encodeur commun aux différentes tâches et une tête de classification spécifique à chacune. Ce modèle sera nommé MT HPS dans la suite de ce mémoire. Après l'aplatissement du tenseur en sortie de l'encodeur, la tête de classification contient une couche de *dropout* et une FC. La couche de *dropout* désactive aléatoirement certaines neurones du réseau après chaque itération, ce qui permet de réduire les risques de surapprentissage. En ce qui concerne l'encodeur, le squelette (*backbone*) du DenseNet-121 a été choisi, puisqu'il s'agit du modèle le plus performant parmi ceux testés pour les *baselines*. Il s'agit d'un CNN comprenant quatre blocs denses et des couches de transition, et prenant en entrée une image de 224 x 224 pixels. Un exemple de l'architecture de ce modèle pour l'évaluation de deux critères est présenté à la Figure 4.9. Ce modèle peut être facilement modifié pour l'évaluation de plus de deux tâches à la fois, en ajoutant le

nombre de têtes de classification nécessaires.

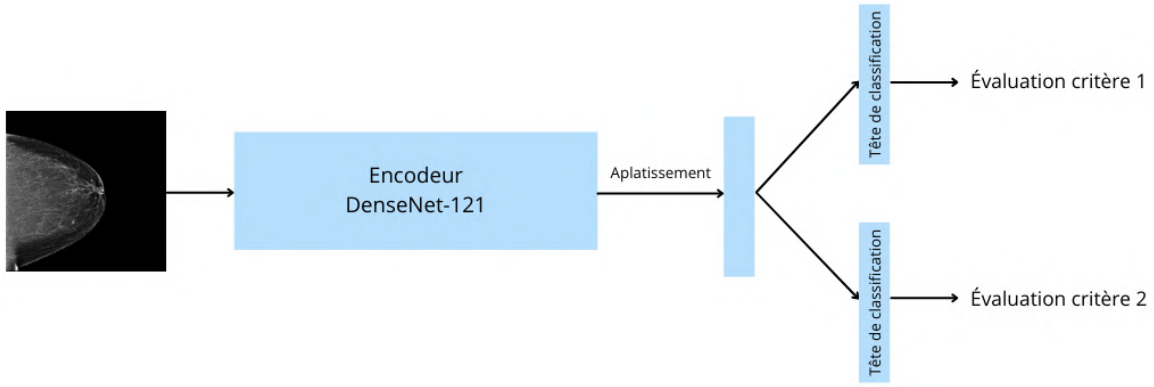


FIGURE 4.9 Architecture du modèle MT HPS pour l'évaluation de deux critères

Modèle MT SPS

Le deuxième modèle élaboré est basé sur une architecture SPS. Bien qu'elle soit moins utilisée et étudiée dans la littérature, il s'agit de la seconde grande catégorie des architectures multi-tâches et mérite d'être considérée. Le modèle comprend des encodeurs et des têtes de classification spécifiques à chaque tâche. Lors de l'entraînement, les paramètres appris sont encouragés à être similaires par le biais d'une régularisation. Ici, la régularisation l_2 a été choisie puisqu'elle est couramment employée dans la littérature [3]. Cette régularisation consiste en l'ajout d'un terme de pénalité à la fonction de coût lors de l'entraînement. La pénalité correspond au carré de la différence entre les paramètres appris par les deux encodeurs. Un exemple de l'architecture de ce modèle pour l'évaluation de deux critères de positionnement est présenté à la Figure 4.10. Comme pour le modèle HPS, ce modèle est facilement modifiable pour l'évaluation de plus de deux critères en ajoutant les encodeurs et les têtes de classification nécessaires. Ce modèle est nommé MT SPS.

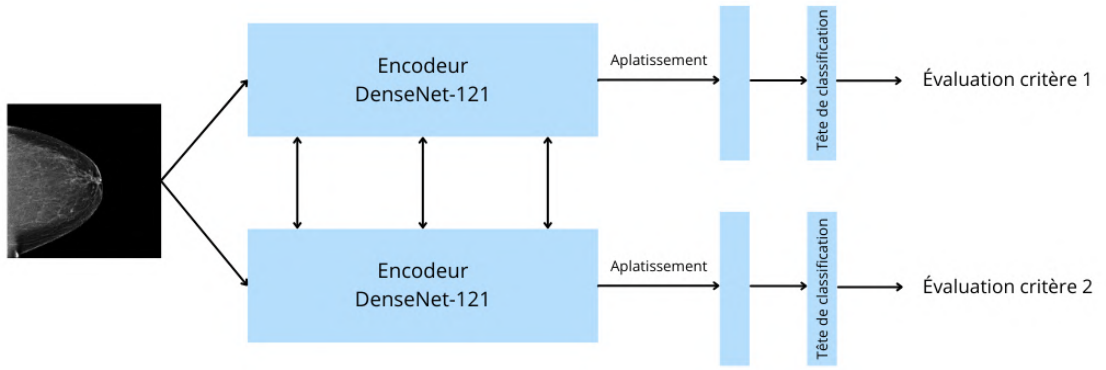


FIGURE 4.10 Architecture du modèle MT SPS pour l'évaluation de deux critères

Modèle ML

Le troisième modèle développé est un modèle multi-labels (ML). Dans ce modèle, les tâches partagent toutes les mêmes couches du réseau, et une classification est effectuée selon plusieurs annotations pour une même image. L'encodeur et la tête de classification sont donc communs à toutes les tâches. Le modèle donne ensuite en sortie les probabilités indépendantes concernant chacun des critères de positionnement. Un schéma de l'architecture du modèle multi-labels pour l'évaluation de deux critères est présenté à la Figure 4.11.

FIGURE 4.11 Architecture du modèle ML pour l'évaluation de deux critères

Modules d'attention

La revue de la littérature a permis de constater l'utilisation courante de modules d'attention dans les modèles multi-tâches. Les mécanismes d'attention permettent à un réseau de neurones de prioriser les informations les plus pertinentes des données en entrée. Nous avons

donc décidé d'introduire deux types de modules d'attention dans le modèle HPS afin d'étudier l'impact de leur intégration au modèle de base. Le premier module d'attention est le *Multi-head Attention module* (MultiHead). Il s'agit du module d'attention original décrit dans l'article de Vaswani et al. [94] : dans ce module, plusieurs têtes d'attention sont concaténées, chacune contenant les scores d'attention calculés en fonction des vecteurs de *query* (Q), *key* (K) et *value* (V). Lorsque les données d'entrée sont des images, les vecteurs Q, K et V correspondent à la multiplication des différentes parties de l'image par les matrices de poids apprises. L'attention MultiHead se calcule à l'aide de la formule suivante :

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(Head_1, ..., Head_n)W^O \\ \text{avec } Head_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{4.3}$$

Ce module d'attention est intégré au modèle HPS dans chacune des têtes de classification, juste avant la FC. Cela pourrait être bénéfique à l'apprentissage, puisque cela permettrait à chaque tête de classification de se concentrer sur les parties importantes de la représentation commune apprise. Ce modèle sera nommé MT MultiHead.

Le second module d'attention testé est un *Cross-Attention module* (CrossAtt). Ce type d'attention reprend le même principe global que l'attention MultiHead, mais en mélangeant les vecteurs Q d'une source avec les vecteurs K et V d'une autre source. Souvent utilisé pour conditionner une modalité par une autre (par exemple : images et texte), il peut être employé dans l'apprentissage multi-tâche afin de forcer le modèle à établir des liens entre les tâches. Dans ce cas, le vecteur K provient d'une tâche, tandis que les vecteurs Q et V proviennent d'une autre tâche. La CrossAtt est calculée sur les projections des vecteurs de caractéristiques, comme présenté à la Figure 4.12. Les caractéristiques originales et d'attention sont ensuite combinées avant les têtes de classification. Ce mécanisme d'attention est inspiré de celui utilisé par Kim et al. [76], et le modèle sera nommé MT CrossAtt pour la suite de ce mémoire.

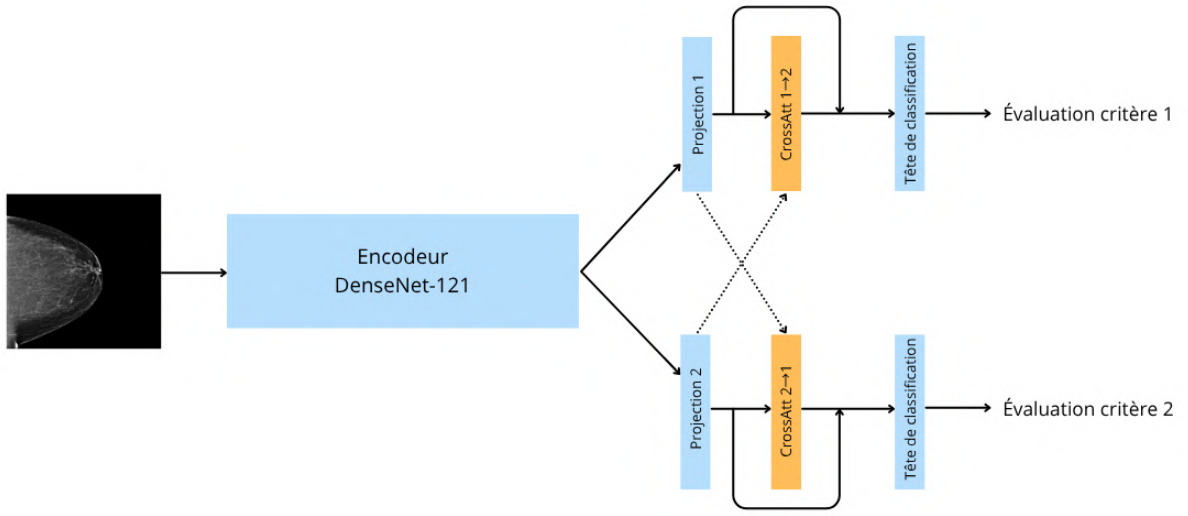


FIGURE 4.12 Architecture du modèle MT CrossAtt pour l'évaluation de deux critères

Le tableau 4.4 présente un résumé des modèles multi-tâches élaborés pour ce projet.

TABLEAU 4.4 Résumé des modèles multi-tâches

Modèles	**Description**
MT HPS	Encodeur commun, têtes de classifications spécifiques aux tâches
MT SPS	Encodeurs et têtes de classification spécifiques aux tâches
ML	Encodeur commun et tête de classification commune aux tâches
MT MultiHead	Modèle MT HPS avec modules d'attentions spécifiques aux tâches
MT CrossAtt	Modèle MT HPS avec modules d'attentions inter-tâches

4.6 Métriques d'évaluation

Cette section présente les principales métriques utilisées dans ce travail pour l'évaluation de modèles de classification. Dans les problèmes de classification binaire dans le domaine médical, les cas dits positifs correspondent aux données anormales, alors que les cas dits négatifs correspondent aux données normales. Dans le cadre de ce projet et pour un critère donné, les cas positifs correspondent aux images dont le positionnement est inadéquat, et les cas négatifs correspondent aux images dont le positionnement est adéquat. Pour la suite de cette section, les termes suivants seront utilisés :

Vrais positifs (TP) : Ensemble des données positives classées comme positives.

Vrais négatifs (TN) : Ensemble des données négatives classées comme négatives.

Faux positifs (FP) : Ensemble des données négatives classées comme positives.

Faux négatifs (FN) : Ensemble des données positives classées comme négatives.

La première métrique est la spécificité. Celle-ci correspond au taux de vrais négatifs par rapport à l'ensemble des négatifs présents dans le jeu de données. Elle met en évidence la capacité d'un modèle à prédire correctement les cas négatifs, sans donner d'informations sur les prédictions des cas positifs. On peut considérer cette métrique comme l'exactitude de la classe négative. La spécificité est donnée par l'équation suivante :

$$Spécificité = \frac{TN}{TN + FP} \tag{4.4}$$

La sensibilité (ou rappel) correspond au taux de vrais positifs par rapport à l'ensemble des positifs dans le jeu de données. À l'inverse de la spécificité, cette métrique permet de faire ressortir la capacité d'un modèle à prédire correctement les cas positifs. Dans le domaine médical, il est important pour les modèles d'avoir une grande sensibilité afin de minimiser les cas de faux négatifs. La sensibilité est calculée à l'aide de l'équation suivante :

$$Sensibilité = \frac{TP}{TP + FN} \tag{4.5}$$

La précision est une métrique représentant le taux de vrais positifs par rapport à l'ensemble des cas positifs par un modèle. Cette métrique permet de mettre en évidence quelle proportion des cas prédits positifs par le modèle le sont réellement. La précision est donnée par la formule suivante :

$$Précision = \frac{TP}{TP + FP} \tag{4.6}$$

Ces trois métriques sont calculées à partir de la classification binaire effectuée par un modèle. Or, les modèles de classification donnent en sortie les probabilités qu'une donnée appartienne à l'une ou l'autre des classes ; un seuil doit ensuite être appliqué à ces probabilités pour obtenir la classification binaire. Dans ce mémoire, nous utiliserons deux métriques indépendantes de seuils, permettant d'évaluer des modèles de classification sur les probabilités en sortie.

D'abord, la Fonction d'efficacité du récepteur (*Receiving Operator Characteristic*) (ROC) est une courbe obtenue en traçant, pour les différentes valeurs de seuil possibles, le taux de vrais positifs (ou la sensibilité) en ordonnées et le taux de faux positifs (ou 1 - spécificité) en abscisse. En calculant l'aire sous cette courbe (ROC AUC), on obtient une valeur représentant

la performance d'un modèle sur la tâche de classification (une aire de 1 indiquant un modèle parfait et une aire de 0,5 indiquant un modèle aléatoire).

Une autre courbe peut être utilisée pour mesurer les performances d'un modèle sans utiliser de seuil fixe : il s'agit de la courbe Précision-rappel (*Precision-recall*) (PR). Cette courbe trace, pour différentes valeurs de seuil, la valeur de la précision en ordonnée et la valeur du rappel en abscisse. L'aire sous cette courbe (PR AUC) peut ensuite être calculée afin d'évaluer un modèle de classification. Cette métrique est plus adaptée aux situations de déséquilibre des classes, puisqu'elle met en valeur la capacité d'un modèle à correctement classer les données de la classe positive, contrairement à la ROC AUC. Un modèle aléatoire devrait obtenir une PR AUC égale à la proportion de données de la classe positive. Par exemple, pour 35% de données positives, une PR AUC de 0,35 correspond à des prédictions aléatoires.

4.7 Protocole d'entraînement et inférence

Pour les modèles *baselines* comme pour les modèles multi-tâches, l'ensemble d'images A a été utilisé pour l'entraînement. De cet ensemble, 20% ont été utilisés pour la validation et 80% pour l'entraînement. D'abord, une recherche d'hyperparamètres a été effectuée afin de déterminer les paramètres optimaux pour chacun des critères. Les hyperparamètres étudiés sont la fonction de perte, le *learning rate* (*lr*) et la *weight decay*.

La fonction de perte est la fonction optimisée durant l'entraînement et qui guide l'apprentissage du modèle : elle mesure la différence entre les prédictions du modèle et la vérité terrain, et a pour but de minimiser cette valeur au cours de l'entraînement. La fonction de perte la plus répandue pour la classification binaire est la fonction d'entropie croisée binaire (*binary cross-entropy*) (BCE). Celle-ci est définie par la formule suivante pour une instance :

$$L(y_i, \hat{p}_i) = -\left[y_i log(\hat{p}_i) + (1 - y_i) log(1 - \hat{p}_i)\right] \text{ [95]} \tag{4.7}$$

avec y_i les vérités terrain et $\hat{p}_i$ les prédictions du modèle. La fonction de perte totale sur le jeu de données est minimisée durant l'entraînement :

$$L_{BCE}(\theta) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{p}_i) \text{ [95]} \tag{4.8}$$

Il est important de prendre en considération le déséquilibre des classes dans le choix de la fonction de perte. En effet, puisque pour presque tous les critères de positionnement à l'étude, les classes positives et négatives sont déséquilibrées, il est crucial d'avoir une fonction de perte

qui met plus d'accent sur les images de la classe minoritaire afin que celles-ci aient aussi un impact dans l'entraînement du modèle. Deux fonctions sont largement utilisées dans ces situations : la fonction de perte BCE pondérée (WCE) et la fonction de perte focale (*Focal loss*). Ces deux fonctions ont été testées dans le cadre de la recherche d'hyperparamètres. La WCE est donnée par la formule suivante :

$$L_{WCE}(y_i, \hat{p}_i) = -\left[\alpha_1 y_i log(\hat{p}_i) + \alpha_0 (1 - y_i) log(1 - \hat{p}_i)\right] \text{ [95]} \tag{4.9}$$

avec α_1 et α_0 les coefficients de pondération spécifiques à chaque classe. La fonction de perte focale est une variante de la perte BCE et est calculée à l'aide de la formule suivante :

$$L_F(p_t) = \alpha(1 - p_t)^\gamma \cdot L_{BCE}(y, p) \text{ [96]} \tag{4.10}$$

avec α et γ les coefficients contrôlant la pondération des classes. Lorsque $\gamma = 0$, la perte focale se réduit à la perte BCE. Pour les modèles MT HPS, MT SPS, MT MultiHead et MT CrossAtt, la fonction de perte à optimiser est la somme des pertes pour chaque tâche. Pour le modèle ML, une seule fonction de perte est utilisée simultanément pour toutes les tâches.

Le *lr* est un hyperparamètre qui contrôle la vitesse d'apprentissage d'un modèle d'IA. Le choix de ce paramètre est très important : un trop petit *lr* peut résulter en un processus d'apprentissage trop lent, alors qu'un trop grand *lr* peut empêcher la convergence du modèle [97]. Lors de la recherche d'hyperparamètres, les valeurs suivantes de *lr* ont été testées : $lr \in \{0.00001, 0.00005, 0.0005, 0.0001\}$.

Finalement, la *weight decay* est une technique de régularisation fréquemment utilisée pour restreindre les valeurs que peuvent prendre les paramètres au cours de l'entraînement, aidant le modèle à apprendre une fonction plus simple et à prévenir le surapprentissage. Ce paramètre est directement intégré dans l'optimiseur. Lors de la recherche d'hyperparamètres, les valeurs suivantes ont été testées : *weight decay* $\in \{0, 0.00001, 0.0001\}$.

Des techniques d'augmentation ont été utilisées lors de l'entraînement ; cela permet d'augmenter artificiellement la taille du jeu de données, de réduire le risque de surapprentissage et d'améliorer la robustesse des modèles. Les augmentations utilisées incluent des retournements horizontaux des images, l'ajustement aléatoire du contraste, l'ajustement de la netteté et l'inversion des valeurs des pixels.

Les modèles ont été entraînés pendant 100 époques, avec une taille de *batch* de 32, et l'optimiseur Adam a été utilisé. Les meilleurs hyperparamètres ont été choisis à partir des résultats de PR AUC sur l'ensemble de validation. Pour obtenir les modèles finaux, une validation

croisée a été effectuée sur cinq blocs (*folds*) de validation : cinq modèles en découlent, validés sur des ensembles différents. Cela permet au modèle de voir des images diversifiées pendant l'entraînement et de s'évaluer sur des ensembles variés. L'entraînement des modèles a été arrêté à l'époque de la meilleure PR AUC sur le bloc de validation. Pour les modèles multi-tâches, la PR AUC combinée des tâches a été considérée. Les processus de validation croisée ont été réalisés trois fois avec des graines aléatoires différentes afin de mesurer la robustesse des modèles à l'initialisation des poids.

Des modèles ensemblistes ont ensuite été élaborés à partir des modèles entraînés sur chacun des blocs de validation. Cette méthode a été utilisée puisqu'elle permet d'obtenir une meilleure généralisation des modèles, ainsi que de meilleures performances [98]. Afin de combiner les modèles, une méthode de moyenne non pondérée des modèles (*unweighted model averaging*) a été utilisée : il s'agit de l'approche la plus répandue dans la littérature pour fusionner les prédictions des modèles. À l'inférence, les images sont données en entrée aux cinq modèles obtenus avec les cinq blocs de validation, et la moyenne des probabilités est calculée pour chaque image afin d'obtenir les probabilités finales. Cette méthode permet d'obtenir de meilleures prédictions, puisque le modèle ensembliste a été exposé à la totalité des images pendant l'entraînement et la validation.

L'ensemble d'images B, soit le sous-ensemble d'images annoté par les trois technologues, écarté jusque là, a été utilisé à l'inférence comme ensemble de test. Afin d'évaluer les performances des modèles, la ROC AUC et la PR AUC ont été privilégiées puisqu'elles sont indépendantes de seuils. Une première évaluation a été effectuée en prenant comme vérité terrain le vote majoritaire des annotations des trois technologues.

Pour tenir compte de la variabilité inter-technologues dans les résultats, nous avons utilisé une méthode inspirée de celle employée par Irvin et al. [99] pour l'évaluation d'un modèle de classification d'anomalies dans des radiographies du thorax. Les auteurs de cet article comparent les courbes ROC et PR du modèle, obtenues à l'aide d'une première vérité terrain, avec les annotations de trois autres évaluateurs par rapport à cette même vérité terrain. Les courbes du modèle sont donc comparées aux points des trois évaluateurs dans les graphes sensibilité/1 - spécificité et précision/rappel. Dans notre cas, nous avons tracé les courbes ROC et PR des modèles par rapport à une première technologue et avons comparé cette courbe avec les annotations des deux autres technologues par rapport à la première technologue, et ainsi de suite pour chacune des trois technologues. Cette méthode permet de déterminer comment se situent les performances d'un modèle par rapport aux autres annotatrices, positionnant les résultats relativement à la variabilité inter-opérateurs.

4.8 Simulation de réduction de dose et robustesse au bruit

Afin de répondre à la seconde question de recherche, une méthode de simulation de réduction de dose de rayons X a été mise en place. Celle-ci est inspirée de la méthode utilisée par Borges et al. [4] décrite dans le chapitre 2. Cette méthode se décompose en les étapes suivantes :

1. Normalisation de l'image et multiplication par un nombre maximum de photons de 1000.
2. Transformation de l'image dans le domaine d'Anscombe. La transformation est définie par la formule suivante :

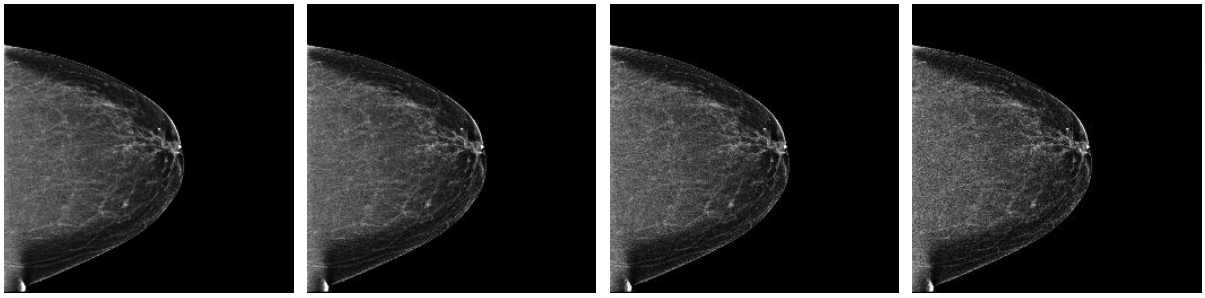
$$A\{g(x,y)\} = 2\sqrt{g(x,y) + \frac{3}{8}} \text{ [4]} \tag{4.11}$$

3. Ajout d'un bruit gaussien (signal-indépendant) à l'image dans le domaine d'Anscombe. Puisque nous n'avons pas accès à des images acquises en faible dose de rayons X pour estimer le bruit, plusieurs intensités de bruit ont été générées. Des variances $\sigma \in \{1, 2, 3\}$ ont été utilisées pour simuler différents niveaux de bruit gaussien.
4. Transformation d'Anscombe inverse pour ramener l'image dans le domaine initial. La transformation inverse est donnée par l'équation suivante :

$$Inv(D) = \frac{1}{4}D^2 + \frac{1}{4}\sqrt{\frac{3}{2}}D^{-1} - \frac{11}{8}D^{-2} + \frac{5}{8}\sqrt{\frac{3}{2}}D^{-3} - \frac{1}{8} \text{ [100]} \tag{4.12}$$

   avec D l'image dans le domaine d'Anscombe. Le bruit ajouté est alors transformé en bruit signal-dépendant (bruit de Poisson).
5. *Clipping* des valeurs d'intensité de l'image pour que celles-ci restent dans la plage d'intensités de 0 à 255.

La Figure 4.13 présente des exemples d'images bruitées obtenues par la méthode décrite ci-dessus.



(a) (b) (c) (d)

FIGURE 4.13 Effet de l'ajout de bruit gaussien dans le domaine d'Anscombe à une image. (a) Image originale, (b) image bruitée avec $\sigma = 1$, (c) image bruitée avec $\sigma = 2$ et (d) image bruitée avec $\sigma = 3$

CHAPITRE 5 RÉSULTATS ET DISCUSSION

Ce chapitre présente les expérimentations et les résultats découlant de la méthodologie décrite dans le chapitre 4. D'abord, les résultats de l'analyse de variabilité inter-technologues seront présentés et discutés. Ensuite, seront présentés les expérimentations et les résultats de l'élaboration des modèles *baselines* et des modèles multi-tâches d'après le processus d'entraînement et d'évaluation des modèles présenté dans le chapitre précédent. Les résultats incluent aussi des comparaisons avec des modèles multi-tâches définis dans la littérature, notamment le modèle développé par Nguyen et al. [86] et le MTAN [74]. Une analyse et une discussion de ces résultats sont présentées afin de mettre en relief la contribution des modèles multi-tâches par rapport aux *baselines*, ainsi que pour replacer les performances des modèles dans le contexte de la variabilité inter-technologues. Finalement, les résultats de l'impact de l'ajout de bruit simulant une réduction de dose de rayons X aux images seront présentés et discutés.

5.1 Variabilité inter-technologue

La variabilité inter-technologue a été analysée sur l'ensemble d'images B, contenant 250 examens annotés communément par les trois technologues. Deux métriques ont été utilisées pour mesurer la variabilité entre les annotations de deux technologues, soient le taux d'accord entre les technologues et le kappa de Cohen. Pour comparer les annotations des trois technologues à la fois, le taux d'accord et le kappa de Fleiss ont été utilisés. Les Figures 5.1 et 5.2 présentent les résultats de variabilité pour les trois technologues à la fois, par critère de positionnement. Ces figures ont été élaborées dans le cadre de l'écriture d'un article sur la variabilité inter-opérateurs [91] par les membres de l'équipe, et sont présentées sous forme traduite dans ce mémoire. Les figures présentant les résultats de variabilité par paires de technologues sont disponibles à l'annexe C.

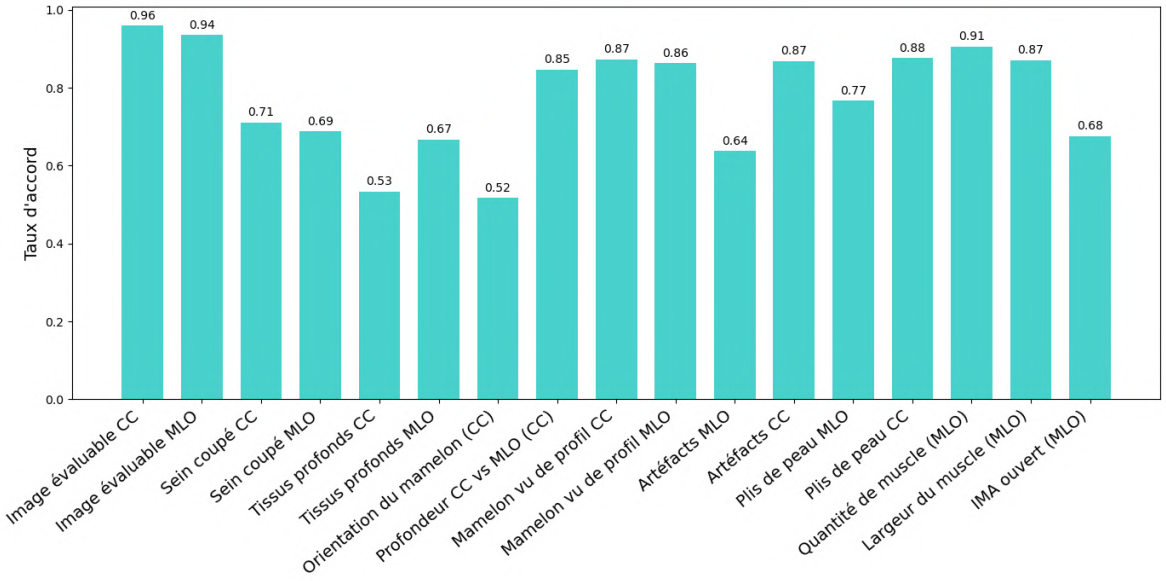


FIGURE 5.1 Taux d'accord entre les trois technologues pour chaque critère de positionnement

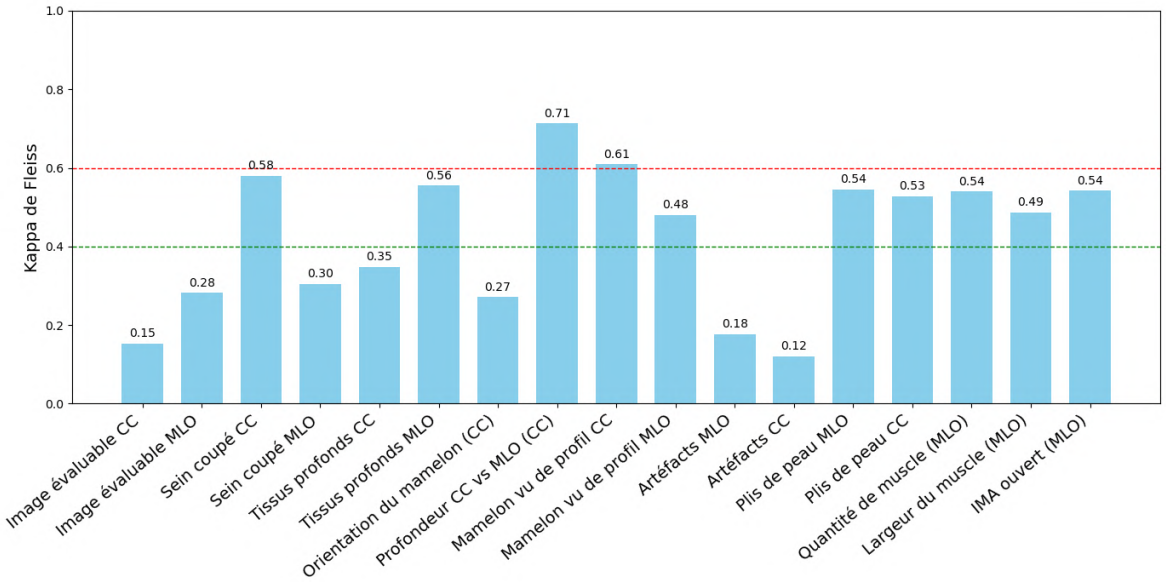


FIGURE 5.2 Kappa de Fleiss pour les trois technologues pour chaque critère de positionnement. Ligne verte pointillée : seuil pour un accord modéré selon Cohen. Ligne rouge pointillée : seuil pour un accord modéré selon McHugh et al. [7]

Le taux d'accord varie entre 0,52 pour le critère de l'Orientation du mamelon et 0,96 pour le critère de l'Image évaluable en CC. Pour le kappa de Fleiss, les valeurs varient entre 0,12 pour le critère des Artéfacts en CC et 0,71 pour la Profondeur CC vs MLO. Dans la Figure

5.2, les lignes pointillées représentent les seuils pour un accord modéré selon Cohen et selon McHugh et al. [7]. On constate que seulement deux critères dépassent le seuil de 0,6 : les critères du Mamelon vu de profil en CC et de la Profondeur CC vs MLO. Le seuil de 0,4 proposé par Cohen est atteint par plus de critères, dont Sein coupé en CC, Tissus profonds en MLO, Mamelon vu de profil en MLO, Plis de peau, l'IMA ouvert et les deux critères concernant la quantité et la largeur du muscle pectoral. Selon Cohen, un kappa de 0,41 à 0,60 indique un accord modéré entre les opérateurs. Par contre, McHugh et al. argumentent que cette interprétation est trop indulgente et permet de considérer un faible accord comme substantiel. Ils suggèrent plutôt qu'un kappa de 0,60 à 0,79 représente un accord modéré, alors que des valeurs en-dessous de 0,60 indiquent un faible accord, réduisant le niveau de confiance pouvant être placé dans les résultats.

Il est important de considérer le taux d'accord et le kappa de Fleiss de façon complémentaire. En effet, le taux d'accord indique la proportion d'images pour lesquelles les technologues étaient unanimes dans leur évaluation, sans tenir compte de la proportion d'images de chaque classe. Donc, pour un critère avec très peu d'images de la classe positive, on peut obtenir un excellent taux d'accord même si les technologues ne sont pas unanimes sur les images positives. En revanche, le kappa de Fleiss prend en compte le déséquilibre des classes puisque son calcul considère l'accord pouvant être obtenu par hasard. Par exemple, pour le critère de l'Image évaluable (en CC et en MLO), le taux d'images positives est très faible et on remarque un taux d'accord élevé, mais un kappa de Fleiss très bas. Cela signifie que, même si pour la grande majorité des images, les technologues sont unanimes, elles sont en désaccord pour la majorité des images de la classe minoritaire, à savoir les images positives. On note la même situation pour le critère des Artéfacts (en CC et en MLO).

Analyse générale

De manière générale, les résultats révèlent une forte variabilité entre les annotatrices, notamment pour les critères considérés dans ce mémoire. Des taux d'accord de 0,71 et 0,69, et des kappas de 0,58 et 0,30 sont obtenus pour le critère Sein coupé en CC et en MLO, respectivement. Le faible kappa obtenu pour le critère Sein coupé en MLO peut s'expliquer par le grand déséquilibre des classes pour ce critère, avec seulement 14,2% d'images positives comparativement à 34,1% pour le même critère en CC. Le critère Sein coupé en CC est celui qui obtient le meilleur taux d'accord et le kappa le plus élevé parmi les critères considérés. Par contre, même pour ce critère, le seuil de kappa de 0,6 pour un accord modéré n'est pas atteint.

Pour le critère Tissus profonds, des taux d'accord de 0,53 et 0,67 sont observés, et des kappas

de 0,35 et de 0,56 sont obtenus pour les vues CC et MLO respectivement. Ces résultats démontrent une plus grande variance entre les annotations pour le critère dans la vue CC que dans la vue MLO. Les valeurs pour les deux vues demeurent toutefois faibles, ce qui pourrait s'expliquer par les nombreuses raisons existantes pour le non-respect de ce critère et le caractère subjectif de son évaluation. Pour le critère de l'IMA ouvert, un taux d'accord de 0,68 est obtenu, indiquant un désaccord entre les technologues pour environ le tiers des images. Finalement, les plus faibles résultats sont obtenus pour le critère de l'Orientation du mamelon : un taux d'accord de 0,52 et un kappa de 0,27. Ces valeurs révèlent une grande variabilité entre les annotations des technologues, peut-être dû au manque de définition d'un seuil d'angle auquel le mamelon est considéré comme désorienté. Bien que nous ayons discuté de ce critère avec les annotatrices en amont du processus d'annotations pour en arriver à un consensus, des différences semblaient persister quant à leur interprétation du non-respect de ce critère.

Dans la littérature, quelques travaux se sont déjà penchés sur la variabilité inter-opérateurs dans l'évaluation du positionnement mammaire. En général, les travaux existants notent une forte variabilité entre les technologues [101–104], ce qui concorde avec les résultats de notre analyse. Dans une étude de 2025, [104] obtiennent des valeurs de kappa allant de 0,085 à 0,701 en évaluant la variabilité inter-opérateurs pour l'évaluation d'une liste de critères semblable à la nôtre. Ceux-ci vérifient des éléments tels que la visibilité du muscle pectoral, la visibilité du mamelon de profil, la présence de plis de peau, la visibilité de l'IMA et la différence de longueur entre la PNL en CC et en MLO. Cette plage de valeurs est très similaire à celle obtenue dans notre analyse, avec des kappas pour l'ensemble des critères allant de 0,12 à 0,71. Cependant, on ne peut faire de comparaison directement entre ces travaux avec le nôtre puisqu'ils ne considèrent pas exactement les mêmes critères de positionnement, et qu'ils utilisent le système d'évaluation *Perfect, Good, Moderate, Inadequate* (PGMI).

5.2 Modèles *baselines*

Dans le cadre de l'élaboration de modèles *baselines* en tâche unique, l'architecture d'un modèle proposé par Brahim et al. [5] a été recréée, et le modèle a été entraîné selon le protocole présenté au chapitre 4. Un modèle DenseNet-121 a aussi été entraîné pour l'évaluation du sous-ensemble de critères de positionnement considéré dans ce mémoire. Les résultats de performance de ces deux modèles sont présentés dans le tableau 5.1. Les détails concernant les hyperparamètres utilisés pour l'entraînement sont consignés à l'annexe D.

TABLEAU 5.1 Résultats de ROC AUC et de PR AUC du modèle de Brahim et al. et du DenseNet-121. Moyennes et écarts-types obtenus sur trois graines aléatoires sur l'ensemble de test (ensemble d'images B, annotations majoritaires). **En gras** : meilleurs résultats.

Modèle	Sein coupé CC		Tissus profonds CC		Orientation du mamelon		Sein coupé MLO		Tissus profonds MLO		IMA ouvert	
	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC
Brahim et al.	0,888 ± 0,003	0,813 ± 0,002	0,767 ± 0,010	0,848 ± 0,007	0,629 ± 0,022	0,438 ± 0,016	0,600 ± 0,009	0,210 ± 0,006	0,661 ± 0,009	0,625 ± 0,015	0,733 ± 0,007	0,619 ± 0,012
DenseNet-121	**0,948 ± 0,003**	**0,919 ± 0,007**	**0,827 ± 0,006**	**0,885 ± 0,005**	**0,801 ± 0,007**	**0,641 ± 0,015**	**0,868 ± 0,009**	**0,513 ± 0,016**	**0,848 ± 0,001**	**0,839 ± 0,006**	**0,940 ± 0,003**	**0,901 ± 0,009**

On constate que le modèle DenseNet-121 performe mieux que le modèle de Brahim et al. sur l'ensemble des critères de positionnement considérés. En effet, le modèle de Brahim et al. est moins profond que le DenseNet-121, ce qui lui procure moins de couches d'abstraction et limite sa capacité d'apprentissage. De plus, le DenseNet-121 semble robuste à l'initialisation, avec des écarts-types de l'ordre de 1% et moins pour tous les critères. Les écarts entre les résultats obtenus pour les différents critères révèlent que l'évaluation de certains semble plus facile que d'autres. Cela peut être dû à plusieurs facteurs, dont la définition du critère, les raisons pouvant mener à un échec du critère, la variance dans les annotations des technologues, les exemples présents dans notre jeu de données et le déséquilibre des classes. Ces résultats sont discutés dans les sous-sections qui suivent.

Modèle de Brahim et al.

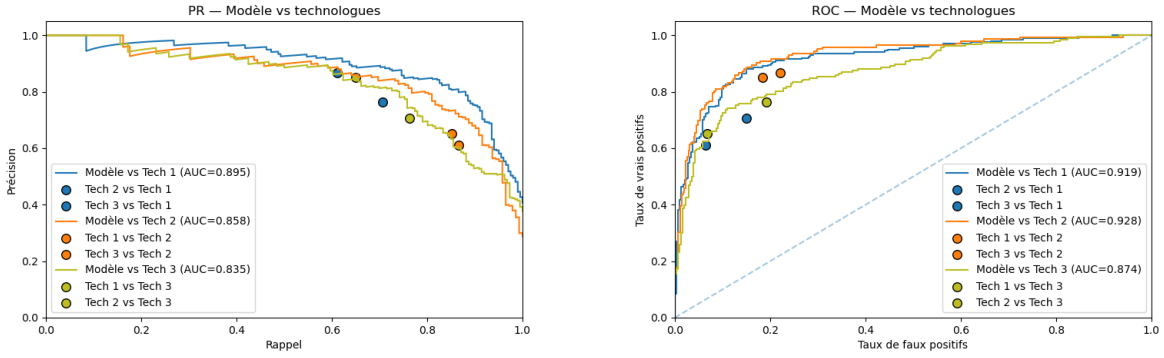
Pour ce modèle, les performances obtenues en entraînant et en testant sur nos données sont beaucoup plus basses que celles présentées dans l'article de Brahim et al. Bien que les critères utilisés dans cet article ne correspondent pas exactement aux mêmes que ceux étudiés dans ce mémoire, le critère Recouvrement de tout le tissu mammaire pertinent a une définition similaire à celle des critères Sein coupé et Tissus profonds. Pour ce critère en CC, les auteurs obtiennent une exactitude de 97%, alors que lorsqu'on applique un seuil optimal aux prédictions du modèle entraîné sur nos données, on obtient 89% et 72% d'exactitude pour les critères Sein coupé et Tissus profonds respectivement.

Plusieurs raisons peuvent expliquer ces différences. D'abord, Brahim et al. ont utilisé un jeu de données équilibré qui ne représente pas la prépondérance des cas négatifs dans un contexte réel. Pour obtenir cet ensemble équilibré, les auteurs ont effectué des augmentations (translations horizontales et verticales) pour générer des images de la classe positive à partir d'images de la classe négative. Selon les descriptions et les exemples donnés dans l'article, les images positives générées par augmentations représentent des cas plus clairs d'échec des critères. Par exemple, pour le critère Recouvrement de tout le tissu mammaire pertinent, on donne en exemple des images où le sein est très clairement coupé, par le haut ou le bas de l'image. De plus, il n'y a aucune mention dans l'article d'une réévaluation des images positives générées par augmentations par des technologues. Dans notre jeu de données, les cas d'images positives sont souvent plus subtils, nécessitant un technologue expert en mammographie pour évaluer les critères. Finalement, les images de notre jeu de données proviennent d'une population différente de celles utilisées par Brahim et al., ce qui peut aussi avoir un impact sur l'apprentissage du modèle.

Modèle DenseNet-121

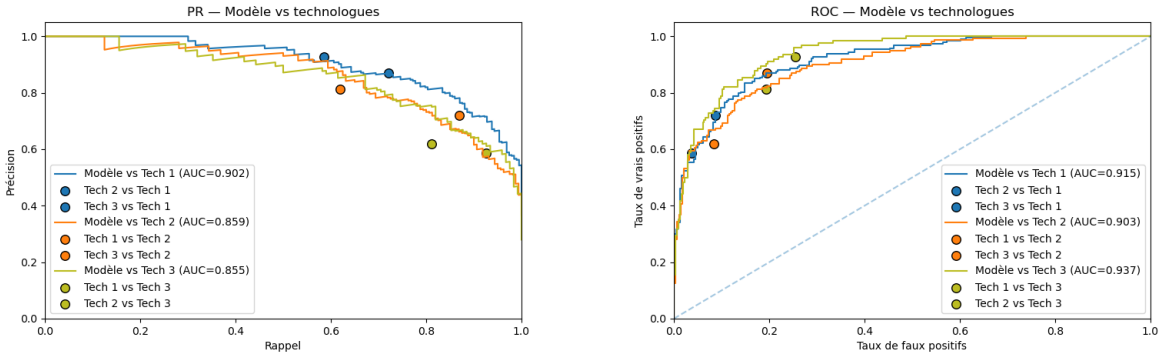
Sein coupé en CC et IMA ouvert : Les meilleures performances sont obtenues pour les critères du Sein coupé en CC (ROC AUC = 0,948, PR AUC = 0,919) et de l'IMA ouvert (ROC AUC = 0,940, PR AUC = 0,901). Il s'agit de deux critères avec un déséquilibre de classes similaire, avec environ 35% d'images positives dans les ensembles d'entraînement et de test. Ces deux critères font aussi partie de ceux ayant les taux d'accord et les kappa de Fleiss les plus élevés parmi les critères étudiés. La moins grande variabilité entre les annotations des technologues pourrait certainement contribuer à faciliter l'apprentissage des modèles et expliquer en partie les performances obtenues. Pour le critère de l'IMA ouvert, de bonnes performances étaient attendues puisqu'il s'agit d'un critère assez bien défini et se basant clairement sur un seul élément de l'image, l'IMA.

Afin de replacer les résultats obtenus dans le contexte de la variabilité inter-opérateurs, les modèles ont été évalués, à l'aide de l'ensemble de test, par rapport aux annotations de chacune des technologues individuellement. Les courbes obtenues ont ensuite été comparées aux points des deux autres technologues par rapport à la technologue utilisée comme vérité terrain. Les figures suivantes montrent les courbes PR et ROC des modèles en comparaison avec les technologues pour les critères du Sein coupé CC et de l'IMA ouvert. Ces résultats ont été obtenus par un seul modèle sur une seule graine aléatoire.



(a) (b)

FIGURE 5.3 Comparaison des performances du modèle DenseNet-121 par rapport aux technologues pour le critère du **Sein coupé en CC**. (a) Courbes PR du modèle par rapport à chacune des technologues et (b) courbes ROC du modèle par rapport à chacune des technologues



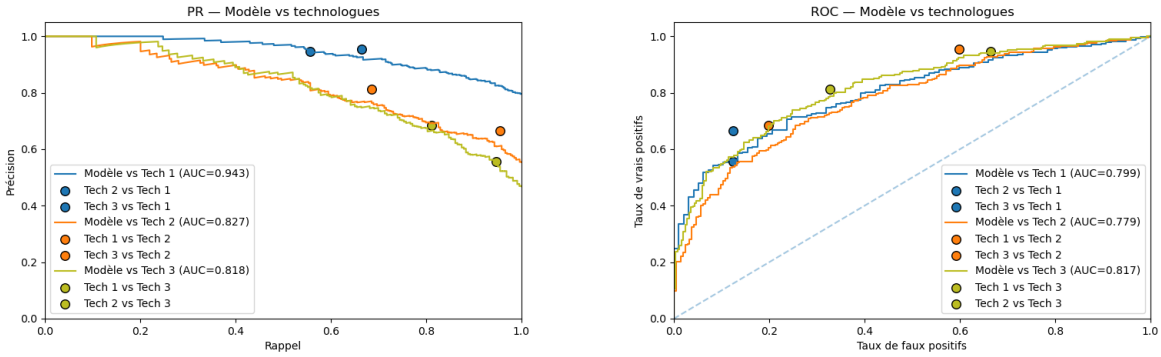
(a) (b)

FIGURE 5.4 Comparaison des performances du modèle DenseNet-121 par rapport aux technologues pour le critère de l'**IMA ouvert**. (a) Courbes PR du modèle par rapport à chacune des technologues et (b) courbes ROC du modèle par rapport à chacune des technologues

Pour ces deux critères, on constate que les courbes des modèles sont au même niveau ou dépassent les points des technologues, sauf pour un point pour le critère de l'IMA ouvert. On peut en conclure que les modèles ont atteint des performances comparables à celles d'un technologue expert. On remarque aussi une assez faible variabilité entre les trois courbes (avec une différence maximale de 6% d'AUC pour la PR AUC du critère Sein coupé en CC), ce qui indique que les modèles performent environ de la même manière, peu importe la vérité terrain considérée.

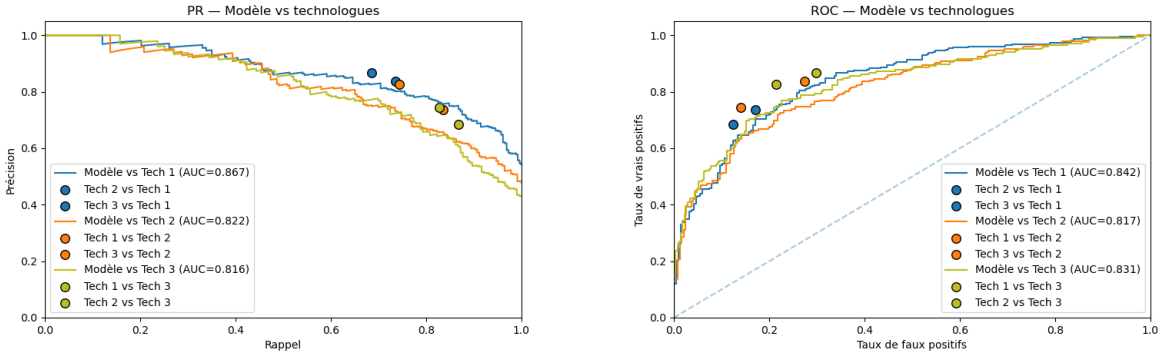
Tissus profonds : Pour ce critère, dans les deux vues, on obtient des valeurs de ROC et PR AUC dans les 80% (ROC AUC = 0,827, PR AUC = 0,885 pour la vue CC et ROC AUC = 0,848, PR AUC = 0,839 pour la vue MLO). Il n'y a pas de déséquilibre des classes pour ces critères, et on remarque même une légère prépondérance des cas positifs dans le jeu de données pour ce critère dans la vue CC. Par contre, comme mentionné dans le chapitre 4, plusieurs raisons peuvent expliquer la non satisfaction de ces critères, ce qui peut rendre l'apprentissage des modèles plus difficile. En effet, puisque les cas positifs peuvent être dus à plusieurs éléments différents des images, il s'agit de tâches plus exigeantes pour le modèle.

Les Figures 5.5 et 5.6 montrent les courbes PR et ROC des modèles en comparaison avec les technologues pour ces deux critères.



(a) (b)

FIGURE 5.5 Comparaison des performances du modèle DenseNet-121 par rapport aux technologues pour le critère des **Tissus profonds en CC**. (a) Courbes PR du modèle par rapport à chacune des technologues et (b) courbes ROC du modèle par rapport à chacune des technologues



(a) (b)

FIGURE 5.6 Comparaison des performances du modèle DenseNet-121 par rapport aux technologues pour le critère des **Tissus profonds en MLO**. (a) Courbes PR du modèle par rapport à chacune des technologues et (b) courbes ROC du modèle par rapport à chacune des technologues

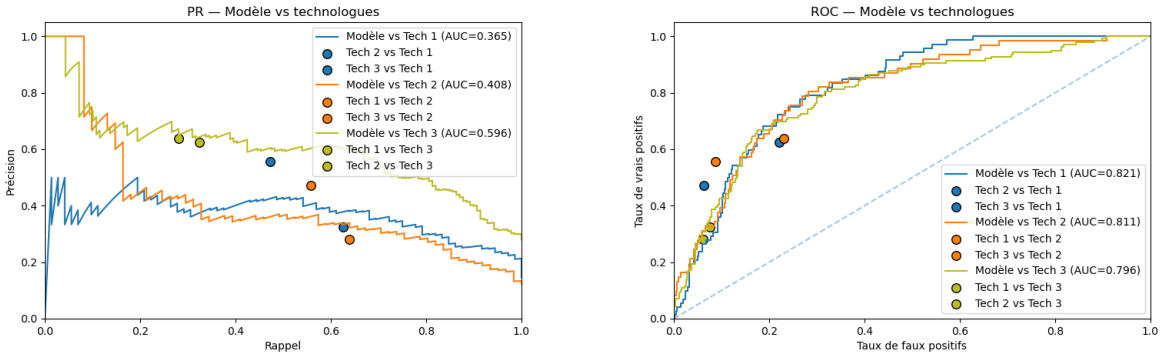
On remarque que les points de comparaison avec les autres technologues se trouvent toujours au-dessus des courbes des modèles. Cela signifie que les performances des modèles n'atteignent pas tout à fait celles que pourrait avoir un quatrième technologue. Les écarts entre les AUCs des courbes du modèle par rapport aux annotations de chaque technologue sont d'au plus 0,12 pour la PR AUC du critère en CC. Pour ce critère, le modèle semble être plus aligné avec les annotations de la technologue 1. Pour le critère en MLO, le plus grand écart obtenu

entre deux courbes est de 0,051, ce qui indique que le modèle performe de façon similaire par rapport à chacune des technologues.

Sein coupé en MLO et Orientation du mamelon : Le critère du Sein coupé en MLO est celui qui obtient les performances les plus faibles (ROC AUC = 0,868, PR AUC = 0,513). L'écart de performances avec le même critère en CC peut être expliqué en partie par le plus grand déséquilibre des classes présent, avec seulement 13 à 14% d'images de la classe positive dans les ensembles d'entraînement et de test pour le critère en MLO. La variabilité entre les annotations des technologues est aussi plus prononcée pour le critère dans cette vue. Ces raisons peuvent expliquer la difficulté du modèle à apprendre une représentation adéquate de l'image. Cependant, considérant que pour ce critère, l'ensemble de test ne contient que 14,2% d'images de la classe positive, la valeur de PR AUC obtenue indique que le modèle performe beaucoup mieux qu'un modèle aléatoire, pour lequel une PR AUC de seulement 0,14 devrait être obtenue.

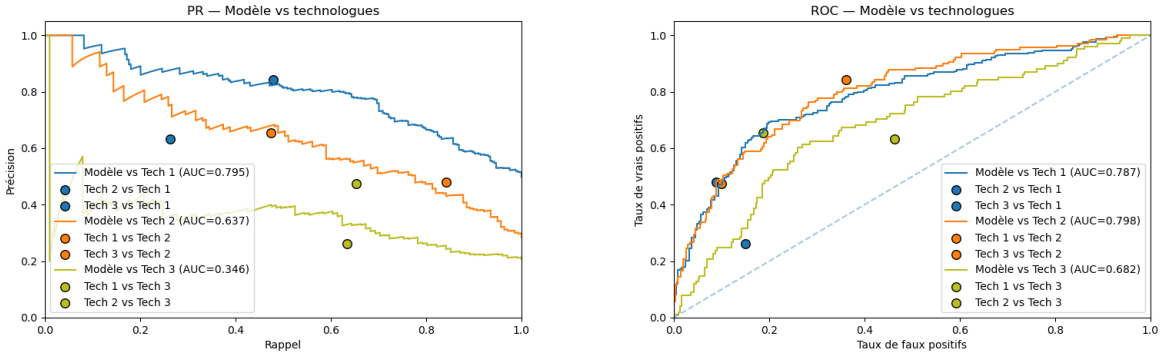
Pour le critère de l'Orientation du mamelon, on obtient une ROC AUC de 0,801 et une PR AUC de 0,641. Ce critère présente un déséquilibre des classes de 28 à 30% pour les ensembles d'entraînement et de test, un taux d'accord de seulement 0,52 et un kappa de Fleiss de 0,27, le plus bas pour les critères considérés dans ce mémoire. Aucune valeur d'angle n'est clairement indiquée dans les lignes directrices des standards de positionnement pour l'évaluation de ce critère, et d'après nos discussions avec les technologues participant au projet, une subjectivité demeure lors de son évaluation. En contexte clinique, les technologues se fient aussi beaucoup aux examens antérieurs et à la morphologie des patient.e.s, considérant que beaucoup d'entre eux peuvent naturellement avoir les mamelons désorientés. Le modèle ne prend pas en compte la morphologie des patientes, ce qui peut expliquer les résultats plus faibles que pour d'autres critères avec un déséquilibre des classes similaire. Malgré les faibles résultats, les prédictions effectuées par le modèle sont meilleures que des prédictions aléatoires, pour lesquelles une PR AUC d'environ 0,30 aurait été obtenue.

Les figures suivantes montrent les courbes PR et ROC des modèles en comparaison avec les technologues pour ces deux critères.



(a) (b)

FIGURE 5.7 Comparaison des performances du modèle DenseNet-121 par rapport aux technologues pour le critère du **Sein coupé en MLO**. (a) Courbes PR du modèle par rapport à chacune des technologues et (b) courbes ROC du modèle par rapport à chacune des technologues



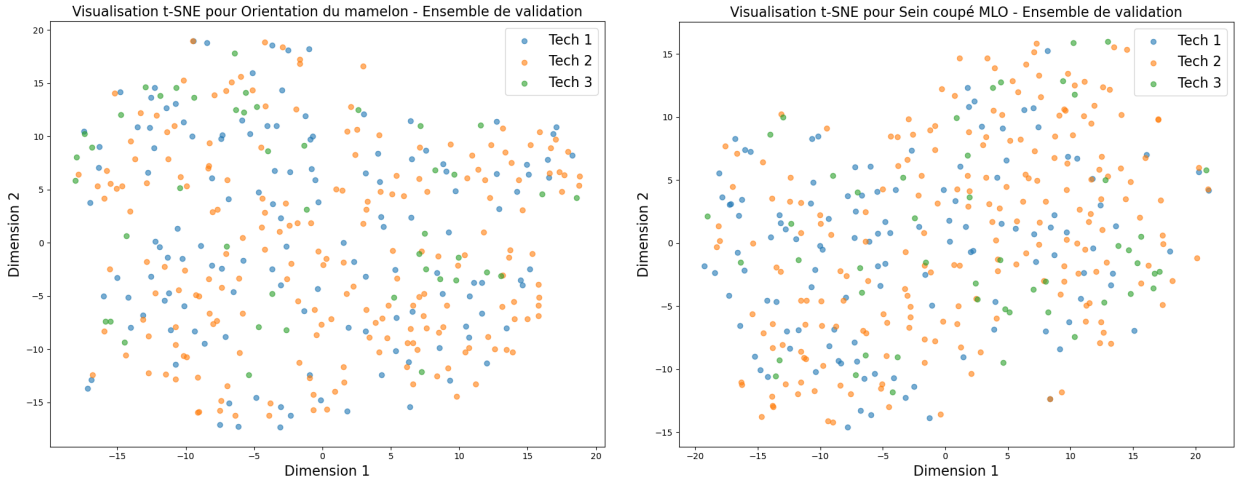
(a) (b)

FIGURE 5.8 Comparaison des performances du modèle DenseNet-121 par rapport aux technologues pour le critère de l'**Orientation du mamelon**. (a) Courbes PR du modèle par rapport à chacune des technologues et (b) courbes ROC du modèle par rapport à chacune des technologues

Ces deux critères présentent les plus grandes variations entre les courbes. En effet, ces critères sont ceux présentant les déséquilibres de classes les plus importants et les plus grandes variabilités inter-technologues. Les écarts entre les courbes (allant de 0,10 à 0,45 pour les PR AUCs) indiquent que les modèles semblent mieux performer par rapport aux annotations de certaines technologues. Pour le critère de l'Orientation du mamelon, en comparant le modèle aux annotations de la technologue 1 comme vérité terrain, une PR AUC de 0,795

est obtenue. Les points des deux autres technologues par rapport à la technologue 1 sont en-dessous ou au même niveau que la courbe, ce qui indique que le modèle performe aussi bien qu'un quatrième technologue. Par contre, les deux autres courbes sont bien en-dessous de la première (PR AUCs de 0,637 et de 0,346), et les points de comparaison avec les autres technologues sont à la fois au-dessus et en-dessous des courbes. Ceci révèle de grands écarts de performances du modèle selon la vérité terrain utilisée, et met en évidence la grande variabilité existante entre les annotations. Le même phénomène peut être observé à la Figure 5.7 pour le critère Sein coupé en MLO : le modèle semble bien performer par rapport aux annotations de la technologue 3, alors que les AUCs par rapport aux deux autres technologues sont beaucoup plus faibles. Ces résultats portent à croire que le manque d'exemples positifs dans le jeu de données, combiné avec la présence de variabilité dans les annotations, rend l'apprentissage difficile pour les modèles.

La grande différence entre les performances des modèles avec chacune des annotatrices comme vérité terrain pour ces deux critères pourrait laisser croire que les caractéristiques obtenues en sortie des modèles sont dépendantes de la technologue ayant annoté l'image. Afin de vérifier cela, une réduction de dimensionnalité par la méthode t-SNE est utilisée pour visualiser la structure des données d'un bloc de validation. Un graphe t-SNE est une représentation visuelle obtenue par projection en deux dimensions des vecteurs de caractéristiques associés aux images. La Figure 5.9 présente les graphes t-SNE des caractéristiques d'un ensemble de validation pour les critères Orientation du mamelon et Sein coupé en MLO. La couleur d'un point correspond à la technologue ayant annoté l'image.



(a) (b)

FIGURE 5.9 Graphes t-SNE pour un ensemble de validation pour les critères (a) Orientation du mamelon et (b) Sein coupé en MLO

Cette figure permet de constater que les vecteurs de caractéristiques ne sont pas regroupés en agrégats par technologues. Les encodeurs des modèles ne semblent donc pas avoir appris à classer les images selon les technologues qui les ont annotées. À la lumière de ces résultats, il est possible d'expliquer les performances plus faibles pour ces deux critères par le déséquilibre des classes et la grande variabilité inter-opérateurs présente dans les annotations.

5.3 Modèles multi-tâches

Pour répondre à la première question de recherche (Q1), plusieurs combinaisons de tâches ont été testées parmi les critères considérés dans ce mémoire. Dans la vue CC, les critères Sein coupé, Tissus profonds et Orientation du mamelon ont été retenus à cause des liens intrinsèques et des corrélations existant entre eux. Ces trois critères ont été combinés par paires et les trois ensemble pour tester les différents modèles multi-tâches présentés au chapitre 4. Dans la vue MLO, les critères Sein coupé, Tissus profonds et IMA ouvert ont été retenus. Comme pour les critères de la vue CC, ces critères ont été combinés par paires et les trois ensemble. Quatre modèles ont donc été entraînés pour chacune des architectures multi-tâches de base présentées (trois modèles pour les combinaisons de deux tâches et un modèle pour la combinaison de trois tâches).

Le Tableau 5.2 présente les résultats de PR et ROC AUC pour les différentes combinaisons de critères dans la vue CC et pour les trois modèles multi-tâches de base, soit MT HPS, MT SPS et ML. Afin d'alléger le texte, les acronymes suivants sont utilisés dans le tableau : Sein coupé - SC, Tissus profonds - TP et Orientation du mamelon - OM.

TABLEAU 5.2 Résultats de ROC AUC et de PR AUC des trois modèles multi-tâches de base pour les différentes combinaisons des critères de la vue CC. Moyennes obtenues sur trois graines aléatoires sur l'ensemble de test (ensemble d'images B, annotations majoritaires). **En gras** : meilleurs résultats, souligné : les deuxième et troisième meilleurs résultats.

Modèle	Sein coupé CC		Tissus profonds CC		Orientation du mamelon	
	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC
DenseNet-121	0,948	0,919	0,827	0,885	0,801	0,641
MT HPS : SC-TP	0,949	**0,924**	0,824	0,891		
MT HPS : SC-OM	0,944	0,909			0,774	0,599
MT HPS : TP-OM			0,815	0,874	0,775	0,618
MT HPS : SC-TP-OM	0,942	0,910	**0,850**	**0,902**	**0,803**	0,642
MT SPS : SC-TP	**0,951**	0,923	0,836	0,896		
MT SPS : SC-OM	0,946	0,918			0,773	0,611
MT SPS : TP-OM			0,824	0,880	0,802	**0,644**
MT SPS : SC-TP-OM	0,936	0,902	0,821	0,879	0,792	0,627
ML : SC-TP	0,948	0,916	0,838	0,896		
ML : SC-OM	0,939	0,903			0,773	0,608
ML : TP-OM			0,815	0,873	0,786	0,626
ML : SC-TP-OM	0,941	0,909	0,831	0,883	0,770	0,605

D'abord, on remarque que ce ne sont pas toutes les combinaisons de critères qui permettent d'améliorer les performances de classification par rapport aux *baselines*. Les combinaisons de critères qui semblent le mieux fonctionner sont Sein coupé et Tissus profonds, ainsi que les trois tâches de la vue CC ensemble (Sein coupé, Tissus profonds et Orientation du mamelon).

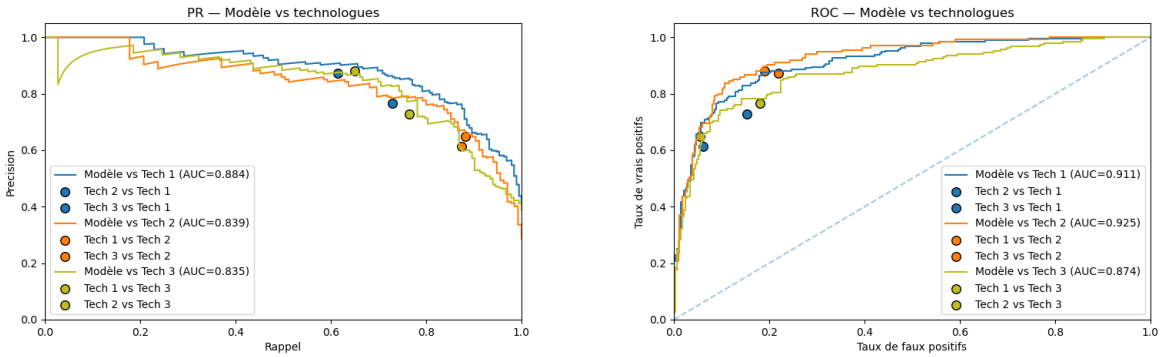
Combinaison SC-TP : Pour la combinaison de Sein coupé et Tissus profonds, on obtient une légère amélioration des AUCs pour les trois modèles multi-tâches par rapport aux *baselines*. Ces augmentations sont d'au plus 1,4% pour la PR AUC de Tissus profonds avec le modèle ML et le modèle MT SPS. Pour le modèle MT HPS, on remarque une amélioration des deux métriques pour le critère Sein coupé, et des valeurs semblables à celles de la *baseline* pour Tissus profonds.

Combinaison SC-TP-OM : Pour la combinaison des trois critères en CC, seul le modèle

MT HPS permet une amélioration des performances. Pour le critère Tissus profonds, des améliorations de 2,3% et 1,7% sont obtenus sur les AUCs par rapport à celles de la *baseline*. Pour les deux autres critères, on obtient des performances similaires à celles des *baselines*.

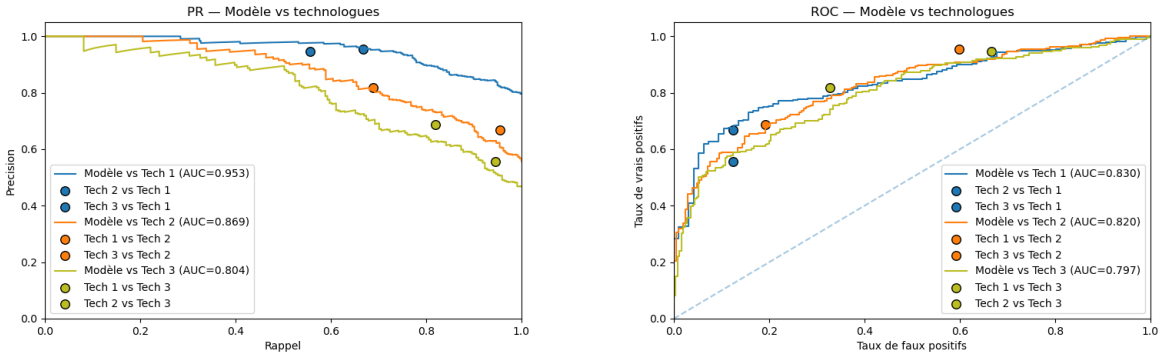
Orientation du mamelon : Pour ce critère, on observe une baisse de performances pour presque toutes les combinaisons de critères et tous les modèles, sauf pour le modèle MT HPS pour la combinaison des trois critères et le modèle MT SPS quand on le combine au critère Tissus profonds. Dans ce cas, on obtient des valeurs de PR et ROC AUC similaires à celles obtenues avec le modèle *baseline*.

Il est important de placer les résultats obtenus en perspective avec la variabilité inter-technologues. Les Figures 5.10, 5.11 et 5.12 présentent les graphes de comparaison entre le modèle MT HPS et chacune des technologues, et ce pour la combinaison des trois critères dans la vue CC. Ces graphes ont aussi été générés pour la combinaison des critères Sein coupé et Tissus profonds, et ceux-ci sont présentés dans l'annexe E.



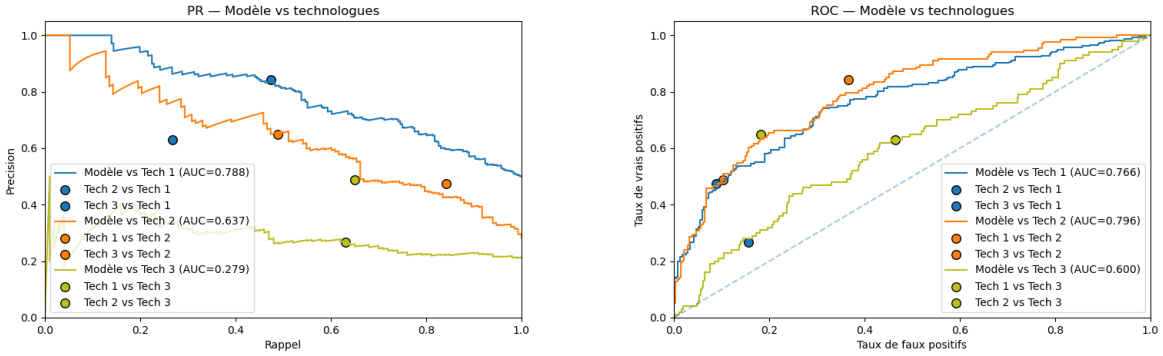
(a) (b)

FIGURE 5.10 Comparaison des performances du modèle MT HPS SC-TP-OM par rapport aux technologues pour le critère du **Sein coupé en CC**. (a) Courbes PR du modèle par rapport à chacune des technologues et (b) courbes ROC du modèle par rapport à chacune des technologues



(a) (b)

FIGURE 5.11 Comparaison des performances du modèle MT HPS SC-TP-OM par rapport aux technologues pour le critère des **Tissus profonds en CC**. (a) Courbes PR du modèle par rapport à chacune des technologues et (b) courbes ROC du modèle par rapport à chacune des technologues



(a) (b)

FIGURE 5.12 Comparaison des performances du modèle MT HPS SC-TP-OM par rapport aux technologues pour le critère de l'**Orientation du mamelon**. (a) Courbes PR du modèle par rapport à chacune des technologues et (b) courbes ROC du modèle par rapport à chacune des technologues

Sein coupé : Pour le critère du sein coupé, les performances du modèle par rapport à chacune des technologues semblent similaires à celles du modèle *baseline*. Dans les deux cas, tous les points des technologues se trouvent en-dessous ou au même niveau que la courbe des modèles, ce qui signifie que le modèle a atteint des performances comparables à celles d'un technologue.

Tissus profonds : Pour le critère des tissus profonds avec le modèle *baseline*, on obtenait des

courbes toutes légèrement en-dessous des points correspondants. Or, avec le modèle MT HPS, les courbes du modèle par rapport à la technologue 1 atteignent et même dépassent les points des technologues, ce qui révèle une amélioration du modèle par rapport aux annotations de cette technologue. En utilisant les annotations des deux autres technologues comme vérités terrain, on remarque que le modèle n'atteint pas tout à fait les performances des technologues, similairement au modèle *baseline*. Par contre, les ROC et PR AUCs augmentent de quelques points par rapports aux courbes obtenues avec le modèle *baseline* en utilisant la technologue 2 comme vérité terrain, ce qui indique une amélioration du modèle.

Orientation du mamelon : En ce qui concerne le critère de l'Orientation du mamelon, le modèle semble être plus en accord avec les technologues 1 et 2, pour lesquelles les courbes du modèle atteignent presque toutes les points des autres technologues. La courbe de la technologue 3 est inférieure aux autres, et elle n'atteint pas les deux points des autres technologues. Environ les mêmes observations pouvaient être faites avec le modèle *baseline*, ce qui indique que l'apprentissage multi-tâches n'apporte pas de bénéfice réel pour l'évaluation de ce critère, qui présente une grande variabilité inter-technologues.

Ces observations permettent de constater une réelle amélioration pour l'évaluation du critère Tissus profonds en CC en utilisant le modèle MT HPS, tout en conservant des performances semblables aux modèles *baselines* pour les deux autres critères.

Le tableau 5.3 présente les résultats de PR et ROC AUC pour les différentes combinaisons de critères dans la vue MLO et pour les trois modèles multi-tâches de base. Afin d'alléger le texte, les acronymes suivants sont utilisés dans le tableau : Sein coupé - SC, Tissus profonds - TP et IMA ouvert - IO.

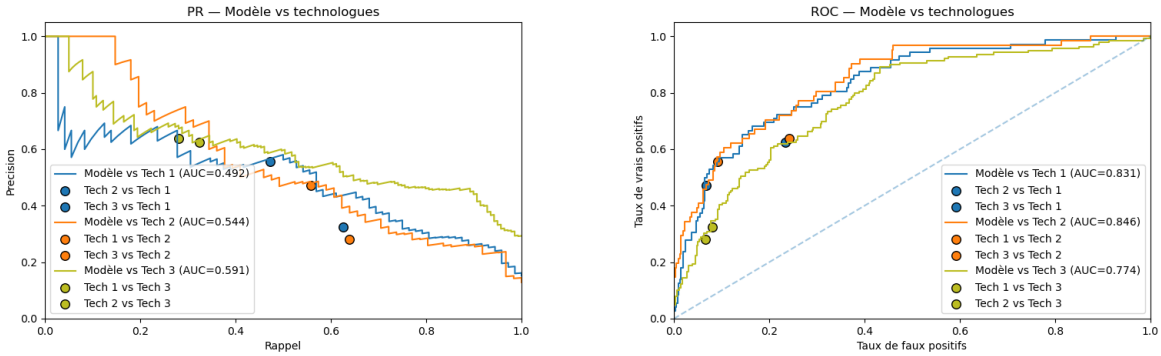
TABLEAU 5.3 Résultats de ROC AUC et de PR AUC des trois modèles mutli-tâches de base pour les différentes combinaisons des critères de la vue MLO. Moyennes obtenues sur trois graines aléatoires sur l'ensemble de test (ensemble d'images B, annotations majoritaires). **En gras** : meilleurs résultats, souligné : les deuxième et troisième meilleurs résultats.

Modèle	Sein coupé MLO		Tissus profonds MLO		IMA ouvert	
	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC
DenseNet-121	0,868	0,513	0,848	0,839	**0,940**	**0,900**
MT HPS : SC-TP	**0,875**	**0,577**	**0,861**	0,842		
MT HPS : SC-IO	0,839	0,480			0,926	0,875
MT HPS : TP-IO			0,838	0,813	0,909	0,842
MT HPS : SC-TP-IO	0,840	0,458	0,844	0,816	0,903	0,827
MT SPS : SC-TP	0,848	0,498	0,855	**0,846**		
MT SPS : SC-IO	0,827	0,479			0,933	0,876
MT SPS : TP-IO			0,844	0,822	0,908	0,839
MT SPS : SC-TP-IO	0,818	0,429	0,838	0,820	0,917	0,860
ML : SC-TP	0,854	0,490	0,821	0,788		
ML : SC-IO	0,853	0,546			0,926	0,873
ML : TP-IO			0,841	0,822	0,915	0,851
ML : SC-TP-IO	0,847	0,493	0,840	0,820	0,906	0,842

Pour les critères de la vue MLO, on observe une amélioration des performances seulement pour la combinaison des critères Sein coupé et Tissus profonds. En effet, pour le modèle MT HPS et cette combinaison de critères, on obtient une augmentation de 6,4% de la PR AUC et de 0,7% de la ROC AUC par rapport à la *baseline* pour Sein coupé, et une augmentation de 2,3% de la PR AUC et de 0,3% de la PR AUC pour Tissus profonds. Les combinaisons de critères incluant l'IMA ouvert ne semblent pas révéler de liens, les performances pour ces combinaisons restant en-dessous de celles obtenues pour le modèle *baseline*. Comme pour les critères en CC, les modèles MT SPS et ML offrent généralement de moins bons résultats que le modèle MT HPS.

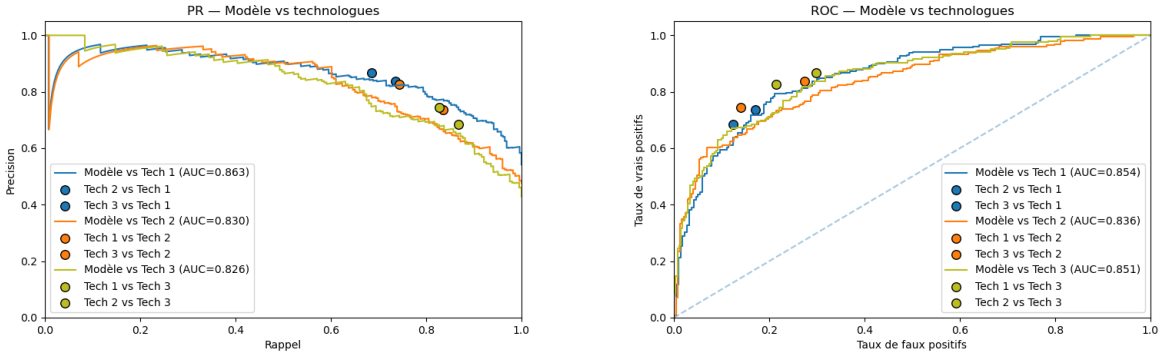
Les Figures 5.16 et 5.14 présentent les graphes de comparaison du modèle MT HPS avec les annotations des différentes technologues pour la combinaison des critères Sein coupé et

Tissus profonds.



(a) (b)

FIGURE 5.13 Comparaison des performances du modèle MT HPS SC-TP MLO par rapport aux technologues pour le critère du **Sein coupé en MLO**. (a) Courbes PR du modèle par rapport à chacune des technologues et (b) courbes ROC du modèle par rapport à chacune des technologues



(a) (b)

FIGURE 5.14 Comparaison des performances du modèle MT HPS SC-TP MLO par rapport aux technologues pour le critère des **Tissus profonds en MLO**. (a) Courbes PR du modèle par rapport à chacune des technologues et (b) courbes ROC du modèle par rapport à chacune des technologues

Sein coupé : La Figure 5.16 montre que tous les points des technologues sont maintenant sous leur courbe respective du modèle, ce qui indique que le modèle a atteint les performances des technologues. Ce n'était pas le cas du modèle *baseline* pour ce critère, ce qui implique une réelle amélioration apportée par le modèle MT HPS et la combinaison avec le critère Tissus profonds.

Tissus profonds : Pour ce critère, les courbes du modèle restent en-dessous des points de technologues, comme pour le modèle *baseline*. Toutefois, on remarque une augmentation des PR et ROC AUCs par rapport à chacune des technologues. Cela indique que, même si le modèle n'atteint pas tout à fait les performances d'un technologue, il y a une amélioration par rapport au modèle *baseline*. Ces résultats démontrent que la combinaison des critères Sein coupé et Tissus profonds dans la vue MLO est bénéfique pour l'apprentissage du modèle.

Modules d'attention et comparaison avec la littérature

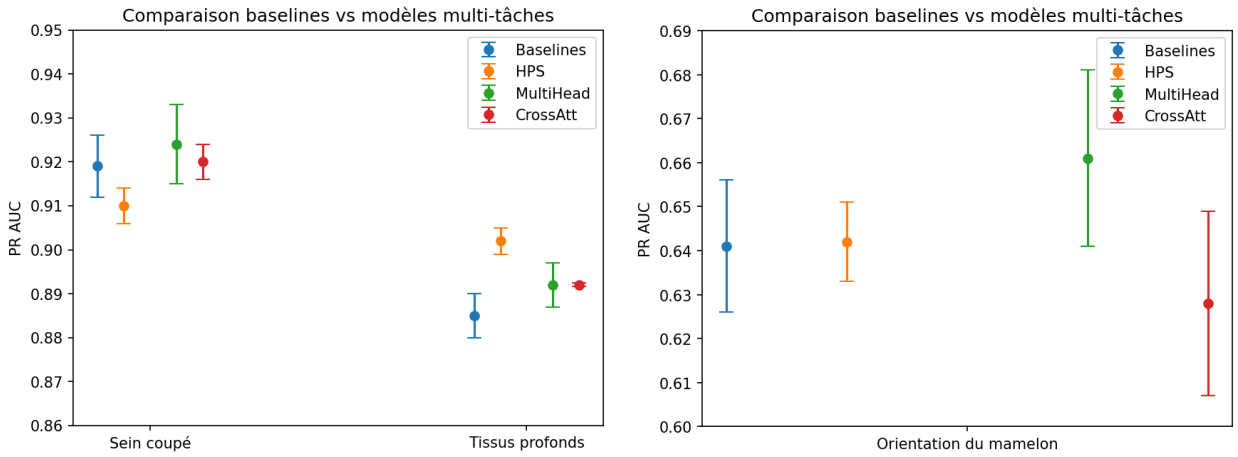
Les modèles variants du modèle MT HPS ont été entraînés et testés pour les combinaisons de tâches suivantes : Sein coupé, Tissus profonds et Orientation du mamelon en CC, et Sein coupé et Tissus profonds en MLO. Ces combinaisons de critères ont été choisies puisque ce sont celles qui offrent les meilleures performances avec les modèles multi-tâches de base. Le tableau 5.4 présente les résultats des modèles MT MultiHead, MT CrossAtt ainsi que ceux de deux modèles de la littérature, soient celui de Nguyen et al. [86] et le MTAN [74], pour la combinaison des trois critères dans le vue CC. Les modèles tirés de la littérature ont été reproduits et entraînés sur notre jeu de données. Les résultats des modèles *baselines* ainsi que du modèle MT HPS sont aussi inclus dans ce tableau à titre de comparaison.

Tableau 5.4 Résultats de ROC AUC et de PR AUC des modèles multi-tâches avec attention et des modèles de la littérature pour la combinaison des critères **SC-TP-OM (CC)**. Moyennes obtenues sur trois graines aléatoires sur l'ensemble de test (ensemble d'images B, annotations majoritaires). **En gras** : meilleurs résultats, souligné : les deuxième et troisième meilleurs résultats.

	Sein coupé CC		Tissus profonds CC		Orientation du mamelon	
	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC
Baseline	0,948	0,919	0,827	0,885	0,801	0,641
MT HPS	0,942	0,910	**0,850**	**0,902**	0,803	0,642
MT MultiHead	**0,950**	**0,924**	0,835	0,892	**0,814**	**0,661**
MT CrossAtt	0,947	0,920	0,830	0,892	0,781	0,628
Nguyen et al.	0,941	0,909	0,823	0,885	0,758	0,596
MTAN	0,883	0,824	0,768	0,844	0,621	0,452

D'après le tableau 5.4, on observe une amélioration des performances par rapport aux modèles *baelines* avec le modèle MT MultiHead pour la combinaison des trois critères de la vue CC.

Les performances sont aussi supérieures à celles du modèle MT HPS pour les critères Sein coupé et Orientation du mamelon. Les modèle MT CrossAtt permet aussi une augmentation des performances par rapport aux modèles *baselines*, mais seulement pour les critères Sein coupé et Tissus profonds. Pour le critère de l'Orientation du mamelon, les résultats restent proches de ceux obtenus pour le modèle *baselines*. Il est important de considérer que les résultats présentés sont une moyenne des résultats de modèles ayant été entraînés sur trois graines aléatoires différentes. La figure suivante montre les résultats de AUC PR des modèles *baseline*, MT HPS, MT MultiHead et MT CrossAtt avec leurs écarts-types.



(a) (b)

FIGURE 5.15 Comparaison des résultats des modèles *baselines* et des modèles multi-tâches avec leurs écarts-types. (a) Résultats pour les critères Sein coupé et Tissus profonds et (b) résultats pour le critère Orientation du mamelon

Pour les critères Sein coupé et Orientation du mamelon, les écarts-types des résultats pour les modèles multi-tâches chevauchent tous les résultats des modèles *baselines*, ce qui confirme que les performances obtenues pour ces critères sont équivalentes à celles des modèles entraînés en tâches uniques. Pour ce qui est des résultats pour le critère des Tissus profonds, seul le modèle MT HPS semble offrir une performance significativement supérieure à celle du modèle *baseline*. De manière générale, le modèle MT HPS offre des performances plus robustes à l'initialisation (écarts-types moins grands) que les modèles intégrant des mécanismes d'attention. Cela peut s'expliquer par une plus grande sensibilité à l'initialisation de modèles plus expressifs, comme c'est le cas des modèles MT MultiHead et MT CrossAtt. L'augmentation du nombre de paramètres de ces modèles par rapport au modèle MT HPS peut aussi affecter leur stabilité à l'entraînement. De plus, le modèle HPS semble offrir le meilleur compromis de performance pour les trois critères de la vue CC.

Le tableau 5.5 présente les résultats des modèles MT MultiHead, MT CrossAtt ainsi que ceux des deux modèles de la littérature pour la combinaison des critères Sein coupé et Orientation du mamelon dans la vue MLO.

Tableau 5.5 Résultats de ROC AUC et de PR AUC des modèles multi-tâches avec attention et des modèles de la littérature pour la combinaison de critères **SC-TP (MLO)**. Moyennes obtenues sur trois graines aléatoires sur l'ensemble de test (ensemble d'images B, annotations majoritaires). **En gras** : meilleurs résultats, souligné : les deuxième et troisième meilleurs résultats.

	Sein coupé MLO		Tissus profonds MLO	
	ROC AUC	PR AUC	ROC AUC	PR AUC
Baseline	0,868	0,513	0,848	0,839
MT HPS	0,875	**0,577**	0,861	0,842
MT MultiHead	**0,876**	0,527	**0,863**	**0,855**
MT CrossAtt	0,855	0,513	0,855	0,850
Nguyen et al.	0,878	0,549	0,854	0,840
MTAN	0,754	0,324	0,774	0,747

On constate que le modèle MT MultiHead permet une amélioration des performances pour les deux critères par rapport aux modèles baselines. Cependant, les performances en termes d'AUC PR pour le critère du sein coupé demeurent plus faibles que celles obtenues avec le modèle MT HPS. Pour ce qui est du modèle MT CrossAtt, les performances obtenues avec ce modèle restent proches de celles des modèles *baselines*. La figure suivante montre les résultats de AUC PR des modèles *baseline*, MT HPS, MT MultiHead et MT CrossAtt avec leurs écarts-types.

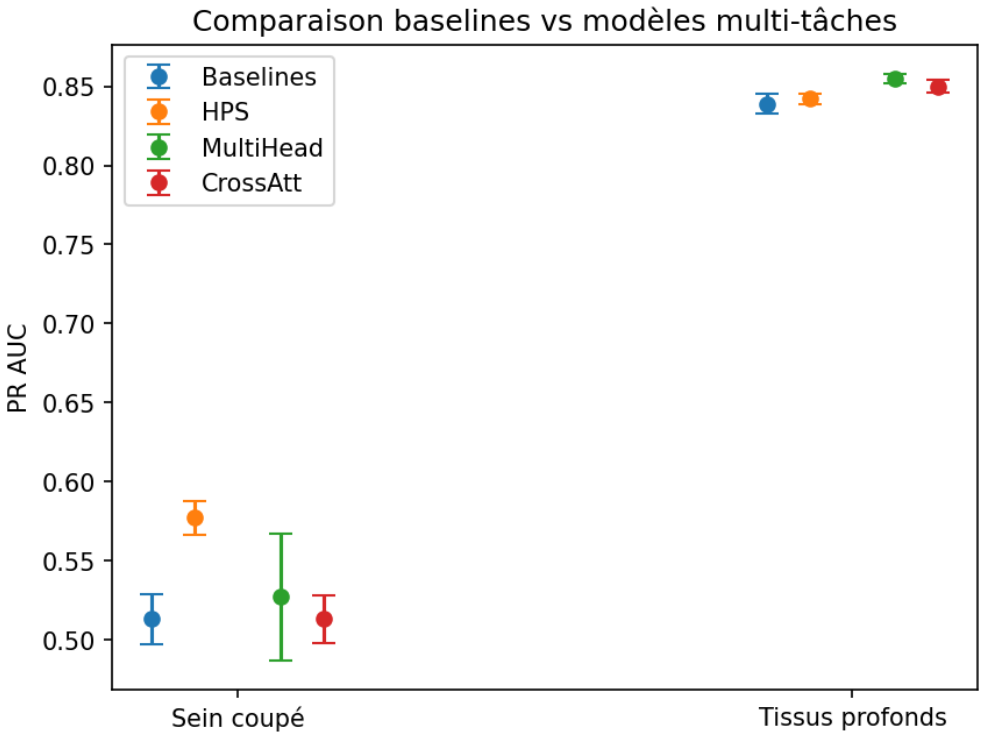


FIGURE 5.16 Comparaison des résultats des modèles *baselines* et des modèles multi-tâches avec leurs écarts-types pour la combinaison des critères Sein coupé et Tissus profonds dans la vue MLO.

D'abord, on constate que, tout comme les résultats de la combinaison des trois critères de la vue CC, les écarts-types des modèles MT MultiHead et MT CrossAtt sont en général plus grands que ceux des deux autres modèles, ce qui démontre leur plus grande instabilité. Le modèle MT HPS est celui qui offre la meilleure valeur d'AUC PR pour le critère Sein coupé. Ce résultat semble significatif puisqu'il n'y a pas de chevauchement avec le résultat pour le modèle *baseline*. Pour le critère des Tissus profonds, les valeurs d'AUC PR se trouvent toutes autour de 0,840 et ont des écarts-types beaucoup plus petits. Le modèle qui offre le meilleur compromis entre les performances des deux critères est le modèle MT HPS, en plus d'être plus robuste à l'initialisation des poids. De par ces observations, il est possible de souligner que les représentations apprises par le modèle en évaluant le critère des Tissus profonds semblent bénéfiques à l'apprentissage pour le critère Sein coupé.

En ce qui concerne les modèles de la littérature, pour les combinaisons de critères dans les deux vues, le modèle de Nguyen et al. obtient des performances en-dessous de celles des modèles *baselines* et du modèle MT HPS. L'écart de performances avec le modèle MT HPS n'est toutefois pas très marqué, ce qui est cohérent puisque les architectures des deux modèles sont

similaires. Enfin, le modèle MTAN est celui qui obtient les valeurs de PR et de ROC AUCs les plus faibles. Dans l'article présentant ce modèle, les gains de performances pour les combinaisons de tâches de classification étaient de plus en plus marqués avec l'augmentation du nombre de tâches effectuées en parallèle. Pour deux tâches, les performances obtenues étaient légèrement inférieures à celles obtenues avec un entraînement en tâche unique. Ceci est cohérent avec les résultats de cette expérimentation, puisque nous ne testons que l'entraînement de deux tâches et de trois tâches à la fois.

Les détails concernant les hyperparamètres utilisés pour l'entraînement des modèles multi-tâches discutés dans cette section se trouvent à l'annexe D.

Discussion générale

L'analyse des résultats des modèles multi-tâches permet de faire ressortir certains éléments. D'abord, on constate que l'apprentissage multi-tâches semble fonctionner pour certaines combinaisons de tâches seulement. En effet, seules les combinaisons de critères ayant de véritables liens entre eux ont semblé bénéfiques à l'apprentissage. La corrélation existante entre les cas positifs des critères Tissus profonds et IMA ouvert dans la vue MLO ne s'est pas avérée correspondre à un lien intrinsèque entre ces critères. Ceci concorde avec le graphe des interdépendances des critères présenté au chapitre 2, qui n'indiquait aucun lien entre ces deux critères. Les combinaisons de critères fonctionnant le mieux pour l'apprentissage multi-tâches sont plutôt les suivantes :

1. Sein coupé et Tissus profonds dans la vue CC
2. Sein coupé, Tissus profonds et Orientation du mamelon dans la vue CC
3. Sein coupé et Tissus profonds dans la vue MLO

Pour ces combinaisons de tâches, le modèle MT HPS est celui qui semble le mieux fonctionner ; celui-ci permet une amélioration des performances pour toutes les combinaisons de tâches ci-dessus. Les représentations des images apprises semblent donc être pertinentes pour toutes les tâches à la fois, au même niveau que les représentations apprises lors des entraînements en tâches uniques. Cependant, il est important de noter que les améliorations de performances observées par rapport aux modèles *baselines* ne sont pas très grandes et ne semblent pas toutes significatives lorsque l'on compare les performances des modèles par rapport aux différentes technologues annotatrices. On peut donc conclure que l'apprentissage multi-tâches peut fonctionner pour certaines combinaisons de tâches, mais n'est pas nécessairement supérieur aux modèles en tâches uniques pour l'évaluation des critères de positionnement.

L'ajout de mécanismes d'attention au modèle MT HPS a aussi permis d'améliorer les performances quant à l'évaluation des critères par rapport aux modèles baselines. Par contre, ces modèles sont moins robustes et ne permettent pas d'atteindre des performances significativement meilleures que les modèles *baselines* en considérant les écarts-types des résultats. Le modèle MT HPS semble demeurer celui qui offre le meilleur compromis entre les performances de toutes les tâches effectuées simultanément.

L'utilisation du modèle MT HPS pour l'évaluation des critères de positionnement Sein coupé CC, Tissus profonds CC, Orientation du mamelon, Sein coupé MLO et Tissus profonds MLO serait bénéfique en contexte clinique puisqu'il s'agit d'un modèle léger. En effet, en utilisant ce modèle, il est possible d'évaluer deux ou trois critères à la fois, au lieu de devoir exécuter un modèle par critère. Puisque ce modèle comprend un seul encodeur pour l'ensemble des critères, celui-ci est aussi moins lourd que le modèle MT SPS, qui comprend un encodeur par critère. Le tableau 5.6 présente, à titre indicatif, la taille des modèles MT HPS et MT SPS ainsi que leur temps d'inférence pour une image sur un GPU 4090 et sur un CPU.

Tableau 5.6 Tailles des modèles et temps d'inférence pour les modèles MT HPS et MT SPS pour l'évaluation de deux critères. Le temps d'inférence considère un modèle ensembliste, donc l'inférence de cinq modèles sur une image.

	Nombre de paramètres (M)	Temps d'inférence sur GPU (ms)	Temps d'inférence sur CPU (ms)
MT HPS	6,69	3,5	199
MT SPS	13,9	5,5	300

Un des besoins en contexte clinique serait d'avoir une rétroaction dans la minute qui suit l'acquisition de l'image afin de pouvoir corriger directement le positionnement et reprendre les clichés mammaires au besoin. De plus, les ordinateurs utilisés en clinique n'ont pas forcément des cartes graphiques aussi performantes que celle utilisée pour tester les modèles. Un temps CPU de 199 ms est amplement acceptable pour une utilisation des modèles en clinique.

Finalement, bien que dans ce mémoire, les métriques présentées pour évaluer les modèles soient indépendantes de seuils, il est possible d'appliquer des seuils sur les prédictions en sortie des modèles pour obtenir une classification binaire des images. En contexte clinique, un seuil variable pourrait être appliqué pour rendre les modèles plus ou moins sévères, en fonction du contexte d'utilisation des modèles ; par exemple, selon le niveau d'expérience du technologue qui l'utilise, ou si l'on se trouve en contexte de formation de nouveaux technologues.

5.4 Simulation de réduction de dose et robustesse au bruit

Afin de répondre à la seconde question de recherche (Q2), les modèles *baselines* ainsi que les meilleurs modèles multi-tâches ont été testés sur des images simulées en basse dose de rayons X. Les tableaux 5.7 et 5.8 présentent les résultats de l'inférence des modèles sur les images bruitées avec différentes valeurs de variance.

Tableau 5.7 Résultats de ROC AUC et de PR AUC des modèles ***baselines*** sur l'ensemble de test (annotations majoritaires) avec l'ajout de bruit de Poisson. Résultats obtenus sur une seule graine aléatoire.

	Sein coupé CC		Tissus profonds CC		Orientation du mamelon		Sein coupé MLO		Tissus profonds MLO		IMA ouvert	
	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC
Normal	0,948	0,921	0,829	0,882	0,811	0,662	0,855	0,494	0,846	0,836	0,943	0,910
$\sigma = 1$	0,948	0,919	0,826	0,880	0,811	0,659	0,860	0,508	0,848	0,842	0,940	0,907
$\sigma = 2$	0,945	0,916	0,824	0,878	0,816	0,668	0,869	0,530	0,854	0,848	0,941	0,906
$\sigma = 3$	0,944	0,902	0,820	0,876	0,815	0,653	0,866	0,545	0,851	0,846	0,937	0,902

Tableau 5.8 Résultats de ROC AUC et de PR AUC des modèles **MT HPS** sur l'ensemble de test (annotations majoritaires) avec l'ajout de bruit de Poisson. Résultats pour les combinaisons de critères **SC-TP-OM (CC)** et **SC-TP (MLO)**, et obtenus sur une seule graine aléatoire.

	Sein coupé CC		Tissus profonds CC		Orientation du mamelon		Sein coupé MLO		Tissus profonds MLO	
	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC
Normal	0,944	0,913	0,844	0,899	0,787	0,639	0,865	0,575	0,861	0,838
$\sigma = 1$	0,944	0,911	0,844	0,898	0,786	0,639	0,864	0,560	0,866	0,846
$\sigma = 2$	0,946	0,913	0,843	0,895	0,788	0,646	0,868	0,560	0,866	0,854
$\sigma = 3$	0,945	0,909	0,839	0,893	0,778	0,630	0,867	0,539	0,874	0,861

En général, l'ajout de bruit gaussien n'entraîne pas de baisses drastiques de performance des modèles. On remarque toutefois que les performances des modèles diminuent un peu plus avec l'ajout de bruit plus important. Pour les modèles *baselines* et pour le bruit maximal ajouté ($\sigma = 3$), on observe une diminution maximale de 2% pour la PR AUC et de 0,9% pour la ROC AUC, tous critères confondus. Pour les modèles multi-tâches, on obtient une diminution maximale de 3,6% pour la PR AUC pour le critère Sein coupé en MLO, et 0,5% de diminution maximale pour la ROC AUC pour le critère Tissus profonds en CC. De plus, on note aussi une augmentation des performances pour certains critères avec l'ajout de bruit gaussien, notamment pour Orientation du mamelon, Sein coupé en MLO et Tissus profonds en MLO. Cela pourrait être expliqué en partie par le fait que le bruit gaussien affûte certains éléments de l'image, ce qui pourrait être bénéfique pour l'évaluation de ces critères.

Le bruit gaussien correspond au bruit qu'on peut observer sur les images acquises en basse dose ; il s'agit donc du type de bruit qu'on retrouve dans les images *pre-shot*. Les résultats obtenus démontrent que les modèles développés, autant les *baselines* que les modèles MT HPS, pourraient être utilisés pour évaluer certains critères de positionnement sur des images *pre-shot* sans engendrer de baisse de performance significative par rapport à des images acquises à dose normale de rayons X. Cela démontre le potentiel de l'utilisation des images *pre-shot* pour la vérification du positionnement mammaire avant l'irradiation complète des patient.e.s.

CHAPITRE 6 CONCLUSION

Ce dernier chapitre conclut ce mémoire en présentant d'abord un résumé des travaux effectués et des résultats obtenus. Ensuite, les limitations de cette étude seront abordées, puis des recommandations et des pistes de travaux futurs seront présentées.

6.1 Synthèse des travaux

Ce travail de recherche a démarré avec la création d'un premier jeu d'images de mammographies (4773) annotées en terme de qualité du positionnement des seins par trois expertes technologues. Ces annotations multiples sur un sous-ensemble du jeu Mammo-Pos ont permis de quantifier la variabilité inter-technologue sur les critères étudiés et, au vue de la variabilité considérable, de souligner l'importance de considérer plusieurs sources d'annotations dans un tel jeu de données. L'étude de variabilité a permis, dans la suite du projet, de produire une balise pour l'interprétation des performances des modèles développés. De plus, le volume de données dans l'ensemble A du jeu Mammo-Pos a permis l'entraînement de différents modèles d'apprentissage profond supervisé pour la classification des images selon les différents critères de positionnement.

Moyennant Mammo-Pos, nous avons adressé la première question de recherche de ce travail, à savoir **Est-il possible de tirer profit des relations entre les critères à l'aide d'une approche par apprentissage multi-tâches afin d'améliorer l'évaluation automatique du positionnement mammaire ?**. Nous avons entraîné et comparé différentes architectures pour chacun des critères à l'étude, soit Sein coupé (CC et MLO), Tissus profonds (CC et MLO), Orientation du mamelon (CC) et IMA ouvert (MLO). Dans un premier temps, en guise de *baselines*, nous avons entrainé un modèle par critère, en prenant en considération le déséquilibre des classes inhérents à nos données. Ensuite, nous avons entraîné et comparé différentes approches multi-tâches et multi-labels, avec et sans modules d'attention, pour tirer profit des relations sémantiques entre les critères. Les résultats obtenus sur notre jeu de données démontrent que l'apprentissage multi-tâches peut fonctionner pour certaines combinaisons de critères de positionnement, mais sans largement dépasser les performances des modèles *baselines* entraînés en tâches uniques. Les combinaisons de critères fonctionnant le mieux sont Sein coupé et Tissus profonds en CC et MLO respectivement, et Sein coupé, Tissus profonds et Orientation du mamelon en CC.

Pour répondre à la seconde question de recherche, à savoir **Est-ce possible de vérifier le positionnement mammaire sur des images acquises à plus faible dose de rayons X, avant l'irradiation complète des patientes ?**, une méthode de simulation de réduction de dose a été appliquée aux images de mammographie de l'ensemble de test. Les résultats de l'inférence des modèles développés sur les images bruitées ont démontré une très faible diminution des performances, confirmant qu'il est possible d'effectuer l'évaluation du positionnement mammaire sur des images simulées en faible dose.

6.2 Limitations

Il est important de considérer les limitations présentes dans le cadre de ce travail. Plusieurs des limitations sont liées au jeu de données Mammo-Pos, tandis que d'autres sont liées directement à l'évaluation des critères dans un contexte d'automatisation.

Tout d'abord, nous identifions des limitations liées au jeu de données Mammo-Pos :

1. Lors de la phase d'annotations, les technologues participant au projet n'avaient pas accès à des écrans à haute résolution comme ceux qu'elles utilisent normalement en clinique. Leurs ordinateurs personnels ont été utilisés pour effectuer les annotations, ce qui a rendu leur expérience différente de ce à quoi elles sont habituées, et qui pourrait contribuer à la variabilité dans les annotations.
2. Certaines annotations demandées étaient différentes de celles que les technologues réalisent habituellement, ce qui pouvait parfois créer de la confusion.
3. La taille du jeu de données étant relativement petite, le nombre d'images disponibles pour l'entraînement des modèles était limité (environ 1500 images), contraignant les modèles pouvant être utilisés et les performances pouvant être attendues.

Ensuite, certaines limitations touchent davantage à l'approche d'entraînement considérée, à savoir par apprentissage supervisé. En effet, au vue du large déséquilibre des classes pour une grande partie des critères de positionnement, ainsi que la variabilité dans les annotations des technologues, les modèles peuvent être biaisés en faveur des classes majoritaires et présenter une sensibilité accrue aux incohérences d'annotation, ce qui limite leurs capacités de généralisation et peut affecter la robustesse de leurses modèles d'apprentissage profond utilisés dans ces travaux ne prennent en entrée que les images de mammographie, sans informations sur les mammographies antérieures ou la morphologie des patient.e.s. Or, en contexte réel, la morphologie des patient.e.s est très importante pour l'évaluation de certains critères de positionnement, notamment celui de l'Orientation du mamelon.

6.3 Recommandations et pistes de travaux futurs

Le projet présenté dans ce mémoire s'inscrit dans un projet plus large visant à développer un système automatique d'évaluation de la qualité du positionnement mammaire. Dans ce contexte, plusieurs travaux futurs peuvent être envisagés.

1. La poursuite de la phase d'annotation avec plus d'images de mammographies serait bénéfique afin d'augmenter la taille du jeu de données. Un jeu de données contenant plus d'images pourrait être utilisé pour entraîner de plus gros modèles et possiblement obtenir de meilleures performances.
2. L'annotation de nouvelles images demandant des ressources monétaires et temporelles, une autre solution aux limitations concernant le jeu de données pourrait être d'explorer des méthodes d'apprentissage peu ou non supervisé. De telles méthodes permettraient l'utilisation d'un grand nombre d'images à l'entraînement, sans avoir besoin d'annotations relatives au positionnement.
3. Prendre en compte la morphologie des patientes dans l'évaluation des critères de positionnement serait idéal afin de rapprocher le plus possible l'évaluation automatique du contexte clinique. Il serait intéressant d'intégrer certaines informations sous forme de métadonnées (informations de morphologie des patient.e.s, informations sur les mammographies antérieures, informations morphologiques) à l'entrée des modèles pour que celles-ci soient prises en compte lors de l'apprentissage.
4. Il serait intéressant de valider les images simulées à faible dose de rayons X à l'aide de véritables images *pre-shot* obtenues des manufacturiers. Ceci permettrait de valider que les modèles développés dans le cadre de ce projet peuvent performer sur de véritables images à plus faible dose, et ainsi permettre de limiter l'irradiation aux patientes.
5. La tomosynthèse mammaire est une technique d'imagerie permettant de reconstruire le sein en 3D par l'acquisition de multiples radiographies à différents angles. Il serait intéressant d'étudier comment l'utilisation de celle-ci en clinique pourrait influencer l'évaluation du positionnement mammaire. Dans un premier temps, les performances des modèles développés dans ce projet pourraient être évaluées sur des images synthétiques CC et MLO générées par tomosynthèse. Ensuite, de futurs travaux pourraient porter sur l'évaluation du positionnement mammaire à l'aide des mammographies et d'images complémentaires de tomosynthèse.

Outre ces pistes d'amélioration, la suite du projet consistera à intégrer ce travail avec ceux des deux autres étudiants travaillant sur le projet global. L'intégration des modèles élaborés et la création d'une interface utilisateur permettra de générer un prototype de solution utilisable en clinique, dans l'optique d'aider les technologues en mammographie dans l'évaluation des critères de positionnement mammaire et ainsi d'améliorer le dépistage du cancer du sein.

RÉFÉRENCES

[1] Institut National Du Cancer, "Anatomie du sein." [En ligne]. Disponible : https://www.cancer.fr/personnes-malades/les-cancers/sein/comprendre-les-cancers-du-sein/anatomie-du-sein

[2] D. Gurve, V. Gauvin, L. Grondin-Reiher, V. Patenaude, K. Morency, N. Lasry, F. Guibault et L. Séoud, "Breast positioning in mammography : consolidated image criteria, their interdependence and trends in automatic evaluation," manuscript submitted for publication.

[3] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017. [En ligne]. Disponible : https://arxiv.org/abs/1706.05098

[4] L. R. Borges, H. C. Oliveira, P. F. Nunes, P. R. Bakic, A. D. Maidment et M. A. Vieira, "Method for simulating dose reduction in digital mammography using the anscombe transformation," *Medical Physics*, vol. 43, n°. 6Part1, p. 2704–2714, 2016.

[5] M. Brahim, K. Westerkamp, L. Hempel, R. Lehmann, D. Hempel et P. Philipp, "Automated assessment of breast positioning quality in screening mammography," *Cancers*, vol. 14, n°. 19, p. 4704, 2022.

[6] R. Wang, J. Li, Chencho, S. An, H. Hao, W. Liu et L. Li, "Densely connected convolutional networks for vibration based structural damage identification," *Engineering Structures*, vol. 245, p. 112871, 2021.

[7] M. L. McHugh, "Interrater reliability : The kappa statistic," *Biochemia Medica*, vol. 22, n°. 3, p. 276–282, 2012.

[8] M. Shafiq et Z. Gu, "Deep residual learning for image recognition : A survey," *Applied Sciences*, vol. 12, n°. 18, p. 8972, 2022.

[9] Société canadienne du cancer, "Breast cancer statistics." [En ligne]. Disponible : https://cancer.ca/en/cancer-information/cancer-types/breast/statistics

[10] World Health Organization, "Cancer du sein." [En ligne]. Disponible : https://www.who.int/fr/news-room/fact-sheets/detail/breast-cancer

[11] Société canadienne du cancer, "Qu'est-ce que le cancer du sein ?" [En ligne]. Disponible : https://cancer.ca/fr/cancer-information/cancer-types/breast/what-is-breast-cancer

[12] N. Thakur, P. Kumar et A. Kumar, "A systematic review of machine and deep learning techniques for the identification and classification of breast cancer through medical image modalities," *Multimedia Tools and Applications*, vol. 83, p. 35849–35942, 2023.

[13] Cleveland Clinic, “Mammogram : When why to get one,” Jan 2026. [En ligne]. Disponible : https://my.clevelandclinic.org/health/diagnostics/4877-mammogram

[14] Gouvernement du Québec, “À propos du programme québécois de dépistage du cancer du sein.” [En ligne]. Disponible : https://www.quebec.ca/sante/conseils-et-prevention/depistage-et-offre-de-tests-de-porteur/programme-quebecois-de-depistage-du-cancer-du-sein/description

[15] S. M. Albeshan, Y. Alashban, F. M. A. Tahan, S. Al-enezi, N. Alnaimy, N. Shubayr et F. Eliraqi, “Mammography image quality evaluation in breast cancer screening : The saudi experience,” *Journal of Radiation Research and Applied Sciences*, vol. 15, n°. 4, p. 100467, 2022.

[16] J. Rouette, N. Elfassy, N. Bouganim, H. Yin, N. Lasry et L. Azoulay, “Evaluation of the quality of mammographic breast positioning : a quality improvement study,” *CMAJ Open*, vol. 9, n°. 2, p. E607–E612, 2021.

[17] E. Gourd, “Mammography deficiencies : the result of poor positioning,” *The Lancet Oncology*, vol. 19, n°. 8, p. e385, 2018. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S1470204518304893

[18] M.-H. Guertin, I. Théberge, M.-P. Dufresne, H. T. Zomahoun, D. Major, R. Tremblay, C. Ricard, R. Shumak, N. Wadden, Pelletier et et al., “Clinical image quality in daily practice of breast cancer mammography screening,” *Canadian Association of Radiologists Journal*, vol. 65, n°. 3, p. 199–206, 2014.

[19] American College of Radiology, “Mammography positioning review guidelines table.” [En ligne]. Disponible : https://learningnetworksupport.acr.org/support/solutions/articles/11000115257-mammography-positioning-review-guidelines-table

[20] N. M. Perry, M. Broeders, C. J. M. de Wolf, S. Törnberg, R. Holland et L. von Karsa, édit., *European guidelines for Quality Assurance in breast cancer screening and diagnosis*. EUR-OP, 2006.

[21] GOV UK, “Guidance for breast screening mammographers.” [En ligne]. Disponible : https://www.gov.uk/government/publications/breast-screening-quality-assurance-for-mammography-and-radiography/guidance-for-breast-screening-mammographers

[22] T. cancer imaging archive, “Cbis-ddsm | curated breast imaging subset of digital database for screening mammography,” 2025. [En ligne]. Disponible : https://www.cancerimagingarchive.net/collection/cbis-ddsm/

[23] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy et D. L. Rubin, “A curated mammography data set for use in computer-aided detection and diagnosis research,” *Scientific Data*, vol. 4, n°. 1, Dec 2017.

[24] MammoImage.org, “Databases.” [En ligne]. Disponible : https://www.mammoimage.org/databases/

[25] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso et J. S. Cardoso, “Inbreast : toward a full-field digital mammographic database,” *Academic Radiology*, vol. 19, n°. 2, p. 236–248, 2012.

[26] H. T. Nguyen, H. Q. Nguyen, H. Pham, K. Lam, L. T. Le, M. Dao et V. Vu, “Vindr-mammo : A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography,” *Scientific Data*, vol. 10, 2022. [En ligne]. Disponible : https://api.semanticscholar.org/CorpusID:247359170

[27] Y. Zheng, “Breast cancer detection with gabor features from digital mammograms,” *Algorithms*, vol. 3, n°. 1, p. 44–62, 2010.

[28] R. Rouhi et M. Jafari, “Classification of benign and malignant breast tumors based on hybrid level set segmentation,” *Expert Systems with Applications*, vol. 46, p. 45–59, 2016.

[29] J. Dheeba et S. Tamil Selvi, “Classification of malignant and benign microcalcification using svm classifier,” dans *2011 International Conference on Emerging Trends in Electrical and Computer Technology*, 2011, p. 686–690.

[30] M. Z. Nascimento, A. S. Martins, L. A. Neves, R. P. Ramos, E. L. Flores et G. A. Carrijo, “Classification of masses in mammographic image using wavelet domain features and polynomial classifier,” *Expert Systems with Applications*, vol. 40, n°. 15, p. 6213–6221, 2013.

[31] N. Gedik et A. Atasoy, “A computer-aided diagnosis system for breast cancer detection by using a curvelet transform,” *Turkish Journal of Electrical Engineering and Computer Sciences*, 2013.

[32] P. Kral et L. Lenc, “Lbp features for breast cancer detection,” dans *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, p. 2643–2647.

[33] M. G. Ertosun et D. L. Rubin, “Probabilistic visual search for masses within mammography images using deep learning,” dans *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2015, p. 1310–1315.

[34] P. U. Hepsağ, S. A. Özel et A. Yazıcı, “Using deep learning for mammography classification,” dans *2017 International Conference on Computer Science and Engineering (UBMK)*, 2017, p. 418–423.

[35] A. Duggento, M. Aiello, C. Cavaliere, G. L. Cascella, D. Cascella, G. Conte, M. Guerrisi et N. Toschi, “An ad hoc random initialization deep neural network architecture for

discriminating malignant breast cancer lesions in mammographic images," *Contrast Media Molecular Imaging*, vol. 1, p. 5982834, 2019.

[36] S. Deep Deb, M. A. Rahman et R. K. Jha, "Breast cancer detection and classification using global pooling," dans *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2020, p. 1–5.

[37] R. Song, T. Li et Y. Wang, "Mammographic classification based on xgboost and dcnn with multi features," *IEEE Access*, vol. 8, p. 75011–75021, 2020.

[38] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride et W. Sieh, "Deep learning to improve breast cancer detection on screening mammography," *Sci Rep*, vol. 9, n°. 1, p. 12495, 2019.

[39] H. M. Frazer, A. K. Qin, H. Pan et P. Brotchie, "Evaluation of deep learning-based artificial intelligence techniques for breast cancer detection on mammograms : Results from a retrospective study using a breastscreen victoria dataset," *Journal of Medical Imaging and Radiation Oncology*, vol. 65, n°. 5, p. 529–537, 2021.

[40] S. Suzuki, X. Zhang, N. Homma, K. Ichiji, N. Sugita, Y. Kawasumi, T. Ishibashi et M. Yoshizawa, "Mass detection using deep convolutional neural network for mammographic computer-aided diagnosis," dans *2016 55th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, 2016, p. 1382–1386.

[41] S. Guan et M. Loew, "Breast cancer detection using transfer learning in convolutional neural networks," dans *2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 2017, p. 1–8.

[42] M. A. Al-antari, M. A. Al-masni, M.-T. Choi, S.-M. Han et T.-S. Kim, "A fully integrated computer-aided diagnosis system for digital x-ray mammograms via deep learning detection, segmentation, and classification," *International Journal of Medical Informatics*, vol. 117, p. 44–54, 2018.

[43] G. Hamed, M. Marey, S. E. Amin et M. F. Tolba, "Automated breast cancer detection and classification in full field digital mammograms using two full and cropped detection paths approach," *IEEE Access*, vol. 9, p. 116898–116913, 2021.

[44] D. Yi, R. L. Sawyer, D. Cohn, J. A. Dunnmon, C. K. Lam, X. Xiao et D. Rubin, "Optimizing and visualizing deep learning for benign/malignant classification in breast tumors," *ArXiv*, vol. abs/1705.06362, 2017. [En ligne]. Disponible : https://api.semanticscholar.org/CorpusID:2390912

[45] K. Loizidou, G. Skouroumouni, C. Nikolaou et C. Pitris, "Breast mass detection and classification based on digital temporal subtraction of mammogram pairs," 2020.

[46] N. Chouhan, A. Khan, J. Z. Shah, M. Hussnain et M. W. Khan, "Deep convolutional neural network and emotional learning based breast cancer detection using digital mammography," *Computers in Biology and Medicine*, vol. 132, p. 104318, 2021.

[47] S. S. Boudouh et M. Bouakkaz, "Advancing precision in breast cancer detection : A fusion of vision transformers and cnns for calcification mammography classification," *Applied Intelligence*, vol. 54, p. 8170–8183, 2024.

[48] Société canadienne du cancer, "Densité mammaire," accessed : 2026-01-20. [En ligne]. Disponible : https://cancer.ca/fr/treatments/tests-and-procedures/mammography/breast-density

[49] J. Lee et R. M. Nishikawa, "Automated mammographic breast density estimation using a fully convolutional network," *Med. Phys.*, vol. 45, n°. 3, p. 1178–1190, 2018.

[50] N. Saffari, H. A. Rashwan, M. Abdel-Nasser, V. Kumar Singh, M. Arenas, E. Mangina, B. Herrera et D. Puig, "Fully automated breast density segmentation and classification using deep learning," *Diagnostics*, vol. 10, n°. 11, p. 988, 2020.

[51] N. Wu, K. J. Geras, Y. Shen, J. Su, S. G. Kim, E. Kim, S. Wolfson, L. Moy et K. Cho, "Breast density classification with deep convolutional neural networks," dans *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, p. 6682–6686.

[52] C. D. Lehman, A. Yala, T. Schuster, B. Dontchos, M. Bahl, K. Swanson et R. Barzilay, "Mammographic breast density assessment using deep learning : Clinical implementation," *Radiology*, vol. 290, n°. 1, p. 52–58, 2019.

[53] R. Sexauer, P. Hejduk, K. Borkowski, C. Ruppert, T. Weikert, S. Dellas et N. Schmidt, "Diagnostic accuracy of automated acr bi-rads breast density classification using deep convolutional neural networks," *European Radiology*, vol. 33, n°. 7, p. 4589–4596, 2023.

[54] P. Casti, A. Mencattini, M. Salmeri, A. Ancona, F. F. Mangieri, M. L. Pepe et R. M. Rangayyan, "Automatic detection of the nipple in screen-film and full-field digital mammograms using a novel hessian-based method," *J Digit Imaging*, vol. 26, n°. 5, p. 948–957, 2013.

[55] M. Moran, "Techniques for automated analysis of mammography positioning failures," 2018. [En ligne]. Disponible : https://epos.myesr.org/poster/esr/ecr2018/C-1089

[56] Y.-C. Zeng, "Automatic detections of nipple and pectoralis major in mammographic images," dans *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, 2018, p. 1–2.

[57] W. B. Yoon, J. E. Oh, E. Y. Chae, H. H. Kim, S. Y. Lee et K. G. Kim, "Automatic detection of pectoral muscle region for computer-aided diagnosis using mias mammo-

grams," *BioMed Research International*, vol. 2016, p. 5967580, 2016.

[58] S. A. Taghanaki, Y. Liu, B. Miles et G. Hamarneh, "Geometry-based pectoral muscle segmentation from mlo mammogram views," *IEEE Transactions on Biomedical Engineering*, vol. 64, nº. 11, p. 2662–2671, 2017.

[59] H. Ture et T. Kayikcioglu, "Accurate detection of distorted pectoral muscle in mammograms using specific patterned isocontours," *IEEE Access*, vol. 8, p. 147 370–147 386, 2020.

[60] S. Saxena, Y. Singh et B. Agarwal, "Boundary detection to segment the pectoral muscle from digital mammograms images," *Int. J. Eng. Adv. Technol.*, vol. 9, nº. 3, p. 1593–1603, 2020.

[61] C. Zhou, H. P. Chan, C. Paramagul, M. A. Roubidoux, B. Sahiner, L. M. Hadjiiski et N. Petrick, "Computerized nipple identification for multiple image analysis in computer-aided diagnosis," *Med Phys*, vol. 31, nº. 10, p. 2871–2882, 2004.

[62] T. Bülow, K. Meetz, D. Kutra, T. Netsch, R. Wiemker, M. Bergtholdt, J. Sabczynski, N. Wieberneit, M. Freund et I. Schulze-Wenck, "Automatic assessment of the quality of patient positioning in mammography," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 8670, p. 24–, 2013.

[63] Y.-C. Zeng, "Detection and recognition of inframammary fold on mlo-view mammograms," dans *2021 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, 2021, p. 1–2.

[64] V. Gupta, C. Taylor, S. Bonnet, L. M. Prevedello, J. Hawley, R. D. White, M. G. Flores et B. S. Erdal, "Deep learning-based automatic detection of poorly positioned mammograms to minimize patient return visits for repeat imaging : A real-world application," *arXiv preprint*, 2020. [En ligne]. Disponible : https://arxiv.org/abs/2009.13580

[65] T. Tanyel, N. Denizoglu, M. E. Seker, D. Alis, E. Cerekci, E. Karaarslan, E. Aribal et I. Oksuz, "Mammographic breast positioning assessment via deep learning," *arXiv preprint*, 2024.

[66] Y.-C. Zeng et Y.-H. Chen, "Skinfold recognition for positioning quality evaluation on mammography," dans *2022 IEEE 11th Global Conference on Consumer Electronics (GCCE)*, 2022, p. 862–864.

[67] Y.-C. Zeng, "Realizing nipple in profile recognition and nipple detection using a single classification," dans *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2023, p. 1500–1505.

[68] ——, "Classification and model interpretation of mammography positioned quality for posterior breast," dans *2023 International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, 2023, p. 45–46.

[69] H. Watanabe, S. Hayashi, Y. Kondo, E. Matsuyama, N. Hayashi, T. Ogura et M. Shimosegawa, "Quality control system for mammographic breast positioning using deep learning," *Sci Rep*, vol. 13, n°. 1, p. 7066, 2023.

[70] P. Hejduk, R. Sexauer, C. Ruppert, K. Borkowski, J. Unkelbach et N. Schmidt, "Automatic and standardized quality assurance of digital mammography and tomosynthesis with deep convolutional neural networks," *Insights into Imaging*, vol. 14, n°. 1, p. 90, 2023.

[71] V. Health, 2023. [En ligne]. Disponible : https://www.volparahealth.com/science/algorithms/volpara-pgmi/

[72] b rayZ, Nov 2025. [En ligne]. Disponible : https://b-rayz.com/technology/#main

[73] Densitas, Jan 2026. [En ligne]. Disponible : https://densitashealth.com/densitas-intellimammo/

[74] S. Liu, E. Johns et A. J. Davison, "End-to-end multi-task learning with attention," dans *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, p. 1871–1880.

[75] Y. Zhao, X. Wang, T. Che, G. Bao et S. Li, "Multi-task deep learning for medical image computing and analysis : A review," *Computers in Biology and Medicine*, vol. 153, p. 106496, 2023.

[76] S. Kim, T. G. Purdie et C. McIntosh, "Cross-task attention network : Improving multi-task learning for medical imaging applications," dans *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 Workshops : ISIC 2023, Care-AI 2023, MedAGI 2023, DeCaF 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 8–12, 2023, Proceedings*, 2023, p. 119–128.

[77] G. Zhang, K. Zhao, Y. Hong, X. Qiu, K. Zhang et B. Wei, "Sha-mtl : Soft and hard attention multi-task learning for automated breast cancer ultrasound image segmentation and classification," *International Journal of Computer Assisted Radiology and Surgery*, vol. 16, n°. 10, p. 1719–1725, 2021.

[78] J. Chowdary, P. Yogarajah, P. Chaurasia et V. Guruviah, "A multi-task learning framework for automated segmentation and classification of breast tumors from ultrasound images," *Ultrasonic Imaging*, vol. 44, n°. 1, p. 3–12, Jan 2022.

[79] M. V. Sainz de Cea, K. Diedrich, R. Bakalo, L. Ness et D. Richmond, "Multi-task learning for detection and classification of cancer in screening mammography," dans

Medical Image Computing and Computer Assisted Intervention – MICCAI 2020 : 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI, 2020, p. 241–250.

[80] R. Shen, K. Zhou, K. Yan, K. Tian et J. Zhang, "Multicontext multitask learning networks for mass detection in mammogram," *Medical Physics*, vol. 47, n°. 4, p. 1566–1578, Mar 2020.

[81] F. Gao, H. Yoon, T. Wu et X. Chu, "A feature transfer enabled multi-task deep learning model on medical imaging," *Expert Systems with Applications*, vol. 143, p. 112957, 2020.

[82] X. Hou, Y. Bai, Y. Xie et Y. Li, "Mass segmentation for whole mammograms via attentive multi-task learning framework," *Phisics in Medecine and Biology*, May 2021.

[83] J. Wang, Y. Zheng, J. Ma, X. Li, C. Wang, J. Gee, H. Wang et W. Huang, "Information bottleneck-based interpretable multitask network for breast cancer classification and segmentation," *Medical Image Analysis*, vol. 83, p. 102687, 2023.

[84] G. Nebbia, D. Arefan, M. Zuley, J. Sumkin et S. Wu, "Multi-task learning to incorporate clinical knowledge into deep learning for breast cancer diagnosis," dans *Medical Imaging 2021 : Computer-Aided Diagnosis*, 2021, p. 34.

[85] Z. Zhou, D. Arefan, M. Zuley, D. Hao, J. Sumkin et S. Wu, "Knowledge-guided multitask learning for breast cancer diagnosis using longitudinal mammogram images," dans *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024, p. 1–5.

[86] T. C. Nguyen, T. D. Nguyen et S. T. T. Nguyen, "Improving multi-task learning for breast cancer detection," dans *Proceedings of the 2024 10th International Conference on Computing and Artificial Intelligence*, 2024, p. 169–173.

[87] A. Foi, M. Trimeche, V. Katkovnik et K. Egiazarian, "Practical poissonian-gaussian noise modeling and fitting for single-image raw-data," *IEEE Transactions on Image Processing*, vol. 17, n°. 10, p. 1737–1754, 2008.

[88] N. Alsaihati, J. Solomon, E. McCrum et E. Samei, "Development, validation, and application of a generic image-based noise addition method for simulating reduced dose computed tomography images," *Medical Physics*, vol. 52, n°. 1, p. 171–187, 2024.

[89] S. Wang et A. Wang, "Simulating arbitrary dose levels and independent noise image pairs from a single ct scan," dans *7th International Conference on Image Formation in X-Ray Computed Tomography*, 2022.

[90] S. Lee, M. Lee et M. Kang, "Poisson–gaussian noise analysis and estimation for low-dose x-ray images in the nsct domain," *Sensors*, vol. 18, n°. 4, p. 1019, Mar 2018.

[91] V. Gauvin, L. Grondin-Reiher, V. Patenaude, N. Lasry, F. Guibault et L. Séoud, "Breast positioning assessment in mammography : an inter-rater reliability study,"

manuscript submitted for publication.

[92] Z. Tao, C. XiaoYu, L. HuiLing, Y. XinYu, L. YunCan et Z. XiaoMin, “Pooling operations in deep learning : From “invariable” to “variable”,” *BioMed Research International*, vol. 2022, n°. 1, p. 4067581, 2022.

[93] G. Huang, Z. Liu, L. Van Der Maaten et K. Q. Weinberger, “Densely connected convolutional networks,” dans *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, p. 2261–2269.

[94] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser et I. Polosukhin, “Attention is all you need,” 2023. [En ligne]. Disponible : https://arxiv.org/abs/1706.03762

[95] J. Terven, D.-M. Cordova-Esparza, J.-A. Romero-González, A. Ramírez-Pedraza et E. A. Chávez-Urbiola, “A comprehensive survey of loss functions and metrics in deep learning,” *Artificial Intelligence Review*, vol. 58, n°. 7, 2025. [En ligne]. Disponible : http://dx.doi.org/10.1007/s10462-025-11198-7

[96] M. Yeung, E. Sala, C.-B. Schönlieb et L. Rundo, “Unified focal loss : Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation,” *Computerized Medical Imaging and Graphics*, vol. 95, p. 102026, 2022. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0895611121001750

[97] V. Dharanalakota, A. A. Raikar et P. K. Ghosh, “Improving neural network training using dynamic learning rate schedule for pinns and image classification,” *Machine Learning with Applications*, vol. 21, p. 100697, 2025. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S2666827025000805

[98] M. Ganaie, M. Hu, A. Malik, M. Tanveer et P. Suganthan, “Ensemble deep learning : A review,” *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105151, 2022. [En ligne]. Disponible : http://dx.doi.org/10.1016/j.engappai.2022.105151

[99] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren et A. Y. Ng, “Chexpert : A large chest radiograph dataset with uncertainty labels and expert comparison,” 2019. [En ligne]. Disponible : https://arxiv.org/abs/1901.07031

[100] M. Makitalo et A. Foi, “Optimal inversion of the anscombe transformation in low-count poisson image denoising,” *IEEE Transactions on Image Processing*, vol. 20, n°. 1, p. 99–109, 2011.

[101] K. Funaro et B. Niell, “Variability in mammography quality assessment after implementation of enhancing quality using the inspection program (equip),” *Journal*

of Breast Imaging, vol. 3, n°. 2, p. 168–175, 2021. [En ligne]. Disponible : https://pubmed.ncbi.nlm.nih.gov/38424823/

[102] N. N. Mohd, Z. Toh, I. C. Isa et Z. N. Mohd, "Variability in mammographic images quality evaluation and predictive factors affecting it," *Journal of Medical Imaging and Radiation Sciences*, vol. 53, n°. 4, Supplement 1, p. S39–S40, 2022, iSRRT conference proceedings. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S1939865422005082

[103] E. Alukić, K. Homar, M. Pavić, J. Žibert et N. Mekiš, "The impact of subjective image quality evaluation in mammography," *Radiography*, vol. 29, n°. 3, p. 526–532, 2023.

[104] T. Santner, C. Ruppert, S. Gianolini, J.-G. Stalheim, S. Frei, M. Hondl, V. Fröhlich, S. Hofvind et G. Widmann, "Pgmi assessment in mammography : Ai software versus human readers," *Radiography*, vol. 31, n°. 5, p. 103017, 2025.

[105] K. Simonyan et A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [En ligne]. Disponible : https://api.semanticscholar.org/CorpusID:14124313

[106] K. He, X. Zhang, S. Ren et J. Sun, "Deep residual learning for image recognition," dans *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, p. 770–778.

ANNEXE A ALGORITHME DE PRÉ-TRAITEMENT DES IMAGES

Cette annexe présente les détails de l'algorithme de pré-traitement des images ainsi que son implémentation.

Le pré-traitement des images est essentiel afin d'obtenir un jeu de données uniforme. D'abord, une binarisation des images est effectuée à l'aide de la fonction *threshold* de la librairie *Open CV* (*cv2.threshold*). Un seuil d'intensité de 5 a été choisi, mais celui-ci a dû être ajusté manuellement pour quelques images pour lesquelles les contrastes étaient différents de la majorité. Ensuite, la détection des contours est réalisée avec la fonction *findContours* de la même librairie (*cv2.findContours*). Le contour le plus long est sélectionné, correspondant au contour du sein, et un masque binaire est créé à partir de l'image binarisée et du contour retenu. On effectue ensuite une dilatation du masque afin de s'assurer que l'entièreté du sein soit comprise à l'intérieur de celui-ci. Celle-ci est réalisée à l'aide de la fonction *cv2.dilate* et en utilisant un noyau (*kernel*) de 100 x 100 pixels. La taille du noyau a aussi été ajustée manuellement pour quelques images présentant des artéfacts. Finalement, on ajuste le masque pour conserver les pixels d'origine dans les coins se trouvant du même côté que le sein ; cette étape est essentielle afin de ne pas masquer des parties du sein qui pourraient être discontinues du contour principal, comme l'IMA. La dernière étape de l'algorithme consiste à redimensionner les images à une taille de 256 x 256 pixels.

Les détails de l'implémentation de l'algorithme sont présentés dans le pseudo-code suivant :

Algorithm 1 Algorithme de pré-traitement des images

Entrée: `img`
Sortie: `resized_img`

```
t ← 5
kernel_size ← 100
img_bin ← cv2.threshold(img, t, 255, cv2.THRESH_BINARY)[1]
contours ← cv2.findContours(img_bin, cv2.RETR_EXTERNAL, cv2.CHAIN_APPROX_NONE)
max_contour ← max(contours, key=cv2.contourArea)
mask ← Initialisation d'un masque de la taille de l'image originale
mask ← Création d'un masque suivant le contour le plus long à l'aide de
  cv2.drawContours(mask, [max_contour], -1, 255, cv2.FILLED)
kernel ← np.ones((100, 100), np.uint8)
dilated_mask ← cv2.dilate(mask, kernel, iterations=1)
masked_image ← cv2.bitwise_and(img, img, mask=dilated_mask)
resized_img ← cv2.resize(masked_img, (256, 256))
retourne resized_img
```

La figure suivante illustre les étapes intermédiaires du pré-traitement d'une image.

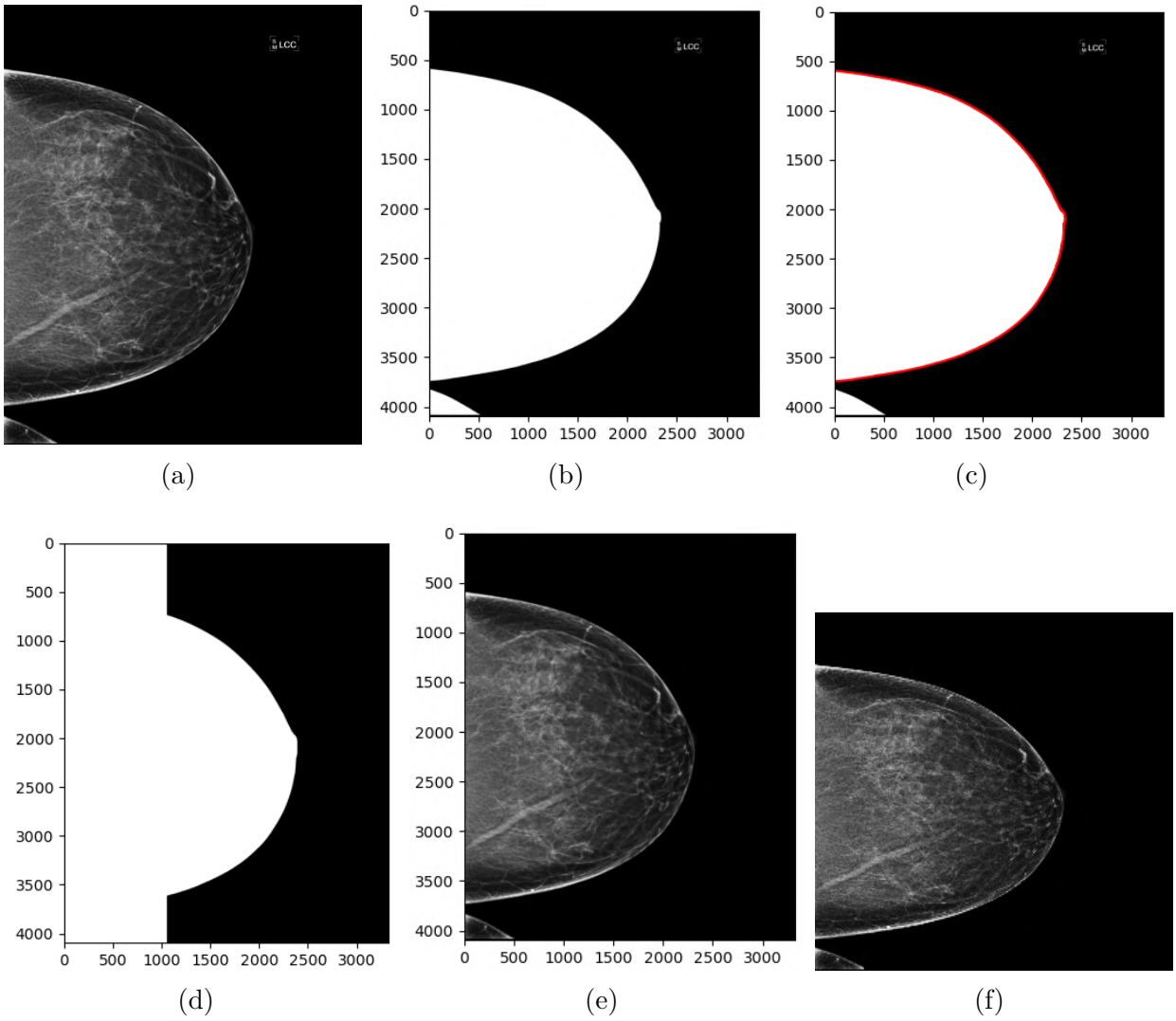


FIGURE A.1 (a) Image originale, (b) image binarisée, (c) détection du plus long contour (en rouge), (d) dilatation du masque et ajout des coins du même côté que le sein, (e) application du masque à l'image et (f) image finale redimensionnée

ANNEXE B CHOIX DU MODÈLE *BASELINE*

Cette annexe présente les résultats préliminaires ayant mené au choix du modèle pour la seconde *baseline*. Deux modèles ont d'abord été testés en plus du DenseNet-121 : le VGG-16 et le ResNet-50. Il s'agit de modèles couramment utilisés dans la littérature pour des tâches de classification d'images dans des domaines variés.

VGG-16 [105] est un des premiers CNNs à avoir été développé et largement utilisé. Il s'agit d'un CNN classique utilisant des filtres de convolution de taille 3 x 3 et prenant en entrée des images de 224 x 224 pixels. Dans l'article présentant le modèle, les auteurs investiguent l'impact de l'augmentation de la profondeur du réseau sur les performances de classification. C'est ainsi qu'ils ont développé deux modèles avec respectivement 16 et 19 couches convolutives : VGG-16 et VGG-19. Les modèles VGG contiennent aussi des couches de *Max Pooling* entre les couches convolutives, ainsi que des FCs pour la classification finale.

Ensuite, le *Residual Network* (ResNet) est un CNN qui a été élaboré par He et al. [106] en 2015. Ce modèle contient les mêmes couches classiques que le VGG (couches convolutives et *Max Pooling*), mais les auteurs introduisent des connexions résiduelles (*skip connexions*) dans le réseau. Ces connexions sont présentes entre les couches convolutives et permettent d'additionner l'entrée à la sortie : on parle alors d'un bloc résiduel. Ce processus facilite l'apprentissage et la descente de gradient, surtout pour des réseaux très profonds. La figure B.1 illustre l'architecture d'un bloc résiduel.

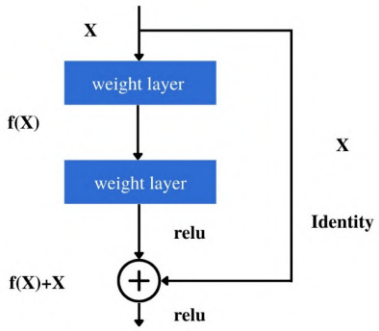


FIGURE B.1 Architecture d'un bloc résiduel [8]

Pour chacun des critères, une recherche d'hyperparamètres a été effectuée sur le *learning rate* (*lr*), la fonction de coût (*focal loss* ou *binary cross entropy loss*) et le *weight decay* afin de déterminer les hyperparamètres d'entraînement optimaux. Ces entraînements ont été effectués sur l'ensemble d'images A et avec un seul ensemble de validation correspondant

à 20% des données. Le reste des images de l'ensemble A a été utilisé pour l'entraînement. L'optimiseur Adam a été utilisé lors de l'entraînement et une taille de *batch* de 32 a été employée. Le tableau B.1 présente les résultats de PR AUC obtenus pour chacun des modèles avec les hyperparamètres optimaux. Cette métrique a été choisie puisqu'elle est indépendante de seuil et est plus appropriée pour des tâches avec des classes déséquilibrées, comme c'est le cas pour la plupart des critères de positionnement.

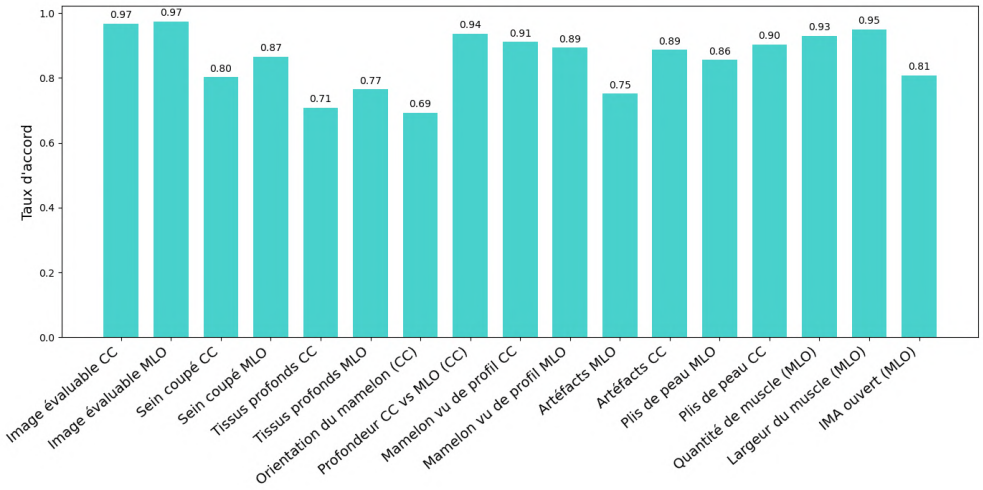
TABLEAU B.1 Résultats préliminaires pour le choix du modèle *baseline*. Les résultats présentés correspondent à la PR AUC.

Modèle	Sein coupé CC	Tissus profonds CC	Orientation du mamelon	Sein coupé MLO	Tissus profonds MLO	IMA ouvert
VGG-16	0,859	0,873	0,661	0,532	0,876	0,830
ResNet-50	0,859	0,868	0,703	0,501	**0,881**	**0,836**
DenseNet-121	**0,867**	**0,886**	**0,777**	**0,550**	0,868	0,833

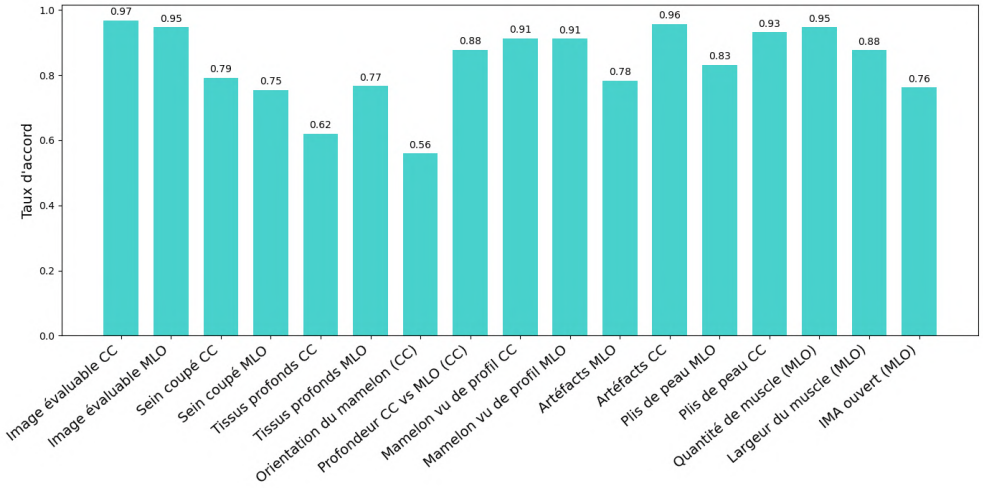
Le DenseNet-121 est celui qui obtient les meilleures performances pour la plupart des critères, c'est pourquoi il a été choisi comme *baseline*.

ANNEXE C RÉSULTATS DE VARIABILITÉ PAR PAIRE DE TECHNOLOGUES

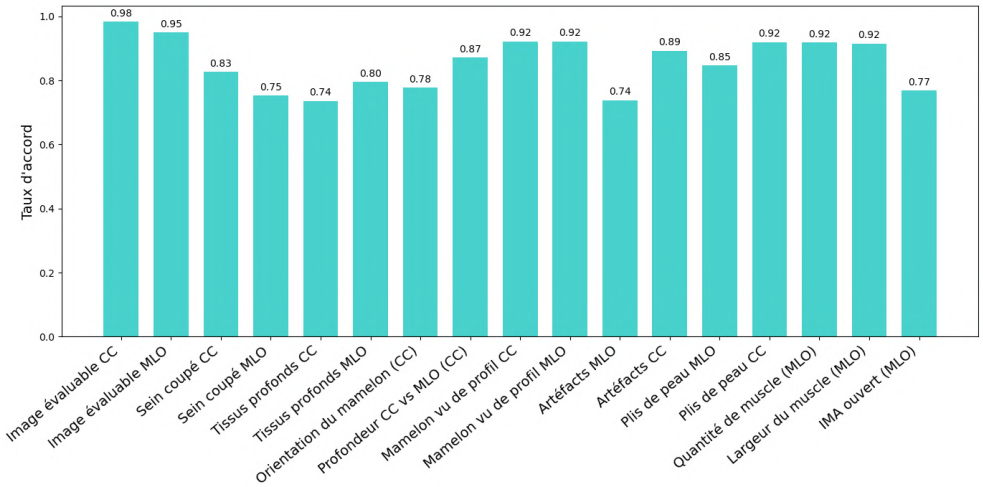
Cette annexe consigne les résultats de variabilité par paires de technologues. La Figure C.1 présente le taux d'accord pour chacun des critères pour toutes les paires de technologues, et la Figure C.2 présente les kappas de Cohen par critères et par paire de technologues.



(a) Taux d'accord entre les technologues 1 et 2.

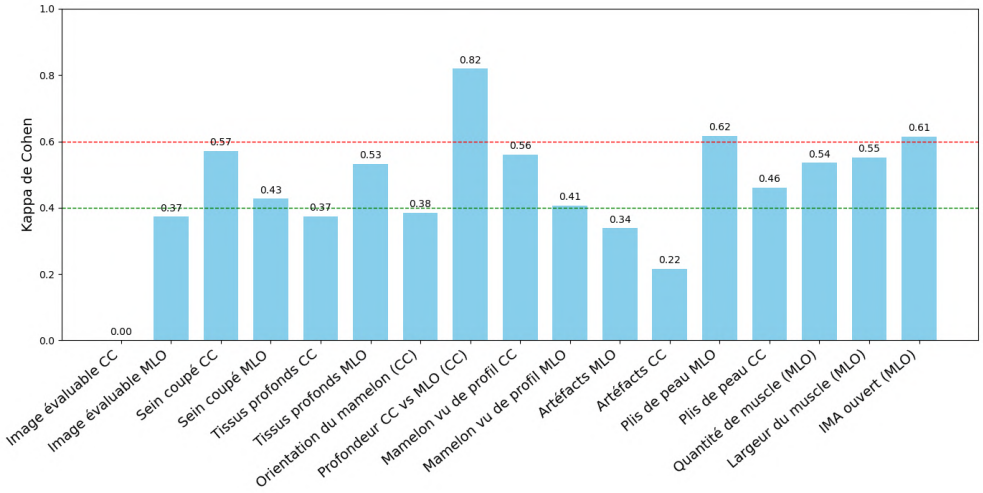


(b) Taux d'accord entre les technologues 1 et 3.

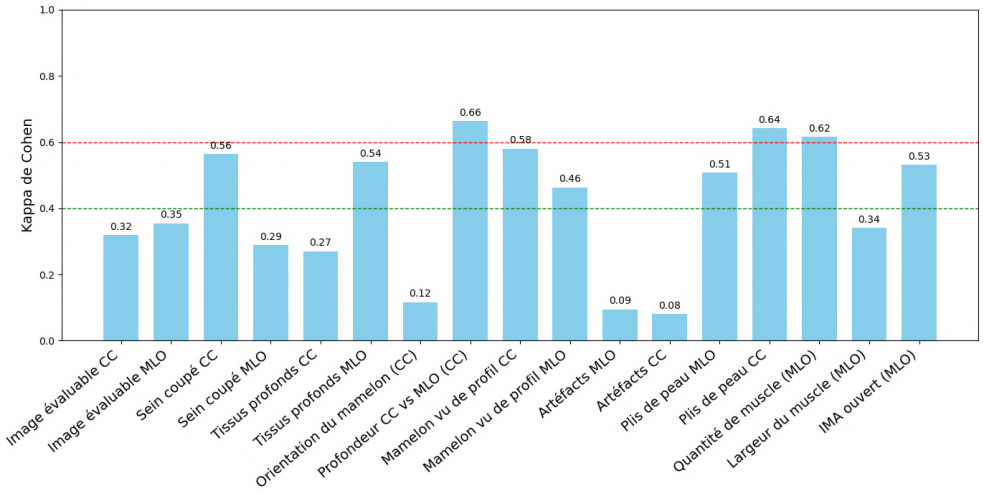


(c) Taux d'accord entre les technologues 2 et 3.

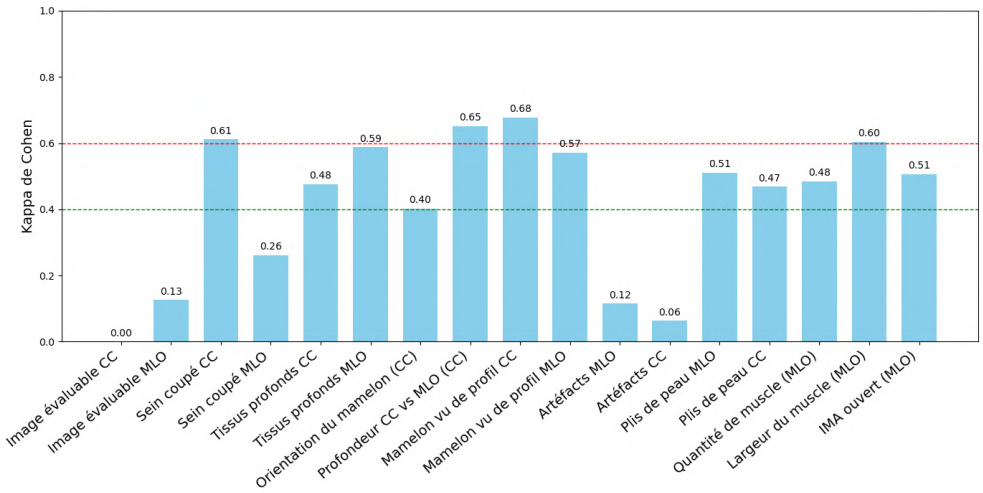
FIGURE C.1 Taux d'accord par paires de technologues.



(a) Kappa de Cohen pour les technologues 1 et 2.



(b) Kappa de Cohen pour les technologues 1 et 3.



(c) Kappa de Cohen pour les technologues 2 et 3.

FIGURE C.2 Kappa de Cohen par paire de technologues. Ligne verte pointillée : seuil pour un accord modéré selon Cohen. Ligne rouge pointillée : seuil pour un accord modéré selon [7].

ANNEXE D HYPERPARAMÈTRES UTILISÉS POUR L'ENTRAÎNEMENT

Cette annexe présente les hyperparamètres utilisés pour les différents entraînements effectués dans le cadre de ce projet dans un souci de reproductibilité. Le tableau D.1 présente les hyperparamètres utilisés pour les modèles *baselines* et le tableau D.2 présente ceux utilisés pour les principaux modèles multi-tâches.

Tableau D.1 Hyperparamètres utilisés pour l'entraînement des modèles *baseline*.

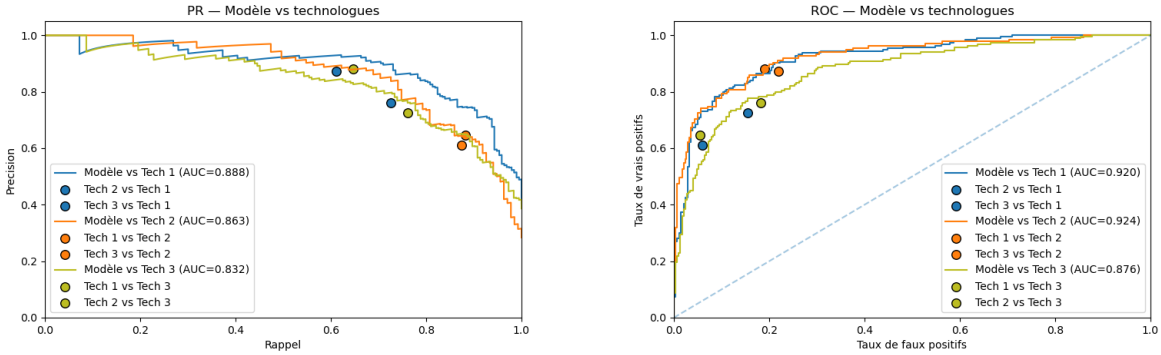
Modèle	Hyperparamètres	Sein coupé CC	Tissus profonds CC	Orientation du mamelon	Sein coupé MLO	Tissus profonds MLO	IMA ouvert
Brahim et al.	*lr*	0,0001	0,00005	0,0001	0,00005	0,0001	0,0001
	Fonction de perte	Focale	Focale	Focale	Focale	Focale	Focale
	weight decay	0,00001	0,00001	0,00001	0,0	0,00001	0,0
DenseNet-121	*lr*	0,0001	0,0001	0,0001	0,0001	0,00005	0,0001
	Fonction de perte	Focale	WCE	WCE	WCE	Focale	WCE
	weight decay	0,0	0,0001	0,00001	0,0	0,0	0,0001

Tableau D.2 Hyperparamètres utilisés pour l'entraînement des modèles multi-tâches.

Modèle	Hyperparamètres	
MT HPS : SC-TP (CC)	*lr*	0,0001
	Fonction de perte	Focale
	weight decay	0,0001
MT SPS : SC-TP (CC)	*lr*	0,0005
	Fonction de perte	WCE
	weight decay	0,0001
MT HPS : SC-TP-OM (CC)	*lr*	0,0005
	Fonction de perte	WCE
	weight decay	0,0
MT HPS : SC-TP (MLO)	*lr*	0,00005
	Fonction de perte	Focale
	weight decay	0,0
MT MultiHead : SC-TP-OM (CC)	*lr*	0,0001
	Fonction de perte	Focale
	weight decay	0,0001
MT MultiHead : SC-TP (MLO)	*lr*	0,0001
	Fonction de perte	WCE
	weight decay	0,0
MT CrossAtt : SC-TP-OM (CC)	*lr*	0,0001
	Fonction de perte	Focale
	weight decay	0,0
MT CrossAtt : SC-TP (MLO)	*lr*	0,00005
	Fonction de perte	WCE
	weight decay	0,00001

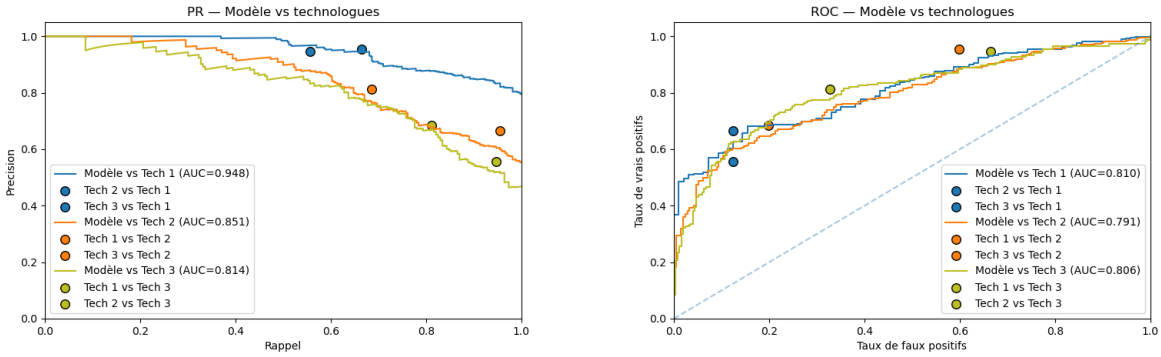
ANNEXE E RÉSULTATS SUPPLÉMENTAIRES DES MODÈLES MULTI-TÂCHES

Cette annexe présente des résultats supplémentaires pour les modèles multi-tâches. Les deux figures suivantes les graphes de comparaison du modèle MT HPS avec les technologues pour la combinaison SC-TP.



(a) (b)

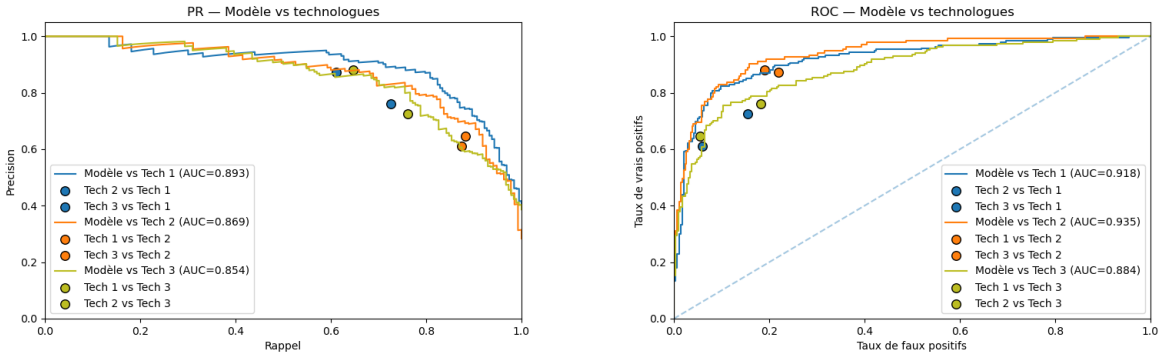
FIGURE E.1 Comparaison des performances du modèle MT HPS SC-TP par rapport aux technologues pour le critère du **Sein coupé en CC**. (a) Courbes PR du modèle par rapport à chacune des technologues et (b) courbes ROC du modèle par rapport à chacune des technologues



(a) (b)

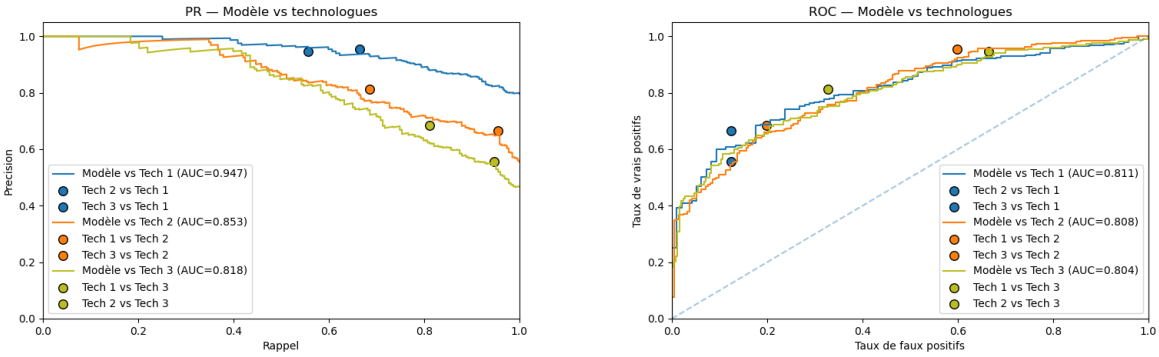
FIGURE E.2 Comparaison des performances du modèle MT HPS SC-TP par rapport aux technologues pour le critère des **Tissus profonds en CC**. (a) Courbes PR du modèle par rapport à chacune des technologues et (b) courbes ROC du modèle par rapport à chacune des technologues

Les deux figures suivantes présentent les graphes de comparaison du modèle MT SPS par rapport aux technologues pour la combinaison SC-TP.



(a) (b)

FIGURE E.3 Comparaison des performances du modèle MT SPS SC-TP par rapport aux technologues pour le critère du **Sein coupé en CC**. (a) Courbes PR du modèle par rapport à chacune des technologues et (b) courbes ROC du modèle par rapport à chacune des technologues



(a) (b)

FIGURE E.4 Comparaison des performances du modèle MT SPS SC-TP par rapport aux technologues pour le critère des **Tissus profonds en CC**. (a) Courbes PR du modèle par rapport à chacune des technologues et (b) courbes ROC du modèle par rapport à chacune des technologues