

Segmentation Anatomique Multiclasse Par Réseaux Convolutifs : étude comparative avec variantes d'architectures *U-Net*

Mehdi Mansour^{1,*}

¹Departement Informatique, ICOM, 5 avenue Pierre Mendès France, 69500 Bron, France

*Corresponding author: mehdi.mansour@univ-lyon2.fr

Abstract

Cette étude compare trois architectures dérivées du paradigme U-Net pour la segmentation multiclasse des poumons et du cœur sur radiographies thoraciques. Nous évaluons le U-Net original, une variante avec encodeur ResNet50 pré-entraîné, et le TransUNet intégrant des mécanismes d'attention. Les expérimentations, menées sur un corpus de 208 radiographies annotées manuellement et augmentées, montrent des performances élevées pour les trois modèles ($mIoU > 0,88$, accuracy pixel-wise $> 0,95$). Le U-Net+ResNet50 obtient les meilleurs résultats ($mIoU 0,8980$), tandis que le TransUNet montre une efficience remarquable avec des performances similaires malgré trois fois moins de paramètres ($mIoU 0,8936$ vs $0,8885$ pour U-Net). L'analyse des IoU par classe (poumon droit > poumon gauche > cœur) révèle l'impact de la variabilité morphologique et des superpositions anatomiques sur la qualité de segmentation. Ces résultats mettent en avant l'apport du transfert d'apprentissage et des mécanismes d'attention pour la segmentation d'imagerie médicale.

Keywords: Segmentation Sémantique; CNN; U-Net; RESNET; TransU-NET; Transfer Learning; Transformers

1. Introduction

La segmentation d'imagerie médicale est une technologie prometteuse dans la médecine moderne, permettant l'extraction précise et automatisée de régions anatomiques ou pathologiques d'intérêt à partir de données d'imagerie complexes. Cette technique constitue désormais un pilier essentiel de la radiologie assistée par ordinateur, révolutionnant le flux de travail clinique en fournissant des analyses quantitatives objectives qui complètent l'interprétation qualitative traditionnelle des radiologues (Rudnicka *et al.* 2024).

Dans le contexte clinique contemporain, la segmentation d'images médicales facilite de nombreuses applications critiques : planification chirurgicale personnalisée, radiothérapie de précision, suivi longitudinal des pathologies, et détection précoce des anomalies subtiles. Les radiographies thoraciques, examen d'imagerie le plus fréquemment réalisé mondialement, bénéficient particulièrement de ces avancées technologiques. La délimitation précise des structures pulmonaires et cardiaques permet en effet de quantifier des paramètres anatomiques essentiels à la prise en charge de pathologies respiratoires chroniques, insuffisances cardiaques, et infections pulmonaires.

L'émergence des architectures d'apprentissage profond, particulièrement les réseaux neuronaux convolutifs (CNN), a considérablement amélioré la précision des algorithmes de segmentation par rapport aux méthodes conventionnelles basées sur des caractéristiques prédéfinies de traitement de l'image. L'architecture U-Net (Ronneberger *et al.* 2015), a marqué un saut qualitatif dans ce domaine en proposant une structure encodeur-décodeur avec connexions résiduelles, spécifiquement optimisée pour les défis propres aux images médicales : la grande variabilité anatomique des sujets, les limites tissulaires parfois impré-

cises, et les corpora d'entraînement souvent limités.

Notre recherche se concentre sur l'évaluation de trois architectures dérivées du paradigme U-Net pour la segmentation multi-classes de trois organes thoraciques: le poumon droit, le poumon gauche et le cœur. Nous analysons l'architecture U-Net originale, une variante intégrant un encodeur ResNet pré-entraîné, et le modèle TransUNet qui incorpore des mécanismes d'attention inspirés des transformers.

Pour cette investigation, nous avons développé un corpus de 208 radiographies thoraciques, labellisées pour localiser les pixels du cœur et des poumons. Notre méthodologie inclut des techniques d'augmentation de données pour enrichir le corpus d'apprentissage. L'évaluation des performances combine des métriques quantitatives traditionnelles de la segmentation sémantique et une analyse qualitative des segmentations générées.

2. Acquisition et Construction du Dataset

A. Radiographies Thoraciques

Notre corpus d'étude est constitué de radiographies thoraciques extraites de la base de données publique NIH ChestX-ray14, mise à disposition par les National Institutes of Health (Wang *et al.* 2017). Cette base de données comprend plus de 100 000 radiographies thoraciques associées à 14 catégories de pathologies pulmonaires.

Pour les besoins de notre étude en segmentation, nous avons sélectionné un sous-ensemble de 300 radiographies selon des critères qui simplifient la modélisation dans un premier temps:

- Exposition aux rayons postéro-antérieur exclusivement, pour que le poumon droit du patient soit toujours du côté gauche de l'image et vice-versa.

2 SEGMENTATION par CV

- Absence d'artefacts susceptibles de polluer les caractéristiques visuelles générales (dispositifs médicaux, bijoux, éléments prothétiques).
- La bonne résolution et la visibilité complète des structures anatomiques qui nous intéressent.

Les radiographies sélectionnées ont été standardisées à une résolution uniforme de 512×512 pixels et converties au format PNG, format plus conventionnel pour les bibliothèques (type OpenCV) qui seront exploitées après.

B. Labellisation

La segmentation des structures thoraciques a été réalisée avec la plateforme *MedSeg*, dont l'interface de labellisation est accessible à l'adresse <https://htmlsegmentation.s3.eu-north-1.amazonaws.com/index.html>.

MedSeg est un outil spécialisé dédié à l'annotation d'images médicales, développé par deux radiologues basés à Oslo, en Norvège. Cette plateforme a été conçue pour intégrer des algorithmes de segmentation automatique basés sur des modèles légers spécialisés ([Sakinis et al. 2019](#)). Cependant, dans le cadre de cette étude, nous avons opté pour une annotation entièrement manuelle afin d'éviter que des biais potentiels de leurs modèles n'interfèrent avec le développement de nos propres modèles.

Des masques ont été créées pour délimiter les structures anatomiques qui nous intéressent: le poumon droit, le poumon gauche et le cœur (fig. 1).

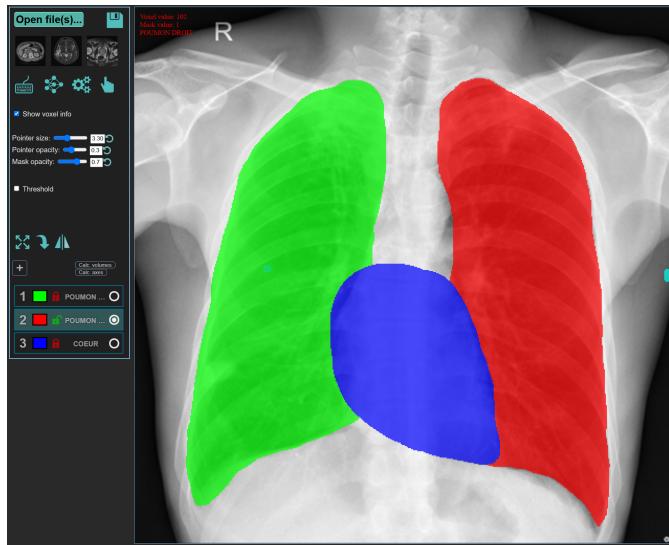


Figure 1 Interface *MedSeg* utilisée pour la création de masques RGB, pour poumons droit (vert), gauche (rouge) et du cœur (bleu).

Les fonctionnalités avancées de l'outil — notamment les paramètres de transparence ajustables, la palette de tailles de pinceaux et les options de superposition — ont permis un travail accessible à un novice, particulièrement aux interfaces organes-organes souvent problématiques.

Pour différencier ces organes tout en gardant les pixels de chacun même si chevauchement, on exploite les 3 channels RGB de l'image:

- Poumon droit : vert [0, 255, 0]
- Poumon gauche : rouge [255, 0, 0]
- Cœur : bleu [0, 0, 255]

Ce processus travail s'est heurté à des difficultés pour certaines images présentant des frontières anatomiques indistinctes. Ces images nécessitant l'expertise d'un radiologue ont été simplement écartées du dataset.

Sur l'ensemble initial d'images collectées, seules 208 ont finalement été labellisées.

C. Stratification et Augmentation des Données

Nous avons partitionné notre corpus de 208 radiographies thoraciques en trois ensembles distincts selon la répartition suivante:

- Ensemble d'entraînement : 70% (145 images)
- Ensemble de validation : 15% (31 images)
- Ensemble de test : 15% (32 images)

Pour l'ensemble d'entraînement, nous avons implémenté un protocole d'augmentation de données en utilisant la bibliothèque Albumentations ([Buslaev et al. 2020](#)). Chaque image originale a été conservée et enrichie par 8 variantes (fig. 2).

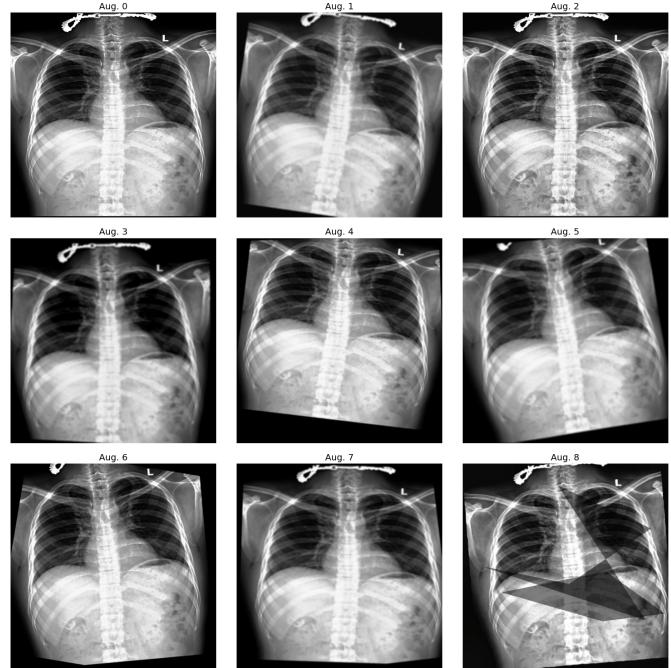


Figure 2 Versions originale (Aug. 0) et augmentées de la radiographie 8 montrant différentes transformations obtenues.

Ces variantes sont générées via un pipeline de transformations probabilistes appliquées de manière identique aux images et à leurs masques correspondants. Ce pipeline comprend:

- Transformations géométriques :
 - Translations ($\pm 8\%$ horizontalement et verticalement, $p=0.7$)
 - Zooms aléatoires ($\pm 10\%$, $p=0.7$)
 - Rotations ($\pm 15^\circ$, $p=0.5$)
 - Transformations affines combinées ($p=0.5$)
 - Déformations élastiques contrôlées ($\alpha=30$, $\sigma=5$, $p=0.2$)
 - Distorsions par grille ($p=0.2$)
- Autres transformations :
 - Ajustements de luminosité et contraste ($\pm 10\%$, $p=0.7$)

- Flou gaussien (noyau 3x5 pixels, p=0.3)
- Compression d'image (JPEG 85-95%, p=0.3)
- Égalisation adaptative d'histogramme (CLAHE, p=0.2)
- Simulation d'ombres (p=0.1)

Chaque transformation est appliquée avec une probabilité spécifique (p), permettant de générer une diversité de combinaisons pour chaque image. Les marges de transformations ont bien sûr été choisies pour rester cohérentes avec le champ des possibles dans le métier. Par exemple pas de rotation excessive de l'abdomen car ça n'existe pas dans les données du terrain. Par ailleurs, une transformation miroir par exemple, n'est pas la bienvenue ici, car nous nous intéressons à des radios prises avec une exposition postéro-anterior (poumon droit à gauche de l'image). L'appliquer aurait inversé les positions des poumons dans l'image et par conséquent pollué l'apprentissage du modèle.

Cette stratégie a augmenté notre ensemble d'entraînement à 1305 images (145 originales et 1160 variantes augmentées), diversifiant ainsi les patterns anatomiques auxquelles les modèles seront exposés durant l'apprentissage.

3. Métriques d'évaluation de la segmentation:

L'évaluation quantitative des performances en segmentation sémantique repose sur l'idée de: "à quel point le modèle a bien classé les pixels de l'image, en les attribuant à la bonne classe".

Diverses approches existent, dont certaines sont très anciennes mais qui restent pertinentes de part leur fondement mathématique universel.

A. Métriques basées sur le chevauchement spatial

A.1. Coefficient de Dice (DSC) Le coefficient de Dice, introduit initialement par Lee R. Dice en 1945 dans un contexte d'écologie ([Dice 1945](#)), constitue l'une des premières métriques largement adoptées en segmentation sémantique. Sa formulation mathématique pour deux ensembles A et B est donnée par:

$$DSC(A, B) = \frac{2|A \cap B|}{|A| + |B|} \quad (1)$$

Dans le contexte d'une segmentation binaire comparant un masque prédit P à un masque de référence (ground truth) G , cette formule devient:

$$DSC = \frac{2TP}{2TP + FP + FN} \quad (2)$$

avec:

- **Vrais positifs (TP):** pixels correctement classifiés comme appartenant à la structure d'intérêt
- **Faux positifs (FP):** pixels incorrectement classifiés comme appartenant à la structure
- **Faux négatifs (FN):** pixels appartenant à la structure mais classifiés à tort comme fond

A.2. Indice de Jaccard (IoU) L'indice de Jaccard, ou Intersection over Union (IoU), a été développé par Paul Jaccard en 1901 ([Jaccard 1901](#)) et est défini par:

$$IoU(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (3)$$

En segmentation binaire:

$$IoU = \frac{TP}{TP + FP + FN} \quad (4)$$

L'IoU est plus sensible aux erreurs de segmentation que le DSC, et la relation entre les deux est:

$$IoU = \frac{DSC}{2 - DSC} \quad \text{et} \quad DSC = \frac{2 \times IoU}{1 + IoU} \quad (5)$$

B. Métriques basées sur la distance

B.1. Distance de Hausdorff (HD) La distance de Hausdorff ([Hausdorff 1914](#)) quantifie la distance spatiale maximale entre les contours des segmentations et est définie par:

$$HD(A, B) = \max\{\sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(a, b)\} \quad (6)$$

Une version robuste, la distance de Hausdorff au 95e percentile (HD95), est souvent utilisée en segmentation sémantique.

B.2. Distance de Surface Moyenne (ASD) La distance de surface moyenne est définie comme:

$$ASD(A, B) = \frac{1}{|S_A| + |S_B|} \left(\sum_{a \in S_A} \min_{b \in S_B} d(a, b) + \sum_{b \in S_B} \min_{a \in S_A} d(a, b) \right) \quad (7)$$

Cette métrique mesure la distance moyenne entre les contours des segmentations.

C. Considérations pour la segmentation multi-classes

En segmentation multi-classes, les métriques sont calculées pour chaque classe k et agrégées selon:

$$mDSC = \frac{1}{K} \sum_{k=1}^K DSC_k, \quad mIoU = \frac{1}{K} \sum_{k=1}^K IoU_k \quad (8)$$

Des approches comme la macro-moyenne (moyenne des indices de chaque classe) et la micro-moyenne (calculer l'indice à partir des moyennes de TP/TN/FN de chaque classe) sont utilisées pour tenir compte de l'importance relative des classes.

D. Considérations pour notre étude

Par manque de temps, notre analyse s'appuie principalement sur le mean IoU, et la moyenne des准确ies des pixels.

4. Modèles implémentés et paramétrage

Cette section détaille les trois architectures de segmentation exploitées. Toutes dérivent du paradigme U-Net, mais chacune présente des approches distinctes pour résoudre la tâche de segmentation.

A. U-Net original

Le U-Net original ([Ronneberger et al. 2015](#)) constitue notre architecture de référence. Il s'agit d'un réseau convolutif en forme de "U" symétrique composé d'un chemin de contraction (encodeur) et d'un chemin d'expansion (décodeur), avec des connexions directes entre les niveaux correspondants (skip connections).

A.1. Architecture L'implémentation (fig. 3) suit fidèlement la proposition originale de Ronneberger et al. (2015) :

- **Encodeur** : Succession de blocs de double convolution (Conv3×3, ReLU, Conv3×3, ReLU) suivis d'opérations de max pooling (2×2), doublant progressivement le nombre de features maps (64, 128, 256, 512).
- **Fond du U** : Bloc de convolution double avec 1024 features maps cette fois.
- **Décodeur** : Succession de convolutions transposées (upsampling), concaténation avec les cartes de caractéristiques de la couche d'encodeur correspondante, puis blocs de double convolution.
- **Couche finale** : Convolution avec un kernel 1×1, suivie d'activation softmax pour générer les probabilités des 4 classes (fond, poumon droit, poumon gauche, cœur).

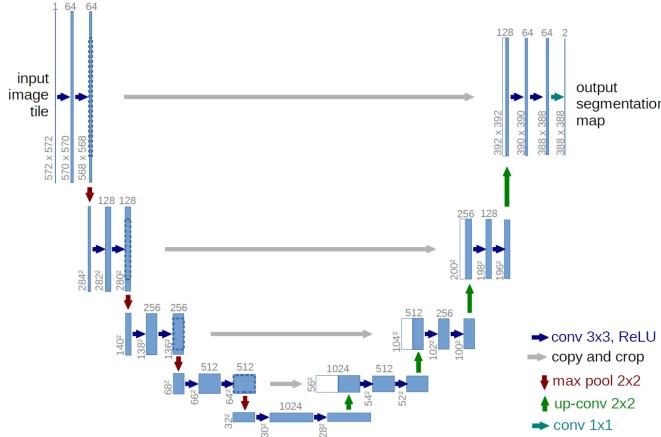


Figure 3 Architecture du U-Net originale avec ses chemins de contraction (gauche) et d'expansion (droite). Les flèches grises représentent les skip connections.

Contrairement à l'implémentation d'origine qui utilisait des images 572×572 avec convolutions sans padding, nous utilisons des images en 512x512 avec un padding de 1 en phase de contraction. Ce padding nous simplifiait le problème de la gestion des tailles et des concaténations entre couches.

A.2. Caractéristiques d'entraînement Le modèle U-Net original présente les spécificités suivantes:

- **Nombre de paramètres** : environ 31 millions, tous entraînables
- **Fonction de perte** : Combinaison de Cross-Entropy et Dice Loss (pondération 0,5), cette dernière étant particulièrement adaptée aux problèmes de segmentation avec déséquilibre de classes
- **Optimiseur** : Adam avec learning rate initial de 1e-4
- **Learning rate schedule** : Réduction sur plateau (facteur 0,5, patience 5 époques)

B. U-Net avec backbone ResNet

Notre deuxième architecture exploite le transfert d'apprentissage en remplaçant l'encodeur classique du U-Net par un réseau ResNet50 pré-entraîné sur ImageNet.

B.1. Architecture Cette variante (fig. 4) combine la puissance d'extraction de caractéristiques des réseaux résiduels profonds avec la capacité de segmentation du U-Net :

- **Encodeur** : ResNet50 pré-entraîné, dont les poids sont gelés pour préserver les caractéristiques apprises sur ImageNet.
- **Décodeur** : Architecture classique de U-Net avec opérations d'upsampling, skip connections et convolutions.
- **Skip connections** : Connectent les blocs résiduels de ResNet aux couches correspondantes du décodeur.

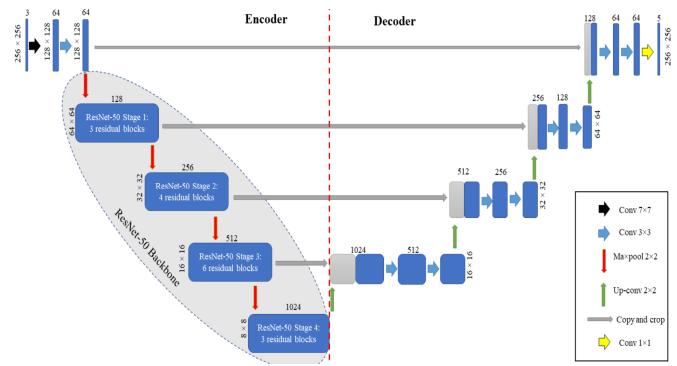


Figure 4 U-Net avec backbone ResNet50. L'encodeur (partie gauche) utilise les blocs résiduels pré-entraînés de ResNet, tandis que le décodeur (partie droite) conserve l'architecture U-Net classique.

Cette approche permet d'exploiter des caractéristiques visuelles générées apprises sur de grands corpus d'images, tout en adaptant le modèle à notre tâche spécifique de segmentation. L'implémentation repose sur la bibliothèque segmentation_models_pytorch, facilitant l'intégration de divers backbones pré-entraînés.

B.2. Caractéristiques d'entraînement Le modèle U-Net avec ResNet50 présente le profil suivant :

- **Nombre de paramètres total** : environ 32 millions soit une taille équivalente à celle du U-NET original
- **Paramètres gelés** : 23,56 millions (encodeur ResNet50)
- **Paramètres entraînables** : 9,15 millions (28% du total)
- **Fonction de perte** : Identique au U-Net original
- **Optimiseur** : Adam avec paramètre filtrant uniquement les poids entraînables

Il était possible d'entraîner tous les paramètres du modèle mais le gel des poids de l'encodeur présentait deux avantages majeurs : réduction du risque de surapprentissage sur notre corpus limité et diminution significative du temps d'entraînement, les mises à jour de gradients n'étant calculées que pour environ un tiers des paramètres.

C. TransUNet

La troisième architecture, TransUNet, incorpore des mécanismes d'attention issus des transformers, suivant le paradigme introduit par (Chen *et al.* 2021) dans leur article "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation".

C.1. Architecture TransUNet (fig. 5) combine CNN et transformateurs pour exploiter à la fois les caractéristiques locales et les dépendances entre caractéristiques distantes:

- **Encodeur CNN initial** : Extraction des caractéristiques locales via une série de blocs convolutifs (filtres : 32, 64, 128, 256, 512)
- **Module Transformer** : Traitement des caractéristiques extraites par l'encodeur CNN à travers 4 blocs transformateur avec mécanisme d'attention multi-têtes (8 têtes d'attention, dimension d'embedding 256)
- **Décodeur CNN** : Structure similaire au U-Net classique avec skip connections depuis l'encodeur CNN

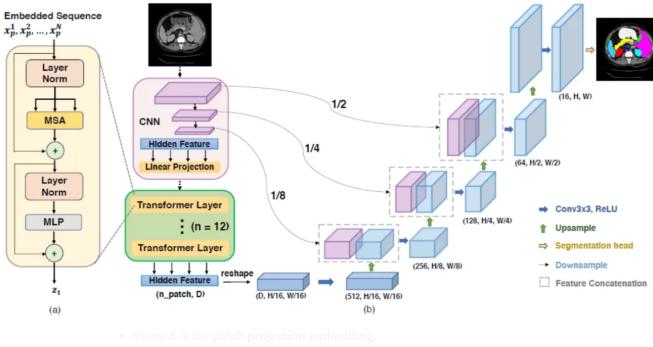


Figure 5 Architecture TransUNet originale (la notre a moins de paramètres) intégrant un module transformateur entre l'encodeur CNN et le décodeur. Les caractéristiques extraites par l'encodeur sont transformées en séquence de tokens, traitées par les blocs transformateur, puis reconvertis en cartes de caractéristiques 2D.

La particularité de cette architecture réside dans l'intégration du mécanisme d'attention, permettant au modèle de pondérer dynamiquement l'importance relative des différentes régions de l'image. Ainsi, le TransUNet est capable de capturer efficacement des contextes globaux tout en conservant les propriétés locales extraites par les CNN traditionnels. Contrairement aux convolutions classiques, dont les champs récepteurs sont fixes, l'attention permet une modélisation explicite et adaptative des dépendances spatiales complexes entre les régions d'intérêt.

C.2. Caractéristiques d'entraînement Le modèle TransUNet implementé est trois fois plus petit en nombre de paramètres que le U-NET original :

- **Nombre de paramètres** : environ 12 millions, tous entraînables
- **Module Transformer** : 4 blocs transformateur, 8 têtes d'attention, dimension d'embedding 256
- **Optimiseur** : AdamW avec learning rate dégressif.

5. Protocole d'expérimentation

A. Stratégie d'entraînement

Les trois modèles ont été entraînés avec un protocole harmonisé pour garantir une comparaison équitable :

- **dataset** : Images 512×512 converties en gray-scale pour les radiographies et les masques, et normalisées. Les masques en RGB seront exploités dans un autre projet pour tester la

segmentation multi-labels. Dans ce projet, de segmentation sémantique, l'appartenance d'un pixel à une classe est donnée en priorité à l'objet le plus en amont de l'image, soit le cœur dans notre cas. Ainsi la conversion en gray scale a pris cette priorité en compte, et par conséquent, lorsque le masque d'un poumon a des pixels communs avec le cœur, ce pixel est attribué au cœur.

- **Fonction de perte hybride** : Cross-Entropy + $0,5 \times$ Dice Loss
- **Nombre d'époques** : 50
- **Learning rate schedule** : Réduction sur plateau avec monitoring de la perte de validation
- **Checkpointing** : Conservation du meilleur modèle selon IoU et selon perte de validation

B. Environnement d'exécution

Tous les entraînements ont été réalisés sur la plateforme GCP, sur une machine virtuelle équipée d'un GPU NVIDIA A100 (40 GB VRAM). Cette infrastructure cloud a permis d'exploiter des ressources de calcul intensives sans nécessiter d'installation matérielle dédiée. L'environnement logiciel repose sur PyTorch 2.3.0 avec CUDA 11.8.

6. Résultats

Cette section présente les résultats obtenus par nos trois architectures de segmentation sur le corpus de radiographies thoraciques labellisées. Nous analysons d'abord le comportement des modèles durant l'entraînement, puis examinons leurs performances comparatives sur l'ensemble de test.

A. Analyse de la convergence et du pas d'apprentissage

Les trois modèles ont été entraînés sur 50 époques avec le protocole énoncé plus haut. Les courbes d'entraînement révèlent des dynamiques d'apprentissage distinctes pour chaque architecture.

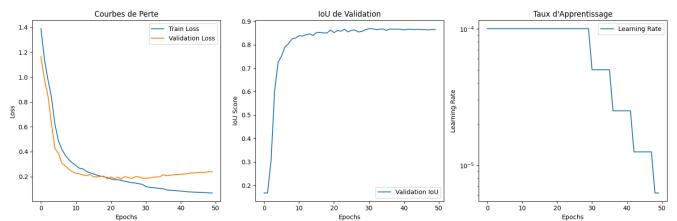


Figure 6 Courbes d'entraînement du modèle U-Net original montrant l'évolution de la perte (gauche), de l'IoU de validation (centre) et du taux d'apprentissage (droite) sur 50 époques.

Le U-Net original (fig. 6) présente une convergence rapide, atteignant un plateau de performance après environ 10 époques. La courbe de perte révèle une nette séparation entre perte d'entraînement et de validation après 20 époques, suggérant l'apparition d'un léger surapprentissage, bien que l'IoU de validation continue à progresser marginalement. Le taux d'apprentissage diminue par paliers, conformément à la stratégie de scheduling.

Le modèle U-Net avec backbone ResNet50 (fig. 7) démontre une dynamique d'apprentissage différente. On observe une convergence encore plus rapide de la perte d'entraînement qui

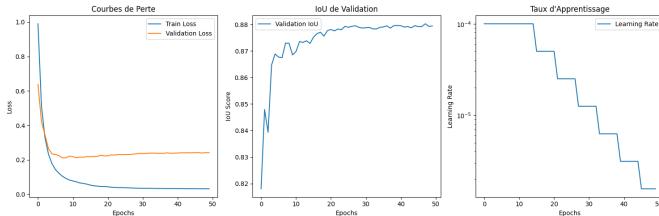


Figure 7 Courbes d’entraînement du modèle U-Net avec backbone ResNet50 montrant l’évolution de la perte (gauche), de l’IoU de validation (centre) et du taux d’apprentissage (droite) sur 50 époques.

atteint des valeurs proches de zéro, tandis que la perte de validation se stabilise autour de 0,25. L’écart significatif entre les deux courbes de perte suggère un sur-apprentissage plus prononcé, malgré le gel des poids de l’encodeur. Cependant, l’IoU de validation continue de progresser graduellement jusqu’à environ 30 époques avant d’atteindre un plateau.

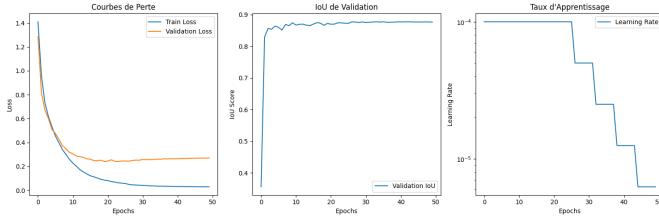


Figure 8 Courbes d’entraînement du modèle TransUNet montrant l’évolution de la perte (gauche), de l’IoU de validation (centre) et du taux d’apprentissage (droite) sur 50 époques.

Le modèle TransUNet (fig. 8) présente un comportement intermédiaire. Sa convergence est rapide, avec un écart qui se creuse entre les courbes de perte d’entraînement et de validation à partir de la quinzième epoch environ. L’IoU de validation est la plus rapide à converger et atteint un plateau après environ 10 époques. Malgré son nombre réduit de paramètres (12 millions contre 31-32 millions pour les autres architectures), le TransUNet parvient à capturer efficacement les caractéristiques pertinentes pour la segmentation.

Pour les trois modèles, le mécanisme de réduction du pas d’apprentissage sur plateau a été déclenché plusieurs fois, comme l’attestent les paliers visibles dans les courbes du learning rate. Cette stratégie a permis d’affiner l’optimisation des poids en fin d’entraînement.

B. Performances sur l’ensemble de test

Les tableaux 1 et 2 donnent un aperçu quantitatif des performances des trois architectures sur l’ensemble de test.

Table 1 Performances globales des architectures sur l’ensemble de test.

Modèle	mIoU	Accuracy pixel-wise
U-Net original	0,8885	0,9564
U-Net+ResNet50	0,8980	0,9605
TransUNet	0,8936	0,9588

Table 2 IoU par classe pour chaque architecture.

Modèle	Background	Poumon D	Poumon G	Cœur
U-Net original	0,9408	0,8963	0,8517	0,8653
U-Net+ResNet50	0,9465	0,9108	0,8539	0,8809
TransUNet	0,9442	0,9019	0,8552	0,8731

L’analyse des performances révèle plusieurs tendances:

- **Performance globale** : Les trois architectures atteignent des performances élevées, avec des scores mIoU (moyenne) supérieurs à 0,88 et une Accuracy pixel-wise dépassant 0,95 sur l’ensemble de test. Ces résultats attestent de l’efficacité des architectures dérivées du U-Net pour la segmentation d’imagerie médicale.
- **Supériorité du modèle avec backbone pré-entraîné** : Le U-Net avec backbone ResNet50 surpasse légèrement les autres architectures sur toutes les métriques, avec un mIoU de 0,8980 et une Accuracy de 0,9605. Ce résultat souligne l’avantage du transfer learning pour cette tâche, même avec un nombre réduit de paramètres entraînables, de l’ordre d’une dizaine de millions dans notre cas.
- **Efficience du TransUNet** : Malgré un nombre de paramètres trois fois inférieur à celui des autres modèles, le TransUNet obtient des performances comparables et surpassé même légèrement le U-Net original (mIoU de 0,8936 contre 0,8885). Ce résultat valide l’efficacité des mécanismes d’attention pour capturer les relations spatiales complexes dans l’imagerie médicale.
- **Variations par classe** : Pour les trois modèles, la segmentation du poumon droit présente les meilleurs scores IoU (0,8963-0,9108), suivie par celle du poumon gauche (0,8517-0,8552) et du cœur (0,8653-0,8809). Cette tendance peut s’expliquer par la variabilité morphologique plus importante du cœur et les superpositions fréquentes entre structures cardiaques et pulmonaires.
- **Class Imbalance** : On note une performance systématiquement supérieure pour la classe d’arrière-plan (background), avec des IoU entre 0,9408 et 0,9465. Cette observation est cohérente avec la distribution déséquilibrée des classes dans les images, où l’arrière-plan représente une proportion significative des pixels.

C. Analyse qualitative des résultats

C.1. Vue d’ensemble des résultats visuellement Les Annexes 1, 2 et 3 (voir fig. 12, fig. 13 et fig. 14) présentent des exemples visuels de segmentations produites par chaque modèle, permettant d’apprécier qualitativement leurs caractéristiques. L’inspection visuelle révèle que les trois architectures produisent des segmentations cohérentes avec l’anatomie thoracique, avec une supériorité du modèle avec backbone Resnet50 dans la précision des contours.

C.2. Difficultés des modèles L’analyse des cas les plus problématiques pour chaque architecture permet d’identifier des motifs d’échecs récurrents et de caractériser les limitations spécifiques des différentes approches de segmentation.

Les figures 9, 10 et 11 présentent les trois cas les plus difficiles pour chaque modèle, la radio 20 (dans l’ordre du dataset test) apparaît dans le top 3 des difficultés des trois modèles. Et les radios 18-20-29 (toujours dans l’ordre du dataset de test) sont

communes aux modèles UNET et TransUNET dans le même ordre de difficulté.

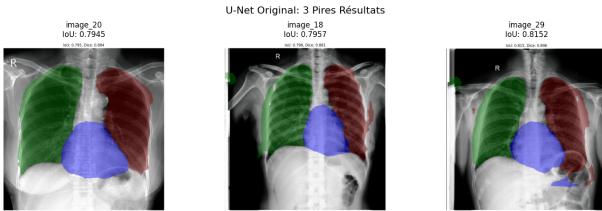


Figure 9 Les trois pires segmentations produites par le U-Net original, présentant des erreurs significatives aux interfaces cardio-pulmonaires et aux régions périphériques des poumons.

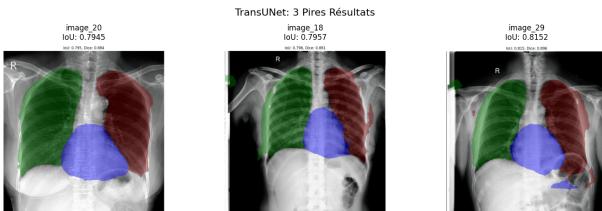


Figure 10 Les trois pires segmentations produites par le TransUNet, montrant des erreurs similaires au U-Net original, malgré l'intégration de mécanismes d'attention.

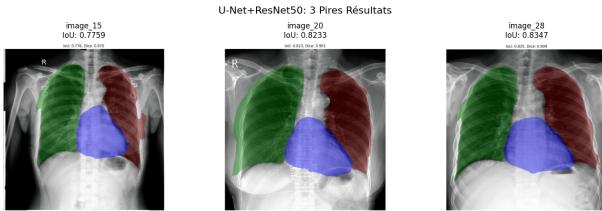


Figure 11 Les trois pires segmentations produites par le U-Net avec backbone ResNet50, qui présente une meilleure performance sur les cas difficiles comparé aux autres architectures.

Caractérisation des erreurs de segmentation L'analyse qualitative des cas difficiles révèle des patterns spécifiques d'erreurs:

- **Erreurs de précision de contours anatomiques** — Les trois modèles montrent des imprécisions aux interfaces de faible contraste, particulièrement à la frontière costale du poumon gauche. Ces erreurs se manifestent par une délimitation imprécise des silhouettes anatomiques et une irrégularité du trait.
- **Erreurs de classification distantes** — Des pixels isolés incorrectement classifiés, pouvant former de petits îlots, apparaissent à distance de l'organe cible. L'exemple le plus flagrant étant la classification en poumon droit (couleur verte) de pixels complètement en dehors du corps du sujet, par UNET et TransUNET.
- **Surestimation de la zone d'intérêt** — En particulier pour le poumon droit (couleur verte), les trois modèles montrent une tendance à déborder de la zone d'intérêt en particulier sur le bord des côtes.

Définitions entre architectures Le U-Net original et le TransUNet présentent des difficultés sur les mêmes cas ('image_20', 'image_18', 'image_29'; IoU: 0.7945, 0.7957, 0.8152), suggérant que les mécanismes d'attention n'apportent pas d'avantage décisif face aux ambiguïtés de contraste. Le TransUNet montre néanmoins une meilleure cohérence globale des segmentations, avec moins d'artefacts isolés.

Le U-Net avec backbone ResNet50 ('image_15', 'image_20', 'image_28'; IoU: 0.7759, 0.8233, 0.8347) démontre une performance significativement supérieure sur les cas difficiles, avec une meilleure précision dans la délimitation des contours anatomiques complexes et une meilleure tolérance aux variations morphologiques.

7. Discussion

L'analyse comparative des trois architectures dérivées de U-Net révèle plusieurs éléments significatifs pour la segmentation multiclasse d'imagerie médicale.

A. Efficacité du transfert d'apprentissage

La supériorité du U-Net+ResNet50 (mIoU 0,8980) confirme l'efficacité du transfert d'apprentissage en segmentation radiologique, malgré la différence entre les domaines source et cible. Trois mécanismes expliquent cette performance:

- Les couches initiales de ResNet50 extraient des primitives visuelles (contours, textures) transférables à l'imagerie médicale
- Le gel des poids de l'encodeur (72% des paramètres) agit comme régularisation, limitant le surapprentissage sur notre corpus restreint
- L'optimisation sélective de 9,15 millions de paramètres améliore l'efficience computationnelle sans compromettre la précision

Ces observations corroborent les résultats de [Raghunathan et al. \(2019\)](#) démontrant l'efficacité du Transfer Learning en imagerie médicale.

B. Avantages des mécanismes d'attention

Le TransUNet, avec seulement 12 millions de paramètres, obtient un mIoU de 0,8936, surpassant le U-Net original (0,8885) malgré une taille réduite. Cette efficience s'explique par:

- La capacité des transformateurs à modéliser simultanément les dépendances locales et distantes, pertinentes pour capturer les relations anatomiques
- La pondération dynamique des caractéristiques par l'attention multi-têtes, optimisant l'exploitation contextuelle des informations
- La complémentarité entre l'extraction locale des CNN et l'intégration globale des transformateurs

Ces résultats quantitatifs démontrent l'avantage computationnel des mécanismes d'attention en segmentation médicale, confirmant les observations initiales de [Chen et al. \(2021\)](#).

C. Gradient de performance par classe anatomique

La hiérarchie constante de performance (poumon droit > poumon gauche > cœur) reflète des contraintes intrinsèques aux données radiographiques:

- Le cœur présente une variabilité morphologique supérieure aux poumons

8 SEGMENTATION par CV

- Le poumon gauche souffre de superpositions avec le médiastin, créant des frontières moins distinctes
- Le poumon droit bénéficie d'un meilleur contraste avec les tissus adjacents

Cette distribution asymétrique des erreurs suggère l'intérêt d'approches spécifiques par classe ou l'intégration de contraintes anatomiques dans les futures architectures.

D. Limitations méthodologiques

Quatre limitations principales affectent la généralisation de nos résultats:

- Corpus restreint (208 radiographies) limitant la diversité morphologique représentée
- Exclusion des cas complexes (dispositifs médicaux, pathologies sévères) réduisant l'applicabilité clinique directe
- Métriques centrées sur le chevauchement spatial (IoU) sans évaluation des distances ou erreurs topologiques
- Absence d'annotations multiples empêchant la comparaison à la variabilité inter-observateur. En effet, ma propre labellisation de la même radiographie était très différente d'un essai à un autre. Face à une radiographie, quand on n'est pas professionnel, le cerveau a tendance à fabriquer des limites et contours subjectifs en fonction de ce qu'il a vu avant ou de l'état de fatigue.

8. Conclusion

Cette étude comparative essaie de quantifier l'efficacité d'architectures dérivées de U-Net pour la segmentation multiclasse .

Le U-Net+ResNet50 démontre une supériorité statistique grâce au transfert d'apprentissage et à l'optimisation ciblée de 28% des paramètres, particulièrement efficace sur un corpus limité.

Le TransUNet établit un compromis performance-complexité optimal (avec 12 millions de paramètres), validant empiriquement l'efficience des mécanismes d'attention pour capturer les relations anatomiques complexes.

La variation systématique des performances par classe reflète l'impact de la variabilité morphologique et des superpositions anatomiques sur la précision de segmentation.

Ces résultats suggèrent trois axes de développement: l'intégration de contraintes anatomiques explicites, l'optimisation spécifique par classe, et l'exploration d'architectures hybrides exploitant conjointement le transfert d'apprentissage et les mécanismes d'attention pour maximiser le rapport performance-paramètres.

Littérature citée

- Buslaev A, Ilyas T, Seferbekov S, Kalinin AA. 2020. Albumentations: Fast and flexible image augmentations. arXiv preprint arXiv:1809.06839..
- Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y. 2021. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.
- Dice LR. 1945. Measures of the amount of ecologic association between species. *Ecology*. 26:297–302.
- Hausdorff F. 1914. *Grundzüge der Mengenlehre*. Veit & Comp.. Leipzig.
- Jaccard P. 1901. Étude comparative de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*. 37:547–579.
- Raghuram M, Zhang C, Kleinberg J, Bengio S. 2019. Transfusion: Understanding transfer learning for medical imaging. In: . volume 32. pp. 3347–3357.
- Ronneberger O, Fischer P, Brox T. 2015. U-net: Convolutional networks for biomedical image segmentation. In: . pp. 234–241. Springer.
- Rudnicka Z, Szczepanski J, Pregowska A. 2024. Artificial intelligence-based algorithms in medical image scan segmentation and intelligent visual-content generation – a concise overview. arXiv preprint arXiv:2401.09857..
- Sakinis T, Milletari F, Roth H, Korfiatis P, Kostandy P, Philbrick K, Akkus Z, Xu Z, Xu D, Erickson BJ. 2019. Interactive segmentation of medical images through fully convolutional neural networks. arXiv preprint arXiv:1903.08205. .
- Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). .

APPENDIX

Annexe 1 : Échantillons de prédictions de U-Net original

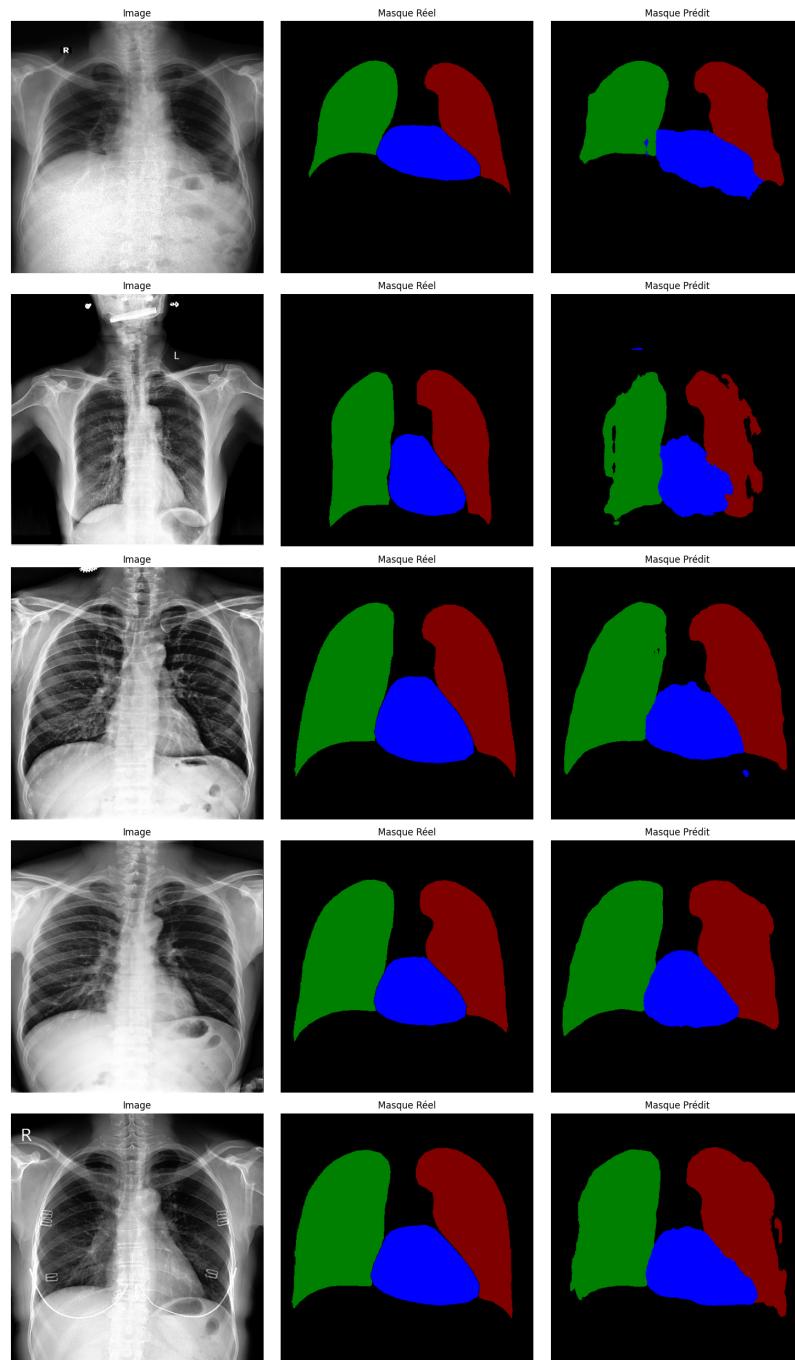


Figure 12 Exemples de segmentations réalisées par le modèle U-Net original. Chaque ligne présente un cas différent. Pour chaque cas, la première colonne montre la radiographie originale, la deuxième colonne présente le masque de référence et la troisième colonne montre la prédiction du modèle. Le poumon droit est représenté en vert, le poumon gauche en rouge, et le cœur en bleu.

Annexe 2 : Échantillons de prédictions de U-Net avec backbone ResNet

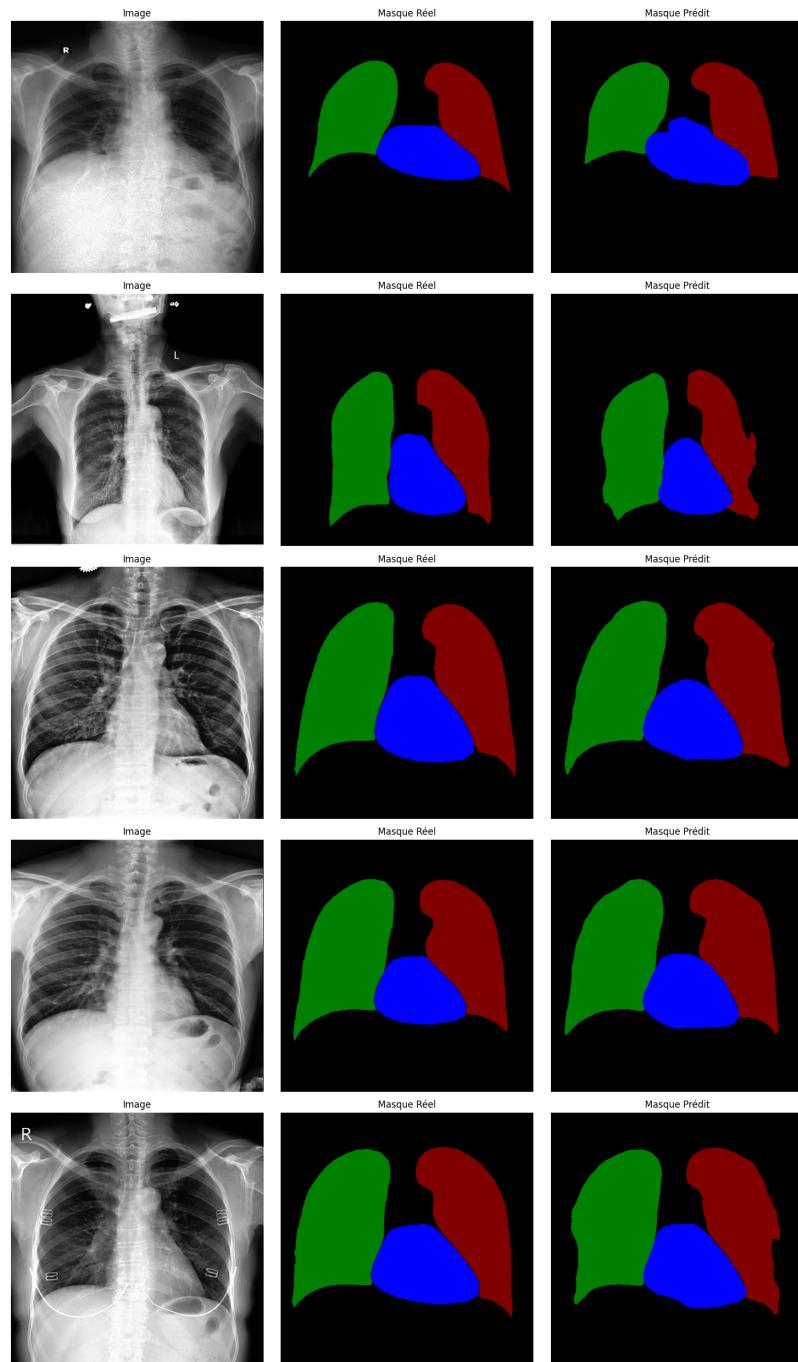


Figure 13 Exemples de segmentations réalisées par le modèle U-Net avec backbone ResNet. Chaque ligne présente un cas différent. Pour chaque cas, la première colonne montre la radiographie originale, la deuxième colonne présente le masque de référence et la troisième colonne montre la prédiction du modèle. Le poumon droit est représenté en vert, le poumon gauche en rouge, et le cœur en bleu.

Annexe 3 : Échantillons de prédictions de TransUNet

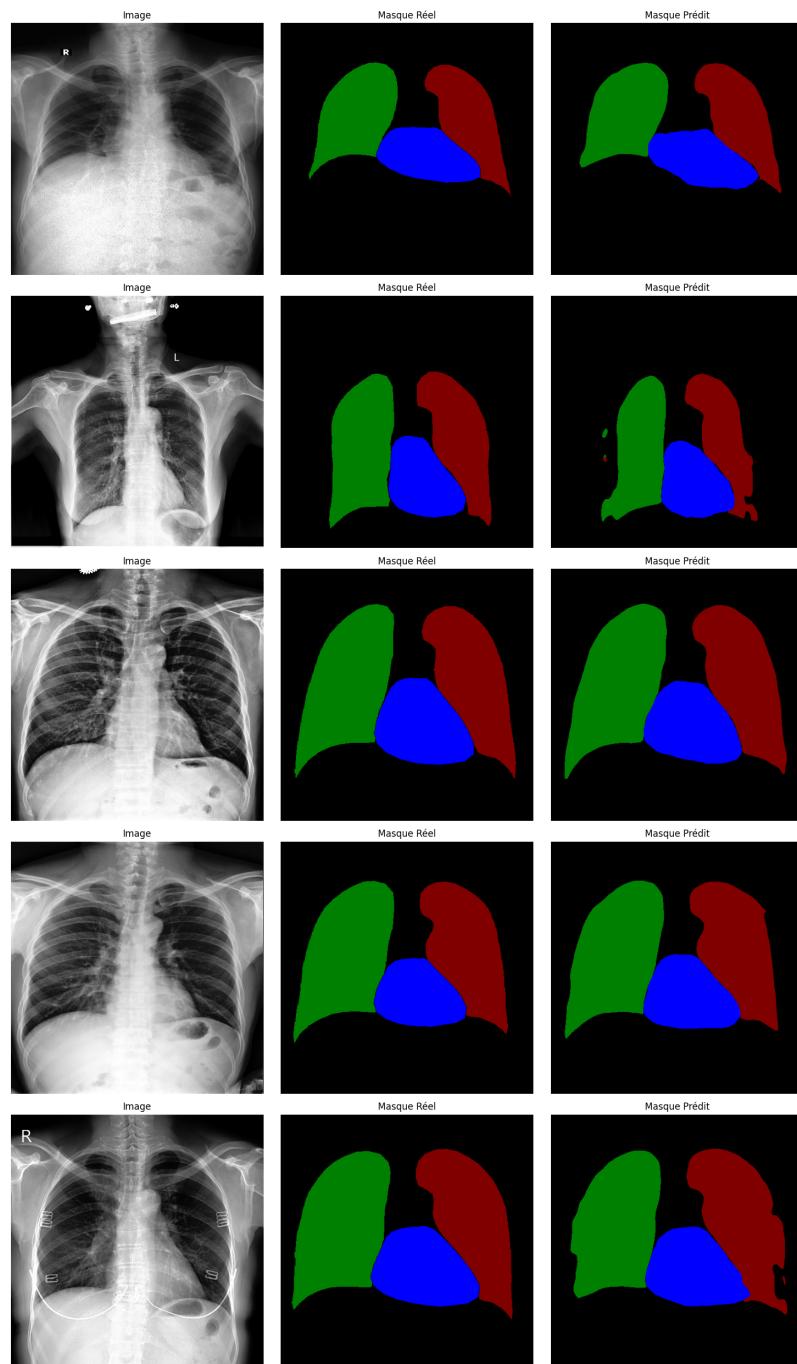


Figure 14 Exemples de segmentations réalisées par le modèle TransUNet. Chaque ligne présente un cas différent. Pour chaque cas, la première colonne montre la radiographie originale, la deuxième colonne présente le masque de référence et la troisième colonne montre la prédiction du modèle. Le poumon droit est représenté en vert, le poumon gauche en rouge, et le cœur en bleu.