CUMUNEL Lucas , LEROUX Tara, LEROY Léo, SIAHAAN–GENSOLLEN Rémy
Encadré par : KIRSCHER Tristan , COUBEZ Xavier

Réduction de l'incertitude en segmentation médicale par méthodes d'ensemble

Note de synthèse de projet de statistique et science des données appliquées

ENSAE

INSTITUT POLYTECHNIQUE DE PARIS

RÉSUMÉ

La segmentation automatique des organes, bien que très utile en imagerie médicale, reste sujette à une forte incertitude, notamment lorsqu'elle repose sur des annotations manuelles potentiellement subjectives. Ce projet présente une évaluation systématique de méthodes d'ensemble de réseaux U-Net pour réduire cette incertitude. Nous entraînons et inférons plusieurs modèles sur des scans tomodensitométriques de patients annotés par différents experts, que nous combinons à l'aide d'une méthode d'ensemble. Ensuite, nous proposons et appliquons à ces prédictions un cadre d'évaluation systématique de leur précision, de leur incertitude aléatoire et de leur incertitude épistémique. Nos résultats indiquent que les méthodes d'ensemble utilisées diminuent significativement les incertitudes des prédictions sans détériorer leur précision.

I Projet

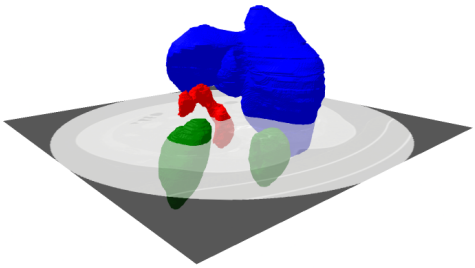

Figure 1 *Segmentation 3D du pancréas, des reins et du foie, accompagnée d'une coupe axiale du scanner abdominal utilisé.*

L'intelligence artificielle révolutionne l'imagerie médicale en automatisant des tâches complexes comme la segmentation d'organes à partir de scanners. Les réseaux de neurones convolutifs, en particulier les architectures U-Net, sont aujourd'hui incontournables pour ces applications. Ils permettent de délimiter avec précision les structures anatomiques, facilitant ainsi le diagnostic, la planification thérapeutique et le suivi clinique. Cependant, la segmentation automatique reste confrontée à une incertitude importante. Cette incertitude provient de la complexité anatomique des structures à segmenter, mais aussi de la variabilité inter-experts : les spécialistes eux-mêmes peuvent diverger sur les contours exacts à tracer. Ces incertitudes sont amplifiées dans les scénarios multi-classes où plusieurs organes sont segmentés simultanément. Une solution explorée par ce projet de statistiques appliquées vise à évaluer l'impact d'une méthode d'ensemble sur les incertitudes des réseaux de neurones entraînés avec des initialisations différentes et des jeux d'annotations différentes pour un même patient. Nous évaluons en particulier l'impact sur les incertitudes aléatoires (causées par les données, dans ce cas les annotations multiples) et les incertitudes épistémiques (engendrées par le modèle).

II Expérience

Ce projet repose sur le jeu de données du challenge CURVAS (*Calibration and Uncertainty for Multi-Rater Volume Assessment in Multiorgan Segmentation*), organisé de mai à octobre 2024. L'objectif était de développer un modèle de segmentation performant, capable de quantifier finement la variabilité inter-experts. Le jeu comprend 90 CT scans de patients anonymisés, chacun annoté par trois experts pour segmenter le pancréas, les reins et le foie. L'entraînement a été réalisé avec le framework `nnU-Net`, basé sur les U-Net classiques, mais adapté automatiquement aux caractéristiques des données médicales. Neuf modèles ont été entraînés : trois par annotateur avec des initialisations différentes (seeds `112233`, `445566`, `778899`). Les modèles ont ensuite été inférés sur les 65 patients de test, générant des sorties softmax combinées en quatre ensembles (un par annotateur, un général). Chaque patient a ainsi 13 prédictions, évaluées en précision, incertitude aléatoire et épistémique.

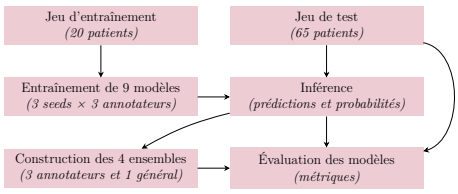

Figure 2 *Diagramme du protocole expérimental*

Pour nos besoins, `nnU-Net` a été modifié : ajout d'un *early stopping* (300 epochs max ou arrêt forcé si aucun progrès n'a lieu pendant 20 epochs) et d'une initialisation déterministe des poids pour assurer la reproductibilité. Les entraînements 3D, réalisés sur GPU via Onyxia (Insee, Genes), prenaient environ 24h par modèle. Les limites de stockage (100 Go/instance) ont imposé de fragmenter les tâches (modèle/patient) et de les traiter séparément. Nous avons donc conçu une interface en ligne de commande (CLI) en `Python` avec `Typer`, centralisant l'ensemble du workflow (entraînement, inférence, envoi vers un espace S3 MinIO alloué par l'Insee). L'application comprend plus de 6 500 lignes de code et gère plusieurs To d'artefacts. Enfin, les calculs de métriques ont été optimisés avec `numba`, compilant certaines fonctions critiques via `llvm`, accélérant jusqu'à 100 fois les opérations lourdes, notamment voxel-à-voxel, et divisant par 4 le temps total d'évaluation.

III Évaluation

Pour mesurer la performance des modèles, nous mesurons les écarts entre les annotations et les prédictions. Ces écarts sont obtenus selon les métriques suivantes : le consensus-based DICE, le CRPS, et la distance de Hausdorff. Pour mesurer l'incertitude aléatoire ici, nous mesurons l'incertitude dans les annotations grâce à la Confiance (ou Uncertainty Assessment) et la NCC.

Les métriques d'incertitude épistémique peuvent se décomposer en deux catégories. La première regroupe les métriques traitant des erreurs de *calibration*, c'est-à-dire de décalage entre la confiance d'un modèle dans une prédiction et sa précision réelle. La deuxième catégorie regroupe quant à elle les métriques traitant de la reconnaissance d'erreur, c'est-à-dire la quantification de la capacité du modèle à reconnaître les erreurs qu'il commet. Pour la première catégorie, nous utilisons l'ECE et l'ACE. La reconnaissance d'erreur peut se faire, elle, à l'aide de métriques comme l'AUROC, l'AURC et l'EAURC. L'entropie est une autre métrique, qui peut être calculée à l'échelle de l'image, ou pour chaque voxel, et nous permet notamment de générer des cartes d'incertitudes comme en figure 3 ci-contre.

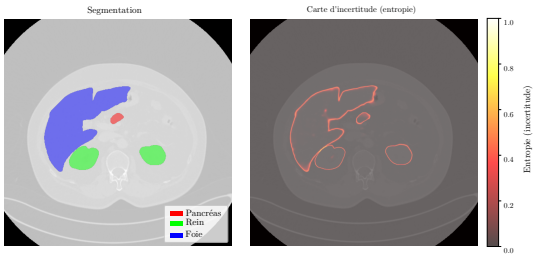

Figure 3 *Carte d'incertitude (à droite) d'une coupe de la segmentation des organes sur le patient 3, selon l'entropie de Shannon des probabilités de sortie du modèle d'ensemble général (sur les 9 modèles initiaux)*

IV Résultats

Malgré nos efforts, certaines étapes n'ont pas pu aboutir. Seuls 61 patients sur 65 ont pu être traités. Une erreur a été détectée trop tard dans le calcul de la NCC, due à une mauvaise conversion des annotations en probabilités. Les autres métriques restent valides. On note certaines valeurs extrêmes (DICE nuls, ECE anormaux), qui pourraient provenir d'erreurs ponctuelles. Elles n'affectent pas la tendance générale des résultats.

Nos résultats indiquent clairement que les modèles d'ensemble permettent de réduire l'incertitude. La figure 4 ci-dessous montre les graphiques « violon » indiquant la distribution de l'ECE moyen (moyenne des ECE sur les 3 annotateurs) avec une boite à moustache classique. Elle indique également chaque point de la distribution, correspondant à chaque patient et modèle (donc 3 fois plus de points pour les modèles individuels que pour les ensemble par annotateur, et 9 fois plus que pour l'ensemble général).

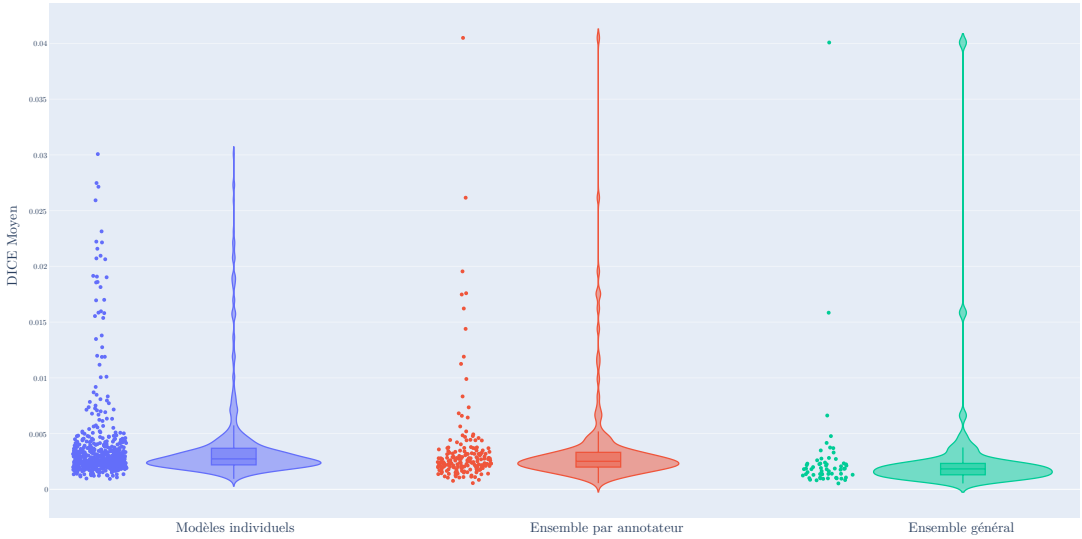

Figure 4 *ECE moyen pour les différents modèles*

On observe une grande réduction dans le nombre de valeurs extrêmes. Les moyennes et médianes de l'ECE diminuent, indiquant une diminution de l'incertitude. Le test U de Mann–Whitney unilatéral (alternative « less ») entre chaque paire de groupes montre que cette diminution de l'ECE est significative.

Concernant la performance de prédiction, on observe que le DICE moyen sur les trois organes de l'ensemble global est supérieur à celui des ensembles par annotateurs, lui-même supérieur à celui des modèles individuels. Ce résultat n'est pas statistiquement significatif, on retiendra donc que les ensemble ne détériorent pas la performance.

V Comparaison avec le benchmark du challenge CURVAS

En mai 2025, les résultats du challenge CURVAS ont été publiés [Riera-Marin et al., 2025]. 7 équipes professionnelles ont proposés des méthodes pour réduire l'incertitude, en utilisant des méthodes différentes d'agrégation des données ou des architectures de réseau de neurones différentes. Voici les principaux résultats, en comparaison avec nos performances.

	Métriques (valeurs moyennes)			
	DSC (%)	Confiance (%)	ECE ($\times 10^{-3}$)	CRPS (cm^3)
MedIG	**94.57 (1)**	**97.87 (1)**	**1.82 (2)**	**8.108 (1)**
PrAEcision	**93.29 (2)**	**97.18 (2)**	**2.22 (3)**	**10.438 (3)**
BreizhSeg	92.60 (4)	**97.17 (3)**	**1.61 (1)**	12.326 (4)
DLAI	**92.72 (3)**	96.23 (4)	3.90 (4)	12.625 (5)
BCNAIM	90.52 (5)	95.88 (5)	6.21 (6)	**9.727 (2)**
CAI4CAI	84.98 (7)	92.10 (7)	4.48 (5)	12.828 (6)
PredictED	85.79 (6)	92.39 (6)	6.64 (7)	25.895 (7)
Modèles individuels	89,36	96,30	3,90	19,77
Ensembles par annotateur	89,62	96,26	3,57	19,33
Ensemble général	89,93	96,26	2,83	16,64

Table 1 *Comparaison entre les performances des équipes participant au challenge CURVAS 2025 et celles de nos modèles.*

Si nous avions participé à cette compétition, nous aurions été classé quatrième pour la confiance et pour l'ECE. Ces résultats sont encourageants pour la technique de l'ensembling général car, par rapport aux équipes professionnelles, nous disposions de moins de temps, moins de ressources, et de moins de données d'entraînement. De plus, nous n'avons pas entrainé nos modèles sur le jeu de validation, qui est composé de 5 patients alors que le jeu d'entraînement en contient 20. Par faute de temps, nous n'avons pas non plus utilisé l'option de cross-validation gérée automatiquement par nnU-Net. Avec plus de ressources, nous aurions souhaité entraîner des réseaux U-Net à encodeurs résiduels (*Residual Encoder U-Net*, ou *ResEnc*).

VI Envergure du projet

Le protocole expérimental s'est révélé ambitieux, tant par sa complexité que par les ressources mobilisées. Travailler avec des architectures comme les U-Net a nécessité un important effort d'auto-formation. L'écart théorique avec les cours de l'ENSAE, ainsi que la prise en main de dépôts de code complexes comme ValUES, ont représenté un véritable défi. Les besoins en calcul et en stockage ont été conséquents : près de 1 080 heures de calcul cumulées (45 jours), avec des tâches longues (entraînement : 24h/modèle ; inférence : 1h/patient/modèle ; ensemble : 2h/patient ; métriques : 1h/patient). Les sorties probabilistes, très volumineuses (3 Go chacune), ont porté notre stockage S3 à près de 2,9 To.

L'ensemble du pipeline n'aurait pas pu être exécuté sans l'accès aux instances de calcul haute performance fournies gracieusement par les plateformes Onyxia du Genes et de l'Insee. Ce soutien a été crucial, tant pour la puissance de calcul que pour la capacité de stockage et de transfert. Un tel projet, réalisé sur des ressources cloud commerciales, aurait représenté un coût financier estimé à environ 760, 50 €, principalement en temps GPU et espace disque.

	Quantité	Par unité			Total		
		Durée (h)	Transfert ↓ (Go)	↑ (Go)	Durée (h)	↓ (Go)	↑ (Go)
Entraînement	9	24	6	1	216	54	9
Inférence	576	1	1.5	4	576	864	2304
Ensemble	64	2	18	16	128	1152	1024
Évaluation	64	1	3.5	0	64	224	0
		Total			**984**	**2294**	**3337**

Table 2 *Coûts en temps, en transferts descendant et montant (donc stockage)*