Quantification de l'incertitude sur des modèles de classification en IA : évaluation comparative de méthodes post-hoc

Paul STEINMETZ Frédérique FROUIN Irène BUVAT

Laboratoire d'imagerie translationnelle en oncologie (LITO), Institut Curie, Inserm U1288, Université Paris Saclay
Rue de la Chaufferie, 91401, Orsay, France

Résumé – La quantification de l'incertitude (QI) en intelligence artificielle est un enjeu clé pour l'intégration des modèles en pratique clinique, notamment pour la classification d'images médicales. Cette étude propose un benchmark de méthodes post-hoc de QI appliquées à un modèle de classification multi-classes d'architecture ResNet18, entraîné en validation croisée (CV) et évalué sur un jeu de données public (organAMNIST). Nous comparons plusieurs approches, dont la réponse softmax maximale (MSR), l'agrégation des prédictions des modèles de la CV (Ens), l'augmentation au test (TTA) et la distance aux données d'entraînement (KNN). Nos résultats montrent que les méthodes simples comme la MSR permettent une bonne identification des erreurs de prédiction, mais que des méthodes plus complexes, notamment KNN_{shap}, offrent de meilleures performances. Enfin, nous explorons des stratégies d'agrégation des différentes mesures d'incertitude et évaluons leur impact sur la performance globale.

Abstract – Uncertainty quantification (UQ) in artificial intelligence is a key challenge for integrating models into clinical practice, for instance for medical image classification. This study presents a benchmark of post-hoc UQ methods applied to a multi-class classification model with a ResNet18 architecture, trained with a 5-fold cross validation setup and evaluated on a public dataset (organAMNIST). We compare several approaches, including Maximum Softmax Response (MSR), ensembling (Ens), test-time augmentation (TTA), and distance to training data (KNN). Our results show that simple methods such as MSR allow for good identification of misclassified cases, but more advanced techniques, particularly KNN_{shap}, outperform simple methods. Finally, we explore different strategies for aggregating uncertainty measures and assess their impact on overall performance.

1 Introduction

Les avancées de l'intelligence artificielle (IA) ont récemment engendré une croissance importante du nombre de modèles d'apprentissage machine proposés, notamment en analyse d'images médicales. En dépit de leurs performances, ces modèles peinent à être adoptés en routine clinique [4] en raison de leur complexité, de leur manque d'interprétabilité et d'adaptabilité à de nouveaux cas d'une part et de l'importance des enjeux associés au diagnostic et choix de thérapie d'autre part. L'identification automatique des cas à haut risque d'erreur pour un modèle de prédiction permettrait d'augmenter la confiance des cliniciens dans les résultats obtenus. De nombreuses méthodes pour représenter et quantifier l'incertitude (QI) des modèles d'IA ont été proposées, notamment pour les tâches de classification à partir d'images biomédicales. On peut séparer ces méthodes en deux catégories [9] : les méthodes intrinsèques modélisant l'incertitude directement dans l'architecture du modèle, et les méthodes post-hoc permettant une estimation de l'incertitude après l'entraînement. Les méthodes post-hoc peuvent être appliquées à n'importe quel modèle entraîné, sans nécessiter de modifications de l'architecture interne. Cependant, la diversité des domaines de recherche et des définitions de l'incertitude (détection d'erreurs, détection de cas hors distribution, quantification d'incertitude prédictive, classification sélective), qui n'abordent pas forcément les mêmes sources potentielles d'erreurs (données bruitées, données provenant d'une distribution autre que celle de l'entraînement) rendent la comparaison des méthodes d'évaluation de la QI difficile. Ainsi, en dépit des revues d'état de l'art des méthodes existantes [3, 6], peu de travaux comparent les performances atteintes par ces différentes approches [5]. Dans cette étude, nous évaluons plusieurs approches post-hoc de quantification de l'incertitude sur un cas d'usage en classification d'images médicales à l'aide d'un réseau de neurones, en utilisant un jeu de données public et des métriques d'évaluation standardisées. Pour pallier les limites de chacune des approches proposées, plusieurs stratégies d'agrégation sont également évaluées.

2 Méthodes

2.1 Evaluation des performances de la QI

Les performances des différentes méthodes de quantification de l'incertitude dans le contexte de tâches de classification sont évaluées en mesurant l'aire sous la courbe ROC AUC_{QI}, qui reflète leur capacité à discriminer les cas bien classés des cas mal classés. Pour chaque approche, le seuil de prédiction qui maximise la précision équilibrée (bACC) sur la détection des cas d'échecs est retenu et en sont déduits les sensibilité et spécificité correspondantes.

2.2 Méthodes post-hoc de QI évaluées

On introduit ici : une instance de test x, la probabilité associée à la prédiction $p(y \mid x)$, les paramètres du modèle de classification θ, les probabilités issues des différents modèles M notées $p_m(y \mid x)$, l'ensemble d'entraînement $\mathcal{D}_{\text{train}}$, l'ensemble de test $\mathcal{D}_{\text{test}}$, l'ensemble de calibration annoté $\mathcal{D}_{\text{cal}}$ et $z(x)$ la représentation latente dans l'avant-dernière couche du modèle. Les données d'entrée requises pour chaque méthode de QI sont résumées dans le tableau 1.

TABLEAU 1 : Données d'entrées nécessaires pour une instance de test x et une classe y.

	MSR	Ens	$\text{TTA}_{\text{RT / RA / GPS}}$	$\text{KNN}_{\text{all / shap}}$
$p(y\|x)$	O	O	O	O
$p_m(y\|x)$	N	O	N	N
$\mathcal{D}_{train}$	N	N	N	O
$\mathcal{D}_{cal}$	N	N	N / N / O	N / O
$z(x)$	N	N	N	O

- **Réponse Softmax Maximale (MSR) :**
  Utilisation de la probabilité maximale de la prédiction du modèle $p(y|x)$ comme mesure de l'incertitude (équation 1) ;

$$U_{\text{MSR}}(x) = 1 - \max_y p(y|x, \theta) \tag{1}$$

- **Agrégation des prédictions des modèles (Ens) :**
  Agrégation des prédictions des différents modèles M issus de la validation croisée et utilisation de l'écart-type des différentes prédictions comme mesure de l'incertitude (équation 2) ;

$$\sigma^{\theta}_{\text{Ens}}(x) = \sqrt{\frac{1}{card(M)} \sum_{m=1}^{M} (p_m(y|x, \theta_m) - \mu^{\theta}_{\text{Ens}}(x))^2}$$

où $\mu^{\theta}_{\text{Ens}}(x)$ est la moyenne des prédictions. (2)

- **Augmentation au test (TTA) :**
  Nous définissons une combinaison d'augmentations au test, P, comme un ensemble de stratégies $\{s_i(\cdot)\}$. Une stratégie $s(\cdot)$ est constituée de $i_{\max}$ transformations d'images $t_j(\cdot)$ appliquées successivement, avec $j \in \{1, \ldots, i_{\max}\}$, où t_j est une opération d'image prédéfinie. Pendant l'inférence, l'incertitude est quantifiée par l'écart-type des prédictions obtenues à partir des différentes stratégies (équation 3).

$$\sigma^{\theta}_{P}(x) = \sqrt{\frac{1}{card(P)} \sum_{s \in P} (p(y|s(x), \theta) - \mu^{\theta}_{P}(x))^2}$$

où $\mu^{\theta}_{P}(x)$ est la moyenne des prédictions. (3)

  Trois méthodes de TTA sont évaluées, pour lesquelles les stratégies utilisées sont définies comme suit :

  (1) TTA_{RT} : transformations de base (rognages, translations, rotations, retournements) aléatoires ;

  (2) TTA_{RA} : transformations de base, élastiques et photométriques, avec amplitudes et occurrences aléatoires (méthode RandAugment [2]) ;

  (3) TTA_{GPS} : optimisation de P sur D_{cal} [8], afin d'identifier les stratégies maximisant AUC_{QI} (cf algorithme 1).

- **Distance à la distribution d'entraînement (KNN) :**
  La distance euclidienne aux 5 proches voisins est calculée par rapport à l'ensemble d'entraînement [1] (équation 4) :

$$\text{KNN}(x) = \frac{1}{5} \sum_{j=1}^{5} d(z(x), z_{NN_j}) \tag{4}$$

  où $z(x)$ et z_{NN_j} sont respectivement les représentations latentes dans l'avant-dernière couche du modèle de dimension L de l'instance de test x et des 5 plus proches voisins issus de l'ensemble d'entraînement.

  Cette distance est calculée après une étape de **Normalisation** : centrage et réduction des variables latentes sur $\mathcal{Z}_{\text{train}}$, représentation latente de $\mathcal{D}_{\text{train}}$ (équation ligne 6 dans algorithme 2)et une **Réduction en composantes principales (PCA)** : application d'une transformation ϕ, en conservant 90% de la variance expliquée sur $\mathcal{Z}_{\text{train}}$ après normalisation.

  Deux variantes sont proposées pour le calcul de la distance KNN :

  (1) KNN_{all} : distance entre $z(x)$ et z_{NN_j}, représentations latentes après normalisation et PCA sur les L dimensions.

  (2) KNN_{shap} : distance entre $z_c(x)$ et z_cNN_j, représentations latentes après normalisation et PCA sur une sélection préalable des variables latentes les plus influentes, effectuée à l'aide des valeurs de Shapley [7] sur D_{cal}. Pour chaque classe c, les 50 variables latentes les plus importantes sont sélectionnées. Pour une instance de test x, seules les 50 variables correspondant à la

Algorithme 1 Optimisation TTA_{GPS} basée sur $\sigma^{\theta}_{P}(x)$

Données : Ensemble de calibration annoté D_{cal}
Ensemble de N stratégies (RandAugment) $\mathcal{P} = \{s_1, \ldots, s_N\}$
Nb transformations/stratégie et amplitude max ($i_{\max}$, A_M)
Nb d'initialisations aléatoires N_{search}
Fonction d'incertitude $\sigma^{\theta}_{P}(x)$
Résultat : Stratégie optimisée d'augmentation P^*.
`// Prédictions sur` D_{cal}
pour $j \leftarrow 1$ *jusqu'à* N **faire**
  Générer une stratégie s_j avec au plus $i_{\max}$ transformations d'amplitude maximale A_M Appliquer s_j sur chaque image $x \in D_{\text{cal}}$ Enregistrer les prédictions $p(y|s_j(x), \theta)$
fin
`// Optimisation gloutonne des stratégies`
Initialiser aléatoirement N_{search} stratégies P_0 parmi les N
pour $k \leftarrow 1$ *jusqu'à* N_{search} **faire**
  $P^{(k)} \leftarrow P_0^{(k)}$ `// Initialisation`
  tant que *une amélioration de* AUC_{QI} *est possible* **faire**
    Sélectionner la transformation s^* qui maximise AUC_{QI} en utilisant $\sigma^{\theta}_{P}(x)$ lorsqu'ajoutée à $P^{(k)}$
    si $AUC_{QI}(P^{(k)} \cup \{s^*\}) > AUC_{QI}(P^{(k)})$ **alors**
      Mettre à jour $P^{(k)} \leftarrow P^{(k)} \cup \{s^*\}$
    fin
  fin
fin
`// Sélection de la meilleure combinaison`
Garder P^* qui maximise AUC_{QI} parmi les N_{search} recherches

Algorithme 2 Quantification d'incertitude basée sur SHAP

Données : Ensembles d'entraînement $\mathcal{D}_{\text{train}}$, et de test $\mathcal{D}_{\text{test}}$.

`// Calcul des valeurs SHAP`

pour $x_i \in \mathcal{D}_{cal}$ **faire**

  Calculer SHAP$(x_i) \in \mathbb{R}^{C\times L}$ où L dimension latente, C nombre de classes

fin

pour $c \in \{1, \ldots, C\}$ **faire**

  `// Sélection des variables influentes`

  Calculer $\bar{s}_c = \frac{1}{|\mathcal{D}_{\text{cal}}|}\sum_{x_i\in\mathcal{D}_{\text{cal}}} \text{SHAP}(x_i)_c \in \mathbb{R}^L$ Sélectionner les 50 variables les plus influentes : $\mathcal{F}_c = \arg\max_{|S|=50} \bar{s}_c(S)$

  `// Normalisation`

  Calculer la moyenne μ_c et l'écart-type σ_c des variables $\mathcal{F}_c$ sur $\mathcal{D}_{\text{train}}$ Standardiser les données :

$$x_i'[\mathcal{F}_c] = \frac{x_i[\mathcal{F}_c] - \mu_c}{\sigma_c}, \quad \forall x_i \in \mathcal{D}_{\text{train}}$$

  `// Réduction de dimension par PCA`

  Appliquer PCA sur $x_i'[\mathcal{F}_c]$ sur $\mathcal{D}_{\text{train}}$, en conservant 90% de la variance Soit $\phi_c : \mathbb{R}^{50} \to \mathbb{R}^{pca_c}$ la transformation PCA obtenue

fin

`// Projection et calcul des distances`

pour $x \in \mathcal{D}_{test}$ **faire**

  Déterminer $c^*(x)$ la classe prédite Standardiser $x[\mathcal{F}_{c^*(x)}]$ avec $\mu_{c^*(x)}$ et $\sigma_{c^*(x)}$ Projeter dans l'espace réduit :

$$z(x) = \phi_{c^*(x)}\left(\frac{x[\mathcal{F}_{c^*(x)}] - \mu_{c^*(x)}}{\sigma_{c^*(x)}}\right) \in \mathbb{R}^{pca_{c^*(x)}}$$

  Trouver les 5 plus proches voisins z_{NN_j} dans $\mathcal{Z}_{\text{train},c^*(x_i)}$ Calculer la distance KNN$_{\text{shap}}$ (équation 4)

fin

classe prédite $c^*(x)$ sont standardisées et projetées dans l'espace réduit. La distance KNN est ensuite calculée entre $z_c(x)$ et ses 5 plus proches voisins dans l'ensemble projeté $\mathcal{Z}_{\text{train},c^*(x)}$ (algorithme 2).

2.3 Agrégation des résultats de QI

Les métriques de QI évaluées étant sur des échelles différentes, l'ensemble des valeurs issues des métriques calculées sur $\mathcal{D}_{\text{test}}$ sont concaténées, puis normalisées globalement à l'aide d'un *Z-score*. Les scores ainsi normalisés sont ensuite agrégés selon trois stratégies : la moyenne, le score d'incertitude maximale sur les métriques incluses ainsi que la prédiction d'un perceptron simple entraîné sur $\mathcal{D}_{\text{cal}}$.

3 Expérimentations

3.1 Cas d'usage : Classification d'organes

La base OrganAMNIST issue de MedMNIST v2 [10] a été choisie comme cas d'usage. Elle contient 58830 coupes scanner (CT) abdominales de taille 28×28 pixels, annotées en 11 classes correspondant aux organes abdominaux. Comme détaillé dans [10], les performances de classification les plus élevées sont obtenues avec un réseau de neurones à convolution de type ResNet18. C'est donc cette architecture qui a été retenue, pré-entraînée sur ImageNet, et utilisé comme modèle de classification. Le jeu de test prédéfini représente un tiers des données (17778 images), les restantes étant réparties en un jeu d'entraînement/validation (80%) et un jeu de calibration (20% : 8211 images). Une validation croisée en cinq plis stratifiés est employée à l'entraînement, sans stratégie d'augmentation des images, générant cinq modèles, dont les prédictions sont moyennées à l'inférence.

Les métriques de performances de la classification des organes sont calculées selon une stratégie un-contre-tous : chaque classe est successivement considérée comme positive face aux autres, et les résultats sont moyennés sur l'ensemble des 11 classes (moyenne macro (écart-types)). Les performances atteintes sont : sensibilité : 0,93 (0,05), spécificité : 0,99 (0,01), bACC : 0,93 (0,03), AUC : 0,995 (0,01).

3.2 Quantification d'incertitude

Les résultats des performances de quantification de l'incertitude des différentes méthodes évaluées sur le jeu de test sont présentés dans le tableau 2.

TABLEAU 2 : performances des méthodes de QI.

	MSR	Ens	$\text{TTA}_{\text{RT / RA / GPS}}$	$\text{KNN}_{\text{all / shap}}$
AUC_{QI}	0,94	0,93	0,50 / 0,73 / **0,91**	0,91 / **0,98**
bACC	0,88	0,87	0,52 / 0,69 / **0,84**	0,86 / **0,95**
sensibilité	0,84	0,83	0,94 / 0,51 / **0,76**	0,79 / **0,91**
spécificité	0,93	0,92	0,10 / 0,88 / **0,93**	0,93 / **0,98**

3.3 Agrégation des résultats de QI

La méthode d'augmentations au test TTA_{GPS} étant plus performante que TTA_{RT} et TTA_{RA}, nous n'avons conservé que cette stratégie pour l'agrégation avec les autres approches. Pour les mêmes raisons, seule l'approche KNN_{shap}, a été conservée. Les différentes méthodes ne nécessitant pas les mêmes données d'entrées, les résultats d'agrégation sont présentés selon deux scénarios différents : MSR + Ens + TTA_{GPS} et MSR + Ens + TTA_{GPS} + KNN_{shap} (tableau 3).

TABLEAU 3 : performances de l'agrégation des méthodes de QI (A : MSR + Ens + TTA_{GPS} ; B : MSR + Ens + TTA_{GPS} + KNN_{shap}) pour les trois approches d'agrégations définies.

	$\text{A}_{\text{(moy / max / percep)}}$	$\text{B}_{\text{(moy / max / percep)}}$
AUC_{QI}	**0,94** / 0,94 / 0,94	**0,98** / 0,98 / 0,95
bACC	**0,88** / 0,88 / 0,87	**0,95** / 0,95 / 0,88
spécificité	**0,92** / 0,93 / 0,90	**0,98** / 0,98 / 0,90
sensibilité	**0,84** / 0,82 / 0,85	**0,91** / 0,91 / 0,86

4 Discussion

La performance atteinte par le modèle de classification d'organes est bonne, avec moins de 7% d'erreurs.

Les différentes méthodes de QI évaluées présentent des performances variables, avec des AUC_{QI} allant de 0,5 à 0,98. Parmi les méthodes ne nécessitant pas l'accès aux données d'entraînement, la méthode MSR, très simple à mettre en œuvre, conduit aux meilleures performances sur notre cas d'usage avec une AUC_{QI} de 0,94. Cette méthode, couramment utilisée, requiert une bonne calibration du modèle [6].

La combinaison des prédictions des différents modèles de la validation croisée permet d'obtenir des performances similaires, avec une AUC_{QI} de 0,93. Bien que nécessitant l'accès aux prédictions des différents modèles, la mise en place d'une stratégie de validation croisée est courante en IA [6].

Les méthodes d'augmentation au test TTA présentent, pour notre cas d'usage, des performances très hétérogènes. L'approche TTA_{RT} faisant appel à des transformations simples n'identifie pas les échecs. Les augmentations plus sophistiquées de TTA_{RA} permettent d'atteindre une AUC_{QI} de 0,73. TTA_{GPS}, bien que nécessitant une optimisation sur un jeu de calibration, permet d'augmenter l'AUC_{QI} à 0,91. Les stratégies d'augmentation appliquées à l'entrainement n'étant pas forcément connues, TTA_{GPS} permet d'adapter les augmentations au test à la distribution des données d'entraînement.

Dans l'hypothèse où les données d'entraînement sont disponibles, l'approche qui estime la distance au jeu d'entraînement dans l'espace latent du modèle après identification des variables latentes les plus influentes sur la classification KNN_{shap}, permet d'atteindre une AUC_{QI} de 0,98 et une spécificité de 0,98 (2% d'erreurs non détectées). Nos résultats soulignent l'importance de l'optimisation des phases de TTA et de KNN.

Plusieurs méthodes d'agrégation sont testées : moyenne, incertitude max, perceptron entraîné sur D_{cal}. Dans ce cas d'usage, l'apport d'un perceptron ne semble pas pertinent. Les seuils de prédiction fixés (maximisant la bACC) permettent d'obtenir un bon équilibre entre le nombre d'échecs correctement prédits tout en limitant le nombre de cas classés comme erreurs à vérifier. Cette stratégie est à adapter en fonction de l'application et des conséquences de possibles erreurs de prédiction.

Le protocole d'évaluation proposé permet une comparaison juste des méthodes de QI (jeu de données et métriques identiques) ce qui est rarement réalisé en pratique, mais plusieurs limitations sont à noter. Les hyperparamètres de la méthode KNN_{shap} (nombre de composantes principales, de variables latentes conservées, de plus proches voisins) ont été choisis empiriquement, et une évaluation systématique de l'impact de ces choix sur les performances reste nécessaire. De plus, bien que les méthodes d'agrégation atteignent les mêmes niveaux de performances que la meilleure méthode incluse dans les deux scénarios testés (3), ces stratégies ne semblent pas apporter de bénéfice dans ce cas d'usage. Enfin, les sources d'erreurs peuvent être variées (données issues de la même distribution / hors distribution) et la métrique d'évaluation des différentes méthodes de QI devrait prendre en compte le niveau de performance du modèle [5], ce qui n'est pas évalué ici. Une étude sur différents cas d'usages (images plus grandes, tâches différentes) reste nécessaire pour établir l'intérêt des différentes approches proposées.

5 Conclusion

Cette étude évalue plusieurs approches de QI associées à un modèle de classification d'images médicales. Les résultats montrent que certaines méthodes simples, comme la MSR, présentent de très bonnes performances. Une méthode originale KNN_{shap}, calculant une distance robuste au jeu d'entraînement dans un sous-espace latent adapté à la problématique posée surpasse les autres approches dans le cas d'usage retenu. Ces résultats suggèrent que les approches de QI peuvent contribuer à faciliter l'usage des modèles d'IA en imagerie médicale.

Ce travail a été réalisé dans le cadre du projet AI.DReAM (2020-2026) financé par BPI France en réponse à l'appel à projets structurants pour la compétitivité (PSPC) du programme d'investissement d'avenir.

Références

[1] C. BERGER, M. PASCHALI, B. GLOCKER *et al.* : Confidence-based out-of-distribution detection : A comparative study and analysis. *In Proc. Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 122–132. Springer, 2021.

[2] E. D. CUBUK, B. ZOPH, J. SHLENS *et al.* : Randaugment : Practical data augmentation with no separate search. *CoRR*, abs/1909.13719, 2019.

[3] L. HUANG, S. RUAN, Y. XING *et al.* : A review of uncertainty quantification in medical image analysis : Probabilistic and non-probabilistic methods. *Med. Image Anal.*, 97:103223, 2024.

[4] E. HÜLLERMEIER, W. WAEGEMAN *et al.* : Aleatoric and epistemic uncertainty in machine learning : An introduction to concepts and methods. *Mach. Learn.*, 110(3):457–506, 2021.

[5] P. F. JAEGER, C. T. LÜTH, L. KLEIN *et al.* : A call to reflect on evaluation practices for failure detection in image classification. *arXiv*, 2024.

[6] B. LAMBERT, F. FORBES, S. DOYLE *et al.* : Trustworthy clinical ai solutions : A unified review of uncertainty quantification in deep learning models for medical image analysis. *Artif. Intell. Med.*, 150:102830, 2024.

[7] S. M. LUNDBERG et S. LEE : A Unified Approach to Interpreting Model Predictions. *In Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[8] A. LYZHOV, Y. MOLCHANOVA, A. ASHUKHA *et al.* : Greedy policy search : A simple baseline for learnable test-time augmentation. *In Conference on uncertainty in artificial intelligence*, pages 1308–1317. PMLR, 2020.

[9] C. TOMANI, S. GRUBER, M. ERDEM *et al.* : Post-hoc Uncertainty Calibration for Domain Drift Scenarios. *arXiv*, 2021.

[10] J. YANG, R. SHI, D. WEI, *et al.* : Medmnist v2 : A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Sci. Data*, 10(1):41, 2023.