Et si les classifieurs profonds oubliaient l'inconnu ? Auto-encodeurs et détection hors distribution

François Thievon[1], Arthur Hoarau[1,2]

[1] CentraleSupélec, Université Paris-Saclay, Metz, France
[2] Loria, CNRS, Université de Lorraine, Nancy, France

Résumé

La détection d'observations hors distribution (OOD) constitue un enjeu central pour le déploiement fiable des modèles d'apprentissage automatique, en particulier dans des systèmes critiques. En apprentissage profond, cette tâche est fréquemment abordée via la quantification d'incertitude épistémique, supposée élevée lorsque le modèle rencontre une entrée supposément éloignée de ses données d'entraînement. Cependant, des résultats récents montrent que de nombreuses méthodes échouent à séparer correctement incertitudes épistémique et aléatoire, ce qui dégrade les performances OOD.
Dans cet article, nous soutenons que cette difficulté provient en partie d'un biais inhérent à la classification : en optimisant la performance discriminative, un classifieur profond peut apprendre une représentation latente qui élimine précisément l'information de densité nécessaire à la détection OOD. Nous illustrons ce phénomène dans un cadre simulé, puis sur une expérience de classification d'images. Nos résultats montrent que, pour une même méthode de quantification d'incertitude, la détection OOD est significativement améliorée lorsque l'on travaille dans l'espace latent d'un auto-encodeur plutôt que dans celui d'un classifieur.

Mots-clés

Quantification d'Incertitude, Détection hors distribution, Classification.

Abstract

Out-of-distribution (OOD) detection is a central challenge for the reliable deployment of machine learning models, especially in safety-critical systems. In deep learning, this task is often addressed through epistemic uncertainty quantification, which is assumed to be high when the model encounters an input that is supposedly far from its training data. However, recent results show that many methods fail to properly disentangle epistemic and aleatoric uncertainty, which in turn degrades OOD detection performance.
In this paper, we argue that this difficulty partly stems from a bias inherent to classification itself : by optimizing discriminative performance, a deep classifier may learn a latent representation that precisely removes the density information required for OOD detection. We illustrate this phenomenon in a controlled simulated setting, and then on an image classification experiment. Our results show that, for the same uncertainty quantification method, OOD detection is significantly improved when operating in the latent space of an autoencoder rather than that of a classifier.

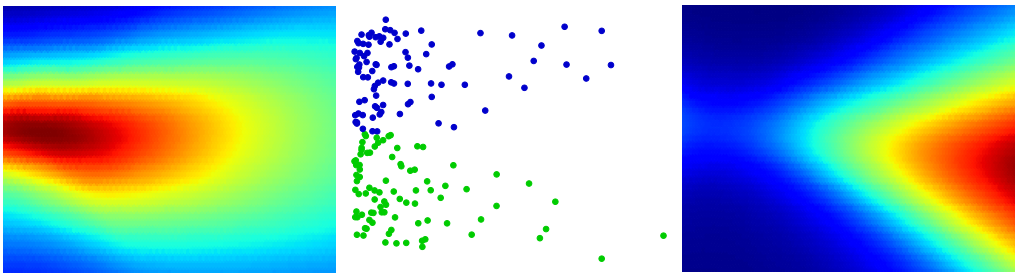
Keywords

Uncertainty Quantification, Classification, OOD Detection.

1 Introduction

La recherche en apprentissage automatique s'est longtemps attachée à améliorer la performance prédictive des modèles. Cependant, avec les avancées récentes en intelligence artificielle (vision, recommandation, modèles de langage, etc.) et son intégration dans un nombre croissant de systèmes parfois critiques, de nouveaux défis émergent. Une attention grandissante est désormais portée aux coûts écologiques, économiques et sociétaux de l'IA [21], ainsi qu'à la demande de modèles plus frugaux. Pour des modèles déjà capables d'atteindre des performances proches de l'optimal, il devient crucial de capturer et de rapporter l'incertitude associée aux prédictions locales les plus sensibles, même si elles sont rares. Une seule défaillance peut être fatale dans des systèmes critiques pour la sécurité, tels que la conduite autonome, les applications médicales ou spatiales. La récente révolution des grands modèles de langage (LLM[1]) et leur démocratisation ont également soulevé des inquiétudes majeures concernant la transparence des prédictions, notamment face à l'excès de confiance que certains modèles affichent dans leurs prédictions. Dans la plupart des applications, des modèles dits "boîte noire" sont déployés, et la recherche vise en partie à rendre plus transparentes les prédictions de ces modèles parfois très complexes.
Dans ce contexte, l'évaluation des modèles d'apprentissage automatique doit aller au-delà de la performance prédictive brute et prendre explicitement en compte l'incertitude et la confiance associées aux prédictions. La détection d'observations hors distribution [17, 13] (détection OOD[2] en phase de test) vise précisément à identifier les entrées pour lesquelles le modèle n'a pas été correctement entraîné [6, 29]. Par exemple, un modèle assistant un médecin et entraîné à catégoriser la pathologie de nouveaux patients devrait être

1. LLM pour *Large Language Model*.
2. OOD pour *Out-of-distribution.*



(a) Incertitude Aléatoire (b) Jeu de données (c) Incertitude Épistémique

FIGURE 1 – Décomposition d'incertitude par apprentissage profond bayésien et approximation de Laplace. L'incertitude aléatoire est présentée en figure 1a, et l'incertitude épistémique en figure 1c, avec un dégradé allant du rouge au bleu, correspondant respectivement à une incertitude élevée et faible.

capable de signaler son incertitude lorsqu'un patient présentant une maladie inconnue (comme le COVID-19 lors de son émergence) se présente. De même, dans un contexte de conduite autonome, la reconnaissance de panneaux de signalisation peut être fortement dégradée par la présence de neige ou de brouillard, absents du jeu d'entraînement.

La quantification d'incertitude est un champ d'étude de l'apprentissage automatique qui s'est en partie attelé à adresser cette question de détection OOD. Dans la littérature, on distingue généralement deux grands types d'incertitude [8] : l'incertitude aléatoire (liée aux données) et l'incertitude épistémique (liée au modèle). L'incertitude aléatoire provient de la nature stochastique du processus générateur des données, tandis que l'incertitude épistémique est associée à un manque de connaissance. La première est généralement considérée comme irréductible, alors que la seconde peut être atténuée par l'acquisition de données supplémentaires. Ces deux types d'incertitude sont illustrés en figure 1, à l'aide d'un modèle d'apprentissage bayésien profond et d'une approximation de Laplace [12], sur un jeu de données synthétique. L'incertitude aléatoire permet d'identifier les zones où la prédiction est intrinsèquement difficile : dans ces régions, les deux classes du jeu de données se chevauchent effectivement. À l'inverse, une incertitude épistémique élevée traduit un manque de connaissance du modèle. Elle correspond à des régions de l'espace des entrées insuffisamment couvertes par les données d'entraînement, pour lesquelles il serait donc risqué de se fier à la prédiction du modèle. C'est précisément ce second type d'incertitude qui est généralement mobilisé en quantification d'incertitude pour détecter des observations hors distribution.

Afin de dissocier l'incertitude aléatoire de l'incertitude épistémique, de nombreuses méthodes ont été proposées ces dernières années [9]. Cependant, cet article de position se veut le plus ouvert possible ; c'est pourquoi nous avons choisi de présenter ces méthodes dans une section dédiée, que nous considérons indépendante du reste de l'article.

L'utilisation de l'incertitude épistémique pour la détection hors distribution (OOD) est largement admise et s'est montrée particulièrement efficace. Cependant, un nombre croissant de travaux se montrent critiques vis-à-vis de cette approche [16], en particulier dans le cadre de l'apprentissage profond. Des travaux récents ont notamment mis en évidence une corrélation (trop) importante entre les composantes aléatoire et épistémique [19] pour la plupart des méthodes considérées.

Dans cet article, nous soutenons qu'il existe un biais inhérent, non pas lié à la méthode de quantification d'incertitude employée, mais à la tâche de classification elle-même, et que ce biais peut en partie expliquer les échecs récents de la détection OOD en apprentissage profond. En conséquence, la comparaison des méthodes de quantification d'incertitude épistémique peut s'avérer secondaire, dès lors que l'information déterminante pour la détection OOD est éliminée en amont par la tâche de classification elle-même. Nous montrons également que ce phénomène peut être partiellement atténué par l'utilisation d'auto-encodeurs, qui préservent mieux la densité dans l'espace latent et permettent ainsi une détection OOD plus fiable. Cet avantage s'accompagne toutefois d'un coût en termes de performances : les représentations latentes issues d'un auto-encodeur sont généralement moins discriminantes que les sorties d'un classifieur. Dans cet article, nous utilisons des modèles simples et n'employons pas, par exemple, d'auto-encodeurs variationnels [4]. Bien que leurs avantages soient nombreux, ils sortent du cadre de ce travail. Ici, c'est la différence de tâche entre classifieur et auto-encodeur qui constitue le point central. En résumé, nous souhaitons attirer l'attention du lecteur sur le fait que la recherche de performances maximales, en particulier en classification, induit un coût rarement explicité : la réduction de la capacité du modèle à détecter les données hors distribution. Autrement dit, les modèles les plus performants ont tendance à "oublier leur ignorance".

L'article est organisé comme suit ; la section 2 introduit les notions de quantification d'incertitude appliquées à la détection OOD. Le reste du papier est volontairement plus accessible et peut être lu indépendamment de cette section. La section 3 présente une mise en évidence empirique du problème, souligné dans cet article, de détection OOD en classification profonde. La section 4 propose une évaluation expérimentale sur un problème de classification d'images. Enfin, la section 5 conclut cet article.

2 Détection hors distribution & Quantification d'incertitude

Dans cette section, nous introduisons la détection hors distribution (OOD) à travers la notion d'incertitude épistémique présentée en introduction. La section 2.1 présente les principales méthodes permettant de capturer et de représenter les incertitudes d'un modèle de classification, tandis que la section 2.2 explicite la quantification d'incertitude épistémique dans les prédictions d'un modèle d'apprentissage automatique.

2.1 Quantification d'incertitude

Dans la littérature, on distingue généralement deux types d'incertitude [8] : l'incertitude aléatoire (liée aux données) et l'incertitude épistémique (liée au modèle). L'incertitude aléatoire provient du caractère stochastique du processus générateur des données, tandis que l'incertitude épistémique est associée à un manque de connaissance. La première est généralement considérée comme irréductible, alors que la seconde peut être atténuée par l'acquisition de données supplémentaires. Afin de dissocier ces deux composantes, de nombreuses méthodes ont été proposées ces dernières années [9]. Ces approches peuvent être regroupées en quatre grandes familles.

Les méthodes bayésiennes calculent une distribution a posteriori sur les paramètres d'un modèle et construisent une distribution de second ordre sur les probabilités de classe via la moyenne bayésienne de modèles (*Bayesian model averaging*) [12, 3, 27] :

$$p(h \mid \mathcal{D}) \propto p(h)p(\mathcal{D} \mid h). \tag{1}$$

La quantification d'incertitude en apprentissage automatique repose traditionnellement sur l'apprentissage bayésien, où $p(h \mid \mathcal{D})$ désigne la densité de probabilité a posteriori d'un modèle $h \in \mathcal{H}$ conditionnellement à un jeu de données $\mathcal{D}$. Intuitivement, la dispersion de $p(h \mid \mathcal{D})$ reflète l'état de connaissance du modèle, et donc son incertitude épistémique : plus cette distribution est concentrée (c'est-à-dire qu'un modèle est nettement plus probable que les autres), moins l'apprenant est épistémiquement incertain quant au fait que h^* soit le modèle optimal. Il est crucial de noter que le fait que h^* soit le meilleur modèle n'implique pas que sa prédiction $\hat{y} = h^*(x)$, ou encore $p_{h^*}(y \mid x)$, soit certaine du point de vue aléatoire. Autrement dit, un modèle peut être épistémiquement confiant tout en reconnaissant que la tâche est intrinsèquement aléatoirement difficile.

Malgré de solides fondements théoriques, l'apprentissage bayésien a également fait l'objet de critiques substantielles, notamment en raison de ses hypothèses parfois restrictives et de la nécessité de recourir à des approximations empiriques afin de rendre l'inférence *tractable*.

Les méthodes d'ensembles capturent l'incertitude épistémique à travers la diversité des estimateurs. Cette diversité peut être obtenue en *randomisant* les données d'entraînement par *bootstrapping*, en variant l'architecture ou les mécanismes de régularisation (par exemple via *MC Dropout* ou *DropConnect* [18]), ou encore en *randomisant* l'optimisation, comme pour *Deep Ensembles* [14]. L'incertitude épistémique est alors élevée lorsque le désaccord entre estimateurs est important, tandis que l'incertitude aléatoire est élevée lorsque l'entropie moyenne des estimateurs est importante. Il a été montré que certaines de ces techniques peuvent être interprétées comme des approximations de la distribution a posteriori de second ordre mentionnée précédemment [12]. Bien qu'intuitive et largement applicable, cette famille de méthodes a également été critiquée quant à sa justification théorique.

L'*Evidential Deep Learning* (EDL) adopte une perspective fréquentiste et vise à minimiser à la fois l'erreur du modèle et l'incertitude épistémique, cette dernière étant représentée par une probabilité de second ordre, sans nécessiter la spécification d'un a priori ou d'un a posteriori sur les paramètres. Dans le cas de la classification multi-classes, la distribution de Dirichlet est généralement utilisée [23, 26]. Malgré des performances empiriques raisonnables, cette famille récente de méthodes a fait l'objet de nombreuses critiques [1]. En particulier, certains comportements attendus se sont révélés impossibles à satisfaire [11], conduisant à la conclusion générale que l'EDL n'est pas adapté à une quantification correcte de l'incertitude épistémique.

Les méthodes fondées sur la densité/distance modélisent typiquement les incertitudes aléatoire et épistémique directement dans l'espace des variables [20, 7]. Ces approches peuvent également construire une distribution de second ordre dans une étape de post-traitement [2]. En pratique, l'incertitude épistémique associée à une prédiction est élevée lorsque sa vraisemblance est faible pour toutes les classes du jeu de données, tandis que l'incertitude aléatoire est élevée lorsque l'instance est susceptible d'appartenir à plusieurs classes. Le principal inconvénient de ces méthodes est que l'incertitude épistémique peut être non bornée ou insuffisamment justifiée théoriquement, bien qu'elles puissent au moins produire un classement cohérent des prédictions.

En raison des limites intrinsèques de la théorie des probabilités classique pour représenter certaines formes d'ignorance et d'imprécision, plusieurs cadres mathématiques alternatifs ont été développés [5]. La théorie des fonctions de croyance, par exemple, permet de représenter différents niveaux d'ignorance. Les probabilités imprécises permettent la construction d'ensembles crédaux, formalisant des bornes inférieures et supérieures pour chaque probabilité. De même, la théorie des sous-ensembles flous fournit un formalisme permettant de représenter des frontières de classes graduelles.

2.2 Incertitude épistémique

Soit (X, Y) une paire de variables aléatoires[3], où X prend ses valeurs dans un espace d'entrée $\mathcal{X} \subseteq \mathbb{R}^r$ et Y dans l'espace des sorties $\mathcal{Y} = \{1, \ldots, K\}$. Les données sont supposées générées selon une distribution jointe inconnue P sur $\mathcal{X} \times \mathcal{Y}$. Un jeu de données d'entraînement est donné par

3. Nous omettons ici la notion de vecteur aléatoire par souci de simplicité.

$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, constitué de n réalisations i.i.d. selon P. Un classifieur est une fonction $h : \mathcal{X} \to \mathcal{Y}$ appartenant à une classe d'hypothèses $\mathcal{H}$ et produisant une distribution conditionnelle $p_h(y \mid x)$.

L'outil optimal [10] pour la détection hors distribution d'une nouvelle instance x est la densité $p(x)$, ou, en classification, la densité conditionnelle $p(x \mid y)$. Une instance présentant une densité conditionnelle très faible pour toutes les classes a peu de chances d'appartenir à la distribution d'entraînement. Cependant, le processus générateur étant inconnu, cette approche est irréalisable en pratique. Pour pallier cette difficulté, il est possible d'utiliser l'incertitude épistémique associée à la prédiction $p_h(y \mid x)$: une incertitude épistémique élevée indique que le modèle a affaire à une entrée pour laquelle il n'a pas été spécifiquement entraîné.

Considérons un prédicteur de second ordre, défini comme un ensemble de modèles $\Delta \subseteq \mathcal{H}$. Dans la pratique, Δ peut correspondre à un ensemble d'estimateurs (méthodes d'ensembles), à un ensemble convexe de probabilités (ensembles crédaux), ou à un espace paramétrique (cadre bayésien). Dans la section 4, Δ correspond à une forêt aléatoire d'arbres de décision.

Pour chaque entrée $x \in \mathcal{X}$, chaque hypothèse $h \in \Delta$ produit une probabilité prédictive $p_h(y \mid x)$, agrégée sous la forme de la prédiction du modèle de second ordre $p_\Delta(y \mid x)$. Une décomposition entropique répandue de l'incertitude s'écrit :

$$\underbrace{H[p_\Delta(y \mid x, \mathcal{D})]}_{\text{Inc. Totale}} = \underbrace{\mathbb{E}[H[p_h(y \mid x)]]}_{\text{Inc. Aléatoire}} + \underbrace{I(y, h \mid x, \mathcal{D})}_{\text{Inc. Epistémique}}, \tag{2}$$

où $H(\cdot)$ et $I(\cdot \mid \cdot)$ désignent respectivement l'entropie de Shannon [25] et l'information mutuelle.

Dans le cas discret des forêts aléatoires [24], l'incertitude aléatoire IA_{ent} est la moyenne des entropies prédictives des M estimateurs h_m :

$$\mathrm{IA}_{ent}(x) = -\frac{1}{M} \sum_{m=1}^{M} \sum_{y \in \mathcal{Y}} p_{h_m}(y \mid x) \log\left(p_{h_m}(y \mid x)\right). \tag{3}$$

L'incertitude totale IT_{ent} correspond à l'entropie de la prédiction moyenne :

$$\mathrm{IT}_{ent}(x) = -\sum_{y \in \mathcal{Y}} \left(\frac{1}{M} \sum_{m=1}^{M} p_{h_m}(y \mid x)\right) \log\left(\frac{1}{M} \sum_{m=1}^{M} p_{h_m}(y \mid x)\right). \tag{4}$$

L'incertitude épistémique IE_{ent} est alors obtenue par différence entre l'incertitude totale et l'incertitude aléatoire :

$$\mathrm{IE}_{ent}(x) = \mathrm{IT}_{ent}(x) - \mathrm{IA}_{ent}(x), \tag{5}$$

et représente l'incertitude liée au désaccord entre les arbres, au sein de la forêt.

Afin de diversifier les mesures, nous considérons également une seconde définition de l'incertitude épistémique fondée sur la variance de la probabilité prédictive [22] :

$$\mathrm{IE}_{var}(x) = \sum_{y \in \mathcal{Y}} \mathrm{Var}\left[p_h(y \mid x)\right], \tag{6}$$

qui, dans le cas des forêts aléatoires, s'écrit :

$$\mathrm{IE}_{var}(x) = \frac{1}{M} \sum_{y \in \mathcal{Y}} \sum_{m=1}^{M} \left(p_{h_m}(y \mid x) - \left(\frac{1}{M} \sum_{i=1}^{M} p_{h_i}(y \mid x)\right)\right)^2, \tag{7}$$

et qui représente la variance entre les prédictions des arbres, au sein de la forêt.

Dans la section 4, nous utiliserons donc IE_{ent} et IE_{var} dans un objectif de détection OOD. La section suivante, quant à elle, ne fait pas encore appel à ces mesures et se concentre sur la mise en évidence du phénomène dans un cadre simulé.

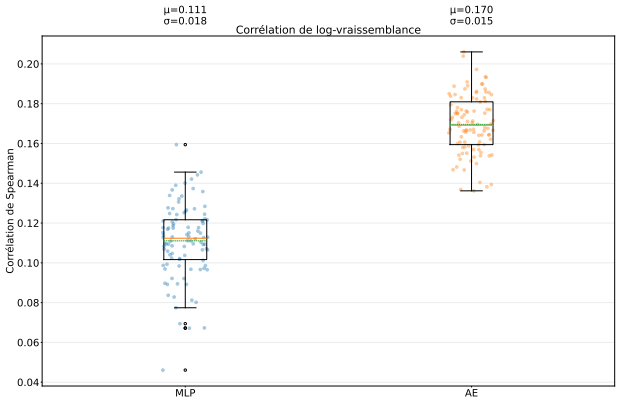


FIGURE 2 – Corrélations de Spearman entre la densité dans l'espace d'entrée et la densité dans l'espace latent du MLP et de l'AE.

3 Quand la classification atteint ses limites

Cette section met en évidence le phénomène central de cet article : en classification profonde, l'information pertinente pour détecter les données hors distribution peut être progressivement éliminée par l'apprentissage, au profit de représentations uniquement optimisées pour la séparation des classes.

En apprentissage automatique classique, de nombreuses approches de détection OOD par quantification d'incertitude (souvent plus robustes que leurs équivalents profonds) opèrent directement dans l'espace d'entrée. Une instance est alors considérée comme hors distribution au sens où elle est atypique dans l'espace des variables. Toutefois, lorsque les données sont complexes ou de grande dimension (images, texte, etc.), travailler directement dans l'espace d'entrée devient rapidement impossible. Les réseaux

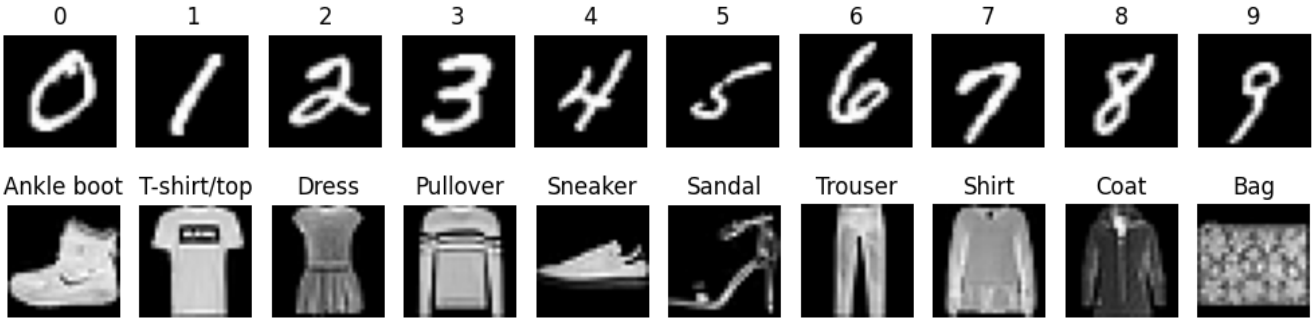


FIGURE 3 – Jeu de données en distribution MNIST (haut) et jeu de données hors distribution Fashion-MNIST (bas).

de neurones profonds[4] jouent alors le rôle d'extracteurs de caractéristiques : couche après couche, ils transforment la donnée afin de produire une représentation de plus en plus abstraite, sur laquelle la prédiction finale devient simple.
Pour illustrer le problème, considérons une classification binaire d'images entre voitures et bicyclettes. De manière schématique, un espace latent intermédiaire peut capturer un concept simple tel que le nombre de roues. La séparation devient alors presque linéaire : moins de deux roues implique une bicyclette, sinon une voiture. Dans ce contexte, une motocyclette, pourtant clairement hors distribution dans l'espace d'entrée, peut devenir *en distribution* dans l'espace latent associé au nombre de roues. Le modèle peut alors prédire avec une confiance élevée qu'il s'agit d'une bicyclette[5].
Ce mécanisme illustre un point clé : en classification, le modèle apprend principalement les invariances et les abstractions nécessaires à maximiser la performance sur la tâche. En conséquence, la représentation peut perdre des informations pourtant cruciales pour détecter des entrées atypiques. La section suivante propose une illustration contrôlée de ce phénomène.

3.1 Illustration simulée

Afin de contrôler explicitement la densité $p(x)$ dans l'espace d'entrée, nous générons les données selon un mélange de gaussiennes :

$$p(x) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(x \mid \mu_k, \Sigma_k), \tag{8}$$

où $\mathcal{N}(x \mid \mu_k, \Sigma_k)$ est la densité d'une gaussienne multivariée et π_k est le *prior* associé à la classe k. Dans nos expériences, la distribution est uniforme sur les classes, i.e. $\pi_k = \frac{1}{K}$. À chaque itération, les moyennes μ_k et les matrices de covariance Σ_k sont tirées aléatoirement. Nous utilisons des matrices de covariance de rang faible et faiblement bruitées[6].

Pour chaque itération, nous générons 5000 données d'entraînement et 5000 données de test, selon 100 variables et 3 classes. Un perceptron multi-couches (MLP) et un auto-encodeur (AE) sont entraînés respectivement à prédire la classe et à reconstruire la donnée d'entrée. Les deux modèles disposent d'un espace latent de dimension 10.

Le jeu de test est ensuite projeté dans l'espace latent du MLP et de l'AE. Nous comparons la log-vraisemblance négative $-\log p(x)$ dans l'espace d'entrée à une estimation de $-\log p(z)$ dans l'espace latent, où z désigne la représentation latente de x. Pour estimer $\log p(z)$, nous utilisons une estimation de densité par noyau gaussien (KDE), avec une largeur de bande fixée à $0.2\sqrt{10}$.

La figure 2 présente les résultats obtenus sur 100 itérations. La corrélation de Spearman entre la densité dans l'espace d'entrée et celle dans l'espace latent est reportée pour le MLP[7] (gauche) et pour l'AE (droite). Lorsque qu'une instance tend à être hors distribution dans l'espace d'entrée, elle est nettement moins souvent associée à une faible densité dans l'espace latent du MLP que dans celui de l'AE.

C'est précisément le phénomène que nous souhaitons souligner : même si un AE classique ne garantit pas la conservation de densité (cf. section 5 pour des alternatives), celle-ci est empiriquement mieux préservée que dans le cas d'un classifieur profond. La section suivante étudie ce phénomène sur un cas concret de détection OOD en classification d'images.

4 Expériences

Dans cette section, nous montrons que l'utilisation de classifieur profond dégrade les performances de détection OOD sur le jeu de données MNIST [15].

Le modèle est entraîné sur MNIST, puis des données hors distribution issues de Fashion-MNIST [28] sont injectées lors de la phase de test (cf. figure 3).

Deux architectures sont étudiées : un perceptron multi-couches (MLP) et un réseau convolutif (CNN). Nous considérons ainsi quatre modèles : deux classifieurs (MLP et CNN) et deux auto-encodeurs (AE-MLP et AE-CNN).

4. Nous parlons ici des réseaux de neurones que nous utilisons dans l'article, à savoir des réseaux convolutifs et des perceptrons multi-couches.
5. Dans cet exemple, un avion à 3 ou 6 roues, OOD dans l'espace d'entrée, resterait OOD dans l'espace de représentation latent. Toute l'information de détection OOD n'est pas perdue dans l'extraction de caractéristiques.
6. Matrices de rang 5.
7. À titre indicatif, l'exactitude moyenne du MLP est de 0.94.

4.1 Protocole expérimental

Le MLP suit l'architecture $784 \to 512 \to 256 \to 128 \to 64 \to 32 \to 10$. Toutes les activations sont des ReLU, à l'exception de la dernière couche, linéaire. Le modèle est entraîné avec une perte d'entropie croisée, un pas d'apprentissage de 10^{-3}, une taille de batch de 128, pendant 4 epochs. Le CNN suit l'architecture $\text{Conv}(32, 3 \times 3) \to \text{Conv}(64, 3 \times 3) \to \text{MaxPool}(2) \to \text{Conv}(128, 3 \times 3) \to \text{MaxPool}(2) \to 128 \times 7 \times 7 \to 256 \to 128 \to 32 \to 10$. Les auto-encodeurs AE-MLP et AE-CNN utilisent le même encodeur que leurs homologues classifieurs, et un décodeur symétrique après le goulot d'étranglement en dimension 32. Ils sont entraînés avec une perte quadratique moyenne, pendant 10 epochs [8].

Les quatre modèles sont entraînés sur les 60 000 images du jeu d'entraînement de MNIST (28×28, normalisées). Pour chacun d'eux, une représentation latente de dimension 32 est extraite. À partir de ces espaces latents (MLP, CNN, AE-MLP et AE-CNN), une forêt aléatoire composée de 200 arbres est entraînée.

Les données MNIST de test (10 000 échantillons en distribution) ainsi que les données Fashion-MNIST de test (10 000 échantillons hors distribution) sont ensuite projetées dans l'espace latent de chaque modèle. La détection hors distribution (OOD) est réalisée dans cet espace à l'aide de la forêt aléatoire, en s'appuyant sur une mesure d'incertitude épistémique calculée selon les équations (5) et (7). Une valeur élevée de l'incertitude épistémique (IE_{ent} ou IE_{var}) est supposée indiquer une instance hors distribution.

4.2 Comparaison de performance

À titre indicatif, le MLP atteint une exactitude de 0.975 et le CNN une exactitude de 0.988 sur le jeu de test MNIST. Les résultats qualitatifs obtenus en sortie des auto-encodeurs AE-MLP et AE-CNN sont présentés en figure 4.

Les figures 5 et 6 présentent respectivement les incertitudes épistémiques calculées par décomposition entropique (5) et par variance (7). Pour chaque modèle, l'histogramme du nombre d'instances de test est représenté en fonction de l'incertitude épistémique, en distinguant les données MNIST (en distribution) et Fashion-MNIST (hors distribution).

Pour les deux classifieurs (MLP et CNN), les distributions d'incertitude associées à MNIST et Fashion-MNIST se chevauchent partiellement très tôt (des instances OOD sont considérées en distribution). Autrement dit, les représentations latentes apprises pour la classification ne permettent pas toujours de distinguer clairement les instances hors distribution. À l'inverse, lorsque la détection est réalisée dans l'espace latent des auto-encodeurs, la séparation entre les deux ensembles est nettement plus marquée, ce qui rend possible un seuillage plus robuste.

Il est toutefois important de noter que, pour les auto-encodeurs, les données OOD et les données en distribution présentent un chevauchement systématique. Par conséquent, il est impossible de distinguer MNIST de Fashion-MNIST sur la seule base de la prédiction du modèle. Néanmoins, nous soutenons que ce phénomène ne traduit pas une défaillance de la détection OOD. Il reflète plutôt la capacité du modèle à identifier, au sein même du jeu de données de référence, des instances qui présentent des caractéristiques OOD. La Figure 7 illustre ce point en présentant les trois instances de test de MNIST pour lesquelles l'AE-MLP manifeste la plus forte incertitude épistémique. Bien qu'appartenant formellement au jeu de données MNIST, ces exemples présentent des caractéristiques clairement OOD.

Ces observations sont cohérentes avec l'hypothèse défendue dans cet article. Les classifieurs apprennent une représentation latente fortement contrainte par l'objectif discriminatif, qui favorise la séparation des classes MNIST mais ne préserve pas nécessairement la structure de densité de l'espace d'entrée. Ainsi, certaines instances de Fashion-MNIST peuvent être projetées dans des régions latentes similaires à celles de MNIST, rendant la détection OOD difficile, même lorsque l'on dispose d'une mesure d'incertitude épistémique raisonnable.

À l'inverse, les auto-encodeurs sont entraînés à reconstruire l'image d'entrée, ce qui les contraint à encoder un spectre plus large d'informations, y compris des facteurs de variation peu utiles pour la classification des chiffres (texture, contours, style d'écriture, structure locale, etc.). Cette préservation plus riche des caractéristiques visuelles semble favoriser une meilleure séparation entre données en distribution et hors distribution dans l'espace latent, et explique la nette amélioration observée en détection OOD.

Enfin, le choix de la mesure d'incertitude épistémique, décomposition entropique (5) ou variance (7), influence peu qualitativement les résultats, et la conclusion reste inchangée.

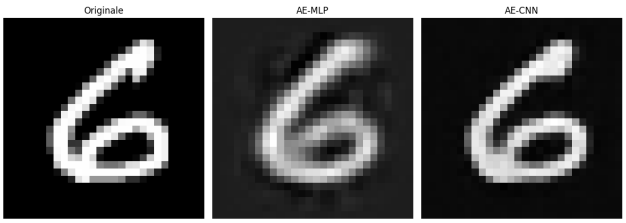


FIGURE 4 – Résultats obtenus en sortie des auto-encodeurs. Chaque modèle encode puis décode l'image d'entrée pour la reconstruire.

5 Discussion & Conclusion

Dans cet article, nous avons montré que la détection d'instances hors distribution par quantification d'incertitude peut s'avérer significativement plus difficile, voire pratiquement impossible, lorsqu'elle est réalisée dans des espaces latents issus de classifieurs profonds, comparativement à des espaces latents issus d'auto-encodeurs. Ce phénomène s'explique par la capacité des classifieurs à apprendre des représentations abstraites optimisées pour la séparation des classes, sans contrainte explicite de conservation de la

8. Le nombre d'epochs est choisi en fonction de la convergence observée.

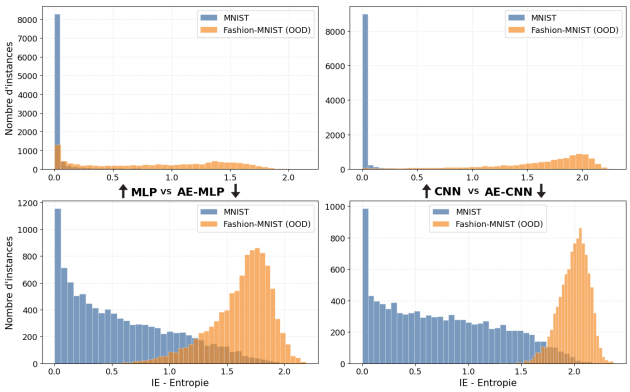


FIGURE 5 – Histogramme de l'incertitude épistémique (IE) selon une décomposition entropique. Les classifieurs sont présentés sur la ligne supérieure et les auto-encodeurs sur la ligne inférieure.

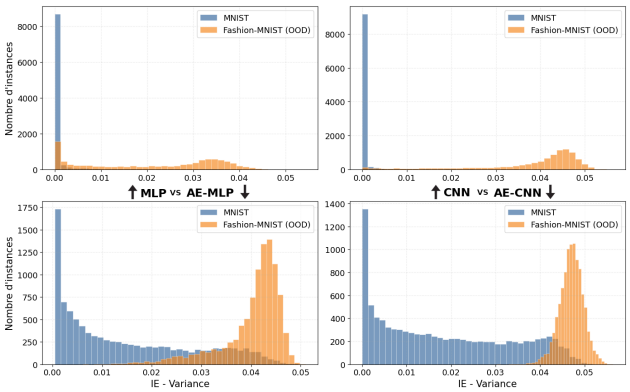


FIGURE 6 – Histogramme de l'incertitude épistémique (IE) selon une décomposition par variance. Les classifieurs sont présentés sur la ligne supérieure et les auto-encodeurs sur la ligne inférieure.

structure de densité dans l'espace d'entrée.

Certaines approches existent pour préserver davantage cette information, par exemple en ajoutant des contraintes sur les distances dans l'espace latent, en utilisant des *normalizing flows*, ou en imposant des contraintes de type *bi-Lipschitz*. Dans tous les cas, ces solutions reviennent à introduire un compromis entre performance de classification et capacité à détecter les données hors distribution, compromis illustré dans nos expériences par la comparaison entre classifieurs et auto-encodeurs.

Nous ne proposons pas de preuve formelle, car nous pensons qu'il n'existe pas de garantie théorique générale : un classifieur peut, dans certains cas, apprendre une représentation latente qui préserve mieux la densité d'origine qu'un auto-encodeur. L'objectif n'est donc pas d'affirmer que les auto-encodeurs sont systématiquement supérieurs, mais de souligner que la détection OOD est intrinsèquement liée au type de représentation appris, et donc à la tâche d'apprentissage.

Pour conclure, nous soutenons qu'évaluer des méthodes de

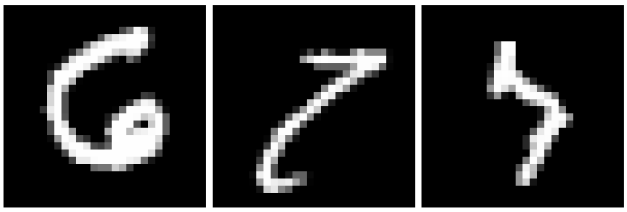


FIGURE 7 – Instances de test de MNIST pour lesquelles l'AE-MLP présente la plus forte incertitude épistémique. Ces instances sont déjà OOD dans le jeu de données témoin.

quantification d'incertitude épistémique en apprentissage profond uniquement à travers des benchmarks de détection hors distribution constitue une approche incomplète. Certaines idées peuvent être pertinentes au niveau de la mesure d'incertitude, mais les performances peuvent être fortement limitées si l'espace de représentation latent a déjà éliminé l'information nécessaire à la détection OOD.

Plus généralement, ce travail vise à attirer l'attention sur un coût rarement explicité de l'optimisation de la performance en classification. Gagner quelques points d'exactitude (*accuracy*) peut se faire au prix d'une représentation plus compressée et plus abstraite, dans laquelle l'information nécessaire à la détection des cas atypiques est partiellement supprimée. Autrement dit, l'amélioration marginale des performances peut réduire la capacité du modèle à reconnaître ses propres limites, et donc à signaler de manière fiable *quand il ne sait pas*.

Remerciements

Nous souhaitons remercier Gauthier Justo et Théo Beaumont, étudiants à CentraleSupélec, pour leur implication dans le projet de fin d'études sur lequel s'appuient les travaux complémentaires présentés dans cet article. Nous remercions également Vincent Lemaire d'Orange Research pour sa participation à la relecture et à la préparation de la version camera-ready de cet article.

Références

[1] Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. On second-order scoring rules for epistemic uncertainty quantification. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2078–2091. PMLR, 23–29 Jul 2023.

[2] Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network : Uncertainty estimation without ood samples via density-based pseudo-counts. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1356–1367. Curran Associates, Inc., 2020.

[3] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1184–1193. PMLR, 10–15 Jul 2018.

[4] P. Kingma Diederik and Welling Max. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4) :307–392, 11 2019.

[5] Zhen Guo, Zelin Wan, Qisheng Zhang, Xujiang Zhao, Qi Zhang, Lance M. Kaplan, Audun Jøsang, Dong H. Jeong, Feng Chen, and Jin-Hee Cho. A survey on uncertainty reasoning and quantification in belief theory and its application to deep learning. *Information Fusion*, 101, 2024.

[6] Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *ICLR*, February 2017.

[7] Arthur Hoarau, Vincent Lemaire, Yolande Le Gall, Jean-Christophe Dubois, and Arnaud Martin. Evidential uncertainty sampling strategies for active learning. *Machine Learning*, 113(9) :6453–6474, June 2024.

[8] Stephen C. Hora. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54(2), 1996. Treatment of Aleatory and Epistemic Uncertainty.

[9] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning : An introduction to concepts and methods. *Machine Learning*, 110 :457–506, 2021.

[10] Joonas, Joonas Rve, Karl Kaspar Haavel, and Meelis Kull. Probability Density from Latent Diffusion Models for Out-of-Distribution Detection. In *ECAI 2025*, pages 5027–5034. IOS Press, 2025.

[11] Mira Jürgens, Nis Meinert, Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Is epistemic uncertainty faithfully represented by evidential deep learning methods ? In *Proceedings of the 41st International Conference on Machine Learning*, International Conference on Machine Learning 24. JMLR.org, 2024.

[12] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision ? In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[13] B. Lakshminarayanan. Practical tutorial on uncertainty and out-of-distribution robustness in deep learning, google research, link : https ://www.gatsby.ucl.ac.uk/b̃alaji/balaji-odsc-talk.pdf, 2024.

[14] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30 : Annual Conference on Neural Information Processing Systems*, pages 6402–6413, 2017.

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

[16] Yucen Lily Li, Daohan Lu, Polina Kirichenko, Shikai Qiu, Tim G. J. Rudner, C. Bayan Bruss, and Andrew Gordon Wilson. *Position : Supervised Classifiers Answer the Wrong Questions for OOD Detection*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 13–19 Jul 2025.

[17] Shuo Lu, Yingsheng Wang, Lijun Sheng, Lingxiao He, Aihua Zheng, and Jian Liang. Out-of-distribution detection : A task-oriented survey of recent advances. *ACM Comput. Surv.*, 58(2), September 2025.

[18] Aryan Mobiny, Pengyu Yuan, Supratik Moulik, Naveen Garg, Carol wu, and Hien Nguyen. Dropconnect is effective in modeling uncertainty of bayesian deep networks. *Scientific Reports*, 11, 03 2021.

[19] Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. Benchmarking uncertainty disentanglement : Specialized uncertainties for specialized tasks. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[20] Vu-Linh Nguyen, Mohammad Hossein Shaker, and Eyke Hüllermeier. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111 :89–122, jan 2022.

[21] Billy Perrigo. Exclusive : The $2 Per Hour Workers Who Made ChatGPT Safer, January 2023.

[22] Y. Sale, P. Hofman, L. Wimmer, Hüllermeier E., and T. Nagler. Second-order uncertainty quantification : Variance-based measures., 2023.

[23] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[24] Mohammad Hossein Shaker and Eyke Hüllermeier. Aleatoric and epistemic uncertainty with random forests. In Michael R. Berthold, Ad Feelders, and Georg Krempl, editors, *Advances in Intelligent Data Analysis XVIII*, pages 444–456, Cham, 2020. Springer International Publishing.

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

[26] Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. Prior and posterior networks : A survey on evidential deep learning methods for uncertainty estimation. *Transactions of Machine Learning Research*, 2023.

[27] Amaan Valiuddin, Ruud Van Sloun, Christiaan Viviers, Peter H.N. de With, and Fons van der Sommen. A review of bayesian uncertainty quantification in deep probabilistic image segmentation. *Transactions on Machine Learning Research*, 2025.

[28] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST : a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv e-prints*, page arXiv :1708.07747, August 2017.

[29] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized Out-of-Distribution Detection : A Survey. *International Journal of Computer Vision*, 132(12) :5635–5662, December 2024.