Analyse de la justesse des modèles de classification binaire : état des lieux dans les systèmes d'aide à la décision clinique

Alex POIRON Marc CUGGIA Sandie CABON

Univ Rennes, CHU Rennes, INSERM, LTSI-UMR 1099, F-35000, Rennes, France

Résumé – Dans le domaine des technologies pour la santé, l'évaluation des systèmes d'aide à la décision basés sur l'apprentissage automatique reste un défi majeur. Cette étude analyse l'évaluation de la justesse de 60 systèmes de classification binaire à travers l'analyse des méthodes de recueil de la vérité terrain et des ensembles de métriques utilisés. Les résultats révèlent que 51.7% des études ne respectent pas la norme ISO 5725-1 :2023 pour le recueil de la vérité terrain et seulement 10% évaluent exhaustivement leurs modèles. Plus de 30 métriques différentes ont été identifiées, avec une évaluation de la calibration présente dans seulement 20% des études. Ces constats soulignent la nécessité de standardiser les pratiques d'évaluation pour renforcer la confiance dans les technologies d'apprentissage automatique appliquées à la santé. Ces travaux constituent une première en proposant une catégorisation des métriques visant à guider dans le choix d'ensemble de métriques pertinent pour l'évaluation des modèles.

Abstract – In the field of health technologies, evaluating machine learning-based decision support systems remains a major challenge. This study analyzes the evaluation of trueness of 60 binary classification systems through the analysis of ground truth collection methods and set of used metrics. Results reveal that 51.7% of studies do not comply with ISO 5725-1:2023 for ground truth collection, and only 10% comprehensively evaluate their models in terms of trueness. Over 30 different metrics were identified, with calibration assessment present in only 20% of studies. These findings underscore the urgent need to standardize evaluation practices to enhance confidence in machine learning technologies applied to healthcare. This work represents a first by proposing a categorization of metrics to guide the selection of a relevant set of evaluation metrics for classification models.

1 Introduction

Un nombre croissant de méthodes basées sur l'apprentissage automatique est développé pour analyser des données images, signal ou tabulaires, créer des systèmes d'aide à la décision ou encore prédire des événements d'intérêt. Ces objectifs concernent de nombreux domaines comme la santé, l'agriculture ou l'environnement. Malgré des performances souvent présentées comme supérieures, la communauté scientifique fait face à plusieurs défis : il reste difficile d'évaluer exhaustivement leurs performances et de définir leurs domaines d'application. Dans le domaine de la santé, bien que le nombre de propositions de systèmes d'aide à la décision à base d'apprentissage ait explosé, peu de ces systèmes sont déployés en pratique clinique. Cela s'explique par la complexité, l'opacité des méthodes et la difficulté d'appréhender en parallèle les multiples aspects nécessaires à leur évaluation. Récemment, pour ces systèmes à hauts risques pour les droits humains, l'AI Act a mis en avant l'importance de critères tels que l'exactitude, la robustesse, la cybersécurité, la qualité des données ou encore la transparence pour garantir la confiance. Étant donné que chaque critère fait l'objet de recherches spécifiques, et que certains manquent de définition et même de méthodes d'évaluation consensuelles, la résolution de ces limitations représente un défi important. Dans cette première étude, nous nous concentrons sur l'évaluation de l'exactitude puisqu'il s'agit d'un élément principal optimisé lors de l'apprentissage. C'est aussi ce critère qui est la base de l'analyse d'autres critères, comme la robustesse ou l'équité. En effet pour évaluer ces derniers, la stabilité des performances en terme d'exactitude sera alors examinée dans divers scénario ou sous-groupes.

L'exactitude peut-être définie comme une combinaison de la "justesse" et de la "fidélité" [1]. La justesse désigne l'étroitesse de l'accord entre la moyenne arithmétique obtenue à partir d'une large série de résultats d'essai et la valeur de référence acceptée ou la valeur vraie. La fidélité désigne l'étroitesse de l'accord entre des résultats d'essai indépendants obtenus dans des conditions déterminées. Dans ce travail, la fidélité n'a pas été étudiée. Elle sera considérée, de par sa définition, comme relative à la robustesse et fera l'objet d'une étude à part entière. Concernant la justesse, de nombreuses métriques ont été proposées pour l'évaluation de la classification [2]. Une seule mesure ne suffit pas à couvrir tous les aspects de la justesse [3]. Aujourd'hui, il reste à définir quels ensembles de mesures choisir pour évaluer rigoureusement un système ? Dans quelle mesure ces ensembles couvrent les différents aspects de la justesse ? Quels en sont ces aspects ?

D'autre part, la définition de la vérité terrain est importante pour évaluer la justesse. Le standard ISO 5725-1 :2023 recense des manières de la récolter [1]. Dans quelle mesure est ce que cela est respecté ? Quelles sont les pratiques courantes ?

Dans cette étude, nous avons réalisé une revue de la littérature afin d'observer les pratiques courantes en technologies pour la santé, un des domaines avec les exigences réglementaires les plus fortes, en se concentrant sur les méthodes de récolte de la vérité terrain ainsi que les ensembles de métriques utilisés. En particulier, nous avons étudié 60 systèmes d'aide à la décision clinique basés sur de l'apprentissage automatique. L'identification des aspects de la justesse à extraire des publications s'est avérée complexe en raison de l'absence d'une taxonomie communément adoptée dans ce domaine. En 2024, nous avions proposé une première analyse que nous consolidons fortement ici [4]. En effet, depuis, une part substantielle de notre travail a consisté à élaborer un cadre adapté pour

identifier à quel aspect de la justesse se réfère les métriques.

L'article est organisé en trois sections. La section 2 décrit la méthode de sélection des articles étudiés, la définition et la justification des informations à extraire et leur catégorisation. Dans la section 3, les résultats obtenus sont présentés. La discussion et la conclusion sont données dans la section 4.

2 Méthodes

Deux questions ont orienté notre analyse de la mesure de la justesse des systèmes d'aide à la décision clinique :

- Quelles sont les stratégies d'annotation utilisées pour capturer la vérité de terrain ?
- Pouvons-nous identifier un ensemble de métrique optimal permettant d'évaluer de manière exhaustive la justesse d'un système d'aide à la décision clinique ?

2.1 Sélection des études d'intérêt

Une revue systématique a été réalisée en suivant les recommandations PRISMA. La requête suivante a été faite sur la base de données PubMed (orientée santé) : ("clinical decision support system") AND ("machine learning" OR "deep learning"). Elle a été réalisée mi-2023 avec un ciblage sur les études publiées en 2022 pour obtenir un ensemble d'études récentes à analyser. Chacune a ensuite été lue et catégorisée en fonction du type de données sur lequel était basé le système d'aide à la décision. Cinq catégories de données ont été considérées : signal, image, tabulaire, vidéo et hybride (combinaison de deux ou plusieurs types de données). Les études ne portant pas sur des systèmes basés sur de l'apprentissage automatique et les revues de la littérature ont été écartées. Enfin, les études ont fait l'objet d'une inclusion proportionnelle, c'est-à-dire qu'une sélection aléatoire a été effectuée en maintenant une représentation proportionnelle des types de données. Il s'agissait de réduire le nombre d'articles à analyser tout en garantissant une représentation des systèmes en fonction des données exploitées.

2.2 Ciblage des informations à extraire

2.2.1 Méthodes de recueil de la vérité terrain

L'identification des différentes manières de définir la vérité terrain est l'un des deux axes clés pour observer l'évaluation de la justesse des systèmes d'aide à la décision dans la littérature [5]. La catégorisation proposée par la norme ISO 5725-1 :2023 a été d'abord considérée [1]. Elle identifie quatre façons de construire la vérité terrain à partir :

- d'une valeur théorique ou établie, fondée sur des principes scientifiques (ISOa) ;
- d'une valeur assignée ou certifiée, fondée sur les travaux d'une organisation nationale ou internationale (ISOb) ;
- d'une valeur de consensus ou certifiée, fondée sur un travail expérimental en collaboration et placé sous les auspices d'un groupe scientifique ou technique (ISOc) ;
- de l'espérance, c'est-à-dire la moyenne d'une population spécifiée de mesures, dans les cas où a), b) et c) ne sont pas applicables (ISOd).

Toutes les pratiques n'entrant pas dans ces catégories, la liste a été complétée par celles rencontrées :

- La méthode de recueil de la vérité terrain n'est pas mentionnée (PC1) ;
- La vérité terrain est recueillie par des experts annotant séparement des parties des données (PC2) ;
- La vérité terrain est recueillie par un seul expert annotant toutes les données (PC3) ;
- La vérité terrain est recueillie par des experts annotant séparement des données et un sous-ensemble des données est annoté par tous et permet d'analyser l'accord inter-annotateur (PC4)

2.2.2 Ensembles de métriques

Il est admis que l'utilisation d'une seule métrique est insuffisante pour une évaluation complète d'un modèle de classification [3]. Cependant, il n'existe pas de lignes directrices établies ni pour guider dans le choix d'un ensemble de métriques adéquat, ni en les reliant aux aspects qu'elles mesurent. Pour analyser la couverture des évaluations faites, nous avons recensé ces aspects, compilant les recommandations de travaux émergents d'experts du domaine de l'apprentissage automatique et des mathématiques.

En classification binaire, deux aspects haut niveau sont clairement identifiés par la communauté pour garantir une évaluation complète des performances : la calibration et la discrimination [6, 7]. La calibration fait référence à la concordance entre les probabilités prédites par le modèle et la distribution réelle du phénomène observé. Un modèle bien calibré produit donc des estimations de risque qui correspondent étroitement aux taux d'événements observés. La discrimination, quant à elle, se réfère à la capacité d'un modèle à distinguer les deux classes qu'il vise à séparer (e.g., individus sains/individus malades, évènement/non évènement). C'est-à-dire, une fois qu'un seuil est appliqué aux probabilités fournies par le modèle pour produire une décision. Les métriques mesurant la discrimination sont majoritairement issues de la matrice de confusion.

Récemment, Canbek et al. ont proposé une représentation similaire au tableau périodique des éléments chimiques PToPI (Periodic Table of Performance Instruments) afin d'agréger visuellement les propriétés de chaque métrique [8]. Ils posent que les nombres de vrais/faux positifs, vrais/faux négatifs sont indivisibles et la base de toutes les métriques. La visualisation proposée précise la complémentarité, la dualité des métriques et le type d'erreur (type I - faux positifs, type II - faux négatifs) qu'elles caractérisent. En effet, des métriques permettent de juger uniquement le comportement du modèle sur la classe positive (e.g., sensibilité) ou la classe négative (e.g., spécificité). Ces deux perspectives sont importantes à considérer parallèlement [7].

De la même façon, les performances peuvent être évaluées soit par les métriques intrinsèques liées aux productions du modèle (e.g., sensibilité, spécificité), soit au regard de la prévalence du phénomène étudié (e.g., valeur prédictive positive, valeur prédictive négative) [7, 8]. Les métriques intrinsèques caractérisent les performances indépendamment du contexte, tandis que les métriques intégrant la prévalence, reflètent la performance du modèle dans une situation opérationnelle spécifique (i.e., le contexte populationnel).

A partir de ces constats, nous avons catégorisé les métriques en quatre catégories principales en fonction de l'aspect d'évaluation de la justesse qu'elles couvrent (calibration, discrimi-

nation globale, discrimination au regard de la classe positive et discrimination au regard de la classe négative). Les métriques ayant trait à la discrimination ont été sous-divisées en fonction de si elles intègrent le regard du modèle ou de la prévalence (Table 1).

TABLE 1 : Catégorisation des métriques

Catégorie	Description
m_{cal}	métriques de calibration reposant sur les probabilités produites par le modèle
m_{disc_g}	métriques de discrimination globales reposant sur : - la performance du modèle $m_{disc_{g/m}}$ - la prévalence du phénomène étudié $m_{disc_{g/p}}$
m_{disc_+}	métriques de discrimination pour la classe positive reposant sur : - la performance du modèle $m_{disc_{+/m}}$ - la prévalence du phénomène étudié $m_{disc_{+/p}}$
m_{disc_-}	métriques de discrimination pour la classe négative reposant sur : - la performance du modèle $m_{disc_{-/m}}$ - la prévalence du phénomène étudié $m_{disc_{-/p}}$

Dans ce travail, nous avons considéré qu'un ensemble de métriques était complet s'il permettait a minima d'évaluer un système avec une une métrique de chaque catégorie (i.e., m_{cal}, m_{disc_g}, m_{disc_+} et m_{disc_-}).

3 Résultats

3.1 Panorama des systèmes étudiés

A la fin du processus de sélection, soixante études portant sur de la classification binaire ont été retenues [1]. Les systèmes exploitent des données tabulaires (64.4%), images (23.7%), hybride (6.7%), signal (3.4%) et vidéo (1.7%). La grande majorité ont pour objectif du support au diagnostic (79.7%). Par exemple, certains systèmes prédisent la mortalité néonatale à partir de données clinico-biologiques ou le cancer du sein à partir d'images histopathologiques. Dans une moindre mesure, ils visent aussi l'amélioration de la sécurité des patients (8.5%), servent de support à l'analyse d'images (7.8%) ou à réduire les coûts d'hospitalisation (5.1%). Les approches rencontrées sont à 57,1% des approches classiques supervisées (e.g., petits réseaux de neurones, régression logisitique, machines à vecteur de support), à 31,1% des approches ensemblistes (e.g., forêt aléatoire, XGBoost) et à 10% des approches par apprentissage profond (e.g., réseaux de neurones convolutionnels, récurents).

3.2 Méthodes de recueil de la vérité terrain courantes

La Figure 1 illustre les pratiques de recueil de la vérité terrain rencontrées. Le premier constat est que plus d'une étude sur deux (51.7%) ne base pas son recueil sur une méthode acceptée par la norme. Pour 33.4% des études, la méthode n'était pas clairement rapportée. La vérité terrain est souvent recueillie par des experts annotant séparement des parties des données (11.6%). Parfois, un seul expert est mobilisé (5%) ou un consensus sur une partie des données est obtenu (1.7%).

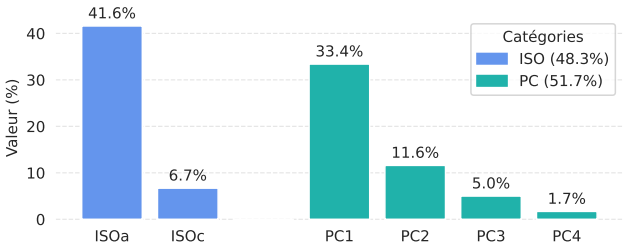

FIGURE 1 : Représentation en pourcentage des méthodes de recueil de la vérité terrain.

3.3 Ensembles et métriques rencontrés

L'analyse des métriques d'évaluation révèle une diversité et une utilisation contrastées. Sur l'ensemble des études, plus de 31 métriques différentes ont été identifiées, avec des ensembles composés de 0 à 11 métriques. Néanmoins, seules 6 études (10%) évaluent leur modèle de manière exhaustive selon les 4 aspects définis. Cette limitation s'explique par deux facteurs principaux : premièrement, 23 études (38.3%) n'utilisent pas plus de trois métriques, et deuxièmement, certains aspects sont sous-représentés. L'évaluation de la calibration n'est présente que dans 12 études (20%), et l'évaluation de la discrimination selon la perspective négative n'est abordée que dans 32 études (53.3%). La Table 2 rapporte les métriques rencontrées selon les catégories définies Section 2.2.2. Elles couvrent plus ou moins toutes les catégories. Il y a une plus grande variabilité pour la calibration (11 métriques) et la discrimination globale (10 métriques). On remarque que peu des métriques (3 métriques) évalue la discrimination d'un point de vue de la classe négative. Aucune métrique tombant dans la catégorie $m_{disc_{g/p}}$, comme la Markedness [8], n'a été rapportée.

4 Discussion

Partant du constat que peu de systèmes à base d'apprentissage automatique sont arrivés au chevet du patient, nous faisons l'hypothèse que c'est en partie à cause de la difficulté de mesurer les performances de ces méthodes.

Dans ce travail, nous nous concentrons sur l'analyse de la justesse. Ils permettent une analyse de la manière dont elle est aujourd'hui évaluée pour les systèmes d'aide à la décision clinique basés sur une classification binaire. Nous avons proposé de l'analyser en la décomposant en deux éléments : la vérité terrain et les ensembles de métriques permettant de mesurer la distance des prédictions à celle-ci. Nous avons construit une catégorisation des pratiques basées sur l'ISO 5725-1 :2023 et les pratiques courantes pour le recueil de la vérité terrain. Et, pour les métriques, nous les avons classées selon quatre aspects principaux : calibration discrimination globale, discrimination au regard de la classe positive et discrimination au regard de la classe négative. Pour celles relevant de la discrimination, nous

[1] liste des références et diagramme PRISMA de la revue exploratoire. https ://gitlab.com/alex.poiron/annexes-gretsi-2025

TABLE 2 : Recensement des métriques observées dans la littérature par catégories (avec distribution en % au sein des 60 études)

Catégorie	N	Liste des métriques rencontrées
m_{cal}	11	Courbes de calibration (13.3%), Score de Brier (6.7%), Perte logarithmique (1.7%), Perte sigmoïde (1.7%), Erreur de calibration absolue (1.7%), Erreur de calibration attendue (1.7%), Indice de calibration attendu (1.7%), Test de Hosmer-Lemeshow (1.7%), Ratio observé/attendu (1.7%), Pente de calibration (1.7%), Différence de risque (1.7%)
m_{disc_G}	10	$m_{disc_{g/m}}$ (8) : AUC-ROC (73.3%), Accuracy (55%), Balanced Accuracy (3.3%), Indice de Youden (3.3%), Ratio de vraisemblance positif (1.7%), Ratio devraisemblance positif (1.7%), Coefficient de similarité de Dice (1.7%), Distance de Hausdorff (1.7%) $m_{disc_{g/p}}$ (2) : Coefficient de corrélation de Matthews (8.3%), Kappa de Cohen (6.7%)
m_{disc_+}	7	$m_{disc_{+/m}}$ (2) : Sensibilité (SE) (70%), Taux de faux négatifs (1.7%) $m_{disc_{+/p}}$ (5) : Valeur prédictive positive (VPP) (50%), Score F1 (28.3%), Aire sous la courbe VPP/SE (13.3%), Score F2 (3.3%), Score F0.5 (1.7%)
m_{disc_-}	3	$m_{disc_{-/m}}$ (2) : Spécificité (45%), Taux de faux positifs (5%) $m_{disc_{-/p}}$ (1) : Valeur prédictive négative (16.7%)

avons effectué une sous-catégorisation en fonction de si l'on se base sur les prédictions du modèle ou sur la prévalence du phénomène.

Nos résultats démontrent plusieurs points critiques. Les méthodes de recueil de la vérité terrain ne sont pas standardisées, ce qui limite fortement la confiance pouvant être portée à ce qui est appris par les modèles. Cela est principalement due à la complexité inhérente à l'expertise nécessaire pour annoter des données de santé. Il est difficile de receuillir l'annotation de nombreuses données par de nombreux experts. La validation de ces pratiques et leur intégration aux standards, notamment PC1 et PC4, est envisageable mais ceci nécessitera de démontrer l'objectivité des modèles. C'est sur cette question que notre travail sera ensuite orienté. Bien que les ensembles de métriques rencontrés soient divers, très peu couvrent tous les aspects identifiés dans la littérature. Il est donc difficile de préciser leurs comportements et le niveau de confiance qu'on peut leur accorder en fonctions des contextes d'application.

La diversité terminologique autour de ce sujet rend l'analyse complexe [8]. Nous avons choisi le terme "justesse" selon la norme ISO 5725-1 :2023, conscients de son caractère potentiellement discutable. L'essentiel demeure : la vérité terrain est primordiale dans l'apprentissage, et mesurer correctement l'écart par rapport à celle-ci reste fondamental.

5 Conclusion

À l'heure où les méthodes d'apprentissage automatique et d'intelligence artificielle connaissent une expansion rapide dans le traitement de l'information, il devient urgent de développer des approches méthodologiques d'évaluation rigoureuses. Notre objectif principal est de fournir des mécanismes de validation permettant de vérifier que les systèmes mesurent effectivement ce qu'ils prétendent mesurer. La multiplication des applications dans des domaines aussi variés que la santé, la finance, ou la recherche scientifique soulève des enjeux fondamentaux. Il est impératif de proposer des lignes directrices précises pour le choix des métriques utilisées. Cette démarche permettra non seulement de quantifier objectivement les progrès réalisés, mais aussi de renforcer la confiance dans ces nouvelles technologies de traitement de l'information.

Références

[1] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION : Exactitude (justesse et fidélité) des résultats et méthodes de mesure — partie 1 : Principes généraux et définitions. Rapport technique ISO 5725-1 :2023, International Organization for Standardization, Geneva, Switzerland, 2023. https://www.iso.org/obp/ui/#iso:std:iso:5725:-1:ed-2:v1:fr.

[2] M. SOKOLOVA et G. LAPALME : A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009.

[3] S. A. HICKS, I. STRÜMKE, V. THAMBAWITA *et al.* : On evaluation metrics for medical applications of artificial intelligence. *Scientific reports*, 12(1):5979, 2022.

[4] A. POIRON, S. CABON et M. CUGGIA : How trueness of clinical decision support systems based on machine learning is assessed ? *Digital Health and Informatics Innovations for Sustainable Health Care Systems*, pages 813–817, 2024.

[5] A. A. H. DE HOND, A. M. LEEUWENBERG, L. HOOFT *et al.* : Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare : a scoping review. npj digital medicine 2022 5 (1) ; 1-13. *this scoping review the authors look at AI-based prediction model (AIMP) using a*, 6, 2022.

[6] Y. HUANG, W. LI, F. MACHERET *et al.* : A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association*, 27(4):621–633, 2020.

[7] G. VAROQUAUX et V. CHEPLYGINA : Machine learning for medical imaging : methodological failures and recommendations for the future. *NPJ digital medicine*, 5(1):48, 2022.

[8] G. CANBEK, T. TASKAYA TEMIZEL et S. SAGIROGLU : Ptopi : A comprehensive review, analysis, and knowledge representation of binary classification performance measures/metrics. *SN Computer Science*, 4(1):13, 2022.