Explication visuelle des systèmes d'intelligence artificielle

CONTEXTE

L'intelligence artificielle (IA) numérique et le *deep learning* ont conduit à plusieurs succès en médecine [2], notamment pour l'aide au diagnostic en imagerie. Dans ce cas, des images annotées et enrichies (par exemple avec le contour des anomalies détectées) constituent une explication satisfaisante pour justifier au médecin les prédictions de l'IA. En revanche, l'explication est plus complexe à produire pour les systèmes de recommandation thérapeutique, c'est-à-dire les systèmes aidant le médecin lors de la prescription d'un traitement, à partir des données patients. En effet, les IA numériques sont des « boîtes noires » dont le fonctionnement est opaque aux yeux du médecin.

Ce manque d'explication est une des principales causes de la réussite bien moindre de l'IA en thérapie : le médecin a besoin de comprendre les recommandations des systèmes d'aide à la décision [11], faute de quoi il ne les suivra pas. De plus, suivre aveuglément les recommandations des IA conduit à un biais d'automatisation [4]. Par conséquent, les systèmes d'aide à la décision thérapeutique existant n'utilisent que rarement l'IA numérique : la plupart d'entre eux reposent sur l'informatisation des guides de bonnes pratiques cliniques sous forme de base de règle [15, 7]. Cependant, cela signifie que ces systèmes ne bénéficies pas des avancées récentes en matière d'IA numérique.

Il serait donc souhaitable de pouvoir expliquer les prédictions de l'IA numérique à un humain (médecin ou patient) en mettant au point des **IA explicables** (IAE, *Explainable Artificial Intelligence*, XAI) [1]. Ce champ de recherche est aujourd'hui en plein essor (numéros spéciaux de revue, ateliers de conférence,...), et le récent rapport Villani sur l'IA [16] recommande « d'ouvrir les boîtes noires » en insistant sur le domaine médical. Plusieurs approches existent, comme l'apprentissage d'un modèle interprétable (symbolique) au voisinage de la prédiction à expliquer [12]. La visualisation est une autre approche envisageable [13], qui peut permettre de visualiser le comportement de l'IA au voisinage de la prédiction, sans avoir besoin de réaliser l'apprentissage d'un modèle.

Dans la littérature, la visualisation a été appliquée à l'IA pour aider les ingénieurs (*data-scientists*) lors de la conception des systèmes intelligents [14, 10] ou pour faciliter l'enseignement de l'IA, mais rarement pour aider l'utilisateur final [5]. De plus, la visualisation est souvent cantonnée à l'explication des systèmes de vision par ordinateur, pour lesquels il est relativement aisé de produire des images. Au contraire, au sein du LIMICS, nous avons réalisé un travail préliminaire lors du projet européen H2020 DESIREE sur la prise en charge du cancer du sein. Nous avons montré la possibilité d'utiliser la visualisation pour traduire graphiquement un raisonnement médical complexe et aider à la décision thérapeutique [8]. Cependant, ce travail reposait sur le raisonnement à partir de cas [3], une technique relativement ancienne et qui n'est pas aussi performante que les techniques d'IA plus récentes (*deep learning*, *boosting*).

QUESTIONS POSÉES

— Comment expliquer visuellement les prédictions issues d'une intelligence artificielle de type « boîte noire » ?
— Comment appliquer ces explications à l'aide à la décision thérapeutique ?
— Quelles sont les limites des explications visuelles ?
— Comment évaluer ces méthodes d'explication ?

OBJECTIFS

Concevoir et évaluer des méthodes visuelles pour expliquer les prédictions produites par une intelligence artificielle de type « boîte noire » (réseau de neurones et *deep learning* ou *boosting* par exemple), dans le cadre de l'aide à la décision thérapeutique.

MÉTHODES

Tout d'abord, une étude bibliographique sera réalisée, en recherchant plus particulièrement : (1) les techniques de visualisation d'information abstraite, (2) les approches d'IA numérique explicable, (3) l'utilisation des IA numériques en médecine et en thérapie.

Ensuite, une méthode sera mise au point pour expliquer des raisonnements « boîtes noires ». L'approche sera la suivante : il s'agira de générer plusieurs entrées, d'obtenir les sorties correspondantes avec l'IA, et de visualiser les différentes entrées et leurs sorties respectives. Appliqué à la thérapie, cela consistera à générer des patients fictifs proches du patient à traiter, à calculer les prédictions de l'IA pour le patient à traiter et les patients fictifs, et visualiser l'ensemble. Ce système permettra non seulement

de voir la recommandation du système pour le patient à traiter, mais aussi comment cette recommandation évoluerait en cas de variation des données en entrée. Par exemple, si le patient avait 5 ans de moins, la décision serait-elle la même ? Ce type d'information est susceptible d'intéresser un médecin car elle lui donne une vue générale et lui permet de choisir une décision alternative, par exemple s'il juge son patient « plus jeune » que son âge.

La première partie de la méthode consistera donc à générer les patients fictifs de manière optimale, c'est-à-dire de sorte à expliquer au mieux le raisonnement de l'IA en montrant les valeurs provoquant le basculement vers une autre décision. Une piste à étudier consisterait à réduire ce problème à un problème d'optimisation, puis à le résoudre à l'aide d'une métaheuristique.

La seconde partie de la méthode consistera à visualiser le patient à traiter et les patients fictifs. Plusieurs techniques de visualisation seront testées, et éventuellement combinées entre elles. Une piste est l'utilisation des boîtes arc-en-ciel [6], une technique pour la visualisation d'ensemble que nous avons proposé récemment au LIMICS et semble prometteuse pour l'explication des raisonnements [8, 9].

À ce stade, un prototype fonctionnel sera réalisé et implémenté. Afin de s'assurer de la nature générique du prototype, plusieurs approches d'IA numériques seront utilisées, par exemple réseau de neurones et *boosting*. Des jeux de données seront utilisés pour tester et évaluer le prototype, en incluant des jeux de données librement disponibles sur Internet et des jeux de données disponibles au sein du LIMICS (par exemple sur la prise en charge du cancer du sein).

L'évaluation de la qualité des explications est une question de recherche à part entière. Plusieurs pistes seront étudiées : L'explication permet-elle de convaincre un médecin ? L'explication permet-elle au médecin de juger si la recommandation est pertinente, ou s'agit-il d'un artefact des données ? L'explication permet-elle au médecin de s'adapter en cas d'imprévus (par exemple dans le cas où le traitement recommandé ne peut pas être prescrit à cause d'une contre-indication, d'une allergie ou du refus du patient) ? L'explication permet-elle au médecin de mieux réaliser l'éducation du patient ?

Enfin, le prototype fera l'objet d'une évaluation auprès d'un groupe de médecins généralistes, par exemple en partenariat avec la SFTG (Société de Formation Thérapeutique du Généraliste), une association avec laquelle nous avons l'habitude de collaborer sur des projets de recherche. L'évaluation qualitative permettra de vérifier que les médecins comprennent bien les explications proposées au point. Une évaluation plus poussée, quantitative, pourra être envisagée selon le temps disponible.

CALENDRIER

- Semestre 1 : recherche bibliographique, mise au point de la méthode pour générer les patients fictifs
- Semestre 2 : mise au point des méthodes de visualisation et sélection de la méthode la plus prometteuse
- Semestre 3 : application sur plusieurs technique d'IA et jeux de données, rédaction d'un premier article
- Semestre 4 : implémentation d'un prototype, évaluation de ce prototype sur un petit groupe de médecins
- Semestre 5 : écriture d'un second article, rédaction de la thèse
- Semestre 6 : finalisation de la thèse et envoi aux rapporteurs

PROPOSITION D'ARTICLES

1. Un article sur la méthode proposée pour expliquer les IA numériques. Cible : une revue en informatique (par exemple *IEEE Transaction in Visualisation and Computer Graphics*) ou une conférence en visualisation (*Information Visualization, EuroVis*) ou en IA (*International Joint Conference on Artificial Intelligence*).
2. Un article de revue sur les méthodes pour évaluer la qualité des explications en santé. Cible : une revue ou une conférence en informatique médicale (*BMC Medical Informatics and Decision Making*, *Medical Informatics Europe*).
3. Un article sur l'application de la méthode d'explication proposée en santé. Cible : revue en informatique ou informatique médicale (*Artificial Intelligence in Medicine*, *Journal of Biomedical Informatics* ou *Journal of the American Medical Informatics Association*).

Directeur de thèse :
Jean-Baptiste Lamy, Maître de Conférences à l'Université Paris 13, HDR, section 27, laboratoire LIMICS
Co-encadrants :
Karima Sedki, Maître de Conférences à l'Université Paris 13 section 27, laboratoire LIMICS

Références

[1] O Biran and C Cotton. Explanation and justification in machine learning : A survey. In *Workshop on Explainable AI (XAI)*, pages 8–13, 2017.

[2] T Ching, D S Himmelstein, B K Beaulieu-Jones, A A Kalinin, B T Do, G P Way, E Ferrero, P M Agapow, M Zietz, M M Hoffman, W Xie, G L Rosen, B J Lengerich, J Israeli, J Lanchantin, S Woloszynek, A E Carpenter, A Shrikumar, J Xu, E M Cofer, C A Lavender, S C Turaga, A M Alexandari, Z Lu, D J Harris, D DeCaprio, Y Qi, A Kundaje, Y Peng, L K Wiley, M H S Segler, S M Boca, S J Swamidass, A Huang, A Gitter, and C S Greene. Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society, Interface*, 15(141), 2018.

[3] N Choudhury and S A Begum. A survey on case-based reasoning in medicine. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 7(8) :136–144, 2016.

[4] K Goddard, A Roudsari, and J C Wyatt. Automation bias : empirical results assessing influencing factors. *Int J Med Inf*, 83(5) :368–75, 2014.

[5] F M Hohman, M Kahng, R Pienta, and D H Chau. Visual Analytics in Deep Learning : An Interrogative Survey for the Next Frontiers. *IEEE transactions on visualization and computer graphics*, 2019.

[6] J B Lamy, H Berthelot, C Capron, and M Favre. Rainbow boxes : a new technique for overlapping set visualization and two applications in the biomedical domain. *Journal of Visual Language and Computing*, 43 :71–82, 2017.

[7] J B Lamy, V Ebrahiminia, C Riou, B Séroussi, J Bouaud, C Simon, S Dubois, A Butti, G Simon, M Favre, H Falcoff, and A Venot. How to translate therapeutic recommendations in clinical practice guidelines into rules for critiquing physician prescriptions ? Methods and application to five guidelines. *BMC Medical Informatics and Decision Making*, 10 :31, 2010.

[8] J B Lamy, B Sekar, G Guezennec, J Bouaud, and B Séroussi. Explainable artificial intelligence for breast cancer : a visual case-based reasoning approach. *Artif Intell Med*, 94 :42–53, 2019.

[9] J B Lamy and R Tsopra. Translating visually the reasoning of a perceptron : the weighted rainbow boxes technique and an application in antibiotherapy. In *International Conference Information Visualisation (iV)*, pages 256–261, London, United Kingdom, 2017.

[10] M Li, Z Zhao, and C Scheidegger. Visualizing neuron activations of neural networks with the grand tour. In *Workshop on Visualization for AI Explainability (VISxAI)*, 2018.

[11] M Lugtenberg, J S Burgers, C F Besters, D Han, and G P Westert. Perceived barriers to guideline adherence : a survey among general practitioners. *BMC family practice*, 12 :98, 2011.

[12] M T Ribeiro, S Singh, and C Guestrin. Why should i trust you ? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[13] R Sevastjanova, F Beck, B Ell, C Turkay, R Henkin, M Butt, D Keim, and M El-Assady. Going beyond visualization : Verbalization as complementary medium to explain machine learning models. In *Workshop on Visualization for AI Explainability (VISxAI)*, 2018.

[14] E M Smith, J Smith, P Legg, and S Francis. Visualising state space representations of LSTM networks. In *Workshop on Visualization for AI Explainability (VISxAI)*, 2018.

[15] B Séroussi, J Bouaud, and G Chatellier. Guideline-based modeling of therapeutic strategies in the special case of chronic diseases. *Int J Med Inf*, 74(2) :89–99, 2005.

[16] C Villani, M Schoenauer, Y Bonnet, C Berthet, A C Cornut, F Levin, B Rondepierre, and S Biabiaby-Rosier. *Donner un sens à l'intelligence artificielle : Pour une stratégie nationale et européenne*. 2018.