Intelligence artificielle explicable pour le cancer du sein : Une approche visuelle de raisonnement à partir de cas

Jean-Baptiste Lamy, Boomadevi Sekar, Gilles Guezennec, Jacques Bouaud, Brigitte Séroussi

▶ **To cite this version:**

Jean-Baptiste Lamy, Boomadevi Sekar, Gilles Guezennec, Jacques Bouaud, Brigitte Séroussi. Intelligence artificielle explicable pour le cancer du sein : Une approche visuelle de raisonnement à partir de cas. Extraction et Gestion des Connaissances (EGC), 2020, Bruxelles, Belgique. hal-03479312

HAL Id: hal-03479312
https://hal.science/hal-03479312

Submitted on 14 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Intelligence artificielle explicable pour le cancer du sein : Une approche visuelle de raisonnement à partir de cas

Jean-Baptiste Lamy*, Boomadevi Sekar**, Gilles Guezennec*
Jacques Bouaud*,***, Brigitte Séroussi*,****

*LIMICS, Université Paris 13, 93017 Bobigny, France,
INSERM UMRS 1142, Sorbonne Universités,
jean-baptiste.lamy@univ-paris13.fr
**School of Computing and Mathematics, Ulster University, United Kingdom
***AP-HP, DRCI, Paris ****AP-HP, Hôpital Tenon, Département de Santé Publique, Paris

Abstract. Dans le cancer du sein, l'intelligence artificielle peut aider les médecins à effectuer le diagnostic et à prescrire le bon traitement. Cependant, la plupart des méthodes récentes (comme l'apprentissage profond) sont des "boîtes noires" qui ne permettent pas d'expliquer les prédictions de machine. Au contraire, les médecins ont besoin de comprendre les recommandations des systèmes d'aide à la décision afin d'y adhérer. Nous proposons une approche visuelle de raisonnement à partir de cas, permettant une visualisation à la fois quantitative et qualitative de la similarité entre les cas. Cette approche a été testée sur 3 jeux de données publics pour le diagnostic et des données réelles pour la thérapie, et présentée à 11 médecins. Cet article est un résumé de: Jean-Baptiste Lamy, Boomadevi Sekar, Gilles Guezennec, Jacques Bouaud, Brigitte Séroussi. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. Artificial Intelligence in Medicine 2019(94):42-53.

1 Introduction

Le raisonnement à partir de cas (RAPC) (Aamodt and Plaza, 1994) est une forme de raisonnement par analogie reposant sur une base de cas antérieur dont la solution est connue. Dans le RAPC, un cas correspond à un problème , et la requête est un nouveau cas pour lequel nous recherchons la solution. Le RAPC suit les étapes suivantes : *rechercher* les cas similaires à la requête dans la base, *réutiliser* les solutions des cas similaires et les *adapter* à la requête, et *apprendre* le nouveau cas en l'intégrant dans la base. Le RAPC connaît un certain succès en médecine (Choudhury and Begum, 2016) : ici, un cas est un patient et une solution est un diagnostic ou un traitement. Le RAPC est une approche "orientée données", contrairement aux approches "orientées connaissances" qui s'appuient sur les guides de bonnes pratiques. Il est particulièrement intéressant pour les patients compliqués pour lesquels les recommandations des guides ne peuvent pas s'appliquer. Le RAPC a notamment été utilisé conjointement à des interfaces d'accès aux images médicales (Müller et al., 2003), les images étant facile à visualiser. D'autres approches ont utilisés la visualisation des données patients, par exemple à l'aide de nuages de points ou de coordonnées parallèles (Falkman G, 2002).

En intelligence artificielle (IA), de nombreuses approches "orientée données", et notamment l'apprentissage profond, souffrent d'un manque d'explicabilité. Elles font de bonnes prédictions mais sont incapables de les expliquer, or les médecins ont besoin de comprendre les recommandations des systèmes d'aide à la décision pour y adhérer (Moxey et al., 2010). Le récent rapport Villani (Villani et al., 2018) recommande d'ailleurs d'"ouvrir la boîte noire de l'IA".

L'intelligence artificielle explicable (*Explainable Artificial Intelligence*, XAI) a pour but d'expliquer les prédictions des IA aux êtres humains (Schoenborn and Althoff, 2019). Le RAPC est intéressant dans ce contexte car les cas similaires peuvent servir d'exemples pour justifier la prédiction. Cependant, la consultation de 2 à 50 cas similaires peut être très laborieuse si la description des cas est complexe, ce qui est souvent le cas en santé. Une solution est l'utilisation de techniques de visualisation.

Nous proposons ici une approche visuelle pour le RAPC (Figure 1), avec un système capable de classer un cas requête de manière automatique (1 sur la figure), ou de manière visuelle (2). La composante visuelle peut être utilisée de manière isolée, ou bien afin d'expliquer le raisonnement automatique (3). La notion "d'explication" est complexe et recouvre différents objectifs; ici, nous cherchons principalement à permettre aux cliniciens de valider les recommandations du système, afin de s'assurer qu'elles correspondent bien à une réalité clinique et non à un « artefact » des données.

Cet article est un résumé en français d'un article paru en 2019 dans la revue *Artificial Intelligence in Medicine* (Lamy et al., 2019a). Ce travail a été réalisé dans le cadre du projet européen H2020 DESIREE (*Decision Support and Information Management System for Breast Cancer*), qui vise à développer une plateforme web pour l'aide à la décision thérapeutique dans le cancer du sein.

2 Description de l'interface visuelle

Nous considérons un système de RAPC avec une base de cas X comprenant plusieurs dimensions et un espace des solutions Y, ainsi qu'une requête q. Chaque dimension peut être réelle, entière, booléenne ou nominale. Nous nous limiterons à un ensemble fini de classes $Y = \{y_1, y_2, ...\}$ avec $2 \leq |Y| \leq 10$. Une couleur spécifique est associée à chaque classe dans Y. L'interface vise à traduire le problème "À quelle classe appartient q ?" en un problème visuel "Quelle est la couleur dominante ?". La Figure 2 montre l'organisation générale de cette interface. Elle comprend deux parties : à gauche un nuage de point montrant des similarités quantitatives sous forme de distances, à droite des boîtes arc-en-ciel montrant des similarités qualitatives : les caractéristiques communes partagées entre la requête et les cas similaires.

Le nuage de point est une projection en deux dimensions arbitraires de la matrice des distances entre les cas. q est représenté par le point blanc au centre de la mire, et les cas similaires par des points dont la couleur indique la classe. Ce nuage de point est généré par une méthode de MDS (*Multi-dimensional Scaling*), modifiée pour utiliser des coordonnées polaires. En effet, sur le nuage, il existe deux types de distance : les distances entre q et un cas similaire, et les distances entre deux cas similaires. Pour le RAPC, les distances impliquant q sont plus importantes. Par conséquent, nous proposons une méthode de MDS qui préserve intégralement ces distances, au détriment des autres. Cette méthode utilise les coordonnées polaires avec q pour origine. Chaque cas similaire est positionné aux coordonnées (l, θ), où l

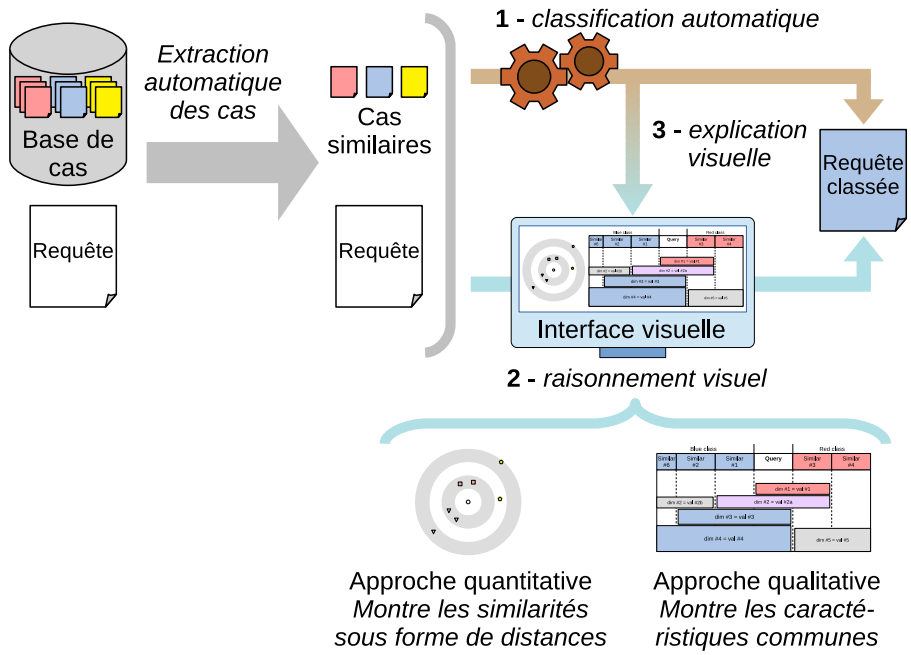

FIG. 1 – *Vue d'ensemble de l'approche à partir de cas proposée.*

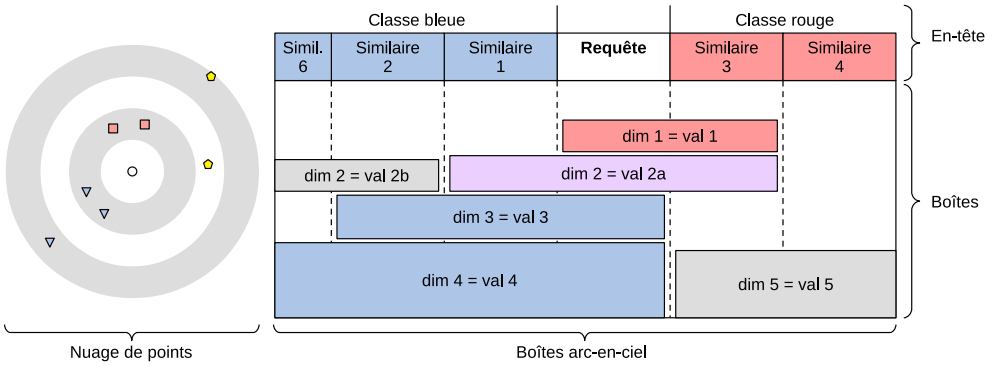

FIG. 2 – *Schéma de l'interface visuelle pour le raisonnement à partir de cas ("dim" : dimension, "val" : valeur).*

est la distance à l'origine et θ l'angle par rapport à l'horizon. Ici, l est déjà connu et figure dans la matrice des distances. Les valeurs de θ sont déterminés par un processus d'optimisation, de telle sorte à minimiser l'écart entre les distances réelles de la matrice et les distances en 2 dimensions sur le nuage de points.

Le nuage de point permet d'identifier le nombre de cas similaires appartenant à chaque classe, et la classe des cas les plus proches de q. Sur la Figure 2, il y a 3 cas similaires appartenant à la classe bleue, et 2 appartenant aux classes rouge et jaune. Les cas les plus proches appartiennent aux classes bleue et rouge. De plus, nous pouvons vérifier que les 3 couleurs semblent bien séparées les unes des autres, ce qui tend à montrer que chaque classe occupe une région bien définie de l'espace du nuage. Dans le cas contraire, le RAPC pourrait ne pas être approprié.

Les boîtes arc-en-ciel (Lamy et al., 2017) et leur version proportionnelle (Lamy and Tsopra, 2019) sont une technique récente de visualisation d'ensembles. Dans notre interface, les boîtes arc-en-ciel affichent une comparaison qualitative d'un sous-ensemble de cas comprenant la requête q et les cas similaires appartenant aux deux classes majoritaires (sur la Figure 2, il s'agit des classes bleue et rouge). Chaque cas est représenté par une colonne. La couleur de l'en-tête de colonne indique la classe du cas, en utilisant les mêmes couleurs que précédemment. De plus, la largeur de chaque colonne représentant un cas similaire est proportionnelle à sa similarité par rapport à q. Cela permet de donner une plus grande place aux cas les plus similaires. Par exemple, sur la Figure 2, l'une des colonnes bleues est moins large que les autres : elle correspond au cas similaire de la classe bleue qui est le plus éloigné de q sur le nuage de points.

Les boîtes colorées représentent des couples (dimension, valeur). Chaque boîte occupe l'ensemble des colonnes pour lesquelles la dimension prend la valeur donnée. Nous utilisons l'information mutuelle (IM) pour sélectionner les couples les plus pertinents par rapport aux classes Y, et la hauteur des boîtes est également proportionnelle à l'IM. Les boîtes sont de couleur grise lorsqu'elles ne contiennent pas q, et sinon leur couleur est la moyenne des couleurs associées aux deux classes, pondérée par le ratio de chaque classe dans la boîte. Par exemple, dans la Figure 2, la boîte "dim 2 = val 2a" est 50% bleu et 50% rouge, conduisant à du violet. Enfin, les boîtes sont empilées, en mettant les plus grandes en bas, et les colonnes sont ordonnées de sorte à mettre côte à côte les colonnes partageant les mêmes boîtes, à l'aide d'un algorithme méta-heuristique.

Les en-têtes de colonne permettent d'identifier rapidement les deux principales classes candidates pour q et le nombre de cas similaires pour chacune, pondéré par leur degré de similarité. Dans la figure 2, il y a 3 cas similaires appartenant à la classe bleue et seulement 2 à la classe rouge. Les boîtes colorées montrent les caractéristiques, c'est-à-dire des couples (dimension, valeur), partagées entre q et les cas similaires. La couleur de la boîte indique la classe vers laquelle la boîte oriente. Notons que la largeur de chaque colonne permet de pondérer chaque cas similaire en fonction de sa similarité. Ici, il y a deux grandes boîtes bleues, une petite boîte violette et une petite boîte rouge. Les boîtes bleues sont plus grandes et plus nombreuses : cela plaide en faveur de la classe bleue. L'utilisateur peut adapter ou atténuer son choix en tenant compte de la pertinence des dimensions et des valeurs indiquées dans les boîtes. Enfin, les zones grises peuvent suggérer des arguments pour ne pas choisir une classe, par exemple la grande boîte grise intitulée "dim 5 = val 5" est un argument pour ne pas choisir la classe rouge, car tous les cas appartenant à cette classe ont la valeur donnée dans

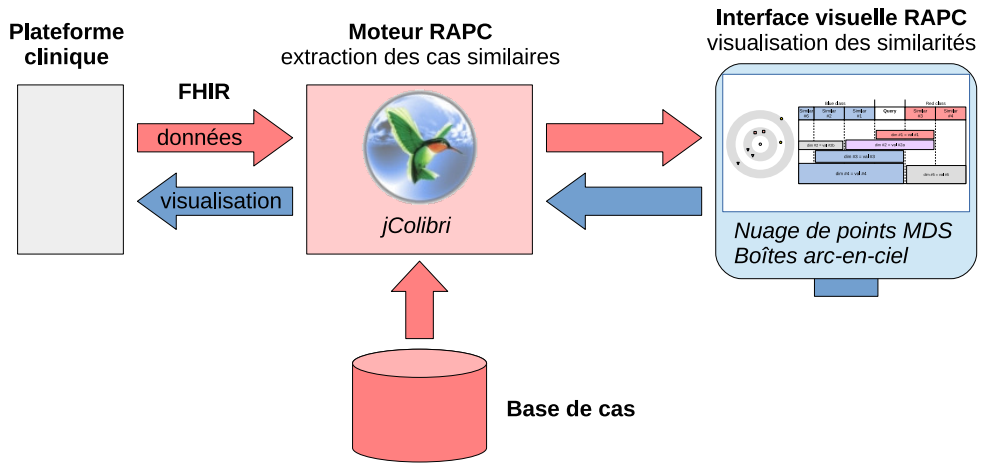

FIG. 3 – *Architecture générale du module RAPC du projet DESIREE.*

la dimension 5, et q ne l'a pas. Cependant, être différent selon une dimension n'implique pas nécessairement qu'il n'y a pas de similitude selon les autres.

L'interactivité est utilisée pour relier entre eux les différents éléments de l'interface : lorsque le curseur de la souris survole un point du nuage, une colonne ou une boîte, les éléments correspondant dans les autres parties de l'interface sont mis en évidence.

En conclusion, l'interface proposée permet au moins trois approches de raisonnement visuel : (1) le nuage de points montre la classe associée aux cas similaires les plus proches, ce qui permet d'effectuer visuellement un algorithme de type *kNN* (k plus proches voisins), (2) le nuage de points et les en-têtes de colonnes des boîtes arc-en-ciel indiquent le nombre de cas similaires et leur similarité, ce qui permet d'effectuer visuellement un algorithme de type *WkNN* (k plus proches voisins pondérés par leur similarité), (3) les boîtes indiquent en quoi les cas similaires sont similaires à la requête et vers quelle classe chaque valeur de caractéristique oriente, ce qui permet une approche qualitative basée sur des couples spécifiques (dimension, valeur). Cette approche qualitative peut également se traduire sous la forme d'un algorithme automatique, que nous avons appelé RBIA (*Rainbow Boxes-Inspired Algorithm*). Pour plus d'information sur la génération de l'interface et sa traduction en algorithme, se référer à l'article original (Lamy et al., 2019a).

3 Application au cancer du sein

Le cancer du sein est l'un des types de cancer les plus fréquents chez la femme en Europe. Il est associé à un taux de survie élevé à 10 ans. Cependant, avec le développement des nouveaux traitements, la gestion de la maladie et de la thérapie devient de plus en plus complexe. En particulier, les cliniciens des unités de soin multidisciplinaires doivent prendre plusieurs décisions délicates, et notamment établir un diagnostic et prescrire un traitement.

Nous avons testé les performances du système proposé sur 3 jeux de données publiques sur le diagnostic du cancer du sein : *Breast Cancer Wisconsin* (BCW), *Mammographic*

Mass (MM) et *Breast Cancer* (BC). Les résultats montrent que l'algorithme RBIA conduit à des résultats similaires aux algorithmes *kNN* et *WkNN* (97,8/77,3/80,8% de réponses exactes pour BCW/BC/MM, contre 97,5/76,6/82,0% pour (*W*)*kNN*), mais possède une explicabilité supérieure. En effet, la visualisation des couples (dimension, valeurs) partagés permet d'expliquer en quoi les cas similaires ressemblent au nouveau patient. Par conséquent, cela montre que les informations affichées dans les boîtes arc-en-ciel sont pertinentes et permettent une bonne classification, bien qu'il s'agisse d'un très petit sous-ensemble des informations disponibles pour les cas similaires.

Nous avons également utilisé un jeu de données simulé et un jeu de données réelles sur la thérapie du cancer du sein en phase initiale. Ce jeu de données réelles est issu de 4 hôpitaux français et espagnols impliqués dans le projet DESIREE. Les données ont été extraites du système de gestion d'information de DESIREE, de manière anonyme, pour 315 patients. En phase initiale, deux types de traitements sont possibles : la chirurgie ou la chimiothérapie (y compris les traitements endocriniens). Dans DESIREE, les données sont structurées à 3 niveaux : patient, côté (deux côtés par patient : gauche et droite) et lésion (zéro, une ou plusieurs lésions par côté), selon une ontologie de domaine au format OWL (Sadki et al., 2018). Les dimensions par côté et par lésion ont été agrégées en ne retenant que la plus mauvaise valeur disponible. Par exemple, si un patient a plusieurs lésions, la dimension "taille de la tumeur" correspond en fait à la taille de la lésion la plus grande. Les données manquantes ont été corrigées en les remplaçant par la valeur la plus fréquente (pour les dimensions catégorielles et booléennes) ou par la moyenne des valeurs observées (pour les dimensions entières et réelles). La dimension "nombre d'avortements" était une exception : les cliniciens ne remplissent ce champ que lorsque la valeur est non nulle. Nous avons donc considéré une valeur par défaut de 0. L'ensemble de données résultant comprend 51 dimensions (22 booléennes, 15 entières, 1 réelle et 13 nominales) et deux classes (avec 118 patients en chimiothérapie et 197 en chirurgie). La distance de similarité utilisée combine une distance Euclidienne classique (pour les paramètres numériques) et une distance sémantique (pour les paramètres qualitatifs), calculée à l'aide d'une ontologie du domaine (voir section 6.3 dans l'article originel).

La plate-forme clinique comporte plusieurs modules d'aide à la décision, ainsi des modules support (par exemple pour l'imagerie médicale). La figure 3 montre l'architecture générale du système proposé. Le module RAPC comprend trois composants : (a) la base de cas, (b) le moteur de RAPC, chargé de retrouver les cas similaires à la requête, et (c) l'interface visuelle présentée ici. Le moteur de RAPC a été implémenté en Java avec JColibri (Recio-García et al., 2014) et l'interface visuelle en Python. L'interopérabilité entre la plate-forme clinique et le module de RAPC est obtenue à l'aide du standard HL7 FHIR (*Fast Healthcare Interoperability Resources*). Cette norme permet l'échange de messages liés à la santé entre services médicaux, y compris les dossiers électroniques des patients. En termes de performances, la génération de l'interface visuelle prend moins d'une seconde.

La figure 4 montre une copie d'écran de l'interface sur le jeu de données réel. Ici, la longueur horizontale couverte par les en-têtes de colonne verte est plus longue que celle d'en-tête rouge, ce qui oriente vers le traitement vert, c'est-à-dire la chimiothérapie. Les boîtes arc-en-ciel fournissent des arguments en faveur de ce traitement. Le principal argument est une taille de tumeur élevée (dans la boîte la plus grande). Au contraire, la faible valeur du "PR result" est un argument qu'un clinicien pourrait prendre en compte pour décider de prescrire une chirurgie.

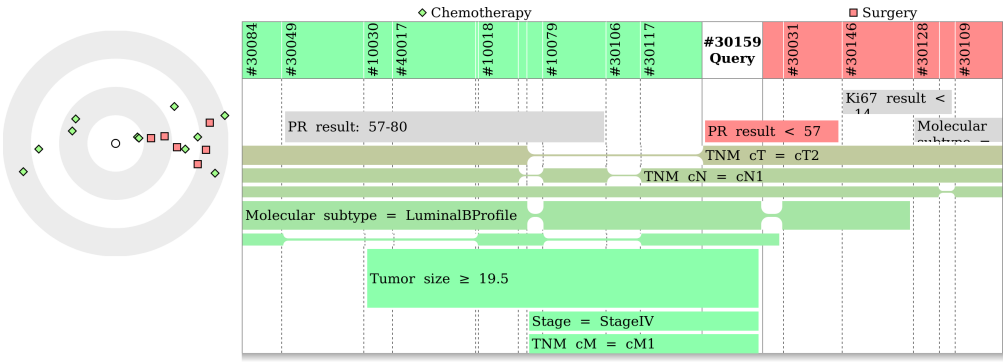

FIG. 4 – *Copie d'écran de l'interface visuelle, sur le jeu de données réelles sur la thérapie initiale du cancer du sein.*

Une étude utilisateur sur 11 médecins montre qu'ils ont été capables d'effectuer le RAPC de manière visuelle (81.8% de réponses correspondant aux réponses attendues), et que, avec cette approche, la confiance de l'utilisateur dans son choix reflète bien la difficulté du cas. L'analyse des commentaires montre que les médecins étaient enthousiastes vis-à-vis d'une approche visuelle. L'approche qualitative est intéressante car elle établit un lien entre les recommandations du système et les caractéristiques du patient. Certains médecins auraient cependant besoin d'une formation avant d'utiliser ce type d'outil.

4 Discussion et conclusion

Nous avons présenté ici une interface permettant d'expliquer visuellement le raisonnement à partir de cas, et nous l'avons appliquée au cancer du sein. Cette interface combine la visualisation de similarités quantitatives (distances) et qualitatives (valeurs partagées). L'interface visuelle permet d'expliquer le processus de raisonnement à l'utilisateur, qui peut l'enrichir en considérant ses connaissances personnelles, et sa traduction en un algorithme automatique permet une meilleure formalisation du processus de raisonnement visuel.

Les perspectives de ces travaux incluent l'application de l'interface à d'autres domaines, son adaptation à d'autres formes de raisonnement au-delà du RAPC, comme le *deep learning* ou le *boosting*, et son extension pour gérer un nombre de classes plus grand. En pratique médicale, le nombre de traitements possibles peut être important car chaque type de traitement se divise en plusieurs sous-types. Il existe ainsi différents types de chirurgies, différents protocoles de chimiothérapies, *etc.* Par la suite, nous avons proposé une version hiérarchique de notre interface (Lamy et al., 2019b), qui permet de diviser une décision complexe avec un grand nombre de classes possibles en plusieurs décisions avec un plus petit nombre de classes. Par exemple, une première interface permet le choix du type de traitement parmi les 4 possibles, puis une seconde interface sur le même principe permet le choix du type de chirurgie (lumpectomie ou mastectomie), puis une troisième permet de choisir le détail (avec ou sans biopsie des nœuds lymphatiques sentinelles axillaires). Les visualisations pourraient également être améliorées, par exemple en montrant les enveloppes des classes sur le nuage de points.

Remerciements

Ce travail a été financé par l'Union Européenne à travers le programme de recherche Horizon 2020, *via* le projet DESIREE H2020 PHC-30-2015, accord de subvention No 690238.

References

Aamodt, A. and E. Plaza (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications 7*(1), 39–59.

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

Falkman G (2002). The use of a uniform declarative model in 3D visualisation for case-based reasoning. In *European conference on case-based reasoning*, pp. 103–117.

Lamy, J. B., H. Berthelot, C. Capron, and M. Favre (2017). 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.

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

Lamy, J. B., B. Sekar, G. Guezennec, J. Bouaud, and B. Séroussi (2019b). Hierarchical visual case-based reasoning for supporting breast cancer therapy. In *International Conference on Advances in Biomedical Engineering (ICABME)*, Tripoli, Lebanon.

Lamy, J. B. and R. Tsopra (2019). RainBio: Proportional visualization of large sets in biology. *IEEE Transactions on Visualisation and Computer Graphics accepted*.

Moxey, A., J. Robertson, D. Newby, I. Hains, M. Williamson, and S. A. Pearson (2010). Computerized clinical decision support for prescribing: provision does not guarantee uptake. *J Am Med Inform Assoc 17*(1), 25–33.

Müller, H., A. Rosset, J. P. Vallée, and A. Geissbühler (2003). Integrating content-based visual access methods into a medical case database. *Stud Health Technol Inform*, 480–485.

Recio-García, J. A., P. A. González-Calero, and B. Díaz-Agudo (2014). jcolibri2: A framework for building Case-based reasoning systems. *Science of Computer Programming 79*, 126–145.

Sadki, F., J. Bouaud, G. Guézennec, and B. Séroussi (2018). Semantically Structured Web Form and Data Storage: A Generic Ontology-Driven Approach Applied to Breast Cancer. *Stud Health Technol Inform 255*, 205–209.

Schoenborn, J. M. and K. D. Althoff (2019). Recent Trends in XAI: A Broad Overview on current Approaches, Methodologies and Interactions. In *Case-Based Reasoning for the Explanation of intelligent systems (XCBR) Workshop*.

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

Résumé

Breast cancer is one of the most frequent cancers in women in Europe. Artificial intelligence can help physicians to make diagnostic and therapeutic decisions for breast cancer. However, most recent artificial intelligence methods (*e.g.* deep learning) are «black boxes» that cannot explain why the machine makes a given prediction. On the contrary, physicians need to understand the rationale of recommendations, in order to follow them. Here, we propose a visual case-based reasoning approach, allowing the visualization of both quantitative and qualitative similarities between cases. This approach was tested on 3 public datasets for diagnosis and one real datasets from the project for therapy, and presented to 11 physicians. This paper is a French summary of: Jean-Baptiste Lamy, Boomadevi Sekar, Gilles Guezennec, Jacques Bouaud, Brigitte Séroussi. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. Artificial Intelligence in Medicine 2019(94):42-53s.