Hierarchical visual case-based reasoning for supporting breast cancer therapy

Jean-Baptiste Lamy
LIMICS, Université Paris 13, Sorbonne
Université, INSERM, 93017 Bobigny, France
jean-baptiste.lamy@univ-paris13.fr

Boomadevi Sekar
School of Computing and Mathematics,
Ulster University, United Kingdom
b.sekar@ulster.ac.uk

Gilles Guezennec
LIMICS, Université Paris 13, Sorbonne
Université, INSERM, 93017 Bobigny, France
gilles.guezennec@univ-paris13.fr

Jacques Bouaud
LIMICS, France
AP-HP, DRCI, Paris, France
jacques.bouaud@aphp.fr

Brigitte Séroussi
LIMICS, France
AP-HP, Hôpital Tenon, Département de
Santé Publique, Paris, France
brigitte.seroussi@tnn.aphp.fr

Abstract—Breast cancer therapy is particularly complex. Case-based reasoning (CBR) is an approach that can support clinicians when prescribing a therapy, and that is able to explain its recommendation to the clinicians. In a previous work, we proposed a visual CBR approach for helping clinicians to choose a treatment between four main categories (*e.g.* surgery, chemotherapy). However, these are broad categories and clinicians need more details about the treatment, *e.g.* several surgeries exist such as lumpectomy. Here, we extend our visual CBR approach for fully supporting the therapy for breast cancer, using a hierarchical approach: first, decide the category, then decide the exact treatment, *etc*.

Index Terms—breast cancer, case-based reasoning, explainable artificial intelligence, XAI.

I. INTRODUCTION

Breast cancer is one of the most common types of cancer that affects women in Europe. It is associated with a high survival rate at 10 years. However, the therapy of the disease is incredibly complex. Many treatments exist, and they can be grouped in four main categories: surgery, chemotherapy, endocrine therapy and radiotherapy. Moreover, a wide range of heterogeneous patient data has to be taken into account by clinicians in the multidisciplinary Breast Units (BUs) when prescribing the treatment, including basic patient characteristics (age, sex), clinical characteristics (either global, per breast or per tumor, since a given patient may have several tumors, in the same breast or in the other one), imaging, genetic markers, previous treatments,... [1].

Machine learning and deep learning recently achieved many successes in diagnostic support [2]. Similar technologies have been applied to therapy, *e.g.* IBM Watson was applied to breast cancer therapy [3]. However, these approaches are "black box" that suffer from a huge lack of explainability: while they may predict the best treatment, they are unable to explain to a human user why their prediction is the best treatment. This is particularly true for neural networks. For medical imaging analysis, an annotated

image, *e.g.* with the contours of the detected anomalies, can make a decent explanation and is usually easy to produce with deep learning. On the contrary, therapy is more abstract by nature, and does not provide images to annotate. However, clinicians need to understand the recommendation of a decision support system. Recently, a review concluded that machine learning is not yet ready for therapy [4]. On the contrary, the recent field of explainable artificial intelligence (XAI) aims at designing intelligent methods that can be explained to a human [5].

The DESIREE European project (Decision Support and Information Management System for Breast Cancer) aims at supporting primary breast cancer therapeutical decision and helping clinicians with the management of the patient data and images. During the project, we develop a web-based platform including several decision-support modules, each of them using a different approach: clinical practice guidelines implementation [6], machine learning, and case-based reasoning (CBR) [7]; the current paper will focus on the latter approach.

CBR solves a new case (*e.g.* a new patient) by retrieving similar older cases with known solutions (*e.g.* treatments), and adapting their solutions to the new case [8]. The new case can then be retained with its solution for future reuse. Contrary to machine learning approaches, CBR does not try to learn a model (*e.g.* a rule set or a neural network) from the case base. A well-known example is the k-nearest neighbors algorithm (kNN). CBR can explain its recommendations by using the older cases as examples.

In a recent journal paper [9], we proposed a visual, fully explainable, CBR approach. It consists of retrieving similar patients and then visualizing their treatments and their similarities with the new patient. Our approach can support the classification of a patient in 2-6 classes, corresponding to treatments (*e.g.* surgery, chemotherapy, *etc*, corresponding to Figure 1-1). However, these are broad treatment categories, *e.g.* several types of surgery can be considered: lumpectomy, mastectomy, quadrantectomy. Then, each surgery can have several modalities, *e.g.* with or without axillary lymph node

dissection. While determining the broad category may help clinicians, they also need support for choosing the exact treatment and its modalities.

In this paper, our objective is to extend the visual CBR approach to fully support the therapy for breast cancer, using a hierarchical approach: first, decide the main category of treatment (as previously done), then decide the exact treatment and finally its modalities. Each step will be achieved using the visual CBR interface.

II. METHODS

A. Visual CBR approach

Our visual CBR approach includes two parts (Figure 1-1). On the right, a scatter plot displays quantitative similarities, *i.e.* distances between patients. The central, white, dot represents the new patient, and the other, colored, dots represent the similar patients. The color of these dots indicates the type of treatment: red for surgery, green for chemotherapy, blue for radiotherapy and yellow for endocrine therapy. The smaller the distance between two dots is, the more similar the two patients are. In Figure 1-1, we can see that the majority of similar patients were treated by surgery, but the most similar ones were treated by chemotherapy, and no similar patients were treated by radiotherapy or endocrine therapy. The scatter plot is a 2D projection of a multidimensional space, and thus some information is lost in the process. We draw it using MDS (Multi-Dimensional Scaling) [10] in polar coordinates; it allows preserving the distance involving the central dot (*e.g.* the new patient), to the detriment of the other distances.

On the left, rainbow boxes display qualitative similarities. Rainbow boxes are a technique we recently introduced for set visualization [11]. This part of the visualization retains only the two major treatment recommendations (here, surgery and chemotherapy). Patients are shown in columns. The central, white, column is for the new patient, and the similar patients are placed on its left or its right, depending on the treatment they received. The same color code as on the scatter plot is used. Column width is proportional to the similarity of an old patient with the new patient (larger columns for more similar patients). Each colored box represents an attribute-value pair shared between a group of patients, for example "tumor size at mammography ≥ 35.5". The box covers the columns corresponding to the patients having this attribute-value pair, *e.g.* here patient #0 (new patient), 2964, 2724, 2789 and 2966 have a tumor size at mammography superior (or equal) to 35.5. A box may have "holes" if the patients it covers are non contiguous; patients are ordered so as to limit the number of holes with a metaheuristic. The color of the box is the weighted mean of the color of the columns it covers. It thus indicates towards which recommendation it orientates: here, the green color suggests that having a high tumor size at mammography orientates towards chemotherapy. On the contrary, boxes that do not include the new patient are gray and represent arguments for not choosing a given recommendation. Finally, the height of the box is proportional to Mutual Information

(MI), a statistical criterion indicating how the box attribute is correlated with the treatment choice.

The interface is interactive: popup labels give more information and display the entire box labels for the smallest boxes in which labels do not fit, and when the mouse is over a patient in the scatter plot or the rainbow boxes, the same patient is highlighted in the other part of the visualization. This interface allows performing visually, at a glance, several reasoning approaches: (1) Using the scatter plot, one can perform visually a kNN or weighted kNN. Here, surgery wins with kNN, but chemotherapy may win with weighted kNN, because similar patients with chemotherapy are closer to the new patient. (2) Using rainbow boxes, one can determine the dominant color in boxes. Here, the green or greenish colors of the boxes suggest that the new patient share more qualitative similarities with old patients treated by chemotherapy.
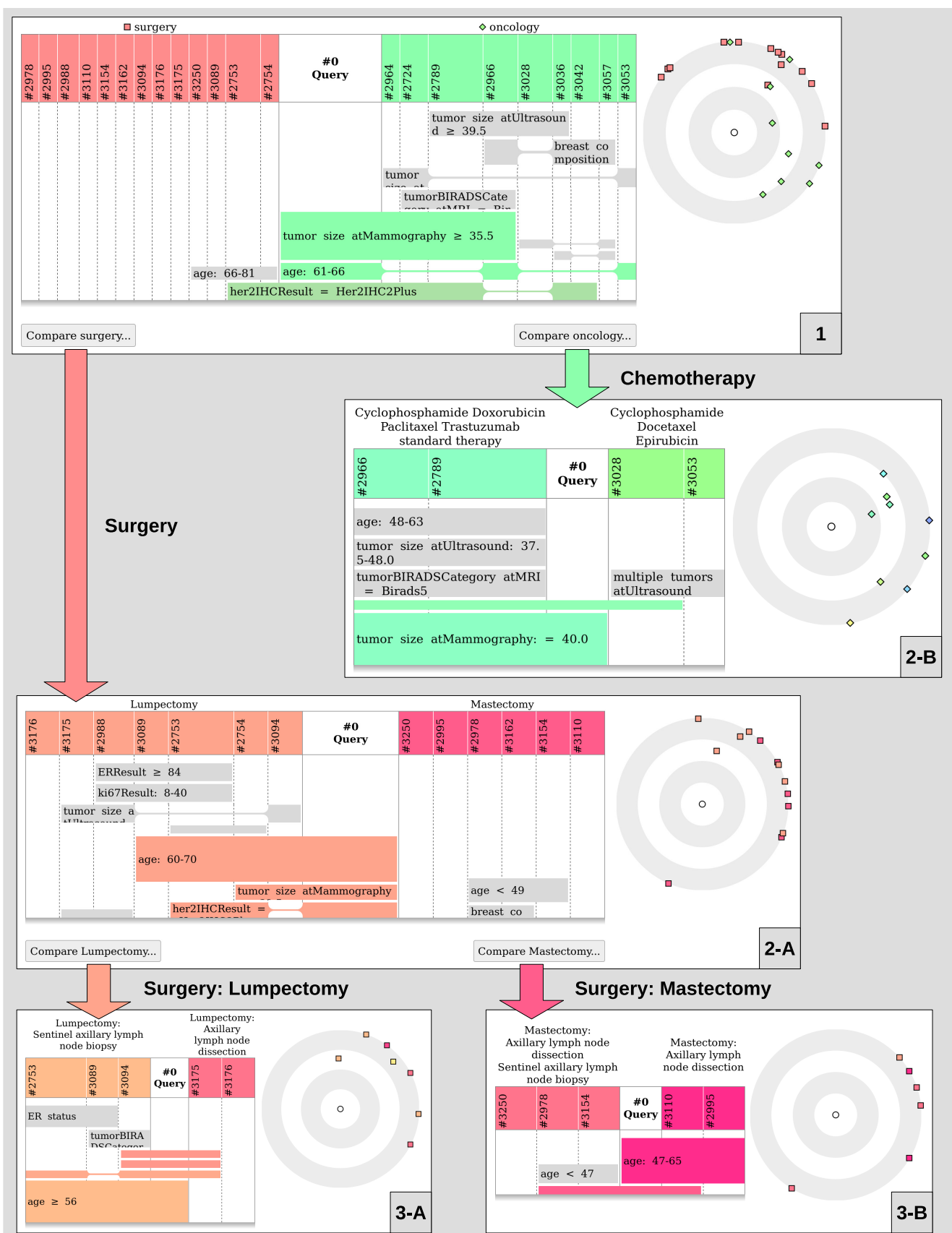
Finally, the clinicians are free to follow the visual recommendations of the system, or to choose another treatment, *e.g.* in figure 1-1, a clinician might consider that tumor size at mammography is not a good criterion for determining the treatment, and opt for surgery.

From a more technical point of view, the system relies on the HL7 (Health Level 7) FHIR (Fast Healthcare Interoperability Resources) protocol for accessing patient data. Similar older cases are stored and queried from a relational database. The visual interface is generated by Python scripts and use web standards (HTML, CSS, Javascript). All dependency files (images, stylesheet, Javascript sources) were directly embedded in the HTML page, to facilitate access in secured environment (some hospitals do not allow Internet access). For more details on this visual approach, please refer to [9].

B. Hierarchical approach

The hierarchical approach we propose here consists of dividing the choice of treatment in several intermediary choices: the main category of treatment, the exact treatment and its modalities (*e.g.* drug dosage). The CBR case base can be modeled as a $(p + 1)$-dimensional database $X = \{x_1, x_2, ..., x_i, ...\}$ with $x_i \in \mathbb{A}_1 \times ... \times \mathbb{A}_k \times ... \times \mathbb{A}_p \times Y$ where $\mathbb{A}_k$ are the attribute dimension spaces and Y is the solution space (*i.e.* the treatment), and the new case as $q \in \mathbb{A}_1 \times ... \times \mathbb{A}_k \times ... \times \mathbb{A}_p$. For the purpose of the hierarchical approach, we consider that the possible treatments in Y are organized hierarchically in a formal ontology (Figure 2), with general categories (*e.g.* Breast Surgical Procedure) and more specific treatments (*e.g.* Lumpectomy).

The first choice will consider broader categories, *i.e.* the solution space is $Y_1 = \{$*Surgery*, *Chemotherapy*, *Radiotherapy*, *EndocrineTherapy*$\}$, and the similar cases X_1 are retrieved using common CBR technics such as the Euclidean distance (we implemented it with JColibri). The two best categories are compared using rainbow boxes, and two buttons, one for each category, allows the user to choose a category. When the user clicks one of the buttons, the CBR process is repeated, producing a new visualization using the same interface. The next choice will consider the subclasses

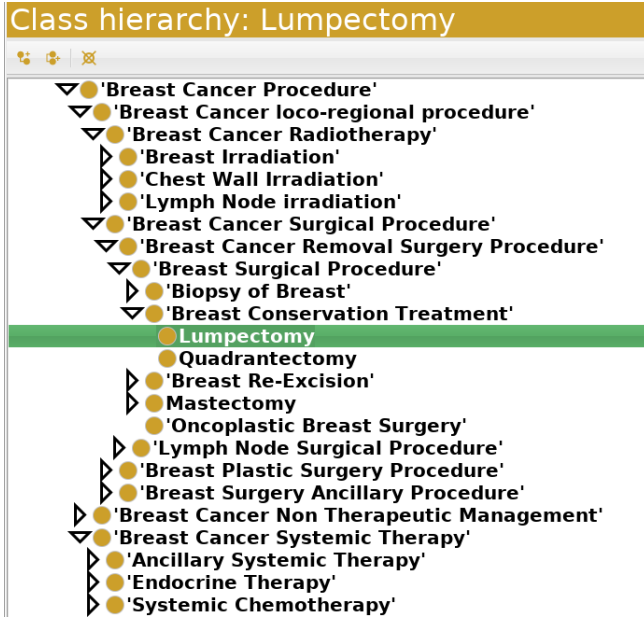Figure 1. Example of hierarchical visual case-based reasoning.

Figure 2. Excerpt of the DESIREE ontology showing the treatment classes.

of the chosen solution $y_1 \in Y_1$ and the subset of similar cases that are associated with the solution y_1, and so on, until the treatment is fully determined. Consequently, for the $i+1^{\text{th}}$ step, the solution space is $Y_{i+1} = \{y' : y' \sqsubseteq y_i \in Y_i\}$, where $\sqsubseteq$ represents the direct subsumption, and the similar cases are $X_{i+1} = \{x \in X_i : x_{p+1} \sqsubseteq y_i\}$, where x_{p+1} is the solution associated with case x.

III. RESULTS AND DISCUSSION

Figure 1 shows 5 screenshots of the hierarchical CBR visual interface. Figure 1-1 shows the initial screen, comparing the broad categories of treatments. Rainbow boxes retain the two major categories. Two buttons at the bottom are proposed to choose each category. When one of the buttons is clicked, the interface is modified to display only the similar cases associated with the chosen category. Colors are automatically created for treatment, which are close to the color associated with the treatment's category (*e.g.* reddish colors for surgeries). Figure 1-2A and 2B show the interface after choosing surgery and chemotherapy, respectively. This process can be applied recursively, *e.g.* for surgery, one can select a precise surgery (*e.g.* lumpectomy) and compare the various modalities.

The main difficulty raised by the hierarchical approach is that the number of similar cases is reduced at each step: in Figure 1, there are 22 similar cases at step 1, 13 at step 2-A, 9 at step 2-B (of which only 4 are displayed in rainbow boxes), and 5 at step 3-A and 3-B. Consequently, in order to have still enough similar cases in the last steps, we need to increase the number of similar cases retrieved at the initial step. However, this may impact the performance of the system: in most CBR systems, there is an optimal number of similar cases to retrieve. Another possibility would be to retrieve new cases at each step, *e.g.* to retrieve the 10 most similar patients at the step 1, and the 10 most similar

patients with surgery at step 2 (possibly including patients not retrieved at step 1).

In the literature, visual approaches for CBR mostly focused on 2D or 3D scatter plots [12], and thus were limited to quantitative similarity: these approaches display the amount of similarity between the new cases and the old ones, but not on which attributes the similarity holds. Another option proposed is parallel coordinates [13].

In conclusion, the proposed hierarchical visual CBR approach can support complex decision-making, such as breast cancer therapy, by dividing the decision process into several, simpler, decisions. The perspective of this work is the verification of the system, possibly using a systematic approach in combination with machine learning for extracting rules and comparing them with guidelines, as we did previously on another decision support system [14], its clinical validation and its application in other domains.

REFERENCES

[1] D. Gu, C. Liang, and H. Zhao, "A case-based reasoning system based on weighted heterogeneous value distance metric for breast cancer diagnosis," *Artif Intell Med*, vol. 77, pp. 31–47, 2017.

[2] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts, "Artificial intelligence in radiology," *Nat Rev Cancer*, vol. 18, no. 8, pp. 500–510, 2018.

[3] S. P. Somashekhar, M. J. Sepúlveda, S. Puglielli, A. D. Norden, E. H. Shortliffe, C. Rohit Kumar, A. Rauthan, N. Arun Kumar, P. Patil, K. Rhee, and Y. Ramya, "Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board," *Ann Oncol*, vol. 29, no. 2, pp. 418–423, 2018.

[4] G. Fang, I. E. Annis, J. Elston-Lafata, and S. Cykert, "Applying machine learning to predict real-world individual treatment effects: Insights from a virtual patient cohort," *J Am Med Inform Assoc*, 2019.

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

[6] B. Séroussi, G. Guézennec, J. B. Lamy, N. Muro, N. Larburu, B. D. Sekar, C. Prebet, and J. Bouaud, "Reconciliation of multiple guidelines for decision support: a case study on the multidisciplinary management of breast cancer within the DESIREE project," in *Proc AMIA Annual Symposium*, 2017.

[7] B. Sekar, J. B. Lamy, N. Larburu, B. Seroussi, G. Guézennec, J. Bouaud, N. Muro, H. Wang, and J. Liu, "Case-based decision support system for breast cancer management," *International Journal of Computational Intelligence Systems*, vol. 12, no. 1, pp. 28–38, 2018.

[8] N. Choudhury and S. A. Begum, "A survey on case-based reasoning in medicine," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 7, no. 8, pp. 136–144, 2016.

[9] 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*, vol. 94, pp. 42–53, 2019.

[10] I. Borg, P. J. F. Groenen, and P. Mair, *Applied multidimensional scaling*. Springer, 2013.

[11] 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*, vol. 43, pp. 71–82, 2017.

[12] F. Klawonn, W. Lechner, and L. Grigull, "Case-centred multidimensional scaling for classification visualisation in medical diagnosis," in *International Conference on Health Information Science*, 2013, pp. 137–148.

[13] S. Massie, S. Craw, and N. Wiratunga, "Visualisation of case-base reasoning for explanation," in *Proceedings of the ECCBR 2004 Workshops*, Madrid, Spain, 2004, pp. 135–144.

[14] J. B. Lamy, A. Ellini, V. Ebrahiminia, J. D. Zucker, H. Falcoff, and A. Venot, "Use of the C4.5 machine learning algorithm to test a clinical guideline-based decision support system," *Stud Health Technol Inform*, vol. 136, pp. 223–228, 2008.