Computers in Biology and Medicine 170 (2024) 108042

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Evidence-based XAI: An empirical approach to design more effective and explainable decision support systems

Lorenzo Famiglini [a,1,*], Andrea Campagner [b,1], Marilia Barandas [d], Giovanni Andrea La Maida [c], Enrico Gallazzi [c], Federico Cabitza [a,b,1]

[a] Department of Computer Science, Systems and Communication, University of Milano-Bicocca, Milan, Italy
[b] IRCCS Istituto Ortopedico Galeazzi, Milan, Italy
[c] Istituto Ortopedico Gaetano Pini | ASST Pini-CTO, Milan, Italy
[d] Associação Fraunhofer Portugal Research, Rua Alfredo Allen 455/461, Porto, Portugal

ARTICLE INFO

Keywords:
Explainable AI
Evidence-based design
Human–AI collaboration
Activation maps
Medical imaging
Human-centered AI

ABSTRACT

This paper proposes a user study aimed at evaluating the impact of Class Activation Maps (CAMs) as an eXplainable AI (XAI) method in a radiological diagnostic task, the detection of thoracolumbar (TL) fractures from vertebral X-rays. In particular, we focus on two oft-neglected features of CAMs, that is granularity and coloring, in terms of what features, lower-level vs higher-level, should the maps highlight and adopting which coloring scheme, to bring better impact to the decision-making process, both in terms of diagnostic accuracy (that is effectiveness) and of user-centered dimensions, such as perceived confidence and utility (that is satisfaction), depending on case complexity, AI accuracy, and user expertise. Our findings show that lower-level features CAMs, which highlight more focused anatomical landmarks, are associated with higher diagnostic accuracy than higher-level features CAMs, particularly among experienced physicians. Moreover, despite the intuitive appeal of semantic CAMs, traditionally colored CAMs consistently yielded higher diagnostic accuracy across all groups. Our results challenge some prevalent assumptions in the XAI field and emphasize the importance of adopting an evidence-based and human-centered approach to design and evaluate AI- and XAI-assisted diagnostic tools. To this aim, the paper also proposes a hierarchy of evidence framework to help designers and practitioners choose the XAI solutions that optimize performance and satisfaction on the basis of the strongest evidence available or to focus on the gaps in the literature that need to be filled to move from opinionated and eminence-based research to one more based on empirical evidence and end-user work and preferences.

1. Introduction

In the age of AI-assisted decision-making, it has become imperative to design systems that not only excel in performance but also consider the human-in-the-loop aspect [1,2] and have been developed with a human-centered approach [3]. This approach requires acknowledging how such systems may impact the community of practice into which they are deployed [4] in terms of either individual or collective performance [3]. The increasing interest in the development of eXplainable AI (XAI) solutions reflects the requirement to make AI systems capable of helping users understand their output and why they produced it [5]. Yet, despite the proliferation of XAI methods and techniques, the development of the XAI discipline has so far been mostly technologically driven, with less attention devoted to user-centric design practices and evaluation techniques that employ user studies [3,6,7].

In this paper, we will first present a simple framework for the interpretation of the strength of the evidence backing a specific *design choice*, or feature or method, in regard to XAI support, from the lowest level (i.e., the informed opinion of some expert or panel) to the highest one (i.e., the statistical analysis of the data from independent studies focused on the evaluation of the same method), inspired by the composite research regarding *evidence-based design* in healthcare [8] and health informatics [9]. Then, we will present a user study that we undertook to compare different design choices, in regard to how to generate activation maps and how to color them, and to collect strong evidence about what solution is better (if any) to inform the design of explainable AI aids, especially in settings that apply such systems in the same diagnostic task: in our case, the detection of vertebral fractures from X-rays.

* Corresponding author.
E-mail address: l.famiglini@campus.unimib.it (L. Famiglini).
[1] Lorenzo Famiglini, Andrea Campagner, and Federico Cabitza contributed equally to this work.

https://doi.org/10.1016/j.compbiomed.2024.108042
Received 24 August 2023; Received in revised form 19 December 2023; Accepted 26 January 2024
Available online 1 February 2024

Table 1
Hierarchy of evidence for AI and XAI empirical studies.

Level 1 (strongest)	Meta-analyses and systematic reviews of randomized controlled trials or experimental studies involving real practitioners
Level 2	Single experimental study (randomized, controlled) with prospective real-world cases considered by real practitioners in real-world settings.
Level 3	Single quasi-experimental study (e.g., nonrandomized, with concurrent or historical controls) involving prospective real-world cases considered by real practitioners in real-world settings.
Level 4	Single experimental study (randomized, controlled) with retrospective real-world cases considered by real practitioners in simulated/laboratory settings.
Level 5	Single quasi-experimental study (e.g., nonrandomized, with concurrent or historical controls) involving retrospective real-world (or simulated) cases considered by real practitioners in simulated/laboratory settings
Level 6	Single quasi-experimental study (e.g., nonrandomized, with concurrent or historical controls) involving simulated cases considered by human participants but not real practitioners in laboratory settings
Level 7	Supervised machine learning train/test studies with external validation (multiple datasets in longitudinal or cross-section/multi-site settings)
Level 8	Supervised machine learning train/test studies with internal validation
Level 9	Consensus opinions of authoritative bodies (e.g., nationally recognized guideline groups with robust peer review processes, notified bodies, standardization organizations)
Level 10 (weakest)	Opinions of recognized experts and case studies

In what follows, we will introduce our main contribution, that is the concept of evidence-based design for the XAI field, and a framework to evaluate XAI solutions along a hierarchy of evidence that we will then apply to interpret the findings of the user study that we will describe in the following sections. Thus, while the user study is a contribution that adds to the growing literature regarding the multiplicity of methods that have been developed and validated regarding activation maps in AI-supported radiology, it can also be seen as an exemplification of how researchers can apply our XAI evaluation framework.

2. Evidence-based XAI and AI design

The hierarchy of evidence (see Table 1) is a design-oriented construct to interpret and rank the relative strength of claims about the effectiveness of design solutions and their superiority with respect to other possible alternatives. Most of the times, presenting an evidence means to give *proof* that a specific design solution works; or, more specifically, that a system implementing a specific solution works better than other systems that do not feature it: a similar definition, although on the more abstract level of design principle, is given by [10]. Most of the times, this is achieved by showing the apparent relationship between adopting a solution (which often instantiates a specific design principle), and some desired effect or positive outcome, such as increased decision accuracy, efficiency or satisfaction.

With design solution or design choice we mean any alternative that is available to the designer of an AI system and that requires their informed choice. As mentioned above, this choice among multiple options can be posed at a high level of design principle or approach, or at a more implementational level, that is, at the level of the individual feature or solution to be integrated into the system, or technique and method to be implemented. Examples of XAI solutions abound, depending on the context [5,11]. Obviously, the choice of which solutions are better, and hence not a waste of time and money (or, worse, potentially harmful) if implemented in some XAI setting, should be grounded on the strongest evidence available; according to the hierarchy of evidence (see Table 1) these come from empirically-grounded studies and well-designed user studies. The resulting evidence of such studies should have some desirable characteristics, reported by Wyatt in [10], such as being specific (i.e., testable), actionable (i.e., converted in implemented features), generic (widely applicable) and, to some extent, novel (i.e., still untested).

The idea of grounding design choices on good quality evidence, and therefore actively and programmatically pursuing this latter, comes from medicine and healthcare. In this broad domain, the need to base practices, interventions, and decisions on empirical research has long been recognized as paramount [12]. To this aim, the responsible practitioner must identify the most effective and appropriate interventions, treatments, and solutions rather than relying on traditional practice or the personal opinions of practitioners, i.e., the so-called *eminence-based* practice [13]. The principles initially advocated by [14] for medicine, in one of the seminal works where the expression *Evidence-Based Medicine* (EBM) was adopted, were gradually extended to a wider range of disciplines and services. Jeremy Hamilton, inspired by the core tenets of EBM, adapted and proposed these principles for application to all fields of design [15], and coined the expression *Evidence-Based Design* (EBD) as an approach that utilizes empirical evidence derived from the systematic collection and analysis of empirical data, research findings, and best practices, to inform the design and development of systems, products, or environments. This evidence is typically derived from rigorous scientific research, which may include quantitative, qualitative, or mixed-methods studies, and which aims to provide a strong foundation for making well-informed decisions throughout the design process. The evidence-based design approach has been particularly influential in fields such as planning, management, and architecture, but its potential for application in computer science is just as promising.

Indeed, in [10], Wyatt introduced the concept of evidence-based health informatics to advocate "that the people designing, developing and implementing health information systems [...] rely on an explicit evidence base derived from rigorous studies on what makes systems clinically acceptable, safe and effective – not on basic science or experts alone". In another piece [16], the same author also proposes an evidence-strength hierarchy to classify the soundness of results regarding the effectiveness of a "generalizable system design principle", and therefore answer research questions like "Will systems based on this generic design principle work better than other systems?". In our contribution we ask the same question about how AI systems can present their output and how they can make it more understandable by providing additional pieces of information that would adequately complement its predictions and advice, that is an XAI function. By incorporating empirical evidence from user research into design, both Hamilton and Wyatt emphasized the potential to create more effective, efficient, and user-centered solutions that reuse effective design principles, and leverage previous user studies to adopt the related best human–AI interaction protocols, to impact human satisfaction and performance positively. As the principles of EBM and EBD continue to gain traction in medicine and architecture, respectively, their potential to shape other interventional disciplines (including IT design and AI development) that can benefit from data-driven decision-making and the rigorous application of research evidence, becomes increasingly clear. This is especially true where those disciplines touch: for instance,

Jin et al. have recently proposed an evidence-based design framework with respect to medical image analysis [17]. Their framework chooses an explanation form based on evaluations concerning understandability and clinical relevance, and the explanation method is based on evaluations of truthfulness and informative plausibility. Also Cabitza et al. advocate for more empirical comparative research where the primary object of study is not so much any vague design principle or the single XAI technique, but rather *the way the whole thing is put together* to allow the human to be supported by the technological system, that is to have human and AI team up, the so-called human–AI collaboration protocol (HAICP) [18].

It is in light of all the above contributions, both those established in the medical and design science literature and the more pioneering ones at the interface with medical AI, that we have produced the hierarchy of evidence in Table 1. As such, this resource is proposed as a heuristic for the designer to make informed and sound design choices, and turn its efforts to *to-implement solutions* for which there is some experimental evidence that they work and that they have a significant and positive effect on the processes they want to improve. But it is also a resource for the researcher, to understand which gaps in the literature exist to fill in or to prove that certain particularly popular solutions actually work. Obviously, each ranking exercise expresses a set of assumptions and preferences of the proposer, and thus has elements of arbitrariness that are reflected in what is considered to be of higher level, that is, to have greater persuasive power. In this sense, our hierarchy is no exception. The relevant dimensions that motivate its ranking are: whether the evidence is empirically-grounded or not (levels 2–8 vs 9 and 10); whether it involves users or not (2–6 vs 7–8); whether these users are real practitioners or not, that is lay human participants (2–5 vs 6); if the study is experimental or quasi-experimental (levels 2 and 4 vs 3, 5, and 6); whether the study is based on prospective or retrospective cases (2 and 3 vs 4 and 5); whether these latter cases are real-world or simulated ones (2–5 vs 5 and 6); whether the study is performed in real-world settings or laboratory ones (2 and 3 vs. 4–6). The relevance of each criterion was assigned having in mind the highest level, which is associated with what derives from a statistical analysis of multiple independent studies in which attempts were made to reduce the effect of confounding factors and increase generalizability, and the lowest level, the simple opinion of those who are considered experts in the relevant community. Therefore, experimental design, the nature of the cases, and the involvement of end users, possibly in their work environment, are the most important criteria in our hierarchy to justify why one outcome should be considered stronger than others.

The careful reader will have also noticed how most of the literature on machine learning in medical (and other) settings is still at level 8, and how few studies still only present external validation of performance estimates of a classifier model (level 7), that is implement a validation procedure with data coming from other facilities than those involved in the training of the model. The hierarchy of evidence that we propose therefore also represents a warning about the fragility (and therefore low usefulness) of certain results and a plea for researchers not to settle for easy results in laboratory settings, but to invest in more user studies. These latter, although more complex, as they require more resources (including the time and availability of real practitioners) and are more expensive and challenging, is what our society really needs now.

3. The user study

In the rest of the paper, we will describe a comparative user study in the mold of others we already undertook to get stronger evidence that a particular type of Pixel Attribution Maps (PAM) (that is heatmaps that highlight the areas in the input image that are crucial to the model's decision-making process [19]), called Class Activation Maps (CAMs) [20], are effective XAI techniques in diagnostic image-based tasks [18,21,22]. While there is still a lively debate about what are the best conditions under which explanations genuinely enhance diagnostic accuracy [1,18,23,24], the application of PAMs, and in particular CAMs, in medical imaging has shown promising outcomes in various studies [25,26]. Our aim is to add to this body of evidence some stronger findings (according to our hierarchy of evidence) on how CAMs should be generated and rendered. To this aim, we focused on the design of CAMs to detect traumatic thoracolumbar (TL) fractures from X-ray images, which is still a difficult task where errors can have a strong impact on patient outcome [21].

In particular, we will not focus on the question of which algorithm produces the best CAMs (e.g., Grad-CAM [27], Grad-CAM++ [28]., Score-CAM [29], PRM [30], or LayerCAM [31]), as this issue is subject to a serious risk of obsolescence in the short term. Rather, we focused on an issue that is usually taken for granted, namely from which layer of the multilayer neural network (i.e., higher-level layers versus lower-level layers) CAMs should be generated to make them more useful to human interpretation, and with which color scheme they should be rendered, to make them more intuitive to the decision makers to whom they are proposed as an XAI method.

In both cases, designers still make their choice (assuming they pose the question at all), according to very low-level considerations, that is what most researchers do or say, or the insight of domain experts. For instance, the idea that the traditional color schema adopted in CAMs, that is the bi-hue (i.e., the blue-red or the green-red) schema, irrespective of the predicted class, might not be the best one, precisely because it does not allow us to distinguish at first glance which class is predicted by the classifier, was suggested to us by a physician involved in the user study reported in [21]. Similarly, CAMs are generally generated from higher-level layers as these latter have conventionally been considered to better retain spatial information and high-level semantics [27,32], even though no evidence, to our knowledge, has been presented in the body of research of XAI to support this statement. For this reason, the user study that we will describe in the next sections will address the following two research questions:

- *R1*. Does utilizing higher-level or lower-level information improve the effectiveness of the related CAMs for detecting TL fractures from X-ray images? In other words, does the layer choice within the DL model employed for CAMs generation have any effect on diagnostic accuracy and decision-making (e.g., in terms of confidence and satisfaction)?
- *R2*. Does the coloring scheme adopted to display CAMs have any effect on diagnostic accuracy and decision-making?

4. Methods

Our study revolves around utilizing XAI principles to ascertain optimal conditions for augmenting diagnostic accuracy in the realm of vertebral X-ray fracture detection. In the following sections, we will describe our methodology, which consists of multiple stages, primarily data preparation, response collection (including model training and map generation), response augmentation, and response analysis.

4.1. Data preparation

The dataset included 630 vertebral cropped X-rays, from 151 trauma patients, collected at the Spine Surgery Centre of the Niguarda General Hospital of Milan (Italy) between 2010 and 2020. The images comprised both positive cases, where fractures were present (302 images), and negative cases without detectable fractures (328 images). Gold Standard CT and MRI images were used by three experienced spine surgeons to annotate these X-rays. A specialized image cropping software was employed to extract multiple cropped images, each showcasing various vertebrae, from each X-ray, as a form of data augmentation. The data was partitioned into training (80%), validation (10%), and testing (10%) sets. Data augmentation techniques were employed on

the training set, which included random horizontal and vertical flips, rotations, and image resizing and normalization.

Two medical specialists curated 18 images, from the test set, that best represented both positive and negative cases, ensuring clarity and a balance of cases with varying diagnostic complexities. These images were used in the creation of the user questionnaire (see Sections 4.2.2 and 4.2.3).

4.2. Response collection

4.2.1. Model training

As also reported in a previous study that used the same dataset [21], in this work, we implemented a transfer-learning approach by utilizing a ResNeXt-50 model that was pre-trained on images from Imagenet. We opted for this approach due to the relatively small size of our training sample: indeed, despite the images used for pre-training come from different domains, previous research has shown that using a transfer-learning-based approach can positively affect generalization [33]. We selected the ResNeXt-50 architecture based on its promising results in previous studies [34]. Compared with the default version of the model, we only modified the architecture by adjusting the number of neurons in the last dense layer (i.e., the classification layer) from 1000 to 2, so as to adapt the model to the binary classification task. To train the model, we used the softmax function as the final activation function and cross-entropy as the loss function.

The classification model (i.e., ResNeXt-50) was then trained[2] to identify TL fractures in vertebral X-rays. Specifically, to contextualize the model to the considered task, we only trained the last block from the network's backbone and the classification layer: the weights of all other layers were kept frozen. In line with the state-of-the-art for transfer learning in imaging models [35], we employed separate learning rates for the network's backbone and the classification layer. For the dense classification layer, we used a learning rate of $lr_g = 2e{-}3$, while we set the learning rate for the fourth block to $lr = 1/6 * lr_g$. We trained the model for 200 epochs and selected the final weights based on the minimum loss value observed on the validation set.

After training the model, we selected the two last convolutional layers, before the dense ones, to extract CAMs. In the following, we will refer to layer 3 (penultimate layer), as *lower-level features*, while we denote the last convolutional layer (i.e., layer 4) as *higher-level features*. This naming was selected so as to more figuratively distinguish between the semantic level of the features extracted by these layers: indeed, as also shown in Figs. 2 and 3, CAMs extracted from layer 3 represent more fine-grained and pixel-level structures (hence, lower-level), while those extracted from layer 4 tend to highlight more focused areas that correspond to regions of interest (hence, higher-level) (see Fig. 1).

4.2.2. Activation maps generation

In technical terms, CAM [20] is an approach to generate weight matrices associated with any given image, such that higher weight values indicate areas of particular relevance for the model's classification. Several algorithms can be used for CAM generation. Our study used the Grad-CAM (Gradient-weighted Class Activation Mapping) approach [27]. This approach builds upon the original CAM method to offer a more refined approach to visual explanations. Grad-CAM is based on the computation of the gradient of the class label scores w.r.t. to the feature maps from any specific convolutional layer. These gradients indicate the importance of each feature map with respect

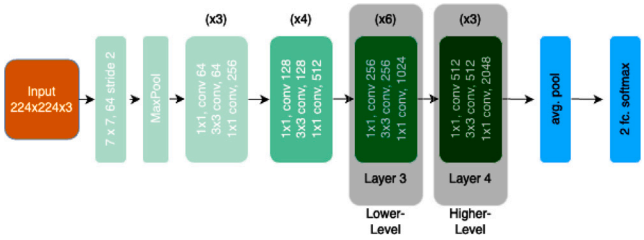

Fig. 1. ResNet-50 neural network adapted for binary classification: key layers highlighted for analysis.

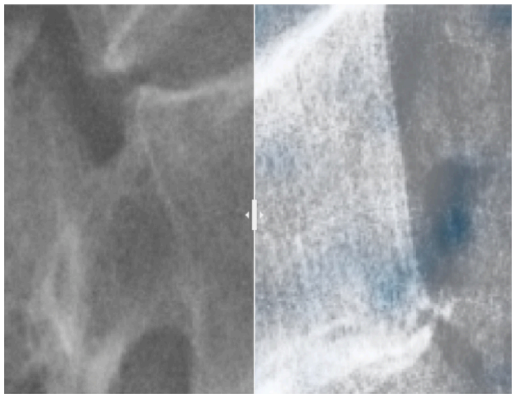

Fig. 2. Example of an X-ray and its corresponding activation map based on lower-level features. In this map, regions shown in darker shades of blue represent areas of higher importance, indicating higher activation values in the context of the X-ray image analysis.

to the target class. The gradients are then globally average-pooled to obtain the weights for each feature map. These weights are subsequently multiplied by their respective feature maps, and the results are summed up to produce the final Grad-CAM heatmap. This heatmap is then resized to match the input image dimensions and overlaid onto the original image to highlight the areas the model considers most important for classification. Compared with other techniques for generating CAMs, Grad-CAM provides a more versatile solution as it can be applied to a wider range of models, including those without fully connected layers or global average pooling layers, making it a valuable tool for visualizing and interpreting model predictions: indeed, as noted by a recent review [36], Grad-CAM is one of the most commonly used explainability techniques. These CAMs were proposed to each doctor as depicted in Figs. 2 and 3.

The image comparison tool employed in this study, as depicted in Figs. 2, 3, facilitated the concurrent analysis of X-ray images and their corresponding activation maps. Each image was rendered at a resolution of 800 × 800 pixels. Users could control a horizontal sliding control to overlay one image onto the other, to get a more in-depth comprehension of the underlying anatomical structures.

To generate a diverse set of CAMs, as anticipated in Section 4.1, two board-certified orthopedic specialists and spine surgeons selected 18 X-ray images from a randomized subset of the test dataset. The chosen images were required to meet the following criteria: (1) sufficiently *clear* for interpretation despite the lower resolution compared to conventional original X-rays, and (2) a balanced distribution of positive and negative examples. Additionally, the selection process aimed to include cases of varying complexity to ensure a comprehensive array of clinical scenarios for evaluation. In particular, the two clinicians considered cases indicative of medium-to-high complexity and such that they were of different diagnostic difficulties.

Four activation maps were produced for each of these selected images, two obtained from the lower-level features, and two from

[2] All pre-preprocessing and model training steps were implemented in Python 3.8, using the PyTorch library (v. 1.10) for implementing the classification model and the Pandas (v. 1.2) and NumPy (1.19) libraries for data processing. The computational environment was a personal computer encompassing 16 GB of RAM and a 6-core CPU, as well as a NVIDIA GTX 1060 Max-Q GPU.

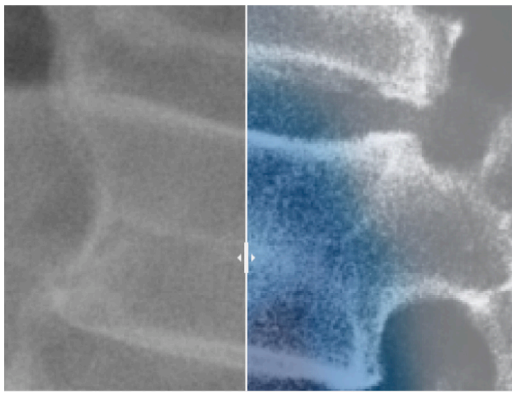

Fig. 3. Example of an X-ray and its corresponding activation map based on higher-level features. In this map, regions shown in darker shades of blue represent areas of higher importance, indicating higher activation values in the context of the X-ray image analysis.

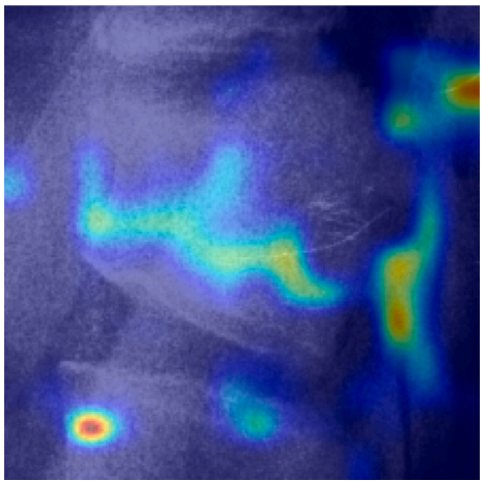

Fig. 4. CAM based on the lower-level features for the traditional color.

the higher-level features. Intuitively, activation maps obtained from lower-level features represent more fine-grained information, while, by contrast, activation maps obtained from higher-level features represent more abstract information. The two activation map images obtained for each layer differed in the coloring scheme: in the *traditional* color scheme, the X-rays were colored using the common red-blue gradient (independent of the class label that the AI model assigned to the X-ray image), denoting importance of the pixels; by contrast, in the *semantic* color scheme the X-rays were colored differently based on the classification of the AI model, with color red used for identifying a possible fracture while color blue used for identifying no presence of fractures, and saturation as a measure of pixel importance that is relevant w.r.t. the predicted class.

4.2.3. Experiment design

To carry out our user studies, we adopted the following study design. To measure the impact of these visual explanations, we considered the AI-first interaction protocol [18]. Furthermore, to collect evidence of sufficient strength, we designed an experimental user-centered study: in particular, our user study can be associated to Levels 4 and 5 in the hierarchy of evidence presented in Table 1, i.e., the highest levels of evidence not involving prospective real-world cases. The user-centered study was structured as follows:

- We created two different physician groups: Group A, associated with the *semantic* coloring scheme of CAM, and Group B, associated with the 'traditional' coloring scheme of CAM. Thus, the analysis of the coloring scheme was performed according to a between-subject study design;
- We showed the clinicians the advice generated by the AI and the activation maps developed (using either lower/higher level features, depending on the image) for each analyzed case. Thus, the analysis of higher-level vs lower-level feature information was performed according to a within-subject study design[3];
- We collected the respondents' definitive diagnosis, their confidence in that diagnosis, and their feedback on the usefulness of the activation maps in aiding the final diagnosis.

As described in the Introduction, this study adopts the proposed principles of EBXAI to investigate the best conditions and modalities for providing XAI support (in the form of CAMs) with the aim of improving diagnostic accuracy (and its utility). To achieve this, we analyzed various scenarios stratified by:

1. *Case Complexity*, represented by the average complexity score given by each clinician on a 1-to −4 ordinal scale;
2. *Presence or absence of a fracture*, as predicted by the AI;
3. *Expertise* of the clinicians, categorized into two levels based on years of experience: novice and less expert physicians, who have worked in the field for less than 5 years, and more experienced physicians, who have worked for at least 5 years.

We also compared the above-defined XAI solutions without any further stratification.

In the user study, an online, multi-page questionnaire (implemented on the LimeSurvey[4] platform) was administered to a cohort of 16 medical professionals, comprising 8 specialized spine surgeons and 8 musculoskeletal radiologists. These participants were tasked with providing diagnoses for an array of clinical cases, utilizing both AI and XAI support tools. As described previously, we divided the 16 physicians into groups A and B as a first step, see Fig. 6. Both groups saw the same cases with the same layers during the trial. The only difference between them was the color of the activation maps, as depicted in Figs. 4 and 5.

For each of the 18 cases, the clinicians were first shown the X-ray image together with the CAM visual aids (in particular, physicians in Group A were shown CAMs with *semantic* coloring, while those in Group B were shown CAMs with traditional coloring), using a Javascript-based comparison element,[5] similar to what is shown in Figs. 2 and 3. The clinicians were also shown, at the same time, the AI model's support: for physicians in Group A, the AI classification was directly represented through the color coding of the CAM, while for those in Group B, the AI classification was provided in textual format. The clinicians were then asked to provide their diagnosis along with an assessment of their confidence level, on 1-to −4 ordinal scale, and the case's perceived complexity and the CAM's perceived utility on 1-to −4 ordinal scale. The scheme adopted in the user study is depicted in Fig. 6.

[3] We performed a Mann–Whitney U Test [37] to compare the two groups of images: We evaluated the perceived level of complexity for each case and examined the presence of significant differences between the two groups: the null hypothesis of equivalence between the two groups of cases could not be rejected ($U = 9454$, $p_{value} = .175$). Thus, any differences in user performance or other outcome measures between the two groups can confidently be attributed to the type of features (low vs. high) used in the CAMs, rather than inherent differences in image complexity.

[4] https://www.limesurvey.org/

[5] https://juxtapose.knightlab.com/

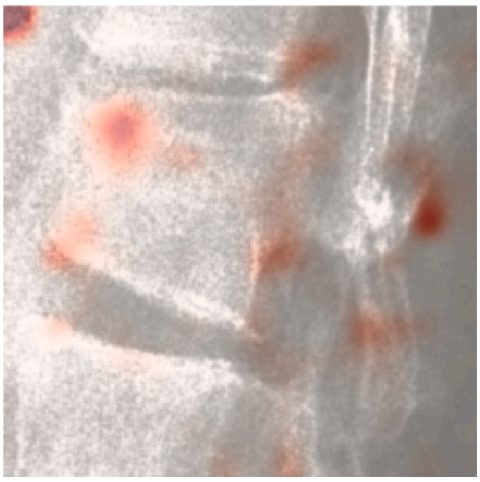

Fig. 5. CAM based on the lower-level features for the semantic color.

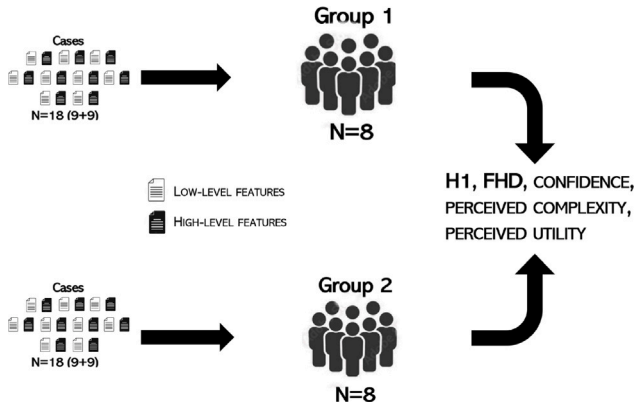

Fig. 6. User Study Workflow: The diagram illustrates the division of participants into two distinct groups for the study. Group 1 engages with traditional coloring explainable AI (XAI) techniques, while Group 2 interacts with semantic coloring XAI support. Each group is tasked with analyzing a set of 18 images, comprising original X-rays and their corresponding Class Activation Maps (CAM). These images are equally divided, with 9 generated from lower-level features and 9 from higher-level features. For each image analyzed, we record the participant's initial judgment (H1) and their final decision after utilizing the XAI support (FHD). Additionally, we gather data on the confidence levels associated with these judgments, the complexity rating of each case, and their perceived utility of the CAM. This workflow is designed to comprehensively assess the impact of different XAI approaches on participants' decision-making processes in image analysis.

4.3. Response augmentation

We employed the bootstrap method to assess the significance of the observed differences. This general, non-parametric statistical technique allows for the estimation of sample distribution properties by resampling the data with replacement [38]. To enhance the reliability of our inferences, we adopted a stratified oversampling approach, which maintains the balance across key variables (namely, the accuracy of the AI and physician, as well as the complexity of the cases) by randomly oversampling within the strata of the groups to be compared. More in detail, as our study involved 16 physicians, we derived a non-oversampled set of 16 accuracy values, one for each physician, under different settings (i.e., lower/higher-level features, CAMs, positive and negative labels, and complexity). Then, given any two sub-groups to compare, these accuracy levels were used to compute the observed statistic as:

$$\Delta_{accuracy} = a\bar{c}c_1 - a\bar{c}c_2, \quad (1)$$

where $a\bar{c}c_i$ is the average accuracy of users in group i, with $i \in \{1,2\}$. Subsequently, during each bootstrap iteration, we computed the empirical statistics for the resampled data and finally used these empirical statistics to estimate the final p-values. The number of iterations was set to 1000, and the number of oversampled and bootstrapped samples was 100. We applied the bias-corrected and accelerated (BCa) method to obtain more accurate confidence intervals [39]. The BCa improves upon the standard percentile bootstrap by accounting for potential bias and skewness in the bootstrap distribution. It does so by applying a bias-correction term that adjusts the confidence interval based on the proportion of bootstrap statistics smaller than the observed statistic and an acceleration term that adjusts the confidence interval by considering the skewness of the distribution. This approach results in a more reliable estimation of confidence intervals, which helps avoid potential over- or under-coverage compared to the standard percentile bootstrap method.

4.4. Response analysis

After collecting all responses, a comparative examination was conducted by applying a hypothesis testing approach. First, we compared the diagnostic accuracy of the clinicians supported by lower-level features CAMs vs. higher-level features CAMs; similarly, we compared the diagnostic accuracy of the clinicians supported by Semantic CAMs vs. Traditional CAMs. Then, we considered the following stratifications: by complexity level (dichotomized at the threshold 2, where cases with perceived complexity lower than 2 were rated as being of low complexity and otherwise of high complexity), by predicted diagnosis (i.e., Positive vs. Negative); and by users' expertise (comparing physicians who had worked for less than 5 years with those who had worked for at least 5 years).

Cohen's D [40] was used as a measure of effect size to assess the impact on accuracy across the stratifications under examination. This measure is defined as $D = \frac{\Delta_{accuracy}}{SD_{pooled}}$, where $\Delta_{accuracy}$ is defined as in Eq. (1), and SD_{pooled} is the pooled standard deviation for the two groups. Since we performed multiple tests for differences, we corrected the p-values for each comparison group using the Benjamini–Hochberg correction method [41] across the entire set of statistical tests (a total of 18) instead of dividing them into distinct groups, so as to control for the false discovery rate.

Finally, we also considered the perceived utility of the explanations as well as the reported confidence on the final decision (FHD). In particular, in regard to the perceived utility, we studied how it varied w.r.t. AI correctness (AI-correct versus AI-incorrect judgments) as well as w.r.t. to the main factors of analysis considered (i.e., CAM coloring and layer). Similarly, we evaluated differences in reported confidence based on the CAM coloring (*semantic* vs. traditional) and layer (lower-level features vs. higher-level features).

5. Results

The involved 16 X-rays board-certified orthopedists interpreted and made a diagnosis for the set of 18 cases selected above, where each case included an X-ray image, the AI advice, and an activation map. This resulted in an aggregate of 576 evaluations by the decision makers involved in our user study.

The results of the statistical analysis are reported in Table 2. In the table, we report the p-values (both adjusted and non-adjusted), the confidence intervals for the observed difference in accuracy, and the Cohen's D effect size. For the lower-level features vs. higher-level features group, a positive Cohen's D value indicates that lower-level features are associated with a higher accuracy compared to higher-level features. Conversely, a negative value means that higher-level features outperform lower-level features in terms of effect upon accuracy. Similarly, a positive Cohen's D value in the Semantic vs. Traditional group implies that Semantic CAMs yield higher accuracy on average than

Table 2
Results of the bootstrapped hypothesis testing for the differences in accuracy between lower-level features and higher-level features XAI support, and Semantic (Group 1) and Traditional (Group 2). The p-values adjustments were performed based on the Benjamini–Hochberg FDR correction. Cohen's D is computed on the original data: a positive (resp., negative) value means that the accuracy of Group 1 (resp., Group 2) was higher than that of Group 2 (resp. Group 1). Bold values indicate a significant difference. The confidence intervals (CI) are calculated using the BCa bootstrap method, which can result in asymmetric intervals. This asymmetry accounts for the skewness in the bootstrap distribution of the statistic. It is important to note that the observed statistic (ΔAcc_{obs}) may not always fall within the confidence intervals due to sampling variability.

Comparison	ΔAcc_{obs}	BCa CI 95%	P-value	P_{adj}	Cohen's D
Lower-Level Features vs. Higher-Level Features					
Overall	.041	[.045, .059]	.027	**.044**	.383
Semantic CAM	.027	[−.001, .086]	.332	.373	.238
Traditional CAM	.055	[.045, .111]	.019	**.037**	.592
AI POS	.134	[.083, .136]	.0001	**.001**	1.113
AI NEG	−.059	[−.102, −.057]	.053	.073	−.313
Low Complexity	.07	[.091, .092]	.001	**.004**	.645
High Complexity	−.019	[−.119, .037]	.638	.672	−.077
Expertise ≤ 5 years	.013	[−.037, .075]	.672	.672	.102
Expertise > 5 years	.069	[.08,.119]	.021	**.037**	.861
Semantic vs. Traditional					
Overall	−.049	[−.054, −.044]	.0001	**.001**	−1.009
Lower-Level Features	−.083	[−.096,−.091]	.002	**.007**	−.828
Higher-Level Features	−.056	[−.083,−.056]	.046	.069	−.5
AI POS	−.055	[−.075, −.051]	.012	**.027**	−.632
AI NEG	−.084	[−.09, −.083]	.003	**.009**	.686
Low Complexity	−.04	[−.075, −.034]	.057	.073	−.456
High Complexity	−.103	[−.123, −.101]	.004	**.01**	−.643
Expertise ≤ 5 years	−.097	[−.098,−.081]	.001	**.004**	−1.2
Expertise > 5 years	−.041	[−.081,−.03]	.147	.176	−.66

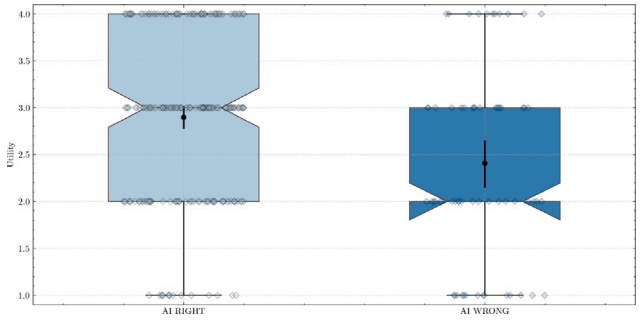

Fig. 7. Utility perceived by physicians stratified by either correct or wrong classification.

their traditional counterparts, whereas a negative value means that Traditional CAMs exhibit better accuracy than Semantic CAMs.

The physicians' perceived utility, based on whether the AI correctly predicted the class, is shown in Fig. 7.

6. Discussion

Our experiment and user study has been aimed at investigating the impact of CAMs on diagnostic accuracy and other psychometric dimensions of a diagnostic task, namely the detection of TL fractures in vertebral X-rays. By adopting what we called an *evidence-based XAI* approach, of which our study represents an exemplar application, our study applies a user-centered approach to the comprehensive evaluation of the effectiveness of human–AI collaboration where CAMs are used to make the AI more transparent, understandable and explainable, that is as XAI method. In particular, by means of a user study producing evidence at the 5th level in the hierarchy reported in Table 1, we explored the effectiveness of CAMs generated from higher-level features (i.e., Layer 4) versus those generated from lower-level features (i.e., Layer 3), and assessed the influence of '*semantic*' versus 'traditional' coloring in aiding diagnosis.

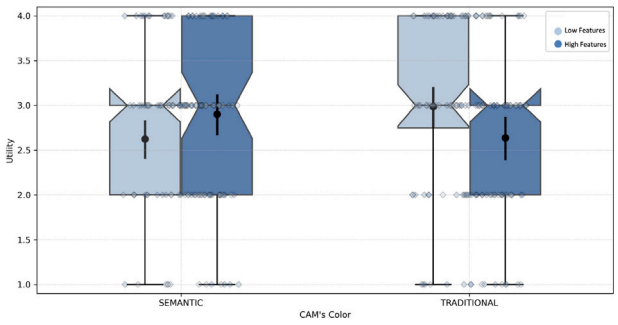

Fig. 8. Utility perceived by the two groups of physicians (Traditional vs. Semantic) stratified by layers.

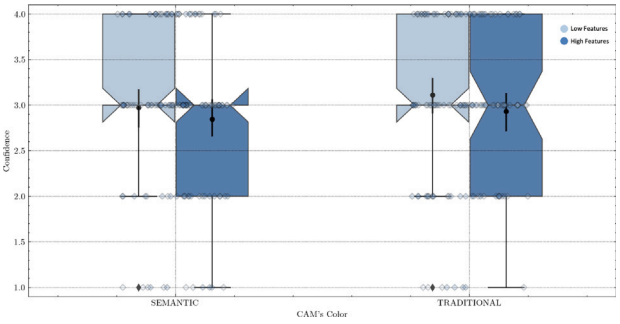

Fig. 9. Confidence perceived by the two groups of physicians (Traditional vs. Semantic) stratified by layers.

6.1. The reference layer for CAM generation

Our first finding regards the fact that CAMs generated from lower-level features are associated with higher diagnostic accuracy than higher-level ones, in several stratification scenarios. This finding is partly surprising since it is a common belief in the XAI literature that CAMs should be generated from higher-level features [42], as this information is "closer" to the final classification and hence considered to be more understandable by human users. By contrast, our results suggest that lower-level features might be more informative and hence helpful in detecting vertebral fractures. In particular, the overall difference in accuracy observed comparing the effect of CAMs generated from either lower-level or higher-level features was significant (p_{adj} = .044, Cohen's D = .383), with lower-level feature CAMs outperforming the higher-level feature ones. Lower-level feature CAMs performed particularly well in cases where AI correctly identified a fracture (AI POS, p_{adj} = .001, Cohen's D = 1.113). A possible explanation for this effect lies in the different anatomical landmarks highlighted in the maps: the lower-level features appeared more "focused" and specific in highlighting critical areas for fracture recognition, while higher-level feature CAMs were perceived as "broader" and less specific in their highlighted zones. To provide a more detailed analysis of this observation, one of the authors identified a sample of cases, extracted from our study, of particular interest.

The first case (see Fig. 10) consists of a peculiar type of fracture, i.e., the *Chance Fracture* or *B1* fracture according to the AOSpine classification; this is a rare pattern, which represents only 1%–2% of all thoracolumbar fractures. In this specific case, the fracture line extends from the anterior part of the vertebral body, where a pathologic inflection can be seen (Fig. 10 A, Yellow Arrow), through the vertebral body and the pedicles. The specific inflection on the anterior part

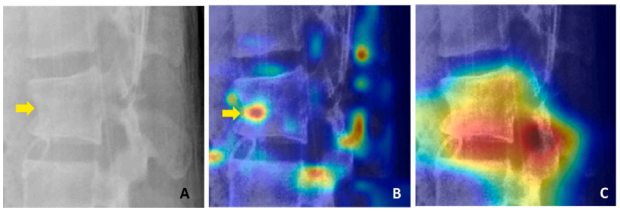

Fig. 10. Visual comparison of X-ray images and associated class activation mappings (CAMs) illustrating varying levels of feature extraction. (A) Original X-ray image depicting the anatomical structures under examination. (B) Corresponding CAM, highlighting lower-level features. (C) The CAM emphasizes higher-level features.

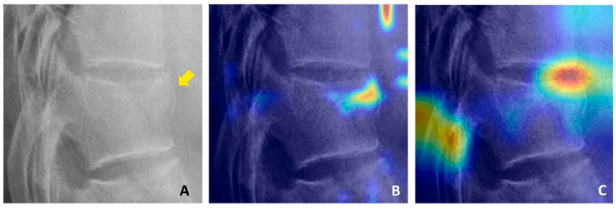

Fig. 11. Visual comparison of X-ray images and associated class activation mappings (CAMs) illustrating varying levels of feature extraction. (A) Original X-ray image depicting the anatomical structures under examination. (B) CAM highlights lower-level features. (C) CAM emphasizes higher-level features.

of the vertebral body was precisely highlighted by the lower-level feature CAMs; on the contrary, the higher-level features CAM shows an overall "broad" and generic activation that mainly highlights the spinous process, which is not involved in this type of fracture (see Fig. 10 C). Admittedly, both maps failed to highlight the fracture line inside the pedicles, which is important for a correct classification of this kind of fracture. However, focusing the observer's attention on the inflection point could be a sufficiently strong clue for correctly identifying the fracture.

The second case example regards a more common fracture pattern, i.e., the A3 AOSpine (26% of cases in our dataset); in this case, the diagnostic difficulty lies in the fact that the anatomical alteration of the vertebral body is minimal, as it can be seen in the X-ray, and hence difficult to identify (see Fig. 11 A). Again, the lower-level features map highlights the correct landmark (see Fig. 11 B), while the higher-level features map highlights a slightly wider and higher area (see Fig. 11 C, i.e., the intervertebral disk). This made the lower-level CAM more indicative and informative towards the correct diagnosis, while the region on the left highlighted by the higher-level feature CAM lacked substantive information.

We also comment on how expertise can be factored in. In the case of physicians with higher expertise, we observed a significant difference in diagnostic accuracy between lower-level feature CAM and higher-level features ones ($p_{adj} = .037$). The observed effect size (Cohen's D) was 0.86, indicating a large effect. This finding suggests that more experienced physicians significantly benefit from CAMs generated from lower-level layers when they are called to detect TL fractures in vertebral X-rays. In contrast, for physicians with lower expertise (≤ 5 years), there was no significant difference between lower/higher-level features ($p_{adj} = .672$), and the effect size was relatively small (Cohen's D = .102). This implies that the reference layer from which to generate the CAMs might not be as influential for less experienced physicians in this diagnostic task. Other factors, such as training or additional guidance, might be required to improve diagnostic performance. For physicians with a higher level of expertise, their ability to leverage the CAMs generated from lower-level features might be an extension of their comprehensive understanding and interpretation of the anatomical intricacies involved in diagnosing TL fractures from vertebral X-rays. In this context, the lower-level features of CAMs, that better emphasize detailed anatomical landmarks, might align more closely with the experienced physicians' cognitive models of the diagnostic task. On the other hand, for physicians with less experience, the lack of significant difference in diagnostic accuracy between lower-level and higher-level feature CAMs might reflect their relative uncertainty in finding the correct fracture pattern. Thus, a map that highlights a broader area where the fracture may lie, is equally effective as a map that better focus on the precise anatomical region.

These considerations bring us to reinforce the point that we asserted in [21], where we advocated a higher acquaintance and familiarity of medical professionals with how machines "see" radiological images. From the present user study we confirm the idea that the ways in which machines perceive X-rays can be highly different from how humans read and interpret them [43]; we called this area of concern, *the semiotics of machine interpretation*. It is quite natural that just as ML models can identify ethnicity, gender, and smoking status from pictures that humans have always considered entirely unrelated to these conditions [44], similarly, ML models may discern characteristics that are indicative of a fracture (for instance, attributable to muscle rigidity or associated microtrauma) that are not immediately apparent to human observation. Consequently, physicians should improve their proficiency in recognizing the patterns that are more correlative of positivity (e.g., fracture presence) as this might be detected by the machine that they use, and highlighted through lower-level CAMs.

As a final remark in regard to the reference layer and type of CAMs, we acknowledge that the above findings, although novel and to some respect even surprising, are not generalizable to other settings: for example, the optimal layer from which to generate CAMs should in general be expected to vary depending on both the considered task, as well as the model architecture and the technique used to generate the CAMs. However, here we want to emphasize a sort of additional finding that those results suggest: the importance of having data scientists and ML programmers explore alternative layer configurations for generating CAMs in different medical imaging applications, so as to consider the inherent nature of the specific medical task being addressed. In other words, the reference layer matters and its choice should be backed by empirical evidence, as the one we statistically drew from our user study.

6.2. The coloring scheme in CAM generation

Our second main finding regards the observation that CAMs adopting a traditional (i.e., red-blue gradient) coloring scheme were consistently associated with higher accuracy than the semantic coloring schema (i.e., red gradient for positive images, blue gradient for negative cases), across all layers. This observation contradicts the conjecture, formulated by the domain experts involved in [21], who advocated for the semantic scheme assuming that this latter would make image interpretation more intuitive and associated with fewer misunderstandings in regard to the classification provided by the machine. Contrary to this idea, which we can consider a sort of expert opinion (Level 10 in Table 1), we collected clear evidence (as mentioned above, Levels 4 and 5 in Table 1) that physicians perceived traditional CAMs significantly more useful and these were actually more effective in helping image interpretation. This finding can be related to the so-called Jakob's Law [45], which is attributed to the physicians' familiarity with conventional color schemes that have been recently used in CAM-aided diagnostic imaging. The overall difference in accuracy between traditional and semantic CAMs was significant ($p_{adj} = .001$, Cohen's D $= -1.009$), with traditional CAMs associated with better performance especially when paired with lower-level features ($p_{adj} = .007$, Cohen's D $= -.828$). Despite the statistical insignificance of rejecting the null hypothesis asserting equivalence in higher-level features between the two groups in terms of predictive accuracy ($p_{adj} = .069$), our analysis elucidated a Cohen's D value of $= -.5$. This result denotes a medium effect size [46]. In this case, the negative sign means that, on average, the group exposed to traditional CAMS exhibited higher diagnostic

accuracy. Furthermore, traditional colored CAMs outperformed semantic colored CAMs for both positive predictions (AI POS, $p_{adj} = .027$, Cohen's D = −.632) and negative ones (AI NEG, $p_{adj} = .009$, Cohen's D = .686). When examining the impact of coloring schemes, we also observed a significant difference in diagnostic accuracy between semantically and traditionally colored CAMs for physicians with lower expertise ($p_{adj} = .004$), with an effect size of −1.2, which suggests a very large effect [46] in favor of traditional CAMs. However, for physicians with higher expertise, there was no significant difference between semantic and traditional colored CAMs ($p_{adj} = .176$), although the effect size was still moderately large (Cohen's D: −0.66). This finding indicates that the choice of the coloring scheme might not play a critical role for more experienced physicians in this context, as they might be more adept at interpreting AI-generated PAMs regardless of the coloring scheme.

6.3. Perceived utility and confidence

When analyzing the utility perceived by physicians, we observed a clear difference in perceived usefulness between cases where AI correctly predicted the class (AI Right) and cases where AI wrongly predicted the class (AI Wrong). As shown in Fig. 7, physicians found the AI and XAI more useful when the model accurately identified the corresponding class. This observation is evident from the non-overlapping confidence intervals of the notches, suggesting a significant difference in perceived utility. The utility perceived by the two groups of physicians (Traditional vs. Semantic), stratified by layers, is illustrated in Fig. 8. On average, higher-level features were perceived as more useful in the Semantic setting, while lower-level features were perceived as more useful in the Traditional setting. However, it is essential to note that, in this case, the median confidence intervals overlap between the various groups, indicating that the differences in perceived utility might not be significant.

We also analyzed the confidence perceived by the two groups of physicians (Traditional vs. Semantic) stratified by feature representations. On average, confidence appeared to be higher in lower-level features. Within the Semantic group, the median confidence interval did not overlap, indicating a significant difference in confidence between the lower/higher-level features. In contrast, the Traditional group exhibited overlapping confidence intervals; however, as depicted in Fig. 9, Layer 3 (lower-level features) still seemed to increase physicians' confidence. These findings on perceived confidence highlight the potential influence of CAMs generated from different layers on physicians' trust in AI-assisted diagnostic tools. The increased confidence observed for Layer 3, particularly within the Semantic group, suggests that CAMs from this layer may be more effective in supporting physicians in their decision-making process.

These results on perceived confidence and utility emphasize the importance of considering the subjective perception of physicians when evaluating the impact of CAMs on clinical decision-making, as an additional form of evidence. Indeed, understanding how physicians perceive the usefulness of different CAMs can inform the design of more effective and user-friendly XAI tools, ultimately contributing to better patient outcomes.

6.4. Limitations

Although we deem the findings of this study original and significant, we must also acknowledge some limitations regarding generalizability. Firstly, the findings we collected in regard to thoracolumbar (TL) fractures in vertebral X-rays, cannot be entirely applicable to other diagnostic tasks, especially across different specialties. Additionally, the relatively small sample size of images, despite being carefully chosen to represent varying complexities and frequencies, and the diversity in expertise levels of the image readers, might limit the broad applicability of our findings within the realm of detecting TL fractures from X-rays.

Furthermore, the dataset partitioning strategy used in our study, with a ratio of 8:1:1 for training, validation, and testing, also presents a potential limitation. This distribution, while considered appropriate for the study's aims, suggests that a more extensive test dataset could have potentially provided more compelling evidence. To mitigate the risk of false negatives and strengthen our statistical claims, we implemented state-of-the-art augmentation techniques (refer to Section 4.3) and calculated effect sizes in addition to reporting p-values. In doing so, future studies might perform a power analysis leveraging the effect sizes identified in this study and benefit from considering a more extensive test dataset to further validate and extend these findings.

7. Conclusion

In this study, we proposed a novel methodological framework to design and evaluate XAI systems that ground on an evidence-based and empirical approach, inspired by evidence-based medicine. The conceptual foundation of our design approach rests on three pillars. (1) Evidence-based methodology: the design of XAI systems should be grounded on solid experimental designs that investigate how the support affects the decision-making process; (2) Hierarchy of evidence: clearly not all experimental designs provide the same evidence strength, and XAI empirical studies should be designed so as to provide the strongest possible form of evidence (according to Table 1) given the considered setting; (3) User-centric evaluation: as emphasized also in Table 1, the human-centered aspect is pivotal in our evidence-based methodology and appropriate evaluation of AI and XAI systems should always include the human users. This approach thus aims to give XAI systems' designers a principled way to create decision support systems and XAI solutions, emphasizing the need to go beyond simple evaluation practices that do not take into account the needs of, and impact on, the final users.

To exemplify our methodological contribution we also presented the results of an experimental study, by which we have elucidated key aspects of Class Activation Maps (CAMs) configuration for enhancing diagnostic accuracy in the detection of thoracolumbar fractures from X-rays. Our findings emphasize the effectiveness of generating CAMs from lower-level features, as well as of utilizing a traditional red-blue coloring scheme. These configurations not only improved diagnostic accuracy but also bolstered physician confidence in their diagnoses. In alignment with our evidence-based framework, our study design was centered on user engagement, and involved real physicians in critical diagnostic tasks, marking a significant advancement in our study. By reaching levels 4 and 5 in the hierarchy of evidence, our study challenges prevalent opinions among domain experts (level 10) and the broader XAI community, particularly concerning the choice of layers for CAM generation. Our approach does not oppose the views of these groups but rather treats them as starting points for more rigorous empirical research. This strategy allows us to isolate and understand the impacts of specific design choices on critical aspects of medical practice, such as diagnostic accuracy and user satisfaction. Consequently, this study not only contributes significantly to the field of effective XAI design but also sets a precedent for future research focused on a human-centered approach in the evaluation of XAI tools in real-world applications.

Moreover, the results of this study also points to the potential of adaptive AI approaches to enhance diagnostic accuracy and decision-making in medical imaging. Adaptive AI tailors the presentation of CAMs based on various factors such as the model's prediction, perceived complexity, or the physician's expertise. This, in turn, offers a more personalized and contextually pertinent visualization, improving the utility of XAI solutions, such as CAMs, enhancing the physician's diagnostic accuracy, and thereby contributing to improved patient outcomes.

Finally, our findings also point out the necessity for further investigations to assess the extendibility of these results across different

imaging modalities and diagnostic tasks. Such exploration should focus on incorporating empirical evidence and user preferences in the design of XAI methods, aimed at configuring better AI systems that strengthen human–AI collaboration thereby enhancing patient care and clinical decision-making.

CRediT authorship contribution statement

Lorenzo Famiglini: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Andrea Campagner:** Conceptualization, Formal analysis, Methodology, Validation, Writing – original draft, Writing – review & editing. **Marilia Barandas:** Writing – original draft. **Giovanni Andrea La Maida:** Data curation, Validation. **Enrico Gallazzi:** Conceptualization, Data curation, Formal analysis, Investigation, Resources, Validation, Writing – original draft, Writing – review & editing. **Federico Cabitza:** Conceptualization, Formal analysis, Investigation, Methodology, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

None Declared.

References

[1] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M.T. Ribeiro, D. Weld, Does the whole exceed its parts? the effect of ai explanations on complementary team performance, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–16.

[2] M. Pacailler, S. Yahoodik, T. Sato, J.G. Ammons, J. Still, Human-centered artificial intelligence: Beyond a two-dimensional framework, in: International Conference on Human-Computer Interaction, Springer, 2022, pp. 471–482.

[3] T.A. Schoonderwoerd, W. Jorritsma, M.A. Neerincx, K. Van Den Bosch, Human-centered xai: Developing design patterns for explanations of clinical decision support systems, Int. J. Hum.-Comput. Stud. 154 (2021) 102684.

[4] F. Cabitza, A. Campagner, C. Simone, The need to move away from agential-ai: Empirical investigations, useful concepts and open issues, Int. J. Hum.-Comput. Stud. 155 (2021) 102696.

[5] F. Cabitza, A. Campagner, G. Malgieri, C. Natali, D. Schneeberger, K. Stoeger, A. Holzinger, Quod erat demonstrandum?-towards a typology of the concept of explanation for the design of explainable ai, Expert Syst. Appl. 213 (2023) 118888a.

[6] U. Ehsan, P. Wintersberger, Q.V. Liao, E.A. Watkins, C. Manger, H. Daumé III, A. Riener, M.O. Riedl, Human-centered explainable ai (hcxai): beyond opening the black-box of ai, in: CHI Conference on Human Factors in Computing Systems Extended Abstracts, 2022, pp. 1–7.

[7] Q.V. Liao, K.R. Varshney, Human-centered explainable ai (xai): From algorithms to user experiences, 2021, arXiv preprint arXiv:2110.10790.

[8] M. Phiri, Design Tools for Evidence-Based Healthcare Design, Routledge, 2014.

[9] E. Ammenwerth, M. Rigby, Evidence-Based Health Informatics: Promoting Safety and Efficiency Through Scientific Methods and Ethical Policy, vol. 222, IOS Press, 2016.

[10] J. Wyatt, Assessing and improving evidence based health informatics research, in: Health Inform., IOS Press, 2010, pp. 435–445.

[11] G. Vilone, L. Longo, Classification of explainable artificial intelligence methods through their output formats, Mach. Learn. Knowl. Extract. 3 (3) (2021) 615–661.

[12] F. Davidoff, B. Haynes, D. Sackett, R. Smith, Evidence based medicine, 1995.

[13] M. Bhandari, M. Zlowodzki, P.A. Cole, From eminence-based practice to evidence-based practice: A paradigm shift, Minnesota Med. 87 (4) (2004) 51–54.

[14] D.L. Sackett, W.M. Rosenberg, J.M. Gray, R.B. Haynes, W.S. Richardson, Evidence based medicine, BMJ: Br. Med. J. 313 (7050) (1996) 170.

[15] D.K. Hamilton, D.H. Watkins, Evidence-Based Design for Multiple Building Types, John Wiley & Sons, 2008.

[16] J. Wyatt, Evidence-based health informatics and the scientific development of the field, in: E. Ammenwerth, M. Rigby (Eds.), in: Evidence-Based Health Informatics: Promoting Safety and Efficiency Through Scientific Methods and Ethical Policy, vol. 222, IOS Press, 2016, p. 14.

[17] W. Jin, X. Li, M. Fatehi, G. Hamarneh, Guidelines and evaluation of clinical explainable ai in medical image analysis, Med. Image Anal. 84 (2023) 102684.

[18] F. Cabitza, A. Campagner, L. Ronzio, M. Cameli, G.E. Mandoli, M.C. Pastore, L.M. Sconfienza, D. Folgado, M. Barandas, H. Gamboa. Rams, Hounds and white boxes: Investigating human–ai collaboration protocols in medical diagnosis, Artif. Intell. Med. 138 (2023) 102506c.

[19] A. Nandi, A.K. Pal, Detailing image interpretability methods, in: Interpreting Machine Learning Models, Springer, 2022, pp. 271–293.

[20] M. He, B. Li, S. Sun, A survey of class activation mapping for the interpretability of convolution neural networks, in: International Conference on Signal and Information Processing, in: Networking And Computers, Springer, 2022, pp. 399–407.

[21] F. Cabitza, A. Campagner, L. Famiglini, E. Gallazzi, G.A. La Maida, Color shadows (part i): Exploratory usability evaluation of activation maps in radiological machine learning, in: Machine Learning and Knowledge Extraction: 6th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2022, Vienna, Austria, August (2022) 23–26, Proceedings, Springer, 2022, pp. 31–50.

[22] C. Natali, L. Famiglini, A. Campagner, G.A.L. Maida, E. Gallazzi, F. Cabitza, Color shadows 2: Assessing the impact of xai on diagnostic decision-making, in: L. Longo (Ed.), XAI 23: Proceedings of EXplainable Artificial Intelligence, the First World Conference on EXplainable Artificial Intelligence, XAI-2023, Springer, Lisbon, Portugal, 2023.

[23] Y. Alufaisan, L.R. Marusich, J.Z. Bakdash, Y. Zhou, M. Kantarcioglu, Does explainable artificial intelligence improve human decision-making? in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 6618–6626.

[24] F. Cabitza, A. Campagner, C. Natali, E. Parimbelli, L. Ronzio, M. Cameli, Painting the black box white: Experimental findings from applying xai to an ecg reading setting, Mach. Learn. Knowl. Extract. 5 (1) (2023) 269–286b.

[25] R. Aggarwal, V. Sounderajah, G. Martin, D.S. Ting, A. Karthikesalingam, D. King, H. Ashrafian, A. Darzi, Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis, NPJ Digit. Med. 4 (1) (2021) 1–23.

[26] M.S. Ayhan, L.B. Kümmerle, L. Kühlewein, W. Inhoffen, G. Aliyeva, F. Ziemssen, P. Berens, Clinical validation of saliency maps for understanding deep neural networks in ophthalmology, Med. Image Anal. (2022) 102364.

[27] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[28] A. Chattopadhay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision, WACV, IEEE, 2018, pp. 839–847.

[29] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, X. Hu, Score-cam: Score-weighted visual explanations for convolutional neural networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 24–25.

[30] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, J. Jiao, Weakly supervised instance segmentation using class peak response, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3791–3800.

[31] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, Y. Wei, Layercam: Exploring hierarchical class activation maps for localization, IEEE Trans. Image Process. 30 (2021) 5875–5888.

[32] H. Jung, Y. Oh, Towards better explanations of class activation mapping, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1336–1344.

[33] A. Ke, W. Ellsworth, O. Banerjee, A.Y. Ng, P. Rajpurkar, Chextransfer: Performance and parameter efficiency of imagenet models for chest x-ray interpretation, 2021, CoRR, abs/2101.06871.

[34] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492–1500.

[35] Z. Shen, Z. Liu, J. Qin, M. Savvides, K.-T. Cheng, Partial is better than all: Revisiting fine-tuning strategy for few-shot learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 9594–9602.

[36] J. Allgaier, L. Mulansky, R.L. Draelos, R. Pryss, How does the model make predictions? a systematic literature review on the explainability power of machine learning in healthcare, Artif. Intell. Med. 143 (2023) 102616.

[37] E.B. Manoukian, Mathematical Nonparametric Statistics, Taylor & Francis, 2022.

[38] P. Shankar, Tutorial overview of simple, stratified, and parametric bootstrapping, Eng. Rep. 2 (1) (2020) e12096.

[39] D. Chen, M.S. Fritz, Comparing alternative corrections for bias in the bias-corrected bootstrap test of mediation, Eval. Health Profess. 44 (4) (2021) 416–427.

[40] L.D. Kyu, Alternatives to p value: Confidence interval and effect size, kja 69 (6) (2016) 555–562, http://dx.doi.org/10.4097/kjae.2016.69.6.555, http://www.e-sciencecentral.org/articles/?scid=1156285.

[41] W.S. Noble, How does multiple testing correction work? Nature Biotechnol. 27 (12) (2009) 1135–1137.

[42] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016.

[43] B.K. Betzler, H.H.S. Yang, S. Thakur, M. Yu, Z. Da Soh, G. Lee, Y.-C. Tham, T.Y. Wong, T.H. Rim, C.-Y. Cheng, et al., Gender prediction for a multiethnic population via deep learning across different retinal fundus photograph fields: Retrospective cross-sectional study, JMIR Med. Inform. 9 (8) (2021) e25165.

[44] J.W. Gichoya, I. Banerjee, A.R. Bhimireddy, J.L. Burns, L.A. Celi, L.-C. Chen, R. Correa, N. Dullerud, M. Ghassemi, S.-C. Huang, et al., Ai recognition of patient race in medical imaging: A modelling study, Lancet Digit. Health 4 (6) (2022) e406–e414.

[45] J. Yablonski, Laws of UX: Using Psychology to Design Better Products & Services, O'Reilly Media, 2020.

[46] S.S. Sawilowsky, New effect size rules of thumb, J. Modern Appl. Stat. Methods 8 (2) (2009) 26.