

# Validating XAI techniques in medical image diagnosis: A venture towards algorithm transparency in a socio-technical system.

JACQUES FÜRST, University of Twente, The Netherlands

Transparency and validation of relevant processes are key factors to consider when developing an algorithm that is to be integrated into a socio-technical system, i.e. a system that views the algorithm and the organization it will be employed in with all relevant surrounding features as one holistic instance, which is yet to be designed. In order to solve the problem of maintaining the transparency of algorithms used in medical diagnosis, it is hence crucial to utilize methods that help clinicians understand the technical processes forming the basis of the information provided by the algorithms. Thus, the application of Explainable Artificial Intelligence (XAI) techniques in this regard is not only advisable and incredibly important to maintain the trustworthiness that is required for the clinicians to adopt the systems in the first place, but is indispensable for the doctor-patient relationship as well. Within this research, four XAI methods were picked for validation through the means of expert interviews with pathologists. The results showed that whilst there is a certain tendency of preferences within the stakeholder population, it should be evaluated on a case-to-case basis which method is most applicable. Furthermore, the integration of the methods is a rather sensible issue. It should thus not only be evaluated which technique to integrate, but how to integrate the method to maximize trust and effectiveness through extensive use. If done effectively, these methods can facilitate the clinicians' workflow significantly.

Additional Key Words and Phrases: Explainable AI, medical image diagnosis, ethical algorithms

## 1 INTRODUCTION

The internal mechanisms of AI algorithms are inherently unclear; they are opaque by design [3]. Thus, their nature can be described as a black box, metaphorically speaking. However, these algorithms are increasingly used to determine information that requires a certain level of transparency, such as a person's credit score, their insurance risk or health status [11]. Specifically, there has been an increasing use of AI technology in the medical sector, due to recent technological progress [39]. Yet, since lives may be at stake in this realm [32], trust and ethical soundness of any algorithm are essential for it to be employed. That is so, since black box models leave much to be desired for clinicians to explain the reasoning behind diagnoses to other stakeholders such as patients and their families. Therefore, it is important to introduce techniques to fill this gap and maintain the trustworthiness and transparency that constitutes the traditional doctor-patient relationship [21].

That is why the field of 'Explainable Artificial Intelligence' (XAI) is so crucial for our society to be able to move forward whilst maintaining this transparency. One may imagine a situation in which a clinician has to make a decision on whether to operate on a patient who might have a brain tumour. On the one hand, the operation

in itself is quite risky, so if the patient does not have a tumour, then it should be avoided at any cost. On the other hand, if they do have a tumour, an operation should definitely be executed since it significantly increases the chances for an extended life span for the patient. If a pathologist would then make their decision based on a classification by an algorithm which they do not have any understanding of, this may lead to disastrous results. If they decide to operate on a person who had no tumour, for instance, the patient would die on the operation table because of a misclassification of the algorithm. It is then going to be very hard for the clinician to justify their decision that was fully based on the trust of a faulty algorithm to the dead patient's family. If they used XAI in this situation, it would increase the chances of them noticing the misclassification due to a wrong area of attention in the MRI slices. Therefore, XAI could be life-saving in situations like these.

For the sake of some arguments, the methods employed in this field shall be categorized as either post-hoc, or ante-hoc. In the post-hoc (explainable AI) approach, a second algorithm is trained by fitting the predictions of the black box algorithm and not the original data. In ante-hoc (interpretable AI) methods, use a 'white-box' method instead of a black box, which is transparent and in an easy-to-digest form [5].

An additional categorization that is relevant for interpretability methods is distinguishing between global and local interpretability. As Salahuddin et al. put it: 'Local explanations identify the attributes and features of a particular image that the DL model considers important for prediction. On the other hand, global explanations aim at identifying the common characteristics that the DL model considers when associating images with that particular class.' [27]. These categories influence the applicability of an XAI method in certain contexts. A global method may not be as much of a fit for explaining an outcome to a patient, for instance. A case-based approach is more advisable under these circumstances.

In the context of this research, we take a socio-technical approach, thus viewing the AI as part of a system that views the algorithm and the organization it will be employed in with all relevant surrounding features as one holistic instance which is yet to be designed [34].

Hence, this research will explore and validate effective XAI techniques that foster the maintenance of trustworthiness and transparency within the field of medical diagnosis. It will do so by the means of a literature review and the consultancy of several experts from the field of XAI itself to determine which methods may be most appropriate to be validated. Finally, clinicians shall be interviewed, and their answers will be analysed to draw conclusions on the relevance of these methods in practice.

### 1.1 Problem statement

This work builds upon the framework developed by van Bruxvoort and van Keulen [34] and extends it by recommending employable techniques that help fulfilling the goals defined in the framework.

TScIT 37, July 8, 2022, Enschede, The Netherlands

© 2022 University of Twente, Faculty of Electrical Engineering, Mathematics and Computer Science.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Specifically, the research shall be focusing on the goal of ensuring the explicability of an algorithm. Clearly, the challenge in this regard lies within the diversity of stakeholders depending on the sector into which the relevant algorithm should be implemented. Patients may have differing knowledge levels and capacities for understanding medical concepts, where clinicians may prefer an explanation based on their individual thought processes and workflows, for instance. Another objective is pointing out methods that can be understood and employed by people who do not necessarily have a technical background.

As the original framework focuses on fraud-detection algorithms, the arguments made in the paper are delimited to the group of stakeholders to whom these apply. This research will thus aim to expand the scope of the framework by researching possible methods to ensure explicability in the medical sector, specifically for disease diagnosis, as mentioned above. It does therefore not only add sector-specific considerations, but gives recommendations on the implementation of transparency, which was not done in the original framework. Moreover, the introduction of three XAI methods utilized less frequently within the medical sector (this excludes Grad-CAM, which is also introduced, but more common in its use) will reveal whether there is a chance to take a novel approach to this conundrum.

## 1.2 Scope

It is noteworthy that the research will focus on interviewing clinicians only. This is due to the fact that they build the metaphorical bridge between the algorithm and all other stakeholders, such as patients and their families or other clinicians interested in the diagnosis. Through their position, they are aware of the needs of the other stakeholders to a certain degree as well, since they have to use the explanation for communicating the reasoning behind the diagnosis to them.

## 1.3 Research questions

The problem statement leads to the following research question:

How can one ensure that an algorithm is sufficiently explicable to pathologists in medical image diagnosis?

To answer it, the following sub-questions will be tackled:

- (1) Which XAI techniques are relevant to be investigated for algorithms used in medical image diagnosis?
- (2) How can the techniques in (1) be explained to clinicians, and, immediately after, be tested on trust and usability?
- (3) Which of the techniques mentioned in (1) are able to maintain the clinicians' trust in the algorithm and effectively facilitate the diagnosis process?

## 1.4 Related work

As to explore the realm of Explainable AI, Google Scholar was utilized as a resource. The terms 'XAI', 'medical sector' and 'diagnosis' yielded the most exploitable results for the initial search for general information concerning the methods which should be introduced

to the clinicians at a later stage of the research. Additionally, the experts consulted during this stage suggested relevant papers.

Generally speaking, there are a number of surveys which address the issue of giving an overview of the range of XAI methods used in the medical sector [21, 32, 39]. The literature review yielded that XAI methods can be split into post-hoc and ante-hoc (or explainable-by-design) methods. From the evaluation done in the surveys, post-hoc methods revealed themselves to be used the most. However, there is some mention of the usage of ante-hoc as well, which built the basis for the choice of including two of these models in the interview evaluation. Additionally, Stiglic and Kocbek [31] mention that there is an increased relevance of visual analytics and their interpretability in the sector, which shall further influence the research done from this point.

Another highly relevant study in this regard was conducted by Tonekaboni and Goldenberg [33]. They asked a range of clinicians from different fields what their understanding of explainability was and which features they would render most useful in this regard. It was found that the most important features are domain appropriateness, potential actionability, and consistency of the model. However, their research focused on the opinion of clinicians only. Arguably, it may take interdisciplinary communication between AI specialists and clinicians to find a common ground.

Furthermore, there appears to be the opinion amongst some scholars that ante-hoc methods may be far more applicable to be employed in the healthcare sector than post-hoc methods [5, 26]. The reasoning in these papers mainly points to the fact that post-hoc methods give the user a wrong sense of understanding through oversimplified and often faulty descriptions of the black box model. Furthermore, they point out that it is hard to find the source of an error in these systems, as the post-hoc explanation is a mere approximation of the actual system, whereas ante-hoc methods point to the actual variables used by the algorithm.

## 2 METHODOLOGY

The whole research was executed in three phases. The phase consisting of the literature research and expert consultation aimed to answer sub-question 1. The interviews and their evaluation then intended to answer sub-questions 2 and 3, respectively.

### 2.1 Literature research and Expert consultation

The first phase entailed researching relevant XAI techniques to gain a deeper understanding of the field. One common post-hoc method of XAI, one that is not as widely used, and two ante-hoc methods were chosen and then investigated. At first, relevant information for choosing these four methods was gathered in close consultation with field experts. This was done to ensure a basic level of effectiveness of the methods before interviewing the relevant stakeholders. For this reason, close contact was maintained with two experts performing research in the field of XAI and one expert who works with medical algorithms. One of the XAI researchers being the supervisor of this research enabled a close communication and weekly exchange of ideas and findings in meetings and text message channels. The other experts were consulted in an ad hoc fashion, as they were not directly involved in the project and thus not available to as

great of an extent. Overall, their knowledge helped tremendously to streamline the research process and incentivize perspective shifts throughout it. In the next step, the methods were chosen based on the previously gathered information and then researched for deeper understanding. Specific applications of the techniques in medical diagnosis were researched in this phase as well.

## 2.2 Interviews and Preparation

The second phase consisted of developing an effective explanation of the methods that can be understood by the pathologists. The interviews then aided to validate the methods by explaining them to a group of three clinicians first and discussing them with the participants immediately after the presentation. Interviews were chosen as a method to gain an understanding of the mental models of the participants [15]; for a deeper understanding of their needs in the setting. As Johs et al. [17] point out, semi-structured interviews are useful for researchers who do not have in-depth knowledge about a specific discipline or group and are inclined to get deep insights from their participants, without having to deal with the boundless qualities of unstructured interviews or the very delimited nature of structured interviews, for that matter. Therefore, making the interviews semi-structured seemed like a valid approach in this context.

The clinicians were picked in a manner such that both their backgrounds and their expertise with algorithms in diagnosis was held as diverse as possible. The group involved some people who had already implemented some of these algorithms themselves, and some who had assisted other researchers with doing so. Therefore, the domain in question was familiar to all of them, just to different extents. The group consisted of one trauma surgeon, one radiologist and one endocrinologist. A number of three participants was chosen due to the scope of the research and the given diversity in the group of participants. They were reached out to via email. It was decided not to interview people who only performed medical research but did not work as pathologists themselves, as they could hardly be classified as users of the to-be-evaluated system.

Before the actual interviews, the interview was practised with a lay participant to see whether the explanation would fit the time frame and how many questions could be asked within one hour, realistically. The slides and question sheet were then adapted accordingly. The participants were sent a short file consisting of one A4-page before the interviews that included a short introduction to the realm of XAI and to each of the methods that would be discussed. This was done to give them the chance to familiarize themselves with the subject before, if they had time and interest to do so.

After a brief introduction, the clinicians were informed about the research topic in general, and asked to fill in the informed consent form. Then, a short presentation of around 20 minutes covering all relevant methods was given to them. Subsequently, the methods were discussed in a semi-structured interview setting. Thus, some guiding questions were prepared, but there was room for follow-up questions to be asked if the flow of the conversation rendered it feasible. All the participants were shown the same examples, as their general medical education was considered sufficient for them to comprehend the methods in these contexts.

The interviews were held via Microsoft Teams and lasted around one hour each. They were recorded for further reference, but not transcribed. The risk of any unexpected happenings was considered very low. To ensure that the informed consent form consisted of all necessary questions, the self-assessment form of the UT ethics committee was consulted, and the research was conducted according to the standards defined in that form.

The interviews mainly measured trust and applicability (or usefulness) of the explanation techniques. Trust was measured, as it is a crucial concern in the context of automation [8, 13, 14, 25]. The questions asked in the interviews were mainly based on the work of Tonekaboni and Goldenberg [33], and Hoffman et al. [15]. Applicability was split into the dimensions of output goodness and satisfaction [15], as well as some factors that clinicians may find advantageous [33]. What should be mentioned is that most of these frameworks assume extensive use of the explanation method before filling in the questionnaire, which is not applicable in this case. Therefore, the questions from the relevant literature had to be rephrased to fit the interview setting of this research.

## 2.3 Evaluation and Results

For the third and final phase, an evaluation of the responses retrieved and drawing conclusions on the employability of the chosen techniques followed. Since the interviews were semi-structured and hence consisted of open questions only, the inter-participant consistency in the answers was low. For the results, interesting facts found during these conversations were pointed out, where every answer was considered as a novel finding on its own, except if the content of two or more answers was very similar.

## 3 CHOSEN METHODS

It was chosen to research two post-hoc methods and two ante-hoc methods in order to be able to compare the two approaches properly in terms of their effectiveness and trustworthiness.

Currently, interpretability in Deep Learning is mainly based on saliency maps [38]. However, high inter-observer variability significantly affects productivity in routine pathology and is especially present in medical centres with diagnostician deficiencies [40]. Other XAI methods may thus possibly introduce a higher level of consistency to this process than the saliency maps, as they give a more detailed interpretation of the outcome. Some scientists even dedicated a whole paper to why post-hoc methods may not be as applicable, even though they are predominantly used in the medical sector as of now [5, 26]. Therefore, the author deemed it interesting to both validate the methods which are used already, but also compare them to methods that are employed less, to see whether these may be even more effective.

Throughout the literature research, the focus was narrowed down to methods utilizing image data based on the recommendations of two of the XAI experts consulted in this research. They pointed out that images are the most commonly used data form in the medical field. Then, the methods were chosen based on the categories defined in the work of Salahuddin et al. [27], which gives a comprehensive overview of current interpretability methods in the field of XAI for medical image data. The survey focuses on deep neural networks,

as these have demonstrated state-of-the-art performance on many medical imaging challenges related to classification, segmentation and other tasks.

In order to ensure sufficient diversity within the range of methods, the author decided to choose two post-hoc methods that were based on different sub-categories within the range of saliency map-producing methods in the paper by Salahuddin et al. [27]. Grad-CAM was chosen based on an expert recommendation, stating that it was used frequently in practice. Similarly, LRP was chosen based on its general popularity as advertised in the literature [3, 12, 19], even though it was not as popular in the medical sector at the point of the research.

For the ante-hoc methods, because they were not used as much yet, the author decided to take two methods that provide the user with diverse outputs. Concept attribution methods were chosen since they emphasize feature importance, a part of the clinical decision-making procedure that is crucial to clinicians [33]. Language description methods, and specifically natural language description, was chosen since they facilitate another part of the diagnosis process, namely report-writing. As Jing et al. [16] point out, report writing can be error-prone to inexperienced clinicians and dreary to experienced ones.

In order to find visualizations for explaining the methods to the clinicians, relevant papers were searched for figures that would represent the output of the method for medical image analysis properly. For the post-hoc methods, the work by van der Velden et al. [35] was consulted. For finding relevant figures for the ante-hoc methods, the work by Salahuddin et al. [27] was used to search for examples.

In the end, each method was exemplified in the form of two cases with an according visualization taken from relevant papers. The images were chosen to be appealing and recognizable, as far as that was possible within the scope of this research. The chosen papers and the associated topics can be found in Table 1.

Table 1. Visualization examples used for the interviews with the clinicians.

Method	Papers	Medical case
Grad-CAM	[37], [24]	Brain tumour, COVID-19
LRP	[23], [9]	Paediatric pulmonary health, Multiple Sclerosis
Language-based methods	[40], [20]	Bladder cancer, breast cancer
Concept attribution methods	[2], [22]	Breast cancer, skin lesion identification

### 3.1 Post-hoc methods

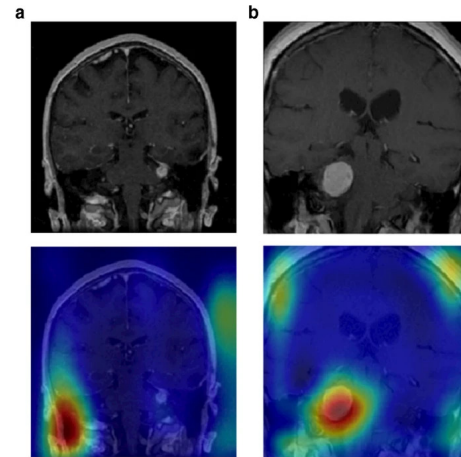


Fig. 1. Grad-CAM applied to slices for brain tumour diagnosis. Obtained from the work by Windisch et al. [37].

**3.1.1 Grad-CAM.** Gradient-weighted Class Activation Mapping (Grad-CAM) is an extension of the classic Class Activation Mapping (CAM) method. It is applicable to a wide variety of CNN model-families. It 'uses the gradients of any target concept (say 'dog' in a classification network or a sequence of words in a captioning network) flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept' [29].

In a nutshell, Grad-CAM uses the gradients of any target concept flowing into a convolutional layer (most often this is the final layer) to produce a bawdy localization map highlighting the relevant regions in the image for predicting the given concept [1]. Using Grad-CAM, it is possible to visually validate that the network is indeed looking at the relevant regions of an image and activating around them. The output of Grad-CAM is a heatmap visualization for a given class label [1]. An example on how to use Grad-CAM for an algorithm performing brain tumour classification can be found in Figure 1.

The advantage of Grad-CAM over traditional CAM is the possibility to retrieve relevance scores for any CNN-based differentiable structure, where CAM can not be applied to networks which use multiple fully-connected layers before the output layer [30].

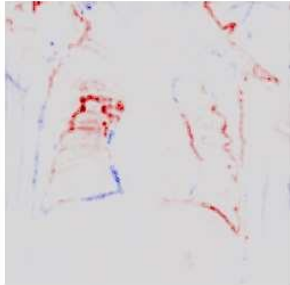
It should be noted that it is best practice to use Guided Grad-CAM to retrieve more fine-grained heat maps [4]. This entails a combination of Grad-CAM with Guided backpropagation. Grad-CAM in itself is quite useful, because it can easily identify classification mistakes due to high-resolution, class-discriminative visualizations [30].

For further reference, the reader is encouraged to consult the paper by Selvaraju et al. [29] in which this method is looked at in detail.

The method was chosen for the research because of its extensive use in medical image diagnosis, as visible in the amounts of examples mentioned in the survey paper by van der Velden et al. [35].



(a) Covid-19 Chest X-ray Image



(b) Covid-19 detection with LRP X-ray

Fig. 2. LRP applied to Covid-19 classification slices. © 2022 IEEE. Reprinted, with permission, from Ggaliwango Marvin and Md.Golam Rabiul Alam [23].

3.1.2 *LRP*. Layer-wise Relevance Propagation (LRP) was first introduced by Bach et al. [6]. It circumscribes a method of back-propagating over the layers of a convolutional neural network to find the most relevant pixels for the classification decision by the model. The activation number of the neurons in each layer determines their individual relevance, where the composition of the network then determines the relevance of the input pixels. The input forms the first layer of the network, where the output classification forms the last layer. The idea is that through a backpropagation over all these layers, the values computed on the way will determine the positive/negative influence of each neuron (or pixel, depending on what is analysed) in the very first layer. The sign before their relevance score determines whether they have positive or negative influence then.

LRP can be used for explaining algorithms with different kinds of visual inputs. It is a heatmap method that has been shown to perform better than previous techniques in terms of explainability and disease-specific evidence [9]

Identifying the relevance of each pixel is particularly useful in the medical sector because it allows for explaining individual classification decisions in a fast and intuitive way without the need for delving deeply into the network structure [9]. Therefore, it can easily be applied to analyses of outputs for physician training or discussion purposes. Furthermore, in heatmaps generated with LRP, group effects in the data occur less [7]. An example of the use of a method using LRP for the explanation of an algorithm performing diagnoses of pulmonary diseases (in this case, Covid-19, specifically) can be found in Figure 2.

To dive deeper into the LRP method, the reader is advised to consult the paper by Bach et al. [6].

In natural images, it has been argued that methods relying on gradients only measure the susceptibility of the output to changes in the input. The results may thus not necessarily correspond to the areas relevant for the network’s decision [7]. This arguably constitutes a clear reason to use LRP instead of gradient-based methods such as Grad-CAM. However, LRP’s quality to point to exact feature details naturally increases the file size of the output of these explanations. Since file size quantifies interpretability in image classification [28], it may thus be less comprehensible by humans than more fuzzy methods with smaller file sizes. Thus, Grad-CAM, which is more fuzzy than LRP by nature, may be more applicable in situations where high interpretability is needed, such as when trying to explain a diagnosis outcome to a patient.

The aforementioned arguments seemed convincing enough to choose this method as a second post-hoc method to discuss, even though it is used less in practice. Anyhow, whether the theoretically postulated advantages and disadvantages would also hold in an interview with practitioners seemed compelling enough to be explored.

### 3.2 Ante-hoc methods

Original ROI image	Visual justification by proposed method	Diagnosis, Margin, Shape	Proposed method	Without $\mathcal{L}_C$
		Benign, Obscured, Oval	<i>There is vague egglike structure overlapped by sharp line pattern.</i>	<i>The boundary of egglike mass is sharply demarcated with clear color transition.</i>
		Malignant, Obscured, Irregular	<i>A lot of sharp lines overlapped blur the complicated mass.</i>	<i>There is no clear demarcation all around the complicated mass form.</i>
		Benign, Obscured, Lobulated	<i>The outline of a bumpy mass is blurred from adjacent tissues.</i>	<i>There are vague lines projecting from inside of uneven mass.</i>
		Malignant, Spiculated, Irregular	<i>The contour of indefinitely formed mass is constructed by many sharp lines projecting outside of that mass.</i>	<i>A complex mass boundary is too blur to define where it is exactly.</i>

■ Correct visual word    ■ False visual word

Fig. 3. Textual and visual justification of a method proposed for breast cancer diagnosis. The textual justification of the proposed method is compared to and the method learned without a visual word constraint. Retrieved from the work of Lee et al. [20].

3.2.1 *Language Description Methods*. In these methods, a combination of both image data and corresponding textual data (either microscopical reports or medical reports) are analysed. Commonly, they provide an output that contains both a heat map and a computer-generated textual report. In order to make sure that the outputs are corresponding to the inputs, the concept of key ‘visual words’, so words which point to specific elements in the images, is used. Notably, the generation of natural language models is often made harder by the limited number of medical report data [20]. Therefore, some scholars use additional relevant sources for their textual data

to ensure a sufficient level of textual diversity [20, 36, 41]. Additional image segments that are key to determining the classification may be added to the output as well [40]. An example of the use of a method generating textual reports circumscribing breast cancer slices can be found in Figure 3.

For further reference, the reader is invited to consult the papers that mention the development of these techniques [20, 36, 40, 41].

A clear advantage of language-based methods is their quality to facilitate the process of medical report-writing [16]. Additionally, they provide accurate descriptions of subtle changes in tissue images that are challenging even for very experienced analysts [41]. Hence, they are very applicable in cases where the medical image itself is hard to understand, even with the help of a heat map. Of course, since medical reports are taken as input data, the interpreter needs to be able to read these structured texts to use the methods in the first place. They may thus be less applicable to situations where the outcome needs to be explained to a patient, as the doctor would probably have to rephrase the textual output in simpler terms.

The additional dimension of adding a computer-generated report led to picking this method as the first ante-hoc method. It provides the clinicians with an explanation of greater detail that is written in their own vocabulary, and thus seemed like a solid choice.

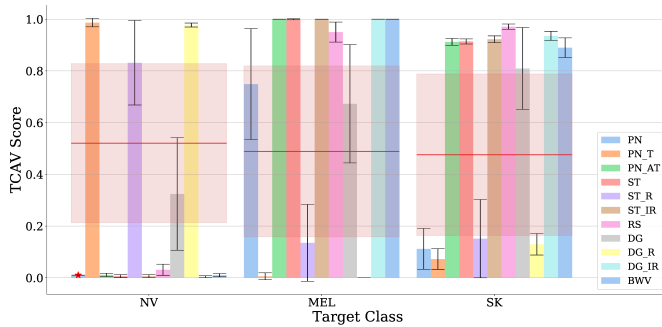


Fig. 4. TCAV scores for skin lesion concepts used in a method for skin cancer diagnosis. © 2020 IEEE. Reprinted, with permission, from Adriano Lucieri, Muhammad Naseer Bajwa, Stephan Alexander Braun, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed [22].

**3.2.2 Concept Attribution Methods.** Concept attribution methods are all based on a methodology created by Kim et al. called 'Testing with Concept Activation Vectors' (TCAV) [18]. The main idea behind this approach is that an algorithm uses certain features to exhibit the classification of every image. TCAV then takes a set of user-defined concepts and looks at how the probabilities for the outcome classification change if the concept in question (i.e. stripes for a Zebra) is removed. The output is a numerical score (TCAV score), that represents the number of pictures in which the classification was positively influenced by the concept from the dataset. This can be applied in the medical sector by using biomarkers as concepts such as the estrogen receptor and the progesterone receptor for breast cancer diagnosis [40]. An example of the use of TCAV scores in the context of skin lesion identification can be found in Figure 4.

For further information, the reader is invited to consult the paper in which the method is introduced [18], or some papers in which the method is applied to medical images [2, 22]. Some scholars took this method even further and applied so-called regression concept vectors to medical image diagnosis [10, 38].

The advantage of the TCAV score is that it effectively measures the ratio of the inputs of a specific class that are positively affected by a certain concept without taking any magnitude into account. As compared to saliency maps or other per-feature metrics, the TCAV score allows for quantitative evaluation of concepts on whole input classes [22]. Of course, the chosen concepts need to be understood by the target group of the explanation in the first place. In the case of patients, this may not always be the case, since the biomarker concepts are usually well-known in expert circles only. Furthermore, TCAV is a global explanation method [18], which makes the case-based explanation commonly needed when talking to a patient harder than with local explanation methods. Therefore, saliency maps (which are produced by LRP or Grad-CAM, for instance) or textual outputs may be more applicable in this case, as these are able to provide this local dimension. Another pitfall of the method is learning a meaningless CAV [18]. This case would render the test and output scores meaningless and produce redundant data only.

Based on the arguments made above, concept activation methods were added to the pool of the techniques investigated in this research. The fact that they give quantifiable evidence of the decisions made by the networks trained, and thus provide a different dimension of explanation for the interviews, manifested their relevance.

## 4 RESULTS

When analysing the interviews with the clinicians, a few overarching themes became apparent.

Foremost, there is no one-size-fits-all when it comes to trying to choose the right XAI method for pathologists in the process of medical image diagnosis.

Clinicians tend to use the explanations (no matter whether post-hoc or ante-hoc) to double-check whether the algorithm misclassified an image, or discuss the outcomes amongst themselves to see whether the method actually performs in the way they want it to perform. Usually, they like to test the algorithm extensively with real-life data that is somewhat dissimilar to the training data to see whether it performs consistently in all scenarios they would want to utilize it in. For these use cases, more elaborate explanations like language-based methods are very applicable, because they give deeper insight into what factors the algorithm based its decision on and use the same terminology as the clinicians themselves.

When trying to justify their decision to patients and their families, doctors tend to prefer heat maps since they are easy to comprehend. One participant mentioned that LRP would be particularly useful for hip fracture detection and its explanation, as it tends to point to contours rather than larger areas like Grad-CAM does. This would make the justification ever so much easier and induce more trust in the patients, because their feeling of understanding the diagnosis would be stronger. Based on this argument, the two post-hoc methods may be more applicable in this use case. However, it is worth mentioning that a lot of ante-hoc methods use heat

maps for their explanation as well. The simple difference is that the complex image sets these methods are applied to are usually hard to comprehend by non-experts, even through a heat map. It was mentioned that concept attribution methods are valuable for justification to patients as well, since they provide numeric outputs and charts.

In order to simplify their workflows, there are several ways in which each of the methods could be employed. It was mentioned that LRP could be used by clinicians for training visual recognition skills. Furthermore, language-based methods would lessen the burden of having to create a full report by themselves, as their output could be used as a basis. Concept attribution methods could, however, help with labelling past tumour data to categorize it more easily. The relevance of embedding any of the methods into the diagnosis process was seen as highly individual and dependent on the type of sickness.

If they think that something was misclassified, pathologists would usually tend to prefer language-based methods as a means of checking where the algorithm went wrong. This is due to the added layer of complexity in the textual output that makes it easier to discuss the explanation with other experts as well. Notably, concept attribution methods were also mentioned as a good method in this regard, because they provide quantifiable relevance for certain illness features. Additionally, they provide a rather unique combination of concept scores for each sickness, rendering them very useful when it comes to classifying sicknesses that are highly similar, naturally. The two post-hoc methods were considered to be useful for error-checking by a more visually prone participant.

The participants were also able to point out some issues that may arise with the methods. Concept attribution methods may be less applicable in the context of fractures, for instance, as these tend to have a binary existence quality in the diagnosis and would hence lack the multi-dimensionality required for this type of explanation method. They may also fail if they gave similar outputs for different sicknesses. For instance, if two concepts displayed very similar TCAV scores for three different sicknesses, these sicknesses would be hard to identify based on the TCAV scores only. More concepts would then have to be added to ensure the uniqueness of the classification explanation. Additionally, it was mentioned that language-based methods may be trained on either convoluted, or even highly personal reports. The former would render their outputs hardly comprehensible even for trained professionals, the latter may invalidate them altogether. The choice of training data would thus be a rather sensitive one in this context.

When it comes to trust, pathologists have mixed opinions about which methods would seem most reliable. Some prefer concept attribution methods as they give a more detailed insight as to which concepts were relevant for the decision-making process, some fancy a combination of LRP and language-based methods in this regard. Evidently, this decision is quite personal and depends both on the background of the pathologist and their individual preference for either visual or textual explanations, for instance. Interestingly, a participant that generally tended to prefer visually-oriented methods mentioned that they would prefer a language-based explanation in case of own uncertainty, as it appears more like the statement of another clinician and thus seems to induce trust from that angle.

Something that did reach high agreement amongst the participants was the fact that consistency and accuracy in the output induced trust, no matter which XAI method would be used. They collectively acknowledged that an added confidence value to the output would be helpful in that regard. This would facilitate the step from the differential diagnosis (all possible sicknesses based on the symptoms given) to the working diagnosis (choice of sickness from all possible ones based on clinicians own judgement), where confidence levels are highly relevant. That would also make it easier to justify their reasoning to other stakeholders, because they could point to the high numeric confidence score of the algorithm as a convincing argument.

## 5 DISCUSSION

Looking at the results, different XAI methods may be effective under varying circumstances. Whilst post-hoc methods are predominantly in use, there seems to be a general agreement amongst pathologists and XAI experts alike to prefer ante-hoc methods due to the added levels of accuracy and complexity. Due to these facts, they seem more trustworthy and simplify the discussion of results with other experts, for instance. However, saliency maps are still needed, at least to explain outcomes to stakeholders with less expert knowledge and for use by clinicians who prefer visual feedback, generally. Hence, combining both approaches in practise seems most applicable.

Furthermore, it is not only important to choose the right method for the right type of diagnosis, but train the algorithm extensively on a sufficiently diverse data set. Pathologists value edge case-centred design in this regard. Standardized data sets should be employed, to ensure the understandability and trust of clinicians in textual outputs, specifically.

### 5.1 Categorization based on interview results

Based on the outcomes of the qualitative data retrieved during the interviews, a rough scheme of when the methods could be best applied was developed. This can be found in Table 2. For each relevant use case, the applicability of the method was evaluated and only use cases specifically mentioned by the participants were chosen.

### 5.2 Limitations

The clearest restraint of this research is its sample size. Interviewing three clinicians only is by far not representative enough to make general statements about this population. Pathologists are an incredibly diverse group not only in terms of expertise in XAI, but in the dimensions of specialization and personal diagnosis style as well.

Furthermore, the amount and diversity of methods discussed could have been more vast. There are a number of other XAI methods that may just be as applicable or even more applicable than the ones discussed. However, both the time constraint of the research and the interview length did not allow for a greater number to be investigated.

Another limitation is the researcher's lack of prowess both in the field of XAI and in conducting interviews. A more experienced professional would probably have been more able to explain these

Table 2. Scheme for Method Application

Method	Use cases
Grad-CAM	expert-to-patient conversation, checking for errors
LRP	expert-to-patient conversation, facilitation of clinician workflows, checking for errors, inducing expert trust in the model
Language-based methods	expert-to-expert discussion, facilitation of clinician workflows, checking for errors, inducing expert trust in the model
Concept attribution methods	expert-to-patient conversation, facilitation of clinician workflows, checking for errors, inducing expert trust in the model

methods to the clinicians and retrieve results with greater insight from the interviews by steering the conversation in the right direction.

### 5.3 Future work

Future research endeavours may look to validate these methods with a larger group of stakeholders. More pathologists need to be consulted to achieve a more consistent overview of which points are relevant to these professionals across different specializations and countries. Furthermore, patients should be consulted for validation purposes as well, because they form a significant chunk of the medical XAI stakeholder group. Their trust is imperative when it comes to whether an expert system will be embedded in the medical practice or not. Additionally, findings like the peculiar applicability of LRP to hip fractures should be validated with patients, since they may only be interesting to the expert eyes of clinicians, after all.

Additionally, future work could look at explicability beyond the medical sector. The methods researched in this work can be applied to many fields and may have different use cases and requirements in each of them.

It may also be of interest to investigate confirmation bias in algorithmic medical diagnosis. After all, doctors are testing the algorithms on their own accord. They check whether it makes decisions based on the same factors as they do. However, the algorithm may discover some relevant points that were unknown to them before, which may be neglected through their methodology.

## 6 CONCLUSION

Noticeably, XAI methods assume high relevance in a vast range of sectors already. In a socio-technical system, their ability to induce trust in different kinds of stakeholder groups is evident from the literature. To substantiate this ability from a user's perspective, a closer look was taken at four of these methods, and their validity for applying them to algorithmic medical diagnosis was established

in this research. In order to address this issue, the following sub-questions were answered:

- (1) Which XAI techniques are relevant to be investigated for algorithms used in medical image diagnosis?

**Answer:** The four relevant techniques and the reasoning behind choosing them can be found in section 3. They were determined by the means of the literature research and the expert consultation in the beginning.

- (2) How can the techniques in (1) be explained to clinicians, and, immediately after, be tested on trust and usability?

**Answer:** The techniques were explained by means of a concise description enhanced with case examples in the beginning of the interviews. Then, a semi-structured interview was held to gain insight into the pathologists' perspectives.

- (3) Which of the techniques mentioned in (1) are able to maintain the clinicians' trust in the algorithm and effectively facilitate the diagnosis process?

**Answer:** Different methods are able to maintain trust under varying conditions, as elaborated upon in section 5. The same holds for the facilitation of the diagnosis process.

Finally, the answer to these questions lead to addressing the overarching research question:

### How can one ensure that an algorithm is sufficiently explicable to pathologists in medical image diagnosis?

Evidently, XAI methods are an effective way of to ensure explicability of algorithms to clinicians in medical image diagnosis. Every technique can contribute to this goal in its very own way, depending on the situational circumstances. Therefore, it is momentous to look into a sensible integration of these methods. Using a combination of several methods may be decisive to maximize their facilitation of the diagnosis workflow. If applied correctly, however, they are able to abate the workloads of even the most conservative of clinicians and enhance the lives of many. After all, a more efficient medical sector would not only enlarge the general quality of life within society, but help citizens to enjoy its benefits over ever-increasing time spans.

## ACKNOWLEDGMENTS

I want to thank both my supervisors, Maurice van Keulen and Xadya van Bruxvoort for all their support and answers to my questions throughout this research project. Furthermore, I want to express my gratitude to Meike Nauta for supporting me with my XAI research as well. Lastly, I want to thank all my interview participants, who were very open to discuss these methods with me and gave me great insight into the clinical perspective on the matter.

**Contact information** For further details concerning this research, please reach out to [j.furst@student.utwente.nl](mailto:j.furst@student.utwente.nl).

**Informed consent** Informed consent was obtained from all individual participants included in the research.

**Permission of use** Permission of use, if needed, was obtained for all those figures included in the paper that were taken from other academic sources and cited as such.



## REFERENCES

- [1] 2020. Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays. *Computer Methods and Programs in Biomedicine* 196 (2020), 105608. <https://doi.org/10.1016/j.cmpb.2020.105608>
- [2] 2021. Determining breast cancer biomarker status and associated morphological features using deep learning. *Communications Medicine* 1, 1 (2021), 1–12. <https://doi.org/10.1038/s43856-021-00013-3>
- [3] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box : A survey on Explainable Artificial Intelligence ( XAI ). *IEEE Access* PP, September (2018), 1. <https://doi.org/10.1109/ACCESS.2018.2870052>
- [4] Meghna P Ayyar, Jenny Benois-Pineau, and Akka Zemhari. 2021. Review of white box methods for explanations of convolutional neural networks in image classification tasks. *Journal of Electronic Imaging* 30, 5 (2021), 050901.
- [5] Boris Babic, Sara Gerke, Theodoros Evgeniou, and I Glenn Cohen. 2021. Beware explanations from AI in health care. *Science* 373, 6552 (2021), 284–286.
- [6] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* 10, 7 (2015), 1–46. <https://doi.org/10.1371/journal.pone.0130140>
- [7] Moritz Böhle, Fabian Eitel, Martin Weygandt, and Kerstin Ritter. 2019. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer’s disease classification. *Frontiers in Aging Neuroscience* 10, JUL (2019). <https://doi.org/10.3389/fnagi.2019.00194> arXiv:1903.07317
- [8] Eric T Chancey, James P Bliss, Alexandra B Proaps, and Poornima Madhavan. 2015. The role of trust as a mediator between system characteristics and response behaviors. *Human factors* 57, 6 (2015), 947–958.
- [9] Fabian Eitel, Emily Soehler, Judith Bellmann-Strobl, Alexander U. Brandt, Klemens Ruprecht, René M. Giess, Joseph Kuchling, Susanna Asseyer, Martin Weygandt, John Dylan Haynes, Michael Scheel, Friedemann Paul, and Kerstin Ritter. 2019. Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. *NeuroImage: Clinical* 24, June (2019), 102003. <https://doi.org/10.1016/j.nicl.2019.102003> arXiv:1904.08771
- [10] Mara Graziani, Vincent Andrearczyk, and Henning Müller. 2018. Regression concept vectors for bidirectional explanations in histopathology. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11038 LNCS (2018), 124–132. [https://doi.org/10.1007/978-3-030-02628-8\\_14](https://doi.org/10.1007/978-3-030-02628-8_14) arXiv:1904.04520
- [11] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *Comput. Surveys* 51, 5 (2018). <https://doi.org/10.1145/3236009> arXiv:1802.01933
- [12] Ji-Hoon Han, Sang-Uk Park, and Sun-Ki Hong. 2021. A Study on the Effectiveness of Current Data in Motor Mechanical Fault Diagnosis Using XAI. In *2021 24th International Conference on Electrical Machines and Systems (ICEMS)*. IEEE, 710–715.
- [13] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.
- [14] Robert R Hoffman, Matthew Johnson, Jeffrey M Bradshaw, and AI Underbrink. 2013. Trust in automation. *IEEE Intelligent Systems* 28, 1 (2013), 84–88.
- [15] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. (2018), 1–50. arXiv:1812.04608 <http://arxiv.org/abs/1812.04608>
- [16] Baoyu Jing, Pengtao Xie, and Eric P. Xing. 2018. On the automatic generation of medical imaging reports. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* 1 (2018), 2577–2586. <https://doi.org/10.18653/v1/p18-1240> arXiv:1711.08195
- [17] Adam J. Johs, Denise E. Agosto, and Rosina O. Weber. 2020. Qualitative Investigation in Explainable Artificial Intelligence: A Bit More Insight from Social Science. (2020). arXiv:2011.07130 <http://arxiv.org/abs/2011.07130>
- [18] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability beyond feature attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *35th International Conference on Machine Learning, ICML 2018* 6 (2018), 4186–4195. arXiv:1711.11279
- [19] Maximilian Kohlbrenner, Alexander Bauer, Shinichi Nakajima, Alexander Binder, Wojciech Samek, and Sebastian Lapuschkin. 2020. Towards best practice in explaining neural network decisions with LRP. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–7.
- [20] Hyebin Lee, Seong Tae Kim, and Yong Man Ro. 2019. *Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis*. Vol. 11797 LNCS. Springer International Publishing. 21–29 pages. [https://doi.org/10.1007/978-3-030-33850-3\\_3](https://doi.org/10.1007/978-3-030-33850-3_3) arXiv:1906.03922
- [21] Jörn Lötsch, Dario Kringel, and Alfred Ultsch. 2021. Explainable Artificial Intelligence (XAI) in Biomedicine: Making AI Decisions Trustworthy for Physicians and Patients. *BioMedinformatics* 2, 1 (2021), 1–17. <https://doi.org/10.3390/biomedinformatics2010001>
- [22] Adriano Lucieri, Muhammad Naseer Bajwa, Stephan Alexander Braun, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. 2020. On interpretability of deep learning based skin lesion classifiers using concept activation vectors. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–10.
- [23] Ggaliwango Marvin and Golam Rabiul Alam. 2022. Explainable Augmented Intelligence and Deep Transfer Learning for Pediatric Pulmonary Health Evaluation. February (2022), 26–27.
- [24] Harsh Panwar, P K Gupta, Mohammad Khubeb, and Ruben Morales-menendez. 2020. A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images. January (2020).
- [25] Vlad L Pop, Alex Shrewsbury, and Francis T Durso. 2015. Individual differences in the calibration of trust in automation. *Human factors* 57, 4 (2015), 545–556.
- [26] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215. <https://doi.org/10.1038/s42256-019-0048-x> arXiv:1811.10154
- [27] Zohaib Salahuddin, Henry C Woodruff, Avishek Chatterjee, and Philippe Lambin. 2022. Transparency of deep neural networks for medical image analysis : A review of interpretability methods. *Computers in Biology and Medicine* 140, October 2021 (2022), 105111. <https://doi.org/10.1016/j.combiomed.2021.105111>
- [28] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. 2021. Explaining deep neural networks and beyond: A review of methods and applications. *Proc. IEEE* 109, 3 (2021), 247–278.
- [29] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2020. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* 128, 2 (2020), 336–359. <https://doi.org/10.1007/s11263-019-01228-7> arXiv:1610.02391
- [30] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. 2016. Grad-CAM: Why did you say that? (2016), 1–4. arXiv:1611.07450 <http://arxiv.org/abs/1611.07450>
- [31] Gregor Stiglic and Primoz Kocbek. 2020. Interpretability of machine learning-based prediction models in healthcare. December 2019 (2020), 1–13. <https://doi.org/10.1002/widm.1379>
- [32] Erico Tjoa and Cuntai Guan. 2021. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems* 32, 11 (2021), 4793–4813. <https://doi.org/10.1109/TNNLS.2020.3027314> arXiv:1907.07374
- [33] Sana Tonekaboni and Anna Goldenberg. 2019. What Clinicians Want : Contextualizing Explainable Machine Learning for Clinical End Use. *MI* (2019), 1–21.
- [34] Xadya van Bruxvoort and Maurice van Keulen. 2021. Framework for assessing ethical aspects of algorithms and their encompassing socio-technical system. *Applied Sciences (Switzerland)* 11, 23 (2021). <https://doi.org/10.3390/app112311187>
- [35] Bas HM van der Velden, Hugo J Kuijff, Kenneth GA Gilhuijs, and Max A Viergever. 2022. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis* (2022), 102470.
- [36] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. 2018. TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-Rays. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2018), 9049–9058. <https://doi.org/10.1109/CVPR.2018.00943> arXiv:1801.04334
- [37] Paul Windisch, Pascal Weber, Christoph Fürweger, Felix Ehret, Markus Kufeld, Daniel Zwahlen, and Alexander Muacevic. 2020. Implementation of model explainability for a basic brain tumor detection using convolutional neural networks on MRI slices. *Neuroradiology* 62, 11 (2020), 1515–1518. <https://doi.org/10.1007/s00234-020-02465-1>
- [38] Hugo Yèche, Justin Harrison, and Tess Berthier. 2019. *UBS: A dimension-agnostic metric for concept vector interpretability applied to radiomics*. Vol. 11797 LNCS. Springer International Publishing. 12–20 pages. [https://doi.org/10.1007/978-3-030-33850-3\\_2](https://doi.org/10.1007/978-3-030-33850-3_2)
- [39] Yiming Zhang, Ying Weng, and Jonathan Lund. 2022. Applications of Explainable Artificial Intelligence in Diagnosis and Surgery. *Diagnostics* 12, 2 (2022). <https://doi.org/10.3390/diagnostics12020237>
- [40] Zizhao Zhang, Pingjun Chen, Mason McGough, Fuyong Xing, Chunbao Wang, Marilyn Bui, Yuanpu Xie, Manish Sapkota, Lei Cui, Jasreman Dhillon, et al. 2019. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nature Machine Intelligence* 1, 5 (2019), 236–245.
- [41] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. 2017. MDNet: A semantically and visually interpretable medical image diagnosis network. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* 2017-Janua (2017), 3549–3557. <https://doi.org/10.1109/CVPR.2017.378> arXiv:1707.02485