Explaining Negative Classifications of AI Models in Tumor Diagnosis

David A. Kelly[1] **Hana Chockler**[1] **Nathan Blake**[1,2]

[1]King's College London, UK
[2]University College London, UK

Abstract

Using AI models in healthcare is gaining popularity. To improve clinician confidence in the results of automated triage and to provide further information about the suggested diagnosis, an explanation produced by a separate post-hoc explainability tool often accompanies the classification of an AI model. If no abnormalities are detected, however, it is not clear what an explanation should be. A human clinician might be able to describe certain salient features of tumors that are not in scan, but existing Explainable AI (XAI) tools cannot do that, as they cannot point to features that are *absent* from the input. In this paper, we present a definition of and algorithm for providing *explanations of absence*; that is, explanations of negative classifications in the context of healthcare AI.
Our approach is rooted in the concept of explanations in actual causality. It uses the model as a black-box and is hence portable and works with proprietary models. Moreover, the computation is done in the preprocessing stage, based on the model and the dataset. During the execution, the algorithm only projects the precomputed explanation template on the current image.
We implemented this approach in a tool, NITO, and trialed it on a number of medical datasets to demonstrate its utility on the classification of solid tumors. We discuss the differences between the theoretical approach and the implementation in the domain of classifying solid tumors and address the additional complications posed by this domain. Finally, we discuss the assumptions we make in our algorithm and its possible extensions to explanations of absence for general image classifiers.

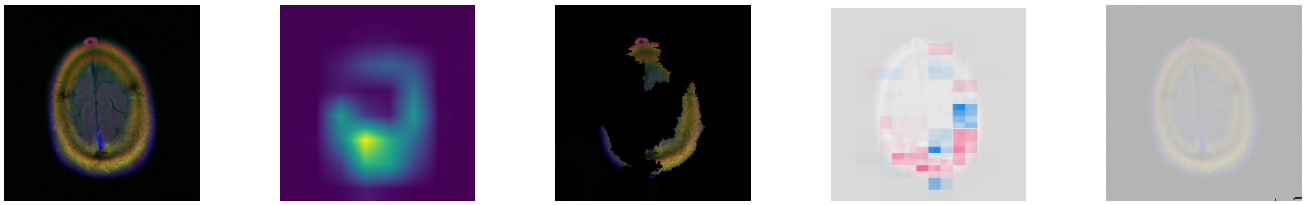
1 INTRODUCTION

There are a plethora of XAI tools which seek to provide an explanation for a label given to an image by a model. These often take the form of a heatmap (or a saliency landscape), which in various ways rank the contribution of the image pixels to a particular model output.

These XAI tools have in common that they seek features which are *local*, and *present* in the input. This is a reasonable strategy, as image classifiers in general domains do not output a classification of absence: if the only object in the image is a cat, the model outputs "cat", but if there is no cat, we would be surprised to get the classification "no cat"—a typical answer would be, for example, "dog", or "chair", depending on what is present in the input.

This is not true for medical imaging. A model for cancer detection in brain MRIs may learn to find features such as clusters of bright pixels, and/or the distortion of morphological features, indicative of certain pathologies. However, pertinent to medical imaging, there is a class of images which are clinically defined in terms of their *absence* of features: images without pathological abnormalities. Here, the classifications "disease" and "no disease" make sense in a way that "cat" and "no cat" do not. In this case, it is precisely the absence of features which defines a healthy scan.

Existing XAI tools are not designed for this task. As illustrated in Figure 1 on brain tumor detectors, in the absence of features the XAI tools typically return irrelevant explanations, even often including areas outside of the brain that are clearly not relevant or useful in providing clinicians with any insight into why the model has determined an image to be healthy. Explainability is crucial in the medical domain. The EU Artificial Intelligence Act makes transparency a regulatory requirement, which is described in terms of explainability: *"Transparency means that AI systems are developed and used in a way that allows appropriate traceability and explainability..."* [Madiega, 2021]. In the medical domain,



(a) Original image (b) GradCAM (c) LIME (d) SHAP (e) REX

Figure 1: A selection of explanations from popular XAI tools for a negative brain tumor classification. All explanations are for the same image with the same model. All displayed tools highlight parts of the image which are clinically irrelevant to a negative tumor classification. Indeed, there is no relevant region which could be highlighted.

this is applicable to *all* images, not just those exhibiting abnormalities. It is particularly important in terms of establishing trust with clinicians who remain legally liable for medical decision making: it is reasonable that they ask for transparent decision making not only for diseased images but also images labeled as healthy [Naik et al., 2022].

There is therefore a gap in the XAI literature addressing the absence of features for explainability. In this paper, we propose an approximation algorithm for constructing explanations of absence based on the formal definition of explanation in the theory of actual causality Chockler and Halpern [2024]. Moreover, by using the definition of partial explanations, we adopt a measure by which the quality of our explanation of absence can be automatically assessed. Our approach uses the model as a black-box and is hence portable and applicable to any, even proprietary, models.

Our algorithm constructs a template for the explanations of absence in the preprocessing stage, based on the model and the dataset. This is done once for a given model and a dataset. Then, during the execution, we project the precomputed template on the current image classified as not having the abnormalities in question. We implemented this approach in a tool NITO[1] and trialed it on a number of medical datasets to demonstrate its utility on the classification of solid tumors. We note that the actual execution step consists of a simple projection, hence does not require any additional computation time or other resources on top of the classifier. We discuss the differences between the theoretical approach and the implementation in the domain of classifying solid tumors and address the additional complications posed by this domain. We then apply the theory of partial explanations to provide a means to automatically quantify the *goodness* of our explanation with respect to a user-provided dataset and show that NITO's explanations have a high ($> 85\%$) sufficiency.

Why do we focus on causal explanations? Causal explanations have the advantage of being based on a rigorous definition that, in particular, ensures minimality and sufficiency for the the desired classification (see Section 3). In our domain of application, this means that in an explanation of a tumor, removing any subset of pixels results in the set of pixels no longer being classified as a tumor. For all intents and purposes this subset of pixels *is* a tumor, with respect the model's decision process. We use this feature in our construction of explanations of absence. Roughly speaking, an explanation of absence of a tumor is a subset of pixels that does not admit this minimal tumor into the image. We output such a subset, explaining why the model decided that no tumor is present.

We implemented our algorithm and quantitative assessment measures and present the experimental results on three different medical datasets: brain tumor MRI, pancreatic cancer, and lung cancer CT images. To the best of our knowledge, there is no baseline to compare against for computing explanations of absence, therefore our experimental effort focuses on the computability and flexibility of our algorithm.

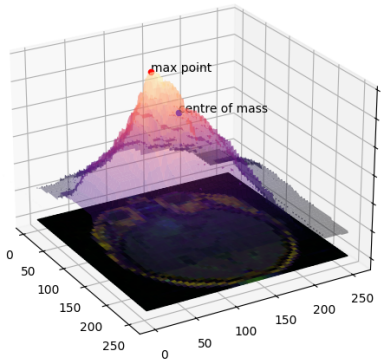
Due to the lack of space, additional theoretical material, additional results, and illustrations are deferred to the appendix. All data and code for reproducibility can be found at `https://figshare.com/s/d3143215218cb2b854af`.

2 RELATED WORK

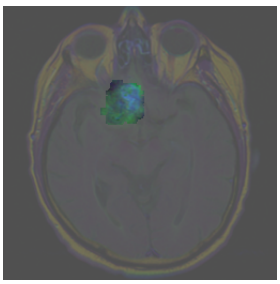
The landscape of XAI tools is large and complex, and each tool is guided by its own definition of explanation, more or less rigorous. At present, there are no post-hoc XAI tools adapted for the absence of features: all tools make the reasonable assumption that the target to be explained is present in the image. For the positive classification use cases, this assumption causes no problems. As we mentioned above, however, this assumption of presence is not very helpful to explain to the clinician why, for instance, a model says that an MRI slice of a brain contains no tumors.

Common XAI tools for medical images include GRADCAM [Selvaraju et al., 2017], LIME [Ribeiro et al., 2016] and SHAP [Lundberg and Lee, 2017]. SHAP adopts a game theoretic approach to find coalitions (subsets) of the image which, by some measure, contribute to a model returning

[1]From the Yiddish phrase "nit do" for "not here".



(a) Responsibility map for a class 1 MRI



(b) The minimal, sufficient pixels to achieve class 1. This is a REX explanation.

Figure 2: Typical output from REX, showing the raw responsibility map in Figure 2a and the extracted explanation in Figure 2b. The explanation is left as original color, with other pixels partially masked out.

a label. When superimposed upon the original image, this draws the eye towards features in that image that can be said to explain the model's output. As Watson et al. [2022] point out, Shapley values are the closest to a de facto standard for XAI, but ambiguities and assumptions [Kumar et al., 2020] muddy the waters of interpretation.

The theory of actual causality presents a precise definition of explanations for image classifiers [Chockler and Halpern, 2024]. REX [Chockler et al., 2024] is a causal explainability tool computing causal approximately minimal explanations. It employs causal reasoning to identify subsets of pixels which are sufficient to reproduce the overall model classification. Unlike more familiar XAI tools, pixels are ranked and also tested for sufficiency against the model itself as oracle. REX has the same limitation as the above mentioned tools in that it assumes the presence of features that can be occluded in some way and that explanations are *local*. Given the underlying theory is the same, it is convenient to utilize the output of REX in our implementation (Figure 2).

Contrastive explanations [Stepin et al., 2021, Chin-Parker and Bradner, 2017] give explanations not in terms of "Why did P happen?", but rather "Why did P happen and not Q?". Human-provided explanations seem to typically be contrastive [Miller, 2019]. Dhurandhar et al. [2018] use pertinent negatives to provide counterfactual explanations for multi-class models. They provide contrastive explanations for MNIST and other common datasets. Their approach, however, still requires *some* features to be present in order to demonstrate that other features are missing. They do not attempt to cover the case where there is a total absence of features required for a given classification. Dhurandhar et al. [2019] extend their approach to use contrastive explanations on structured data.

It has been suggested that large language models (LLMs) provide intrinsic explainability [Kroeger et al., 2023]. For instance, Med-Gemini-M 1.5, a LLM for medical data, can take an image input and return a text output [Saab et al., 2024]. This is similar to how reports are given in the medical domain, in which clinicians describe pertinent anatomical features. Ostensibly, this includes healthy images, providing an explanation of absence in a way consistent with current practice. However, LLMs are known to hallucinate and even if the end result is correct (*i.e.*no disease), there is no reason to believe their "reasoning", in the form of text, aligns with relevant clinical features (or their absence).

3 BACKGROUND ON ACTUAL CAUSALITY

While the need for explanations is recognized almost universally, there is no definition of explanation even close to universal acceptance [Miller, 2019]. We use a definition provided by the theory of actual causality. This definition has a number of useful properties which we use in our method. Actual causality was first introduced in Halpern and Pearl [2005]. The reader is referred to that paper and to Halpern [2019] for an updated overview and more information on actual causality (see also the supplementary material for the formal definition of explanation in the general case). Below we give an informal introduction to the theory and simplified definitions suitable for the case of image classification. The definition of an *actual cause* is based on the concept of *causal models*, which consist of a set of variables, a range of each variable, and structural equations describing the dependencies between the variables. Actual causes are defined with respect to a given causal model, a given assignment to the variables of the model (a context), and a propositional formula that holds in the model in this context.

Actual causality extends simple counterfactual reasoning Hume [1739] by considering the effect of *interventions*, which are changes of the current setting. Roughly speaking, a subset of variables $\vec{X}$ and their values in a given context is an actual cause of a Boolean formula φ being True if there exists a change in the values of other values that cre-

ates a counterfactual dependency between the values of $\vec{X}$ and φ (that is, if we change the values of variables in $\vec{X}$, φ would be falsified). The formal definition by Halpern and Pearl [2005] and in its modifications, the latest of which is by Halpern [2015], are far more complex due to the potential dependencies between the variables and considering causes of more than one element. In our setup, where we are only interested in singleton causes and in interventions only on the input variables, all versions of the definition of (a part of) an actual cause are equivalent under the assumption of independence between the input variables. This assumption is far from trivial, and we discuss its implications in Section 7.

In the context of image classification, following Chockler and Halpern [2024], we take endogenous variables to be the set $\vec{V}$ of pixels that the image classifier gets as input, together with an output variable that we call O. The variable $V_i \in \vec{V}$ describes the color and intensity of pixel i; its value is determined by the exogenous variables. The equation for O determines the output of the neural network as a function of the pixel values. As mentioned above, we assume that there are no dependencies between the feature variables, thus, the causal network has depth 2. While, in general, this assumption is not true in practice, in the context of MRI and CT scans it is reasonably accurate, as tumors can appear in most parts of an affected organ. Assuming independence makes the algorithms much simpler.

Chockler and Halpern [2024] proved that for a causal model corresponding to an image classifier $\mathcal{N}$, the following definition is equivalent to the definition of explanation in actual causality.

Definition 1 (Explanation) $\vec{X} = \vec{x}$ *is an* explanation *of* $O = o$ *iff the following conditions hold:*

EX1 *Setting* $\vec{X}$ *to* $\vec{x}$ *results in the classification* $O = o$ *for all images in the dataset;*

EX2 *For all images* $\mathcal{I}$ *in which* $\vec{X} = \vec{x}$ *and* $O = o$*, at least one conjunct* $X = x$ *in* $\vec{X} = \vec{x}$ *is a (part of) an actual cause of* $O = o$*; in other words, there exists a (possibly empty) set of variables* $\vec{Y}$*, a value* x'*, and a set of values* $\vec{y}'$ *such that setting* X *to* x' *together with setting* $\vec{Y}$ *to* $\vec{y}'$ *results in* $O \neq o$*;*

EX3 $\vec{X}$ *is minimal, that is, no subset of* $\vec{X}$ *satisfies the conditions above.*

"Folded" in Definition 1 is the definition of an *actual cause* of $O = o$, which, using the notation in EX2, would be $(\{X\} \cup \vec{Y} = \{x\} \cup \vec{y})$. The notion of *responsibility* quantifies actual causality and is defined for $X = x$ as above as $1/(|\vec{Y}| + 1)$, where $\vec{Y}$ is the smallest set satisfying EX2.

To facilitate a dialog between the clinician and the AI system, we also use the definition of a *partial explanation* for image classifiers by Chockler and Halpern [2024].

Definition 2 *[Partial Explanation]* $\vec{X} = \vec{x}$ *is a* partial explanation *of* $O = o$ *with goodness* (α, β)*, where* $\alpha, \beta > 0$*, relative to a set of images* $\mathcal{K}$ *if the following conditions hold:*

PEX1 *setting* $\vec{X}$ *to* $\vec{x}$ *results in the classification* $O = o$ *for all images in the dataset with probability at least* β*;*

PEX2 *the probability of* $\vec{X} = \vec{x}$ *to be a (part of an) actual cause of* $O = o$ *in an image* $\mathcal{I}$ *in the dataset is at least* α*;*

PEX3 $\vec{X}$ *is minimal.*

4 EXPLANATIONS OF ABSENCE

Consider an AI model $\mathcal{N}$ that classifies medical images as having or not having solid tumors. We start with a theoretical analysis with simplifying assumptions and then discuss whether these assumptions hold for real AI models and the implications of relaxing them.

4.1 THEORETICAL FOUNDATIONS

Recall that we assume independence between the pixels of the image. We now add the assumption that tumors are equally likely in all areas of the scan.

Lemma 1 *Under the assumptions above,* $C_{\mathcal{N}}$ *can only detect tumors based on the number of pixels with values (color and intensity) matching those of tumors, that is, the size of a potential tumor on an image.*

Proof. The proof is based on the observation that due to the assumptions, the effect of changing each pixel in an input image is the same. Hence, $\mathcal{N}$'s decisions rely only on the number of the pixels with values matching those of tumors, that is, the size of a potential tumor. ∎

The following lemma explains why responsibility maps are useless for explanations of absence, as illustrated in Figure 1. As only REX uses a formal definition of *responsibility* for its pixel ranking map, we use this definition in the lemma.

similar.

Lemma 2 *If none of the pixels in an input image* $\mathcal{I}$ *have values consistent with a tumor, the responsibility of each pixel of an input image* $\mathcal{I}$ *for the negative classification of* $\mathcal{N}$ *is the same and is equal to* $1/k$*, where* k *is the size of a smallest tumor recognized by* $\mathcal{N}$*.*

Proof. The proof is based on the observation above that under our simplifying assumptions, $\mathcal{N}$ can only use the size of a candidate tumor to decide whether to classify $\mathcal{I}$ as having a tumor. Hence, by EX2 of Definition 1, the responsibility of each pixel $(X = x)$ for the negative classification of $\mathcal{I}$ by

$\mathcal{N}$ is the same and is $1/k$, where k is the size of a smallest set of pixels required to change the negative classification to a positive one (aka "there is a tumor"). ∎

Corollary 1 *The responsibility of all pixels of an input image $\mathcal{I}$ classified as having no tumors for its classification is not an informative measure for explaining the classification.*

Based on Lemma 1, the following construct is an explanation of absence of tumors in an input image $\mathcal{I}$ according to Definition 1.

Definition 3 *[Absence grid] For an image $\mathcal{I}$ classified as not having tumors by an AI model $\mathcal{N}$, a subset of pixels $\vec{G} \subseteq \mathcal{I}$ and their values $\vec{G} = \vec{g}$, is an* absence grid *for $\mathcal{I}$ and $\mathcal{N}$ if:*

AG1 *$\vec{G} = \vec{g}$ is a grid of clusters of pixels;*

AG2 *The distance between any two clusters in $\vec{G}$ is smaller than the size of a smallest tumor recognized by $\mathcal{N}$;*

There exists a cluster of pixels $C \subseteq \mathcal{I}$ in the explanation that is (a part of) an actual cause of classifying $\mathcal{I}$ as not having tumors; that is, all pixels in C have the value that is incompatible with being a part of a tumor, and there exists another set of pixels $T \subseteq \mathcal{I}$ such that changing the values of $C \cup T$ changes the classification of $\mathcal{I}$ to having a tumor, but changing the values of C alone does not change the classification (of "no tumor").

AG3 *$\vec{G} = \vec{g}$ is minimal.*

It is easy to see that an absence grid is an explanation for the negative classification of I by $\mathcal{N}$, according to Definition 1. We also note that the location of the pixels on an absence grid as defined in Definition 3 depends only on $\mathcal{N}$ and is independent of $\mathcal{I}$ and of the dataset. The only thing that depends on $\mathcal{I}$ is the values of these pixels. Therefore, an absence grid can be constructed *in advance* and projected on a given image $\mathcal{I}$ to get an explanation of absence of tumors.

4.2 APPLYING THE THEORY TO PRACTICE

In practice, the assumption that solid tumors are equally likely in all areas on the scan does not quite hold, as tumors are more likely to appear in some areas than in others. Moreover, tumors might be non-homogeneous, which makes it harder to measure their size. Indeed, while the assumption of independence of pixels is a good approximation in this domain, an AI model $\mathcal{N}$ might also take into account an outline of a suspected tumor, rather than just its size, to decide whether there is a tumor on the scan.

An absence grid defined in Definition 3 is, thus, impossible to construct precisely; in particular, the size of a smallest tumor may depend on the location on the scan and its shape. We therefore construct an approximation of this grid instead, as defined below.

Definition 4 *[Partial Absence Grid] For an image $\mathcal{I}$ classified as not having tumors by an AI model $\mathcal{N}$ and a dataset $\mathcal{K}$, a subset of pixels $\vec{G} \subseteq \mathcal{I}$ and their values, $\vec{G} = \vec{g}$ is a* partial absence grid *with goodness (α, β), where $\alpha, \beta > 0$ for $\mathcal{I}$ and $\mathcal{N}$ in context $\mathcal{K}$ if:*

PAG1 *$\vec{G} = \vec{g}$ is a grid of clusters of pixels, such that all pixels have values incompatible with tumors;*

PAG2 *The distance between any two clusters in $\vec{G}$ is smaller than the smallest explanation, $\vec{X} = \vec{x}$, of a tumor in the set $\mathcal{K}$, recognized by $\mathcal{N}$,*

There exists a cluster of pixels $C \subseteq \mathcal{I}$ in the explanation that is (a part of) an actual cause of classifying $\mathcal{I}$ as not having tumors; that is, all pixels in C have the value that is incompatible with being a part of a tumor, and there exists another set of pixels $T \subseteq \mathcal{I}$ such that changing the values of $C \cup T$ changes the classification of $\mathcal{I}$ to having a tumor, but changing the values of C alone does not change the classification (of "no tumor").

PAG3 *$\vec{G} = \vec{g}$ is minimal.*

Note that, in particular, all pixels in $\vec{G}$ must have values inconsistent with a tumor. This may take the form of healthy value interpolation, or a neutral interpolation value. We discuss this choice in more detail in Section 6 and compare the results against taking an out-of-distribution neutral value.

For a partial absence grid $\vec{G} = \vec{g}$ as defined in Definition 4, let $0 \leq \alpha, \beta \leq 1$ be such that:

- probability of $\vec{X} = \vec{x}$ to be a (part of an) actual cause of $O = o$ in an image $\mathcal{I}$ in the dataset is at least α;
- setting $\vec{X}$ to $\vec{x}$ results in the classification $O = o$ for all images in the dataset with probability at least β.

We are now ready to state our main result.

Theorem 1 *A partial absence grid is a partial explanation. That is, for an image $\mathcal{I}$ classified as not having tumors by an AI model $\mathcal{N}$ and a dataset $\mathcal{K}$, the partial absence grid $\vec{G} = \vec{g}$ in Definition 4 is a* partial explanation of absence of tumors *in $\mathcal{I}$ wrt $\mathcal{N}$ and a set of contexts $\mathcal{K}$ with (α, β)-goodness, where α and β are as defined above.*

The proof follows from Definition 2.

Theorem 1 allows us to quantitatively assess the quality of explanations of absence, constructed as in Definition 4. It is important to note that a partial absence grid is always defined with respect to a particular model and dataset: a

Algorithm 1 NITO$(x, \mathcal{N}, U, \delta, r, Pr)$

INPUT: input image x, model $\mathcal{N}$, set of images U, density δ, radius r, *Pr* (a probability distribution over U)
OUTPUT: a gridded image x', a tuple (α, β)

1: $\varphi \leftarrow \mathcal{N}(x)$
2: $\mathcal{K} \leftarrow$ neighborhood of $x \in U$
3: **if** $\mathcal{K} = \varnothing$ **then**
4: **return** $x, 1, 0$
5: **end if**
6: $\mathcal{E} \leftarrow$ causal explanations of $\mathcal{K}$
7: $e_{\mathcal{E}} \leftarrow$ find smallest explanation $\in \mathcal{E}$
8: grid $\leftarrow$ calculate_grid$(e_{\mathcal{E}}, \delta, r)$
9: **if** grid $= \varnothing$ **then**
10: **return** $x, 1, 0$
11: **end if**
12: $\beta \leftarrow 0.0$
13: **for** $s \in \mathcal{K}$ **do**
14: $o \leftarrow \mathcal{N}(s)$
15: $o' \leftarrow \mathcal{N}(grid(s))$
16: $o \neq o' \; ? \; : \beta = \beta + Pr(s)$
17: **end for**
18: **return** grid$(x), (1, \beta)$

bad model and/or a poor dataset would produce low-quality explanations. However, with the definitions provided above, we emphasize that there is no sense in which an explanation can be "wrong": a human might disagree with the output, but a causal explanation is sufficient to reproduce the original class. In Section 7 we discuss how these explanations can support a dialog between an AI model and a clinician.

5 NITO

Algorithm 1 is the main NITO algorithm. The set $\mathcal{U}$ is the full dataset, containing scans from different slices of the organ under examination. For a given input x (a single image), we define $\mathcal{K}$ as all scans *similar* to x. The similarity can be defined parametrically; for this implementation, we define it as images of the same slice, for example the same part of the brain for MRIs of brains. The probability distribution over $\mathcal{K}$ is a parameter and is assumed uniform by default. As the dataset grows with each new examined patient, the set $\mathcal{U}$ grows as well, with more similar scans to the current input. In the future, the concept of similarity can be extended to include external parameters affecting the scans, such as biological sex, age, etc. We note that if $\mathcal{K}$ is empty, we cannot say anything useful about the current input image x. In what follows, we assume class 0 to mean "no tumor", and class 1 to mean "tumor".

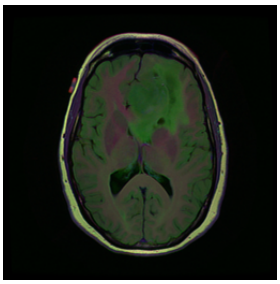
The algorithm first calculates an approximation of a smallest explanation *for a tumor* in $\mathcal{K}$ by executing REX (see Appendix B for an overview of REX). Note that an explanation of an object is typically smaller than the object itself, representing just a part of it [Chockler et al., 2024]: if an explanation of a tumor cannot be made to fit between the grid then the tumor itself certainly cannot. Note that we assume a reasonably regular shape of explanations: it is possible in theory to have tumors that fit neatly between the nodes of the grid, thus avoiding detection, while being overall quite large, however in practice tumor shapes are reasonably close to convex.

Size and color of the grid The procedure on line 8 of Algorithm 1, calculate_grid$(x, e_{\mathcal{E}}, \delta, r)$, receives the input image x, a smallest explanation $e_{\mathcal{E}}$ of a tumor in the set $\mathcal{K}$, and optionally the density δ and radius r, and constructs a grid with each node being of radius r and the distance δ between the nodes over the image x. If δ and r are not given, the procedure calculates them using a binary search, constructing the *least dense* grid that satisfies Definition 4. We assume convexity of explanations, hence the calculation of δ is based on the size of a minimal bounding box for $e_{\mathcal{E}}$. The radius r of each node is determined as the smallest that satisfies PAG2 of Definition 4 and is, again, determined using a binary search.

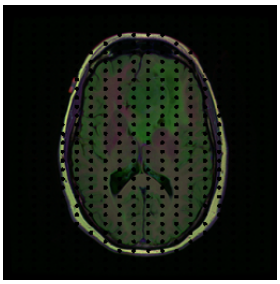
We also allow the option of receiving δ and r from the user, in case a user with domain knowledge is able to provide hints to the algorithm. An example is shown in Figure 4. The grid pixel values should be the same as in x, but in practice this often does not provide enough contrast, due to the quality and color scheme of the image. We therefore also allow the option of having out-of-distribution neutral values, and we compare these approaches in Section 6.

Calculating α and β The probability that an image in $\mathcal{K}$ has both $\vec{X} = \vec{x} \wedge \varphi$ is given by α. For simplicity, let us say that φ means "tumor". $\vec{X} = \vec{x}$ then will be a set of pixels and their values. Even if we ignore the spatial dimension of of pixels, the chances of $\vec{X} = \vec{x}$ being the same in a different image for which φ holds is very slim. Explanations are rarely identical between images. If we take into account the spatial dimension as well, then α is likely to be near 0 on any dataset. If we are interested in $\vec{G} = \vec{g}$, our absence grid, then we can approximately control the value of α if either $\mathcal{K}$ does not contain any healthy slices (in which case $\alpha = 1$), or we use a neutral, constant, masking value for the grid. In either case, α does not reveal as much information about our grid as we would like.

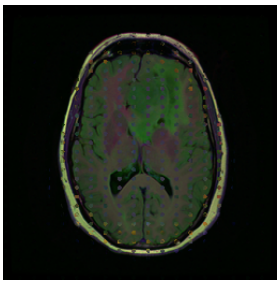
We present a simple visual example of a β-goodness calculation (Figure 6), with the absence grid only partially applied for clarity. Causal explanations, by virtue of their minimality, tend not to be robust. Hence, the explanation of the smallest tumor is itself a *partial* explanation. Future work will examine the usefulness of explanation *vs.* partial explanation of tumor on grid calculation from the point of



(a) An MRI slice with a tumor

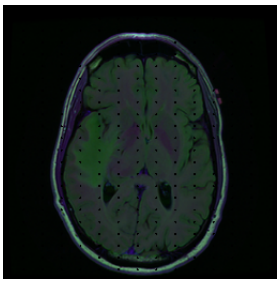


(b) The grid in black

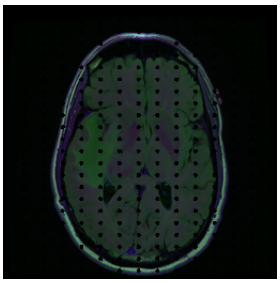


(c) A grid in real colors

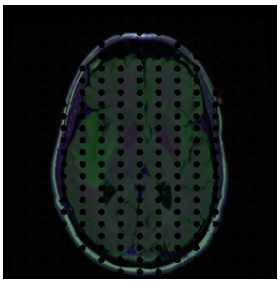
Figure 3: A MRI slice with tumor (upper right quadrant of brain) with two grid overlays: with nodes in black Figure 3b, and in the original color of the image Figure 3c. Note that the grid does not respect physical features of the brain. In practice, having the grid in the original colors of the image does not provide sufficient contrast to change the classification, whereas having the grid in black (or other high-contrast color) does.



(a) Insufficiently dense grid for class 0



(b) A denser grid, but still insufficient



(c) A sufficient grid for class 0

Figure 4: The grid is refined by search (here shown over an entire slice for clarity). The grids in Figure 4a and Figure 4b are insufficiently dense to change the class of the image from 1 to 0. The final grid, sufficient to change the class, is in Figure 4c.

view of the clinician.

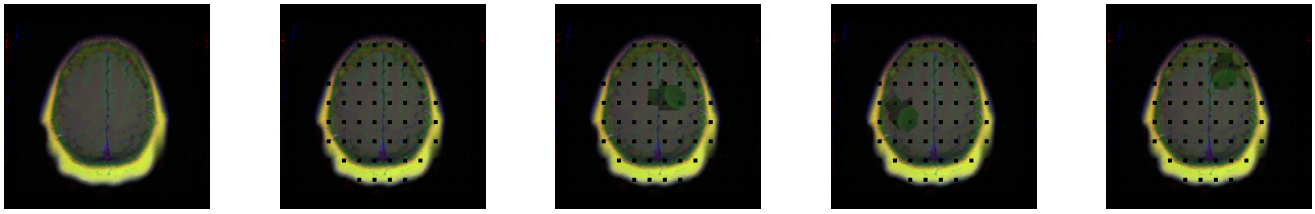
Quantitative evaluation of partial explanations The α value is the fraction of images in $\mathcal{K}$ classified as "no tumor" and for which $\vec{X} = \vec{x}$, represented as probability. In our experiments, there was only one healthy image per slice (the explanandum), thus $\alpha = 1$. We discuss general computation of α in Section 5. β measures the effect of setting $\vec{G}$ to $\vec{g}$ in $\mathcal{K}$. For example, let $|\mathcal{K}| = 5$, where $\vec{G} = \vec{g}$ is an absence grid for 3 out of all 5 images. Assuming a uniform distribution over these images, we have $\beta = 0.6$, as 3 images change their classification under $\vec{G} \leftarrow \vec{g}$.

The definition of β assumes equal weight for all elements in $\mathcal{K}$. However, given that $\mathcal{N}$ is an AI model, we can weight the calculation by the model's confidence in the classification. We consider two options for this weighting: β_a denotes the parameter weighted by the confidence over all classifications in $\mathcal{K}$ when $\vec{G} \leftarrow \vec{g}$, and β_p considers only the confidence of images where changing $\vec{G}$ to $\vec{g}$ does not change the classification (otherwise 1).

Dialog with a clinician Given a stable, unchanging, dataset, the grid is entirely computable off-line. Changes to the dataset mean that the grid needs to be recomputed for the appropriate slices only. This is unlikely to occur during deployment. As a result, NITO supports a dialog with a clinician in real time. Figure 7 illustrates the following scenario. An input MRI slice is classified as class 0 by the model (Figure 7a) but with relatively low confidence. A clinician might choose this slice for closer examination. There is an area, highlighted in red (Figure 7b), of slightly increased density. The clinician can take the grid for the appropriate slice and superimpose it over the suspicious area. The suspicious area fits reasonably neatly inside the grid (Figure 7c). As the grid was calculated from the smallest explanation in the dataset at that slice, this indicates that, from the model's perspective, there is no evidence of a tumor present. Of course, a clinician may disagree, in which case there is an argument for retraining or refining the model by adding this image to the training data.

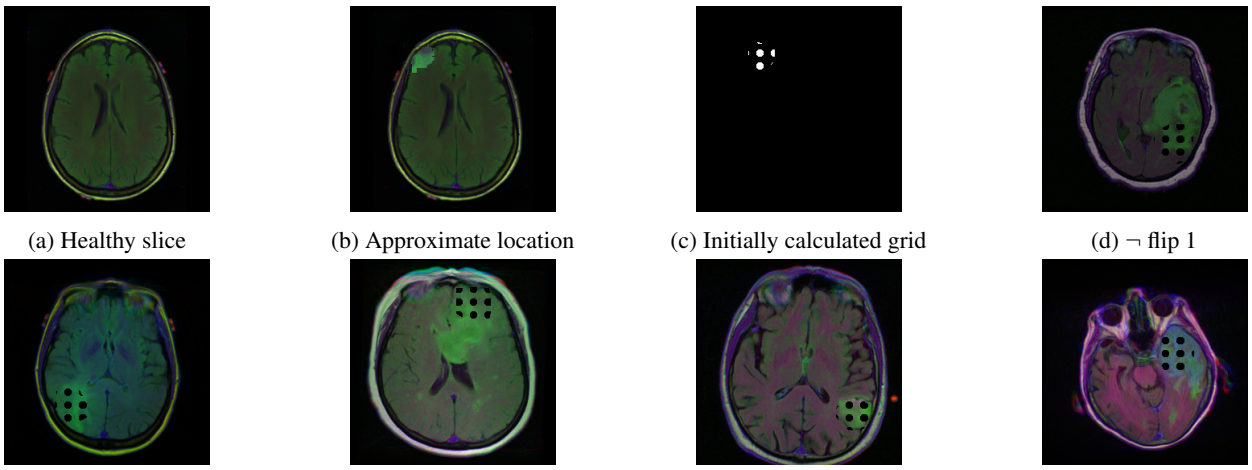
6 EXPERIMENTAL RESULTS

We implemented the NITO algorithm and present the results of evaluation of the explanations of absence produced by our implementation on three different models over three different, publicly available, datasets: one MRI and two CT scans. While there is nothing in our definition that is model or dataset dependent, we are interested to see both the variance



(a) class 0 (b) absence grid (c) excluded tumor (d) excluded tumor (e) excluded tumor

Figure 5: The grid prevents an explanation of a tumor from fitting into the image. As a causal explanation is approximately minimal and is recognized as class 1 by the model, this means that, modulo the dataset, no tumor can be present in the image. We cannot lose parts of the explanation to make it fit, as a smaller collection of pixels will no longer be recognized as tumor by the model.



(a) Healthy slice (b) Approximate location (c) Initially calculated grid (d) $\neg$ flip 1

(e) Flip 1 (f) $\neg$ flip 2 (g) Flip 2 (h) Flip 3

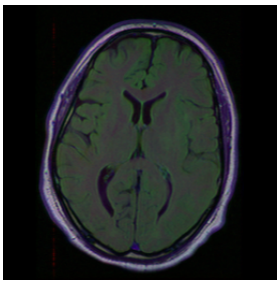
Figure 6: A example of a grid, approximate location of the explanation superimposed on the healthy brain (Figure 6b) and $\mathcal{K}$. $\mathcal{K}$ consists of Figures 6a and 6d to 6h. In this example, only Figure 6a is in the set $(\vec{X} = \vec{x})\varphi \subseteq \mathcal{K}$, so $\alpha = 1.0$. Of the five slices where we apply the intervention (d - h), 3 change classification, so $\beta = \frac{3}{5}$ or 0.6.

in β-goodness on real-world data and the effect of different masking values. The model for the MRI data is a pretrained CNN based on the ResNet50 architecture [Legastelois et al., 2023, Blake et al., 2023]. Brain magnetic resonance imaging (MRI) data was obtained from The Cancer Imaging Archive, as published by Buda et al. [2019] and publicly available on Kaggle[2]. 3,929 slices were extracted from 110 scans, each slice either containing tumor or having no tumor. As they were gathered from five distinct US institutions, the instrumentation and acquisition protocols may have varied. The data for the lung and pancreatic cancer dataset was obtained from 'The Medical Segmentation Decathlon' challenge – a publicly available dataset designed to be more difficult than many existing publicly available medical datasets [Antonelli et al., 2022]. From this, 17,657 slices were extracted from 96 CT scans and 26,719 slices from 420 CT scans respectively. From these slices, we chose 4000 healthy images uniformly at random for evaluation. Both datasets were included in the challenge for the small size of the tumors. For both datasets, a ResNet18 model was trained. For all three datasets, we created a causal explanation database using REX. All explanations were saved to sql database for efficient querying. All the models were trained as binary classifiers (tumor or no tumor). These models are not designed to be clinically useful: our goal is to generate and automatically assess the quality of explanations for absence. Hence, we did not attempt to

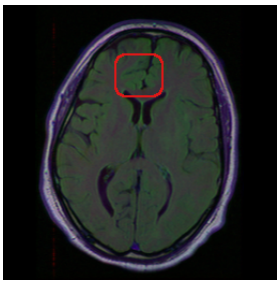
[2] https://www.kaggle.com/datasets/mateuszbuda/lgg-mri-segmentation

Datasets	Masking Values					
	0			real		
	β	β_p	β_a	β	β_p	β_a
Brain	0.87	0.87	0.85	0.44	0.43	0.4
Lung	0.96	0.94	0.85	0.96	0.93	0.83
Pancreas	0.89	0.88	0.62	0.81	0.79	0.66

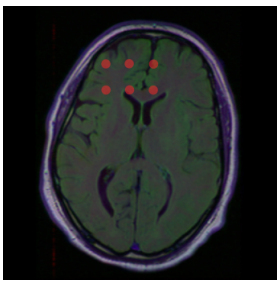
Table 1: β-goodness of grids over three different databases, using two different grid color values.



(a) A class 0 brain



(b) A suspicious area



(c) A local grid application

Figure 7: Suppose a clinician finds an area of a brain suspicious (Figure 7b). With the pre-calculated grid at the appropriate slice, the clinician can superimpose the grid over the area of interest. Here, the suspicious area sits is able to fit inside the grid, so it falls below the model's information requirements. This explains why the model gave class 0, but does not necessarily indicate health in the patient.

optimize the performance of our models nor make them generalizable to out-of-distribution data. Explanations of absence is not computationally expensive. All experiments were run on an Ubuntu 20.04 server with an Nvidia A40 GPU. With the explanations cached in advance, an individual grid calculation and β-goodness evaluation takes in the order < 1 second.

Table 1 summarizes the results. A masking value of 0 performs well on all datasets and models. The real values paint a more mixed picture. Interestingly, the model which performs least well on real values in the brain dataset. As this is the only dataset in true color, this suggests that the model is more sensitive to exact pixel values than models for lung and pancreas cancers. Both lung and pancreas datasets are CT data, treated as pseudo-RGB for the purposes of REX.

The parameters β, β_p, and β_a assess the quality of provided explanations *wrt* the model and the dataset. β does not take model confidence into account, hence it just shows the fraction of inputs in $\mathcal{K}$ for which superimposing the partial absence grid changes the classification from 1 to 0. $\beta \leq 1$, and it is lower than one due to the approximations of minimal explanations and the assumption location independence. On images where superimposing the grid does not change the classification, there must be sufficient information left in the image for the model to still classify it as positive.

Considering model confidence has a significant effect on some of the models. While β_p, on these datasets, is generally similar to β, β_a indicates that the pancreas model is less confident about its predictions, as can be seen by the relatively low $\beta_a \approx 0.64$, compared to β on both masking values. This β_a may be of more use for models returning low confidence classifications.

7 DISCUSSION

While our theoretical definitions are based on the assumptions of pixel independence and equal probability of tumors, the NITO algorithm circumvents them by finding minimal explanations of tumors on real scans for computing the absence grid. The experimental results show that NITO computes high quality explanations on different datasets. An important assumption on which all XAI tools rely is *locality of explanations*. This is also why in this work we evaluated NITO on solid tumors: other abnormalities can be distributed over the image, for example as a texture or a general change in size. Tumors are a good example of objects with local explanations and hence also lend themselves to computable explanations of absence.

We note that as NITO relies on the size of the dataset to compute minimal explanations of tumors, its accuracy depends on the quality of the dataset. An important advantage of NITO is its efficiency. The computation of a smallest explanation is done as a preprocessing step, and requires only one additional REX call on adding a new image to the dataset. For an image classified as "no tumors", NITO only needs to superimpose the precomputed grid on top of the image, without any computation.

We modeled a possible dialog with a clinician based on the NITO output, but we did not address the problem of *false negatives*, that is, images that are classified by a model as healthy, despite containing tumors. This is a crucial problem, and we will address it in future work. Finally, an extension of this work to general images is highly non-trivial, as the pixel independence assumption does not hold for general images.

Acknowledgements

The authors were supported in part by CHAI—the EPSRC Hub for Causality in Healthcare AI with Real Data (EP/Y028856/1).

References

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

Nathan Blake, Hana Chockler, David A. Kelly, Santiago Calderon Pena, and Akchunya Chanchal. Mrxai: Black-box explainability for image classifiers in a medical setting, 2023. URL https://arxiv.org/abs/2311.14471.

Mateusz Buda, Ashirbani Saha, and Maciej A Mazurowski. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in biology and medicine*, 109:218–225, 2019.

Seth Chin-Parker and Alexandra Bradner. A contrastive account of explanation generation. *Psychonomic Bulletin & Review*, 24:1387–1397, 2017.

Hana Chockler and Joseph Y. Halpern. Explaining image classifiers, 2024.

Hana Chockler, David A. Kelly, Daniel Kroening, and Youcheng Sun. Causal explanations for image classifiers, 2024.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: towards contrastive explanations with pertinent negatives. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 590–601, Red Hook, NY, USA, 2018. Curran Associates Inc.

Amit Dhurandhar, Tejaswini Pedapati, Avinash Balakrishnan, Pin-Yu Chen, Karthikeyan Shanmugam, and Ruchi Puri. Model agnostic contrastive explanations for structured data. *ArXiv*, abs/1906.00117, 2019. URL https://api.semanticscholar.org/CorpusID:173990728.

Joseph Y. Halpern. A modification of the Halpern–Pearl definition of causality. In *Proceedings of IJCAI*, pages 3022–3033. AAAI Press, 2015.

Joseph Y. Halpern. *Actual Causality*. The MIT Press, 2019.

Joseph Y. Halpern and Judea Pearl. Causes and explanations: A structural-model approach. Part I: Causes. *British Journal for the Philosophy of Science*, 56(4), 2005.

David Hume. *A Treatise of Human Nature*. John Noon, 1739.

Nicholas Kroeger, Dan Ley, Satyapriya Krishna, Chirag Agarwal, and Himabindu Lakkaraju. Are large language models post hoc explainers? *arXiv preprint arXiv:2310.05797*, 2023.

I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with shapley-value-based explanations as feature importance measures. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5491–5500. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/kumar20e.html.

Benedicte Legastelois, Amy Rafferty, Paul Brennan, Hana Chockler, Ajitha Rajan, and Vaishak Belle. Challenges in explaining brain tumor detection. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*, pages 1–8, 2023.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 4765–4774, 2017.

Tambiama Madiega. Artificial intelligence act. *European Parliament: European Parliamentary Research Service*, 2021.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

Nithesh Naik, BM Hameed, Dasharathraj K Shetty, Dishant Swain, Milap Shah, Rahul Paul, Kaivalya Aggarwal, Sufyan Ibrahim, Vathsala Patil, Komal Smriti, et al. Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? *Frontiers in surgery*, 9:266, 2022.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *Knowledge Discovery and Data Mining (KDD)*, pages 1135–1144. ACM, 2016.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *International Conference on Computer Vision (ICCV)*, pages 618–626. IEEE, 2017.

Ilia Stepin, Jose M Alonso, Alejandro Catala, and Martín Pereira-Fariña. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021.

David Watson, Limor Gultchin, Ankur Taly, and Luciano Floridi. Local explanations via necessity and sufficiency: Unifying theory and practice. *Minds and Machines*, 32, 03 2022. doi: 10.1007/s11023-022-09598-7.

A FORMAL DEFINITIONS OF CAUSES AND EXPLANATIONS

The material in this section is largely taken from Chockler and Halpern [2024], and the reader is referred to that paper for more context.

Causal models capture the way some variables causally influence others. This influence is modeled by a set of *structural equations*. The variables are typically split into two sets: *exogenous* variables, whose values are determined by factors outside the model, and the *endogenous* variables, whose values are determined by the exogenous variables. The structural equations describe how these values are determined. We also assume acyclicity. In other words, given the values of exogenous variables, we can propagate these values according to the structural equations and get a complete valuation of all variables in the model.

Formally, a *causal model* M is a pair $(\mathcal{S}, \mathcal{F})$, where $\mathcal{S}$ is a *signature*, which explicitly lists the endogenous and exogenous variables and characterizes their possible values, and $\mathcal{F}$ defines a set of *(modifiable) structural equations*, relating the values of the variables. A signature $\mathcal{S}$ is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is a set of exogenous variables, $\mathcal{V}$ is a set of endogenous variables, and $\mathcal{R}$ associates with every variable $Y \in \mathcal{U} \cup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for Y (*i.e.*, the set of values over which Y *ranges*). For simplicity, we assume here that $\mathcal{V}$ is finite, as is $\mathcal{R}(Y)$ for every endogenous variable $Y \in \mathcal{V}$. $\mathcal{F}$ associates with each endogenous variable $X \in \mathcal{V}$ a function denoted F_X (*i.e.*, $F_X = \mathcal{F}(X)$) such that $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} - \{X\}} \mathcal{R}(Y)) \to \mathcal{R}(X)$. This mathematical notation just makes precise the fact that F_X determines the value of X, given the values of all the other variables in $\mathcal{U} \cup \mathcal{V}$.

The structural equations define what happens in the presence of external interventions. Setting the value of some variable X to x in a causal model $M = (\mathcal{S}, \mathcal{F})$ results in a new causal model, denoted $M_{X \leftarrow x}$, which is identical to M, except that the equation for X in $\mathcal{F}$ is replaced by $X = x$.

We can also consider *probabilistic causal models*; these are pairs $(M, \Pr)$, where M is a causal model and $\Pr$ is a probability on the contexts in M.

The dependencies between variables in a causal model $M = ((\mathcal{U}, \mathcal{V}, \mathcal{R}), \mathcal{F})$ can be described using a *causal network*(or *causal graph*), whose nodes are labeled by the endogenous and exogenous variables in M, with one node for each variable in $\mathcal{U} \cup \mathcal{V}$. The roots of the graph are (labeled by) the exogenous variables. There is a directed edge from variable X to Y if Y *depends on* X; this is the case if there is some setting of all the variables in $\mathcal{U} \cup \mathcal{V}$ other than X and Y such that varying the value of X in that setting results in a variation in the value of Y; that is, there is a setting $\vec{z}$ of the variables other than X and Y and values x and x' of X such that $F_Y(x, \vec{z}) \neq F_Y(x', \vec{z})$.

We call a pair $(M, \vec{u})$ consisting of a causal model M and a context $\vec{u}$ a *(causal) setting*. A causal formula ψ is true or false in a setting. We write $(M, \vec{u}) \models \psi$ if the causal formula ψ is true in the setting $(M, \vec{u})$. Finally, $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\varphi$ if $(M_{\vec{Y}=\vec{y}}, \vec{u}) \models \varphi$, where $M_{\vec{Y} \leftarrow \vec{y}}$ is the causal model that is identical to M, except that the equations for variables in $\vec{Y}$ in $\mathcal{F}$ are replaced by $Y = y$ for each $Y \in \vec{Y}$ and its corresponding value $y \in \vec{y}$.

A standard use of causal models is to define *actual causation*: that is, what it means for some particular event that occurred to cause another particular event. We briefly review the relevant definitions below.

The events that can be causes are arbitrary conjunctions of primitive events (formulas of the form $X = x$); the events that can be caused are arbitrary Boolean combinations of primitive events. an arbitrary formula ϕ.

Definition 5 *[Actual cause]* $\vec{X} = \vec{x}$ *is an* actual cause *of* φ *in* $(M, \vec{u})$ *if the following three conditions hold:*

AC1. $(M, \vec{u}) \models (\vec{X} = \vec{x})$ *and* $(M, \vec{u}) \models \varphi$.

AC2. *There is a a setting* $\vec{x}'$ *of the variables in* $\vec{X}$*, a (possibly empty) set* $\vec{W}$ *of variables in* $\mathcal{V} - \vec{X}'$*, and a setting* $\vec{w}$ *of the variables in* $\vec{W}$ *such that* $(M, \vec{u}) \models \vec{W} = \vec{w}$ *and* $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}]\neg\varphi$*, and moreover*

AC3. $\vec{X}$ *is minimal; there is no strict subset* $\vec{X}'$ *of* $\vec{X}$ *such that* $\vec{X}' = \vec{x}''$ *can replace* $\vec{X} = \vec{x}'$ *in AC2, where* $\vec{x}''$ *is the restriction of* $\vec{x}'$ *to the variables in* $\vec{X}'$.

To define explanation, we need the notion of *sufficient cause* in addition to that of actual cause.

Definition 6 *[Sufficient cause]* $\vec{X} = \vec{x}$ *is a* sufficient cause *of* φ *in* $(M, \vec{u})$ *if the following four conditions hold:*

SC1. $(M, \vec{u}) \models (\vec{X} = \vec{x})$ *and* $(M, \vec{u}) \models \varphi$.

SC2. *Some conjunct of* $\vec{X} = \vec{x}$ *is part of an actual cause of* φ *in* $(M, \vec{u})$*. More precisely, there exists a conjunct* $X = x$ *of* $\vec{X} = \vec{x}$ *and another (possibly empty) conjunction* $\vec{Y} = \vec{y}$ *such that* $X = x \wedge \vec{Y} = \vec{y}$ *is an actual cause of* φ *in* $(M, \vec{u})$.

SC3. $(M, \vec{u}') \models [\vec{X} = \vec{x}]\varphi$ *for all contexts* $\vec{u}' \in \mathcal{R}(\mathcal{U})$.

SC4. $\vec{X}$ *is minimal; there is no strict subset* $\vec{X}'$ *of* $\vec{X}$ *such that* $\vec{X}' = \vec{x}'$ *satisfies conditions SC1, SC2, and SC3, where* $\vec{x}'$ *is the restriction of* $\vec{x}$ *to the variables in* $\vec{X}'$.

The notion of explanation builds on the notion of sufficient causality, and is relative to a set of contexts.

Definition 7 *[Explanation]* $\vec{X} = \vec{x}$ *is an* explanation *of* φ *relative to a set* $\mathcal{K}$ *of contexts in a causal model* M *if the following conditions hold:*

EX1. $\vec{X} = \vec{x}$ *is a sufficient cause of* φ *in all contexts in* $\mathcal{K}$ *satisfying* $(\vec{X} = \vec{x}) \wedge \varphi$. *More precisely,*

- *If* $\vec{u} \in \mathcal{K}$ *and* $(M, \vec{u}) \models (\vec{X} = \vec{x}) \wedge \varphi$, *then there exists a conjunct* $X = x$ *of* $\vec{X} = \vec{x}$ *and a (possibly empty) conjunction* $\vec{Y} = \vec{y}$ *such that* $X = x \wedge \vec{Y} = \vec{y}$ *is an actual cause of* φ *in* $(M, \vec{u})$. *(This is SC2 applied to all contexts* $\vec{u} \in \mathcal{K}$ *where* $(\vec{X} = \vec{x}) \wedge \varphi$ *holds.)*
- $(M, \vec{u}') \models [\vec{X} = \vec{x}]\varphi$ *for all contexts* $\vec{u}' \in \mathcal{K}$. *(This is SC3 restricted to the contexts in* $\mathcal{K}$.*)*

EX2. $\vec{X}$ *is minimal; there is no strict subset* $\vec{X}'$ *of* $\vec{X}$ *such that* $\vec{X}' = \vec{x}'$ *satisfies EX1, where* $\vec{x}'$ *is the restriction of* $\vec{x}$ *to the variables in* $\vec{X}'$. *(This is SC4).*

EX3. $(M, u) \models \vec{X} = \vec{x} \wedge \varphi$ *for some* $u \in \mathcal{K}$.

The requirement that the first part of condition EX1 as given here holds in all contexts in $\mathcal{K}$ that satisfy $\vec{X} = \vec{x} \wedge \phi$ and that the second part holds in all contexts in $\mathcal{K}$ is quite strong, and often does not hold in practice. We are often willing to accept $\vec{X} = \vec{x}$ as an explanation if these requirements hold with high probability. Given a set $\mathcal{K}$ of contexts in a causal model M, let K_ψ consist of all contexts $\vec{u}$ in $\mathcal{K}$ such that $(M, \vec{u}) \models \psi$, and let $\mathcal{K}(\vec{X} = \vec{x}, \varphi, \text{SC2})$ consist of all contexts $\vec{u} \in \mathcal{K}$ that satisfy $\vec{X} = \vec{x} \wedge \varphi$ and the first condition in EX1 (i.e., the analogue of SC2).

Definition 8 *[Partial Explanation]* $\vec{X} = \vec{x}$ *is a* partial explanation *of* φ *with goodness* (α, β) *relative to* $\mathcal{K}$ *in a probabilistic causal model* $(M, \Pr)$ *if*

EX1′. $\alpha \leq Pr(\mathcal{K}(\vec{X} = \vec{x}, \varphi, SC2) \mid \mathcal{K}_{\vec{X}=\vec{x}\wedge\phi})$ *and* $\beta \leq Pr(\mathcal{K}_{[\vec{X}=\vec{x}]\phi})$.

EX2′. $\vec{X}$ *is minimal; there is no strict subset* $\vec{X}'$ *of* $\vec{X}$ *such that* $\alpha \leq Pr(\mathcal{K}(\vec{X}' = \vec{x}', \varphi, SC2) \mid \mathcal{K}_{\vec{X}'=\vec{x}'\wedge\phi})$ *and* $\beta \leq Pr(\mathcal{K}_{[\vec{X}'=\vec{x}']\varphi})$, *where* x' *is the restriction of* $\vec{x}$ *to the variables in* $\vec{X}$.

EX3′. $(M, u) \models \vec{X} = \vec{x} \wedge \varphi$ *for some* $u \in \mathcal{K}$.

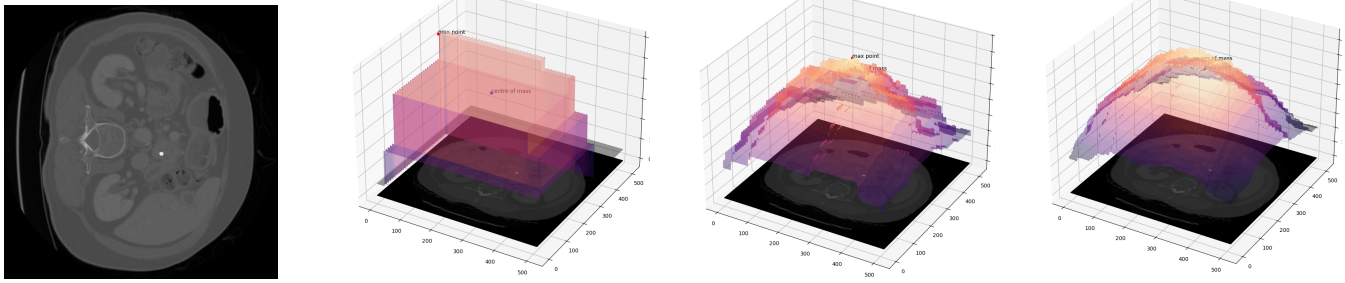
B EXPLANATIONS IN REX

REX is a causal explainability tool that produces a *responsibility landscape*. From this landscape, it extracts causal explanations: sets of pixels, possibly disjoint, that are sufficient to reproduce the original model classification. We show a typical example in Figure 2 and another for lung data in Figure 9. The tool itself is available at `https://github.com/ReX-XAI/ReX`. While the full algorithm is rather complex, broadly speaking, REX creates mutants of an initial input image by subdividing it into 4 *superpixels*. These superpixels are created by random partitioning. The model is queried on all combinations of these superpixels, with "non-active" superpixels set to a masking value (by default 0). The causal responsibility is calculated for these combinations. Combinations with non-zero responsibility are further broken down into more (smaller) superpixels and the process repeated. Once superpixels reach a predefined size limit, the algorithm quits. This procedure is repeated many times to avoid the issue of a poor initial partitioning. The effect of multiple iterations is to smooth the final responsibility map. The map then provides a pixel ranking from which REX greedily extracts an explanation. Pixels are added into an initially blank image, from highest responsibility to lowest, until the pixels are sufficient to obtain the same class as the initial class prediction.

C NEUTRALITY OF NEUTRAL GRID

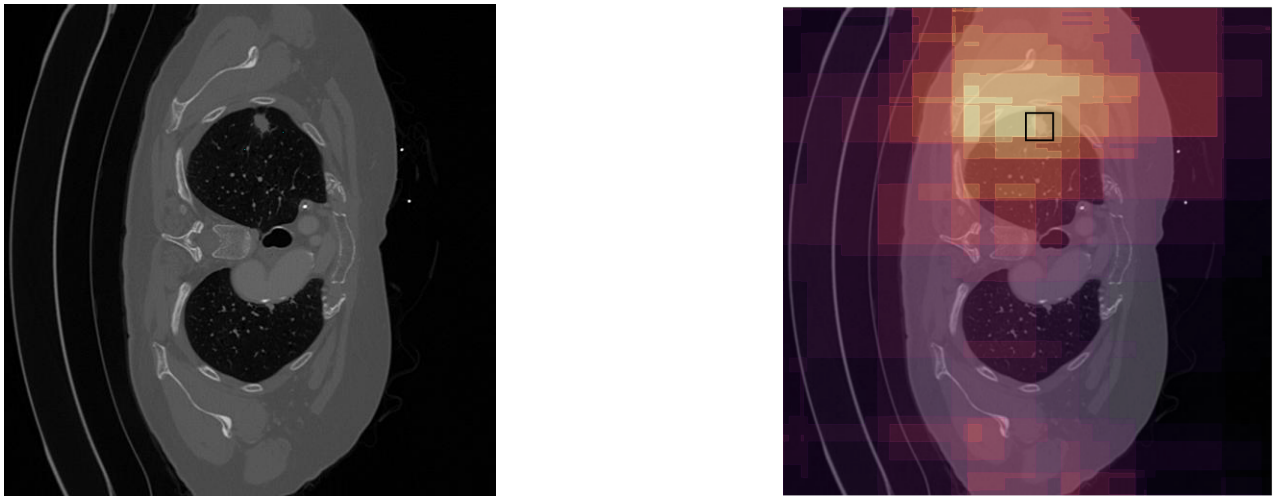
The effect of the user-colored grid is not entirely neutral on the confidence of the model on a given image. On the brain MRI data, for example, in a small number of cases (15) in the brain data, the calculated grid actually changes the classification from negative to positive. In general, the model confidence on these 15 images is low, with a mean value of 0.83, with the lowest confidence for the no-tumor classification being just 0.54. It is, of course, possible to set a β-goodness as a target rather than as a byproduct of the grid. We envisage this being the actual use case for clinicians. If a user required an explanation of absence to have a β of 1 then it would simply be the case of changing the density and radius of the grid until this is achieved. This procedure could form part of a dialogue between user and model, strengthening trust in the model, or revealing its weaknesses.

Figure 10 shows a representative sample of these flipped classifications. Further investigation is required to discover why the grid changes the class in this year. Likely this is due to a disruption, by the grid, of some learned concept.



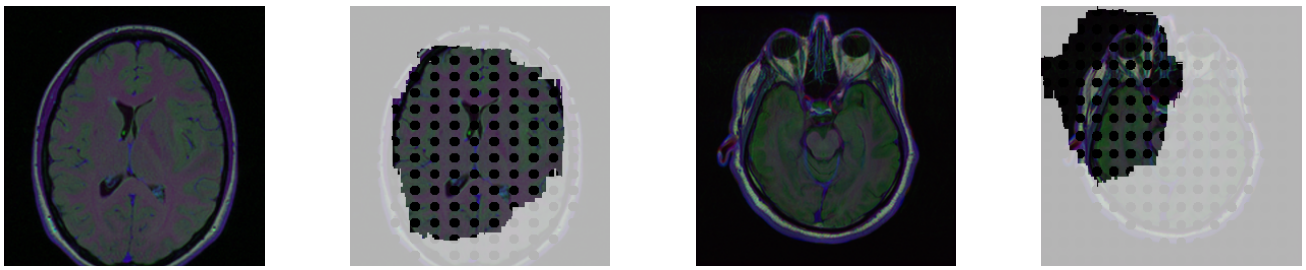
(a) A pancreas CT slice (b) 1 iteration (c) 10 iterations (d) 30 iterations

Figure 8: The smoothing of the responsibility map over multiple iterations, here shown on a slice from an CT image of a pancreas (Figure 8a). REX extracts explanations using the responsibility pixel ranking.



(a) A lung CT slice with a tumor (b) Heatmap of responsibility

Figure 9: REX also produces heatmaps of the responsibility map. We have manually marked the location of the lung tumor in Figure 9b. The heatmap includes the tumor, but seems to be localizing slightly to the left of the main lump.



(a) A healthy brain with 0.54 confidence (b) The grid changes the classification to 1 (c) A healthy brain with 0.99 confidence (d) An explanation for a healthy brain

Figure 10: A selection of images and their actual causal explanations where the grid changed the classification from no-tumor to tumor. To the human eye, at least, there is no obvious reason why the grid has had the effect of a counterfactual. The explanations are unusually large.