Essi Pietilä

METHODS AND METRICS FOR EVALUATING EXPLAINABLE ARTIFICIAL INTELLIGENCE IN HEALTHCARE DOMAIN

Bachelor's thesis
Faculty of medicine and health technology
Pedro Moreno-Sánchez
April 2023

ABSTRACT

Essi Pietilä: Methods and Metrics for Evaluating Explainable Artificial intelligence in Healthcare Domain
Bachelor's thesis
Tampere University
Biotechnology and biomedical engineering
April 2023

Sometimes hard-to-interpret black-box artificial intelligence models might not induce trust in their users, particularly in the healthcare domain. Explainable artificial intelligence has been developed as a solution to this mistrust so that the specialists of different fields could better understand the processes that have led to the solutions offered them by the artificial intelligence model. Explainability is thought to give the specialists chance to evaluate feasibility of the algorithms and to encourage them to use AI in the decision-making process.

Importance of explainability is particularly evident in the healthcare domain where the applications of artificial intelligence are used for example for treatment decisions of patients and for large-scale decisions about healthcare infrastructure. These decisions impact not only the health and welfare of a single patient but even larger communities. Decisions in healthcare domain require preciseness from the tools and responsibility from the decision makers. When the European regulations about the patients' rights for explanation about the decisions considering them, it becomes clear that explainability is needed from artificial intelligence.

To promote trust in specialists by using explanations we must be able to evaluate and validate the explanations with accurate metrics. At the moment, there are no standardised metrics or methods for evaluating explainable artificial intelligence and the field consensus is that rigorous study is needed to construct some. This thesis aims to find the state-of-art methods and metrics used for evaluating explainable artificial intelligence models, discuss their feasibility for healthcare and give basis for further studies to build unified set of metrics that can be used for validating new models. Overall, 54 metrics and methods were found and summarised in tables.

Keywords: Explainable AI, XAI evaluation, explainability, faithfulness, robustness, fairness, understandability, framework

TIIVISTELMÄ

Essi Pietilä: Mittarit ja metodit selitettävän tekoälyn arviointiin terveydenhuollon alalla
Kandidaatin tutkielma
Tampereen yliopisto
Bioteknologia ja biolääketieteen tekniikka
Huhtikuu 2023

Perinteiset vaikeasti tulkittavat tekoälymallit eivät aina herätä luottamusta käyttäjissään, varsinkaan terveydenhuollon alalla. Ratkaisuksi tähän on kehitetty selitettävä tekoäly, jotta eri alojen ammattilaiset voisivat paremmin ymmärtää prosesseja, jotka ovat johtaneet tekoälyn heille tarjoamiin ratkaisuihin. Selitettävyyden tarkoitus on antaa tekoälymallien käyttäjille mahdollisuus algoritmien ja niiden tuottamien ratkaisujen arviointiin ja niiden oikeellisuuden varmistamiseen rohkaisten heitä käyttämään tekoälyä päätöksenteossa.

Selitettävyyden tärkeys korostuu terveydenhuollossa, sillä tekoälyä hyödyntäviä työkaluja käytetään esimerkiksi potilaiden hoitoa tai yleisesti sosiaali- ja terveydenhuoltoa koskeviin kysymyksiin, jolloin tarjotulla ratkaisulla voi olla suuri merkitys niin yksittäisen potilaan kuin kokonaisten ihmisryhmien terveyteen ja hyvinvointiin. Terveydenhuollossa päätökset vaativat tarkkuutta työkaluilta sekä vastuuta päättäjiltä, kuten lääkäreiltä. Kun huomioidaan myös Euroopan unionin asettama potilaan oikeus selitykseen häntä koskevaan päätökseen johtaneista tekijöistä, on ilmeistä, että terveydenhuollossa käytettävän tekoälyn on oltava selitettävää.

Jotta luottamusta tekoälyyn voitaisiin tehokkaasti parantaa selitettävyyden avulla, on oltava keinoja arvioida selitysten paikkansapitävyyttä ja ymmärrettävyyttä objektiivisesti. Tällä hetkellä erilaisia mittareita ja metodeja on paljon, mutta yhtenäisiä käytänteitä tai standardeja selitettävyyden arviointiin ei ole. Tässä työssä kerätään yhteen ja esitellään nämä mittarit ja metodit. Arviointimenetelmät myös kerätään tiivistävin taulukoihin jaoteltuna sen mukaan, mitä mittaavat ja niiden soveltuvuutta terveydenhuollon työkalujen arviointiin arvioidaan. Tämän työn tarkoituksena on koota ajantasainen tieto selitettävän tekoälyn arviointimenetelmistä, jotta myöhemmät tutkimukset voisivat kehittää päteviä selitettävän tekoälyn arviointiprosesseja. Jatkotutkimusten pitäisi olla mahdollisia, sillä erilaisia mittareita ja metodeja esitellään 54 kappaletta ja useimmat niistä ovat hyvin käyttökelpoisia.

Avainsanat: Selitettävä tekoäly, tekoälyn arviointi, selitettävyys, luotettavuus, vakaus, reiluus, ymmärrettävyys, viitekehys.

TABLE OF CONTENTS

1. INTRODUCTION 1

1.1 Explainable artificial intelligence 2

1.2 Explanation types 3

2. METHOD OF THE LITERATURE REVIEW 5

3. METRICS FOR EVALUATING EXPLAINABLE ARTIFICIAL INTELLIGENCE 6

3.1 Domain independent metrics with human in the loop 6

3.2 Domain independent metrics without human in the loop 8

3.2.1 Metrics for evaluating explanation robustness 8

3.2.2 Explanation faithfulness 9

3.2.3 Metrics for evaluating understandability 15

3.2.4 Miscellaneous metrics 15

3.3 Frameworks for evaluating XAI 18

4. METHODS AND METRICS FOR EVALUATING EXPLAINABLE ARTIFICIAL INTELLIGENCE IN HEALTHCARE DOMAIN 21

4.1 Methods to evaluate XAI for healthcare applications with expert in the loop 21

4.2 Metrics to evaluate XAI for healthcare applications without expert in the loop 23

5. DISCUSSION 24

5.1 Summary tables of the metrics and methods 24

5.2 From a jungle of metrics to valid frameworks 49

5.3 Limitations of the review 50

6. CONCLUSION 51

REFERENCES 52

ABBREVIATIONS AND MARKINGS

AI	Artificial intelligence
AI HLEG	High-Level Expert Group on Artificial Intelligence
ALTAI	Assessment List for Trustworthy Artificial Intelligence
DoX	Degree of explainability -metric
EU	European Union
GDPR	General Data Protection Regulation
IROF	Iterative Removal of Features -metric
LIME	Local Interpretable Model-Agnostic Explanations
MSFI	Modality Specific Feature Importance -metric
MUsE	Model Usability Evaluation -method
SAGAT	Situation Awareness Global Assessment Technique -method
SAFE-AI	Situation Awareness Global Assessment Technique -method
SCS	System Causability Scale -metric
XAI	Explainable artificial intelligence

1. INTRODUCTION

Artificial intelligence (AI) is used in various fields of modern society. AI algorithms can effectively and accurately provide solutions to many difficult problems and therefore ease the life of their users. However, the algorithms are often complex, and user cannot always tell what the reasoning behind the given solution is. Thus, on fields with high risks and need of precision, such as healthcare, AI solutions are yet to be adopted as the users are generally not trained in AI and might be reserved towards it. [1] Healthcare professionals need to be able to understand and trust the solutions offered to them in order to use them in their work but this can be very challenging when it comes to unintuitive “black-box” AI models such as those using deep learning.

Explainable AI (XAI) has been developed to answer the need of understanding the opaque AI models. XAI means AI models for which it is possible to generate an explanation about the reasons that have led to its predictions. The explanation is supposed to be understandable to the model’s audience, ie. the patient does not have to be able to understand the explanation a doctor gets from a decision support system as long as the doctor can explain its effects to the patient [1]. Evaluating the received prediction is easier due to the explanation and therefore mistakes in the final decisions can be reduced, user’s trust in the system can be pro-moted and the applications become easier to approach. One of the main goals of XAI is to make AI models more trustworthy. [1], [2] This makes use of AI more feasible in situations with high stakes.

Explainability of artificial intelligence is particularly important in healthcare as the field is tightly regulated. For example, the General Data Protection Regulation (GDPR) of the European Union (EU) guarantees the patient a right to explanation about the grounds leading to the decisions made about her or him. On the other hand, EU’s Ethics Guidelines for Trustworthy AI direct the developers to make AI models explainable. [3] Furthermore, the high responsibility medical professionals bear in their work makes it paramount that they can trust the system used to support their decision making. Without explainability, clinical decision support system using highly technical algorithms such as deep neural networks can feel unapproachable as they are very difficult to understand without training. [1]

Since trustworthiness and facilitating adoption of AI based tools in highly regulated or otherwise precision requiring fields are some of the main goals of XAI, it is important that the explanations can be evaluated and validated as well. Assessing the validity of the explanations can be difficult for an average user and therefore standardised or at least generally approved metrics would be an important addition to the field. At the moment there is an obvious lack of standardisation and validation of metrics and many studies call for standardisation of metrics for evaluating XAI. While there are several methods of constructing an explanation, there are even more ways to evaluate them. While researchers hurry to develop as rigorous and credible metrics as possible, there remains a need for studies gathering the metrics together and deciding what should be measured and how. [1]

This thesis reviews the state-of-the-art of different methods and metrics used for evaluating explainable AI. The aim of it is to find out if some metrics are especially suitable for the purpose and to collect information about the current metrics to support becoming studies aiming at validating metrics for evaluating XAI. Introduction continues to give background information about XAI after which domain independent metrics and methods are addressed followed by those developed particularly for healthcare applications. Finally, observations, most potential metric and future challenges and prospects are discussed.

1.1 Explainable artificial intelligence

Explainable artificial intelligence means AI models which can produce an explanation about the logical process leading to their outcomes [1]. Explainability of artificial intelligence can be implemented in many ways. Although explainability sometimes refers also to intrinsically understandable, so-called interpretable AI models, in this thesis explainable AI refers to black-box AI models for which explainability is implemented post hoc. These explanations should reveal the model’s reasoning process and what are the factors that led to the outcome.

The need to explain a model varies according to its complexity. While some models are intrinsically interpretable thus needing little explanations, opaque models like deep neural network models might require explanation methods to become explainable to a user who is not a specialist in artificial intelligence. [1] Explanation design is heavily dependent on the audience as the explanation should be detailed enough to provide the needed information, but not include too much information. In short, the users’ needs and expected knowledge and goals of the model have to be considered in XAI design. [1], [2] Explainable AI in general has many goals which may vary according to the context but typically XAI aims to increase users' trust to AI and make AI based solutions more accessible, make the models more faithful and fair and ensure the model's informativeness, transferability and causality [1]. Particularly in healthcare users need to be sure that their tools fulfil clinical and legal requirements, they are easy enough to use correctly without causing too heavy additional workload and that they provide the information the users need. This can be achieved by adding explainability to black-box AI models. [4] Furthermore, explainability gives the user chance to evaluate the model's validity and facilitate trust to AI based tools [2].

Despite the benefits of explainability, not all AI models are made explainable. This is due to a trade-off between explainability and model accuracy. Sometimes it is not essential to know the reasoning behind the model's results as long as they are as accurate as possible. These situations could be such where quick decisions have to be made according to plenitude of data or when precision is essential, but the explanation would not produce relevant extra insights for the users. In these cases, the trade-off between explainability and accuracy makes it wiser to opt for the accuracy. [1] The need for explanations should be considered in the model's development phase by developers and users.

1.2 Explanation types

Explainability can be implemented in many different methods and there is no single explainability method that would suit all AI models [5]. Explainability can be implemented in many methods. The explanations can be designed to suit one model or be model agnostic methods and they can either explain the whole model (global explanation) or a specific prediction (local explanation). [1], [6] For example, Local Interpretable Model-Agnostic Explanations (LIME), is a model agnostic explainability method that is based on simplifying the decision at hand. [1]

Different algorithm types often require different explainability techniques and methods. For instance, models working with images are better explained with saliency maps than linguistic explanations and multimodal classification tasks might become clearer through simplifying explainability techniques rather than decision trees. Therefore, many different techniques have been developed to fit the different needs and to help understand the similarities and differences between the explainers, they can be grouped. Categorisation of the explainability techniques can be done in many ways and Barredo Arrieta et al. use a taxonomy of dividing them first into categories according to the technique used, then the method and finally according to the data used in the intended purpose of the explainer. For example, techniques can include explanations by simplification, explanations by example, visual explanations, linguistic explanations, local explanations and feature relevance. and so on. Explanation methods can include for instance sensitivity analysis or localisation maps. [1]

Another way to categorise explainability models is to use the three levels described by Sanneman and Shah. They include goals of XAI into the categorisation and divide XAI into XAI for perception, ie. those that explain what the system does and what decisions it has made, XAI for comprehension, ie. why it made the decisions and how this relates to the model’s goals and finally to the third level, XAI for projection, ie. explanations that reveal the upcoming decisions, how to model would act in another similar instance or what should change to end up in a different decision. [7]

Despite the group an XAI method falls into, it should be sound, accessible and achieve its goals. To ensure this, explainability methods should be rigorously evaluated. [1], [2], [6]

2. METHOD OF THE LITERATURE REVIEW

This thesis is primarily a literature review. It is conducted by surveying papers in Science Direct with keywords “explainable AI”, “interpretable AI”, “evaluating”, “validating”, medical” and “healthcare”. The search query used provided approximately 250 results at the time of collecting the sources and the abstracts of the results were examined. Papers that included aforementioned keywords and seemed to include relevant information were chosen to closer review. Some papers were also collected from Google Scholar. In total, this phase of the paper collection process, 43 documents were kept for closer review.

When the selected 43 papers were read more closely, those proposing metrics or methods were used as reference in this thesis and those referring to metrics or methods from other researchers were only kept if they contained relevant background information or good points for discussion. This meant keeping 13 papers. The rest of the papers were either left completely unused or their references were used to find the articles originally introducing metrics or methods for evaluating XAI. The remaining 40 references in this thesis were found as described from the references of the original 43 papers. The most generic mathematical evaluation methods such as area-under-the-curve were excluded from the thesis as they were considered to be common knowledge.

3. METRICS FOR EVALUATING EXPLAINABLE ARTIFICIAL INTELLIGENCE

Although the main focus of the thesis is on the healthcare applications of XAI, it is important to find out what general domain metrics there are as they can be used to evaluate healthcare specific XAI models as well. There are some universal aspects of explanations that should be evaluated despite the application as they work as sanity checks for the explanation. These include information about how easily the explanation changes according when the input changes, how accurately the explanation has recognised the most important features, accuracy of the counterfactuals and the model fairness. These aspects are evaluated in healthcare applications typically with general domain metrics.

3.1 Domain independent metrics with human in the loop

Despite being developed to enhance user experience and trust to artificial intelligence, methods and metrics for evaluating explainable AI often do not consider users or domain experts. While the vast majority of found metrics evaluate XAI mathematically, there are some metrics that do require users in the evaluating process. Probably the most used method of them is Likert scale with task specific questions.

Likert scales are surveys that consist of questions with numerical answer options. **Likert scale** is a very modifiable evaluation method as the amount and contents of the questions as well as the answer scale can be tailored for the study. For example, Chakraborti et al. use five-point scale for their surveys but Hoffman et al. opt for three point scale in their questionnaire.[2], [8] **Likert scales** are a good tool for conducting user studies but it is essential to carefully design the questions to measure the desired property and also to evaluate the scale validity [2], [7]. Using **Likert scale** as an evaluation method means a large enough user study is needed and researchers should also ensure diversity of the survey participants for the results to be worthwhile. Therefore, the method requires resources and is not especially lightweight evaluation method. [9]

Likert scales are also used in the **System Causability Scale (SCS)** proposed by Holzinger et al. which is used to evaluate the perceived explainability and ensure that the XAI system fits its purpose. The **SCS** consists of 10 questions:

1. “I found that the data included all relevant known causal factors with sufficient precision and granularity.”
2. “I understood the explanations within the context of my work.”

3. “I could change the level of detail on demand.”
4. “I did not need support to understand the explanations.”
5. “I found the explanations helped me to understand causality.”
6. “I was able to use the explanations with my knowledge base.”
7. “I did not find inconsistencies between explanations.”
8. “I think that most people would learn to understand the explanations very quickly.”
9. “I did not need more references in the explanations: e.g., medical guidelines, regulations.”
10. “I received the explanations in a timely and efficient manner.”

The users evaluate the XAI systems with a five point scale from strongly disagree to strongly agree and the results can be evaluated by summing the ratings of each question and dividing this by 50. Holzinger et al. test the **SCS** with a healthcare application so it is implementable in healthcare domain as well and also quite efficient way to conduct a user survey. [10]

Sanneman & Shah introduce the **Situation Awareness Framework for Explainable AI (SAFE-AI)** for evaluating if the explanation provides all the information the user needs. Sanneman and Shah divide XAI in three levels as explained in 1.2 and propose means for evaluating XAI models in each level and finally exhibit an adapted version of the **Situation Awareness Global Assessment Technique (SAGAT) test** introduced by Endsley in [11] to evaluate the information provided by XAI and situation awareness in general. The evaluation comprises of questions that are tailored for the type of XAI in question. Answers are collected in simulated use cases by stopping the situation momentarily to ask question. The **SAFE-AI** modification of **SAGAT** is best used in the development phase of an XAI model as it reveals what information the users need and if the model induces trust. **SAGAT** is a very user-centric evaluation framework utilising the knowledge of human factors throughout. [7]

Ribeiro et al. evaluate explainability along the same lines as **SAGAT** with their **user study of the Anchors** explainability method. The user study is aimed to users with some understanding of AI and during it the users are asked to predict the AI model’s behaviour first with random test cases and then with 10 instances after seeing explanations for the previous. The users only make predictions if they are confident that they are correct and otherwise refrain from answering. The result is a percentage of correct predictions on the round after seeing the explanations. [12]

3.2 Domain independent metrics without human in the loop

Most metrics for evaluating XAI do not include users or field experts in the evaluating process. These metrics often focus on measuring easily quantifiable features of the explanation and are often mathematical equations that compare two or more explanations given by the explanation method [13]. Not including humans in the evaluation makes the process quicker and potentially more objective and requires less resources [14].

3.2.1 Metrics for evaluating explanation robustness

Robustness is a quality of an explanation that describes how well the explanation holds when minor changes are made in the input. This is an important property as similar instances logically tend to need similar explanations and in real-life instances, particularly in healthcare, the data may be noisy. If an explanation is robust, it is not sensitive to noise and is able to detect the underlying reasons for the predictions rather than explaining the noise.

Sensitivity is a metric that reveals how the explainability method reacts to small changes in the input. Sensitivity of an explanation can be considered either a benefit or a defect of a model and as Alvarez-Melis and Jaakkola point out, it partly depends on the purpose of the explanation. Notwithstanding, Alvarez-Melis and Jaakkola as well as Bhatt et al. argue that generally low sensitivity of an explanation is desirable as thus the explanation ignores unstable noise producing similar explanations for similar instances making the model more stable and robust. [15], [16] Bhatt et al. also expand the concept of sensitivity to calculate maximum sensitivity and average sensitivity to give more detailed understanding of the explainer. Sensitivity is calculated as the distance between explanations for an instance and its perturbations. [16]

Input invariance by Kindermans et al. further expands the previous sensitivity metrics. **Input invariance** measures what effect noise in the input has on the explanation. Main idea is that if noise in input does not affect the AI model's attributions, it should not affect the attributions of the explanation either. As are many of the other metrics, **input invariance** is developed to be used for saliency maps. Stability is an important aspect of an explanation, but Kindermans et al. admit that their implementation depends on the reference point chosen for the evaluation. Thus, the metric does not always produce comparable results. [17]

Dasgupta et al. take a slightly different angle to measuring explanation robustness. They propose a metric called **consistency** to evaluate how probably instances with the same explanation are given the same prediction. Particularly **global consistency** ie. consistency calculated for the entire model is interesting and consistency of a model indicates that explanations are not randomly generated. **Consistency** is a model agnostic metric and can be valuable tool for evaluating method stability. [18]

Local Lipschitz continuity is a variation of Lipschitz criterion proposed by Alvarez-Melis and Jaakkola to further evaluate explanation robustness. It is an improved version to the usual and more global Lipschitz continuity as it focuses on the input points closer to the reference input rather than all, even significantly different inputs. However, calculating **local Lipschitz continuity** for black-box models and their explainers can be computationally expensive and challenging. [19]

Stability as presented by Yuan et al. in their not peer-reviewed article, is once again a metric where difference between an explanation for one instance and its perturbation are compared. It is important to notice with this and the previous similar metrics that the prediction is assumed to be the same in both explanation cases. **Stability** is intended to be used with graph neural networks. [20]

Final metric for robustness is **reiteration similarity** by Amparore et al. which tests if the explainability method produces same explanation for reiterations of the same instance. **Reiteration similarity** is calculated as Jaccard similarity for the set of explanations for an instance. **Reiteration similarity** is a crucial property of an explainability model to ensure that the explanations are valid. [21]

3.2.2 Explanation faithfulness

For the explanation to be credible, it is essential that the explanation is faithful, that is the right properties are explained and feature importances are correct. Measuring these properties is very popular particularly when it comes to saliency maps as explanation but with other techniques too. Typically features are removed from the input in relevance order derived from the explanation and different calculations are conducted based on the output.

The first metric of this kind is **fidelity** introduced by Pope et al. in [22]. Pope et al. define **fidelity** as the loss of accuracy of an AI model if features with saliency values higher than 0.01 are removed. This produces a fidelity score between 1 and 0, 1 being the highest, which means that the saliency map explaining the AI model has selected extremely relevant features. In graph neural network explainer's **fidelity** describes how well the explanation corresponds to the function at a node [23]. Despite being originally designed to evaluate graph neural networks, **fidelity** should be applicable to other models as well due to its generalisable nature. [22]

Completeness by Sundarajan et al. and **summation-to-delta** by Shrikumar et al. in turn, ensure that the feature importances of an image explanation or a graph network are correct by checking that the total attribution in an image explanation corresponds to the difference between prediction F at instance x and the determined baseline. This is a sanity check for explainers based on feature importance and desirably the value should be low. [24], [25]

Rather similar metrics to fidelity are **sensitivity-n**, **local concordance** and **continuity**. These measure the difference in the output accuracy when n features are removed. **Sensitivity-n** by Ancona et al. and **local concordance** by Amparore et al. generalise this concept by developing metrics that consider only the n most important or otherwise interesting features and calculate the loss in accuracy when obtaining only them [21], [26]. **Continuity** by Montavon et al. on the other hand quantifies how continuous an explanation function is. This is done by calculating the strongest variation over all inputs. [27]

Focus by Arias-Duart et al. measures faithfulness in a very similar way to for example **fidelity** with the exception that the labelled images are divided into N, typically four, mosaics that act as the points of reference to evaluate how well the pixel relevances match the reality. [28]

Non-sensitivity by Nguyen & Martínez on the other hand ensures that zero importance is only assigned to features that do not affect the model accuracy [29]. These metrics are important in evaluating meaningfulness of an explanation. They prevent the user from favouring explanations that fit their own reasoning rather than the model’s actual function and therefore are valuable tools in XAI evaluation. Moreover, it can be argued that the general metrics like **sensitivity-n** and **local concordance** should be favoured over the simpler ones as they can be used more versatilely, and they provide more detailed information. [21], [26]. Figure 1 shows an example where **local concordance** and **fidelity** is calculated.

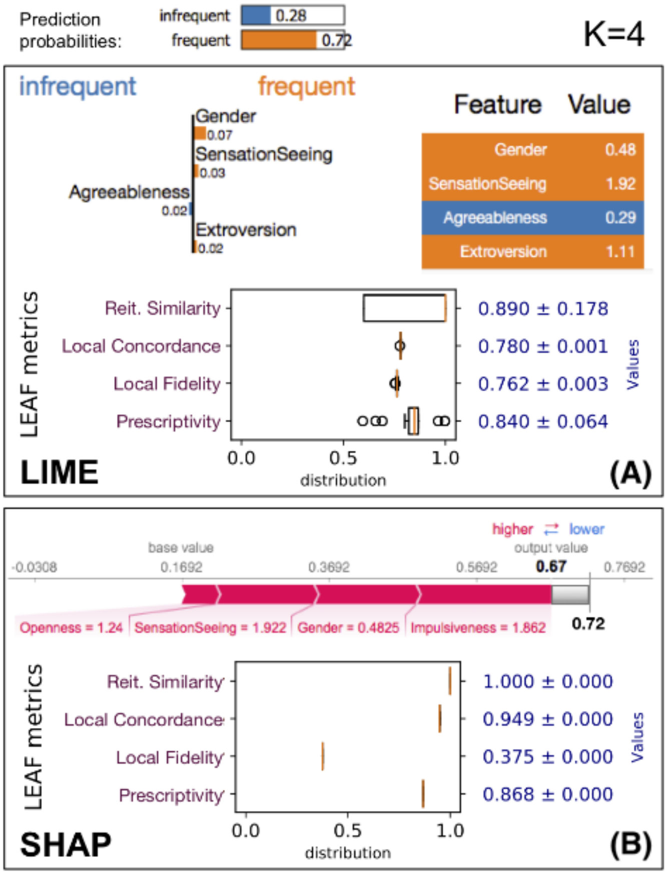


Figure 1: LIME and SHAP explanations for drug abuse likeliness evaluated by reiteration similarity, local concordance, local fidelity and prescriptivity. Figure is created by Amparore et al. ***[21]***

Faithfulness, or **faithfulness correlation** as called in Hedström's **Quantus** toolkit, proposed by Arya et al. is very similar to **fidelity** [5], [30]. However, for **faithfulness**, only S most important features are considered and the correlation of the change of prediction accuracy when the features are set to input baseline and number of the features used is calculated. This, as the other faithfulness metrics, is an important sanity check of an explanation. Even so, it comes with problems as the feature baseline can be defined in several ways and it also may be difficult to aggregate overall faithfulness with this metric. [5]

Region perturbation by Samek et al. and **Iterative Removal of Features (IROF)** proposed in Rieger and Hansel's not peer-reviewed paper are otherwise same as fidelity but work with image segments rather than independent pixels. This is a significant benefit in terms of healthcare related classification problems as often the images contain pixels that are very dependent on each other. [31], [32] **Region perturbation** and **IROF** are designed for neural networks but should be generalisable like fidelity.

Also Montavon et al. argue for removing segments of an image instead of pixels in their paper constructing grounds for a metric that is called **selectivity** by Hedström et al. in their Quantus library, which will be discussed later. [27], [30] **Selectivity** is much like **IROF** but has been published earlier and is based on a pixel-wise relevance measurement proposed by Bach et al. and often referred to as **pixel-flipping** [27], [30], [33]. The principal logic in both of these is that most relevant pixels or segments of an image are selected according to the explanation map and removed by masking them. After this, the perturbed image is used as an input and the effect on the prediction function f(x) is plotted. Finally area-under-the-curve is calculated with low result marking a good explanation. [27], [33] Figure 2 gives an example of **pixel-flipping.**

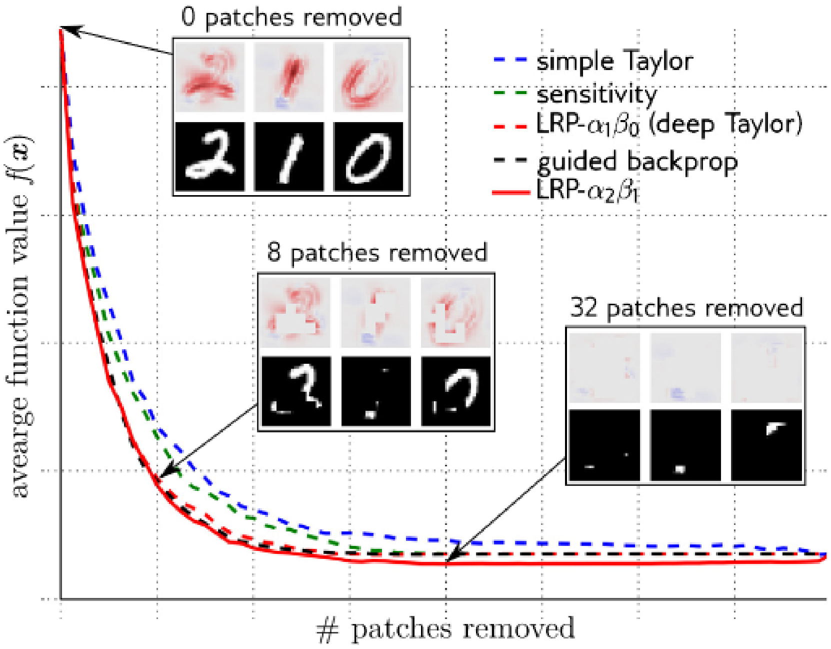


Figure 2: Graph by Montavon et al. to illustrate effect of pixel-flipping on prediction accuracy. ***[27]***

Very similar but reverse to the above metrics is **faithfulness**, proposed by Alvarez-Melis and Jaakkola, implemented by Arya et al. in a not peer-reviewed article and referred to as **faithfulness estimate** by Hedström [5], [26], [30]. **Faithfulness** also measures correctness of feature importance but instead of removing pixels or features, they opt for adding them. After getting the feature importance vector θ and the predictions for the iterated inputs p, **faithfulness** φ is calculated using Pearson's correlation:

$$\varphi = -\rho(\theta, p). \tag{1}$$

Greater φ is evidence of a faithful explanation. [5]

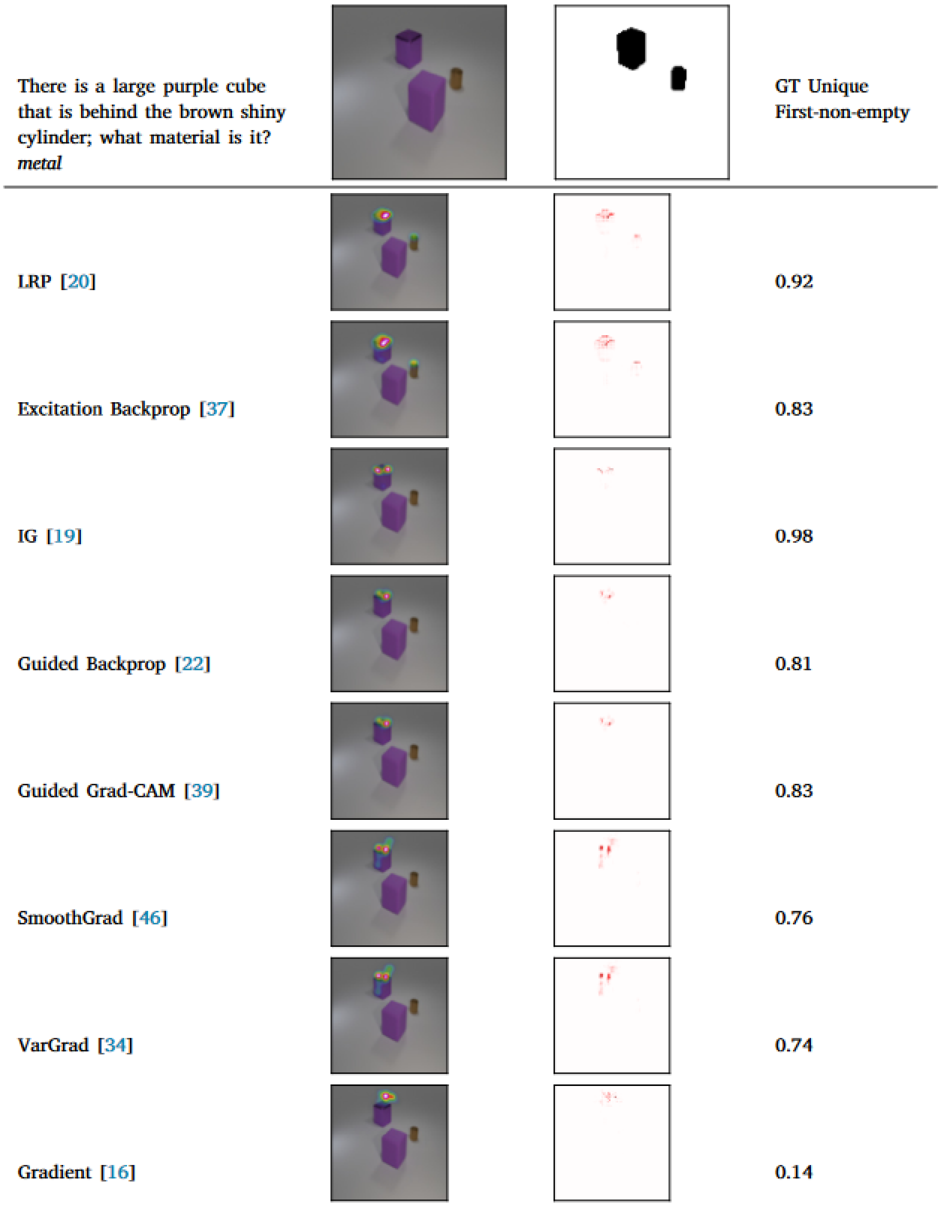
The same principle is followed in **monotonicity**, which checks that feature importances are detected correctly by calculating Spearman's correlation coefficients for all attributions that have a relevance value and are added in increasing relevance order. If the prediction accuracy increases monotonically, then the feature importances conform monotonicity. [5], [29]

To take feature interactions into consideration, Nguyen and Martínez also propose **effective complexity**, which is also calculated for the k most important features but with equation 2:

$$k^* = argmin_{k\in\{1,\dots,N\}}|M_k|\ s.t. E(l(y^*, f_{-M_k})|x^*_{M_k} < \epsilon \tag{2}$$

where M_k is the set of k most important features, f_{-M_k} is restriction function for the less important features of model f given the fixed values of the features in M_k and ϵ is a chosen tolerance. This is an improvement to many previous metrics as it does not assume that the different features are independent from each other.[29]

Other faithfulness measuring metrics are **relevance mass accuracy** and **relevance rank accuracy** introduced by Arras et al. and **pointing game** proposed by Zhang et al. which determine how well the explanatory saliency map fits to the ground truth. **Relevance mass accuracy,** as demonstrated in figure 3, and **pointing game** measure the ratio of relevance values at each pixel of the saliency map and relevance values at each pixel of the ground truth. [34], [35] **Relevance rank accuracy** instead measures how much of the highest intensity pixels of the saliency map is located at the same site as the ground truth. [34] Benefit of **fidelity**, **relevance rank accuracy**, **relevance mass accuracy** and **pointing game** is that they present the ability of the explanatory heatmap to pick the most important features of the classifier as a simple value between 0 and 1. However, for evaluating instances with large segments of image, the results are less useful as the denominator is close to the nominator [35].



There is a large purple cube that is behind the brown shiny cylinder; what material is it? *metal*			GT Unique First-non-empty
LRP [20]			0.92
Excitation Backprop [37]			0.83
IG [19]			0.98
Guided Backprop [22]			0.81
Guided Grad-CAM [39]			0.83
SmoothGrad [46]			0.76
VarGrad [34]			0.74
Gradient [16]			0.14

Figure 3: Relevance mass accuracy calculated according to explanations created with different methods for a CLEVR-XAI prediction. Cropped image from Arras et al. ***[34]***

Further metric for evaluating relevance of a saliency map is **attribution localisation** proposed by Kohlbrenner et al. **Attribution localisation** is very similar to **relevance mass accuracy** but weighs the relevance inside the explanation’s bounding box to total relevance in image ratio with the size of the bounding box to total image size ratio. As in previous metrics, higher attribution score indicates a better explanation as the explanation has succeeded to find the object rather than the background. [36]

The model agnostic **sufficiency** metric proposed by Dasgupta et al. and the not peer reviewed **target class validity** for counterfactual explanations by Mahajan et al. are crucial metrics for evaluating how convincing the explanations are. Both of these metrics test if an explanation given for a prediction holds in other cases where the feature used in it is present. [18], [37] That is, if a feature x is used to explain prediction π, then other instances where feature x is present should get labelled as π to be **sufficient**, even if the instance was explained differently [18]. Similarly, if an explanation says “If you smoked, you would have been predicted to be in risk of a lung cancer”, **target class validity** is achieved if the prediction would change to being in risk of lung cancer if the parameters given for the model were the same except for smoking. **Target class validity** of a model is calculated as a percentage of counterfactuals that were classified into the target class. [37] These metrics ensure that the explanations are faithful and that the features used as explanations are actually relevant. At least **sufficiency** can also be verified by humans [18].

3.2.3 Metrics for evaluating understandability

Understandability of an explanation is an integral part of explainability. Even if the explanation was faithful and robust, it cannot be explainable if the user is uncapable of understanding it. Although understandability is often measured with users in the loop, there are some metrics that can be used to evaluate understandability without involving users.

Continuous proximity and **categorical proximity** proposed by Mothilal et al. are metrics designed to evaluate understandability of counterfactual explanations. They are based on the assumption that counterfactual that are similar but not identical to the original instance would be the most useful to the user and make the explanation most understandable. They are calculated as the negative vector distance between the instance and its counterfactual’s features. The only difference between these two is the method of calculating the distance. [38] **Continuous proximity** is calculated for continuous features as average distance and median absolute deviation can be calculated as well. **Categorical proximity** however, is calculated as the mismatch of the categorical values between the instances and the counterfactuals. [37], [38]

Complexity is a metric introduced by Bhatt et al. to check how many features of the input are included in the explanation. Although including all the features would result in a more accurate explanation, Bhatt et al. argue that it would make the explanation more difficult to understand. [16]

3.2.4 Miscellaneous metrics

One of the metrics not having humans in the loop but still measuring explainability is **Degree of Explainability (DoX)** introduced by Sovrano and Vitali. To evaluate the **DoX**, first different “explanation archetypes” must be defined. These are derived from social sciences and they are questions such as who, where why, what next, what if and so on.

When the archetypal questions for a model are defined, the DoX can be calculated according to equation 3.

$$DoX = \frac{\sum_{q \in Q} R_{D,q,A}}{|Q|} \tag{3}$$

where q is an archetype question to an aspect a of all aspects A, Q is the set of all possible archetypes q, and $R_{D,q,A}$ is the explanatory illocution per archetype which **DoX** is an average of over the set Q. The **DoX** does require some resources from the evaluators and due to the different possible archetypes, meaningful comparison of DoX is difficult between XAI models. [14]

Another widely agreed aspect of a good explanation is that explanations for different instances or even in different classes should not be the same. **Contrastivity** by Pope et al. and not peer reviewed **separability** by Honegger et al. both propose their own implementations of this. **Separability** does not make a difference between the outcomes of the prediction unlike **contrastivity**, which requires binarized saliency maps. **Separability** and **contrastivity** are also designed for different explanation methods **separability** comparing feature importances and **contrastivity** heat maps. [22], [39] Even so, the principle can be applied for other explanation methods as well. **Contrastivity** is calculated as:

$$\frac{d_H(m_0, m_1)}{m_0 \vee m_1}, \tag{4}$$

where d_H is the Hamming distance between saliency maps m_0 and m_1 and is normalized by the total amount of atoms in the saliency maps. Again, higher **contrastivity** and **separability** indicate better explanation. [22]

Sparsity by Pope et al. describes how the explanation covers the graph. **Sparsity** is calculated by dividing the amount of nodes considered in at least other of the maps m_0 and m_1 by the total amount of nodes in the graph $|V_0|$ and subtracting this from one:

$$1 - \frac{m_0 \vee m_1}{|V_0|}. \tag{5}$$

High **sparsity** is important if the graph to be explained is very large as a very local explanation might miss some important aspects of the classifier. Then again, as **complexity** points out, too comprehensive explanation is not good either. [22]

Amparore et al. propose **prescriptivity** to evaluate how well a local linear explainer reveals the least change required in the input that would change the prediction. This is an interesting explainability metric as local linear explainers are not always expressive with this aspect unlike for example counterfactual explanation. **Prescriptivity** is calculated as

$$l\left(\frac{1}{C}\cdot|f(x')-g(x')|\right) \qquad (6)$$

where $l(\cdot)$ is a hinge loss function, $C = \max\left(\mathrm{y}', 1-\mathrm{y}'\right)$, y' is a boundary, often $\frac{1}{2}$ but could be any, f(x) is the prediction function and g(x) is the explanation function. [21]

Model parameter randomisation test proposed by Adebayo et al. evaluates how much an explanation method using saliency maps depends on the AI model's parameters. This is done by measuring the difference between outputs of the explanation method when same instance is given for a properly trained AI model and to an untrained AI model of which parameters are picked randomly. If the explanation method is sensitive to the model parameters, the output saliency maps will have significant differences. [40]

The data randomisation test proposed by Adebayo et al. is a method to reveal if the explanation depends on instances and their labels. In **data randomisation test** the difference between explanations produced for same instance and created for a model trained with random labels and a model trained with real labels is calculated. If the explanations are similar, it can be assumed that the explanation method does not base its output on the relationship of the instance and label. In other words, if we want an explanation about the reasons leading to the prediction, an explanation should be affected by randomisation of labels. [40]

Fairness and bias are some of the most common concerns connected to AI and while explainability in itself should tackle the problem at some level, it is important to be able to measure this as well. **Statistical parity** is a metric introduced by Dwork et al. and it is originally used to measure group fairness of AI algorithms. **Statistical parity** is calculated by finding Earthmover distance for μ of two different groups and if the distance is lower than a selected threshold, the model conforms **statistical parity**. **Statistical parity** exposes bias against different groups in algorithms by showing how the outcomes vary according to demographic group of test subjects. This is an important metric as AI models are prone to bias as the training sets often fail to depict the whole population therefore producing less accurate outcomes for some demographic groups. This is especially true in the field of healthcare and thus it is essential to consider statistical parity in model development. [41] Although statistical parity is originally developed to be used for AI algorithms instead of explanations, it can be applied to evaluate explainability as well [23], [41]

3.3 Frameworks for evaluating XAI

In addition to several metrics for evaluating explainable AI, there are also frameworks combining the different metrics and methods into one evaluation process. Frameworks are a useful addition to the field, as their developers have collected the most important metrics and established convincing means of utilising them. However, anyone desiring to evaluate XAI models with a particular framework needs to make sure that it measures the desired property of XAI. As with metrics, there are frameworks for different purposes, for example measuring usability like Dieber and Kirrane, reliability with multiple quantitative metrics like Agarwal et al. or for rigorous testing with **Quantus** of Hedström et al. [23], [30], [42].

One domain independent XAI evaluation method is **Model Usability Evaluation (MUsE)** proposed by Dieber and Kirrane. MUsE evaluates usability of explanation methods with user interviews of which questions are based on usability standard ISO 9241-11:2018 and related guidelines. The system was tested with LIME graphs and the users were academics with at least some understanding about AI. The evaluation questions are grouped into three subsets measuring the explanations effectivity of achieving interpretability, the resources required to reach interpretability and if the explanation method is satisfactory in the given context. The subquestions are open questions and the framework does not in itself provide clear instructions on how to evaluate them, but different metrics described in this thesis could be used for some. **MUsE** is developed to evaluate usability of XAI models used on tabular data and the output is natural text. Hence, it might not be the most effective evaluation method for all purposes but provides valuable information about usability in certain applications. While Dieber and Kirrane admit that **MUsE** is not enough for rigorously evaluating and validating XAI models, it is a useful tool for evaluating user experience which is an important aspect of XAI. **MUsE** is originally designed to evaluate LIME explanations but is model agnostic and can be used for other explainability methods as well. [42]

Agarwal et al. combine four metrics to evaluate an explanation reliability and faithfulness for graph neural networks [23]. First, faithfulness is calculated as **fidelity** by Pope et al. ie. determining the loss of accuracy when only most important features are preserved [22], [23]. The second metric derived from previous studies is **stability** from Yuan et al. measuring the distance between two similar explanations [20]. Final metric from other studies is **group fairness preservation** defined by Dwork et al. as **statistical parity** and tweaked to apply to graph neural network explainers. This is done by calculating **statistical parity** of a node and its perturbations in a neural network and subtracting the **statistical parity** of their explanations. The smaller the absolute value of the subtraction is, the better the explanation preserves group fairness. [23], [41] Finally, **counterfactual fairness preservation** is calculated.

Counterfactual fairness preservation is a metric introduced by Agarwal et al. to evaluate how well the explanation tolerates minor perturbations in the input. The main principle is familiar from many previous metrics: if the changes in input do not affect the AI model, they should not affect the explanation either. [23]

The combination of metrics Agarwal et al. use is effective in evaluating an explanation quantitatively. It is a useful framework for evaluating explanations as it covers many of the important aspects of a good explanation: fidelity, robustness, sensitivity and fairness. The framework should also be applicable beyond graph neural networks as the metrics, although here designed for graph neural networks, are general and mostly derived from completely model agnostic sources. On the other hand, Agarwal et al. do not consider the user aspect of the explanation thus failing to make the framework fully comprehensive in terms of evaluating explainability. [23] Adding an expert-involving metric to the framework could make it quite effective.

Another framework for evaluating explainability of AI has been proposed by the High-Level Expert Group on Artificial Intelligence (AI HLEG) established by the European Commission. This framework is a transparency subsection of the **Assessment List for Trustworthy Artificial Intelligence (ALTAI)** consisting of two questions. Explainability is assessed according to **ALTAI** framework by evaluating following questions:

1. “Did you explain the decision(s) of the AI system to the users?”
2. “Did you continuously survey the users if they understand the decision(s) of the system?”

This is certainly a general domain framework containing also many evaluation tools for other aspects of trustworthy AI even though explainability is left quite incomprehensive. [43]

Z-inspection by Zicari et al. is a rigorous framework for evaluating trustworthy AI and it has been proven to work also in healthcare domain. As explainability is one key aspect of trustworthy AI, it is also included in the **Z-Inspection** framework, but it can be evaluated with any metrics or methods that the evaluators feel suitable for the purpose. The **Z-Inspection** is a qualitative analysis and could well be utilised in evaluation process of any AI system that is intended to be trustworthy and explainable. [44]

Finally, although not being a framework as such, it is worth to mention **Quantus** toolkit by Hedström et al. **Quantus** is a Python library where many of the metrics discussed in this thesis are implemented and ready to be used for evaluating XAI. **Quantus** has grouped the metrics according to what they measure and the metrics are shortly described in the initial comments of the implementation. [30] This should be a useful toolkit for XAI developers to evaluate their models but does not replace the need of expert-in-the-loop metrics.

4. METHODS AND METRICS FOR EVALUATING EXPLAINABLE ARTIFICIAL INTELLIGENCE IN HEALTHCARE DOMAIN

Many metrics and methods that were not originally developed for evaluating XAI in healthcare suit well for medical domain. However, sometimes it is important to evaluate the XAI model in a more domain specific manner, particularly because medical experts have valuable specialist information about the use cases that computer scientists might lack. Thus, involving healthcare professionals at least in the evaluation process can lead to better results. [45] This section introduces evaluation methods particularly intended to be used in healthcare applications of XAI.

4.1 Methods to evaluate XAI for healthcare applications with expert in the loop

As in general domain, user interviews about a model's explainability are a valid option for explainability evaluation in medical domain. Involving users or field experts in the evaluation process gives the developers hands-on feedback on the performance of the explainability method [13]. One example of expert surveys is one conducted by Schoonderwoerd et al. in [4]. The **DoReMi** method of Schoonderwoerd et al. is intended to be used throughout the development process and it includes several surveys to assess how well the users understand a clinical decision support system and what should be improved. This survey consists of mainly Likert scales with which usability and understandability of explanations are evaluated. **DoReMi** is developed to be used for clinical decision support systems but is likely to be applicable in other domains as well. Figure 4 depicts the steps of the method. [4]

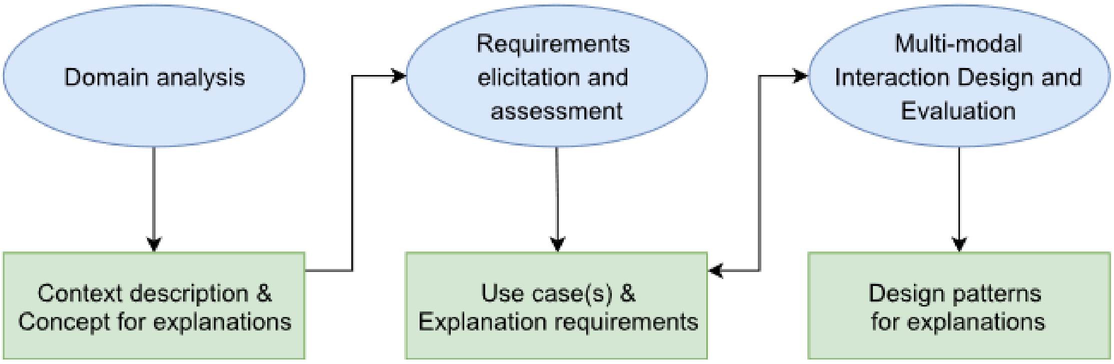

Domain analysis
Requirements elicitation and assessment
Multi-modal Interaction Design and Evaluation
Context description & Concept for explanations
Use case(s) & Explanation requirements
Design patterns for explanations

Figure 4: Steps of the DoReMi user survey method. Figure is created by Schoonderwoerd et al. ***[4]***

Another metric suitable for healthcare XAI validation is **Trustworthy Explainability Acceptance** introduced by Kaur et al in [3]. **Trustworthy Explainability Acceptance** includes a group of experts in the evaluation process. **Trustworthy Explainability Acceptance** is calculated by first evaluating the Euclidean distances between explanations provided by an expert X_i and the explainability model $Y_{i.}$ When the distance, which is in range [0,1], is calculated, explainability acceptance A_e of the system can be calculated as

$$A_e = 1 - d_{X^Y} \tag{7}$$

A_e is also a number in range [0,1], 1 being the best acceptance and 0 the worst. The experts also rate their trust T_e to the explanation to weigh the **explainability acceptance**. The **Trustworthy Explainability Acceptance** is calculated as an average of each experts' trust multiplied by explainability acceptance. Finally, Kaur et al. propose to calculate confidence of the Trustworthy Explainability Acceptance c_{Tw_A} as

$$SE_{Tw_a} = \frac{\sqrt{\Sigma_e (Tw_A T_e)^2}}{n} \tag{8}$$

$$c_{Tw_A} = 1 - 2\left(SE_{Tw_A}\right). \tag{9}$$

Trustworthy explainability is a tuple of **trustworthy explainability acceptance** and its confidence. [3]

Modality-Specific Feature Importance (MSFI) introduced by Jin et al. is a metric developed for XAI models that use saliency maps as explanation method and that are used in clinical context. It measures explanation plausibility and faithfulness. Prior to calculating the **MSFI**, a user study is required to find the clinical requirements for the case and to prioritise the modalities present in the classification problem. Shapley values are calculated for each modality according to their importance. Secondly, the section of the saliency map that overlaps the ground truth mask is calculated for all modalities and this ratios are weighted by their corresponding modality importance values. The output is a number in range [0,1], one being the ideal situation. **MSFI** is robust metric as it is not sensitive to the size of the ground truth mask or signal strength of the saliency map. [46]

Clinical Explainability Failure and **Explainability Failure Ratio** are good examples of model agnostic metrics with experts in the loop. These metrics introduced by Venugopal et al. are developed for saliency maps in healthcare applications but are applicable to other fields as well. **Clinical Explainability Failure** is detected with a two-step test. First a model's output is compared to ground truth to see if it recognises the correct features.

If the bounding box inserted on the saliency map does not match the ground truth and an expert fails to understand why, the model is deemed to have made a **Clinical Explainability Failure**. **Explainability Failure Ratio** is then calculated as the amount of Explainability Failures divided by the total amount of explanations. [47]

4.2 Metrics to evaluate XAI for healthcare applications without expert in the loop

Insertion and **deletion** by Hu et al. are metrics designed specifically for medical x-ray imaging although arguably they could be used for saliency maps created for other purposes as well. **Insertion** and **deletion** are metrics that measure the explanation's sensitivity to changes in the input images combining **pixel-flipping** and **fidelity** like metrics producing a similarity score as an output. For **insertion** new pixels are added to a masked out input image and for **deletion** they are masked by standard grey. The pixels are added or removed in relevance order that is got from the original explanatory saliency map. Similarity score *s* is calculated as

$$s(f_q, f_{\hat{r}}) = \max\left(\frac{\boldsymbol{f_q} \cdot \boldsymbol{f_{\hat{r}}}}{|\boldsymbol{f_q}||\boldsymbol{f_{\hat{r}}}|}\right), \tag{10}$$

where $\boldsymbol{f_q}$ is the feature vector of the original query image and $\boldsymbol{f_{\hat{r}}}$ the feature vector for the perturbed images. Finally, area-under-the-curve is calculated for the perturbation sets and higher results are desired from insertion and lower from deletion as these show that the explanation has located the most relevant features. **Insertion** and **deletion** are relevant metrics and unlike most metrics designed for healthcare applications of XAI, they do not require a lot of human resources. [48]

5. DISCUSSION

One of the main findings of this thesis is that there are many different metrics and methods for evaluating XAI and most of them suit assessing healthcare applications very well too. To help quick review of the metrics they are collected in alphabetical order and grouped and summarised in tables.

5.1 Summary tables of the metrics and methods

Table 1 contains all general domain expert-in-the-loop methods for evaluating XAI that were found. Most of the methods are very general and therefore suit any field including healthcare but to highlight one, **Likert scales** could be discussed in more detail. They are very modifiable and easy to fill, and overall **Likert scales** are good for measuring confidence and trust in the system if the questions are formed well. It is hard to quantify user trust without involving them in the evaluation and therefore methods like **Likert scales** are needed. Benefit of **Likert scales** is that it is a quantitative metric thus making it easier to compare results of different systems.

Table 2 summarises metrics for evaluating the explanation's resistance to small perturbances. This measure is based on the widely accepted axiom that similar explanations should be given to similar instances with the same prediction. Robustness is an essential aspect of an explanation as very different explanations for very similar instances can degrade users' trust in the system and also indicates that the explanations might not be accurate. [15], [16] This is the one of the metric groups that is so popular that many scientists have developed their own metrics while agreeing that this is something worth to measure from XAI models. As robustness can be measured quite reliably with general domain metrics for any field of application, some of these metrics are certainly useful for healthcare applications as well. For example, **input invariance** could be a valuable metric to be included in evaluating XAI saliency maps for healthcare.

Explanation faithfulness is evaluated with even more metrics than robustness. Table 3 contains those found for this thesis and most of them are essentially the same. The most common logic within these metrics is that most relevant features of an explanation are set in relevance order, after which they are either removed or inserted iteratively and the effect of removal or insertion in the prediction accuracy is calculated. These are good metrics to evaluate credibility and faithfulness of the explanation, especially when it is based on feature importances as with saliency maps or decision trees. However, many of these metrics do not take in account that the actual object to be detected may take up quite large portion of the whole image, making the results of simpler metrics less useful. This problem is taken into account in **attribution localisation** which makes it stand out slightly. Even so, in especially in medical image analysis, investigating segments instead of separate pixels likely produces better results and for example **IROF** is a metric that both works with segments and is tested with XAI systems in medical context [31].

Rather smaller group of metrics is one for those assessing understandability of explanations. As the **proximities** and **complexity** are for different types of explanations they are slightly difficult to compare. Understandability of an explanation is a key requirement to achieve explainability and therefore it should be measured when evaluating XAI. In healthcare domain this is particularly important as the decisions to be made have importance to patients' health and therefore the AI systems should be properly understood. These metrics are summarised in table 4.

Not only are there many grouping options for metrics but also many metrics that do not obviously belong to any larger group. These metrics are summarised in table 5. Some important metrics are found from these too. For example, **contrastivity** and **separability** are relevant to establish reliability of the explanation and should also be considered in healthcare applications too.

Fairness is another less measured quality of an explanation despite arguably being an important one. Interestingly no metrics were found that would have been particularly targeted for healthcare use. Even so, the general domain metrics can be used in detection of bias even in healthcare applications. As only one metric, **statistical parity**, explicitly measures explanation fairness, it is included in table 5.

Possibly the most useful tools for someone desiring to evaluate XAI – be it in healthcare domain or in another context – are frameworks. Frameworks establish a process that ensures proper evaluation of an XAI model and eliminates the need of every developer to come up with valid and rigorous evaluation processes. There are existing frameworks and guidelines for evaluating different aspects or combinations of aspects of metrics. These aim to evaluate explanations more rigorously than single metrics and often include experts in the evaluation process as well. Tables identified in this thesis are listed in table 6.

Table 7 shows the metrics designed for evaluating healthcare XAI with expert in-the-loop. When evaluation methods tailored for healthcare applications of XAI are developed, experts are often involved in the evaluation processes. This is sensible not only because explainability is most surely achieved when it is verified that the users understand the explanations but also since healthcare applications of XAI may have specialities that can be better identified by experts. All the methods described can be valuable in assessing healthcare XAI. While the need to choose suitable methods for the intended problem, the **MSFI** and **clinical explainability** are good for evaluating faithfulness of saliency methods and **DoReMi** and **Trustworthy explainability acceptance** are model agnostic methods to evaluate expert perspective and trust.

Only few metrics without experts in the loop were tailored particularly for healthcare use, probably because many general domain metrics suit well healthcare applications and vice-versa. The identified metrics in this category are summarised in table 9. **Deletion** and **insertion** are good faithfulness metrics for healthcare, but as they only consider individual pixels in the evaluation process, a metric deleting or inserting segments could turn out more useful for this purpose [31].

With a good combination of different methods and metrics evaluation of XAI can be made rigorous and worthwhile. Thorough evaluation is essential to ascertain validity of the explainability model and that the end users truly understand the explanations. The existing frameworks and guidelines combined with these metrics could be utilised in the validation processes and even harmonised into standards.

Table 1: General domain metrics and methods with expert in the loop

Metric	Input datatype	Inputs	Logic	Output	Benefits and challenges	Explainability method	Source
Anchor user study	Text	Generated explanations, users' predictions	Users are asked to predict the function of a model based on previous explanations. They are asked to only answer if they're confident, otherwise "I don't know". High percentages indicate goodness.	Percentage of accurate predictions	+ Actual users' understanding is measured - Laborous - Users need experience on machine learning	Model agnostic	Ribeiro et al. [12]
Likert scales	Integer	Users' answers on a numerical scale to question tailored for the study.	User fills a questionnaire with options on a numerical scale. Questions can be tailored for the study.	N answers to the survey question, typically on scale 1-5.	+ Easy to modify + Tailormade questions + User friendly - Requires lot of resources and answerers - Data quality depends on the question and sample	Model agnostic	Many, eg. Antoniadi et al. [9], Chakraborti et al. [8], Hoffman et al. [2], Holzinger et al. [10] and Sanneman

							and Shah. [7]
SAFE-AI modification of the SAGAT test	Text	Users answers in a frozen simulation situation	Users are surveyed to find out their informational needs to enhance explainability	Information on how well an explainability method conforms users' needs for information	+ User centric - Large-scale studies require a lot of resources	Model agnostic	Sanneman & Shah, Endsley [7], [11]
System Causality scales	Integer on scale 1-5	Users answers to the question on five-point scale	Evaluates system causality with specified Likert scale	A number in range [0,1]. The closer to 1, the better the system.	+ Applicable for healthcare	Model agnostic but aimed for healthcare applications	Holzinger et al. [10]

Table 2: Metrics for evaluating robustness

Metric	Input datatype	Inputs	Logic	Output	Benefits and challenges	Explainability method	Source
Consistency	Not specified	Predictions of similar instances that produce same explanations	instances with same (good) explanation should have same prediction	Probability distribution	Not specified in the document	Model agnostic	Dasgupta et. al [18]
Input invariance	Image	Explanations with and without added noise so that the noise has not affected the model's attributions	Explanation should not be affected by noise if the outcome has not been affected either.	Distance between the input explanations.	+ Effective way to measure robustness, takes sensitivity further - Chosen reference point affects the result making comparability difficult	Saliency methods (model specific)	Kindermans et al. [17]
Local Lipschitz continuity	Number, prediction function	Explanations close to local point x_0	Lipschitz criterion but implemented only locally	Local Lipschitz value	Not specified in the document	Model agnostic	Alvarez-Melis & Jaakkola [19]
Reiteration similarity	Not defined	Explanations for an instance x	Calculates the Jaccard similarity for all explana-	A number in range [0,1]	+ Very important sanity check	Model agnostic	Amparore et al. [21]

			tions for an instance. The more similar the explanations, the more robust method.				
Sensitivity	explanation function, prediction function	Explanations of slightly different instances that are classified in the same group	If an explainer has low sensitivity, explanation is not significantly affected by noise or other small perturbations	Distance between an explanation and its perturbations.	- Requires that the model works similarly in both instances used in evaluation	Neural networks (model specific)	Bhatt et al. [16], Alvarez-Melis & Jaakkola [15]
Stability	Graph	Node and its perturbation and their corresponding explanations	Count distance between explanations of similar instances.	Distance between explanations	+ Tells if perturbations that do not affect the model outcome affect the explanation	Graph Neural Networks (model specific)	Yuan et al. [20]

Table 3: Metrics for evaluating XAI faithfulness

Metric	Input datatype	Input	Logic	Output	Benefits and challenges	Explainability method	Source
Attribution localisation	Image	Relevance inside the bounding box, relevance in the image, size of image and the size of the bounding box.	Relevance inside the bounding box to total relevance is calculated and then multiplied by size of image to size of bounding box ratio.	A number, no range but higher is desirable.	+ Useful even when ground truth is large compared to the total image	Saliency methods (model specific)	Kohlbrenner et al. [36]
Completeness	Number	Attribution for output F and a baseline	Checks that amount of attributions in a heatmap is valid	Optimally 0 but essentially the difference between attributions to output F at instance x to a baseline x'.	Not specified in the document	Deep neural networks (model specific)	Sundarajan et al. [24]
Continuity	Image	Explanations for all inputs	Maximum variation of the explanation function is calculated over the input domain	Strongest variation	- Only applies if prediction function is continuous.	Deep neural networks (model specific)	Montavon et al. [27]

Effective complexity	Number	Features, their importance and a tolerance	Similar to other faithfulness metrics but the equation takes into account feature interactions	A number describing effective complexity. The lower the better.	+ Not affected too much by feature interactions.	Model agnostic	Nguyen and Martínez [29]
Faithfulness / faithfulness estimate	Number	Prediction accuracies when most important features are added.	Faithfulness is calculated as Pearson’s correlation of prediction accuracy and addition of most important features. High correlation is an indicative of a good explanation	Pearson’s correlation of amount of the most important features used and the prediction accuracy.	Not specified in the document	Feature importance methods (model specific)	Alvarez-Melis and Jaakkola, Arya et al. [5], [19]
Faithfulness / faithfulness correlation	Number	Prediction accuracy with decreasing amount of features	Higher correlation of drop in prediction accuracy and reduction of features indicates a better explanation	Correlation score	- Baseline can be defined in many ways - Difficult to aggregate	Feature importance methods (model specific)	Bhatt et al. [16]

Fidelity	Number	Accuracy before and after removing most important features.	The higher the difference in accuracy the higher the fidelity	A number in range [0,1]	+ Reveals easily if the explanation can recognise most important features	Feature importance methods (model specific)	Pope et al. [22]
Focus	Image	Most important features in mosaics	Works as fidelity but image is replaced with mosaics	number in range [0,1]	+ No need for pixel-wise labelling due to image division into mosaics.	Saliency methods (model specific)	Arias-Duart et al. [28]
Iterative Removal of Features (IROF)	Image	Accuracy before and after removing most important image segments.	Works as fidelity but image segments are removed instead.	A number in range [0,1]	+ Avoids the problems caused by related pixels through image segmentation. - Segments have to be selected separately.	Saliency methods (model specific)	Rieger & Hansen [31]
Local concordance	Number	Approximation g of a model and a conciseness constraint.	Like fidelity but only features with importance higher than	A number in range [0,1]	Not specified in the document	Local linear methods (model specific)	Amparore et al. [21]

			the conciseness restraint are considered.				
Monotonicity	Number	Spearman’s correlation coefficients for all features and prediction accuracies when features are added in increasing relevance order.	Features are added in an increasing relevance order. If prediction accuracy increases monotonically, the explanation is sound.	Boolean value, where true means monotonic explanation.	+ Relevant sanity check. - Only applicable for individual features.	Feature importance methods (model specific)	Nguyen & Martínez [29]
Non-sensitivity	Number	Features with zero importance and features that the model’s functioning is not dependent of.	Symmetric difference between the input sets is calculated. Only features that do not affect the prediction should be given relevance value 0.	Symmetric difference, preferably low.	- Only applicable for individual features	Feature importance methods (model specific)	Nguyen & Martínez [29]

Pixel-flipping	Image	Prediction accuracies when most important pixels are masked.	Works like fidelity but single pixels are masked and area under the prediction accuracy curve is calculated.	Area under the curve	- Not very useful if pixel importances depend on surrounding pixels	Saliency methods (model specific)	Bach et al. [33]
Pointing game	Image	Amount of explanation's pixels inside ground truth, total amount of pixels.	The ratio of pixels inside the ground truth bounding box and total pixels is calculated. High scores tell that explanation is sound.	A number in range [0, 1]	- Trivial when ground truth's bounding box is large.	Saliency methods (model specific)	Zhang et al. [35]
Region perturbation	Image	Accuracy before and after removing most important image segments.	Works as fidelity but image segments are removed instead.	A number in range [0,1]	+ Avoids the problems caused by related pixels through image segmentation. - Segments have to be selected separately.	Saliency methods (model specific)	Samek et al. [32]

Relevance mass accuracy	Image	Relevance value at pixels, set of pixels within the ground truth mask, number of pixels inside it and the total number of pixels.	Explanation mass within ground truth over the total mass.	A number in range [0, 1]	Not specified in the document	Saliency methods (model specific)	Arras et al. [34]
Relevance rank accuracy	Image	Pixels of a saliency map and ground-truth bounding pixels	Ratio of pixels of a saliency map that lie inside the ground truth.	A number in range [0, 1]	Not specified in the document	Saliency methods (model specific)	Arras et al. [34]
Sensitivity-n	Image	Explanations of slightly different instances that are classified in the same group.	If an explainer has low sensitivity, explanation is not significantly affected by noise or other small perturbations	Drop in prediction accuracy in range [0,1]	+ Only most important features are considered	Gradient-based attribution methos (model specific)	Ancona et al. [26]
Selectivity	Image	Prediction accuracies when most important image segments are masked.	Works like fidelity but image segments are masked and area un-	Area under the curve	+ Better for healthcare-like situations where	Saliency methods (model specific)	Montavon et al. [27]

			der the prediction accuracy curve is calculated.		pixes’ importances depend on the surroundings.		
Sufficiency	Text	Property pi of an explanation, instance x for which pi is included in its explanation, instance x' which also includes property pi and their predicted classes.	If a property is used to explain a prediction for instance x, then instance x' with the same feature should be classified similarly even if the explanations were different.	Probability distribution.	+ Important sanity-check for counterfactual image explanations.	Counterfactual methods (model specific)	Dasgupta et al. [18]
Summation-to-delta	Image	Attribution for output F and a baseline	Checks that amount of attributions in a heatmap is valid	Optimally 0 but essentially the difference between attributions to output F at instance x to a baseline x’.	Not specified in the document	Saliency methods (model specific)	Sundarajan et al. [24]

Target class validity	Text	Target classes and predicted classes of counterfactuals.	Calculates the percentage of counterfactuals which are predicted to belong to the targeted class.	Percentage of valid counterfactuals	+ Important sanity-check	Counterfactual methods (model specific)	Mahajan et al. [37]

Table 4: Metrics for evaluating understandability

Metric	Input datatype	Input	Logic	Output	Benefits and challenges	Explainability method	Source
Categorical proximity	Number	Categorical values of instance and their counterfactuals.	Counterfactuals alike the instance should be easier to understand	Mismatch between the inputs	Not specified in the document	Counterfactual methods (model specific)	Mothilal et al. [38]
Complexity	Number	Number of features covered in explanation, total amount of features	Including all features in the explanation makes it too difficult to understand	Ratio of used features and total features	- Determining how much is too much may be difficult	Feature importance methods (model specific)	Bhatt et al. [16]
Continuous Proximity	Text	Average distance and median absolute deviation of the continuous features	Counterfactuals alike the instance should be easier to understand	A number, no range	Not specified in the document	Counterfactual methods (Model specific)	Mothilal et al. [38]

Table 5: Miscellaneous metrics

Metric	Input datatype	Inputs	Logic	Output	Benefits and challenges	Explainability method	Source
Contrastivity	Image	Binarised saliency maps of same instance and the number of atoms in the maps in total	Different instances should not get similar explanations which is checked by calculating distance between maps for different classes.	Hamming distance between the input maps divided by the amount of atoms.	Not specified in the document	Saliency methods (model specific)	Pope et al. [22]
Data randomisation test	Image	Explanations for a model trained with randomised data labels and a model trained with the original data	Explanation should be sensitive to the relationship between an instance and its label.	Similarity of the explanations, preferably low.	Not specified in the document	Saliency methods (model specific)	Adebayo et al. [40]

Degree of explainability	Text, number	Set of explanandum aspects, archetypes, details of support material, method to define the details for each aspect and the function to calculate a detail's relevance to an archetypal question about an aspect	DoX score for each archetype	A number describing the degree of explainability for each archetype.	- Complex - Requires relatively lot of resources - Does not produce comparable results due to different archetypes	Model agnostic	Sovrano & Vitali [14]
Model parameter randomization test	Image	Explanation for model with randomised parameters and explanation for the original model	Explanation created by a saliency method should be different for a trained model and an untrained model.	Similarity of the explanations, preferably low.	Not specified in the document	Saliency methods (model specific)	Adebayo et al. [40]

Prescriptivity	Number, explanation function, prediction function	Boundary y’, explanation function, prediction function,	Hinge loss function evaluates the change required.	Ability of a local linear explainer to explain the smallest change required to alter the prediction class.	Not specified in the document	Local linear explainers (model specific)	Amparore et al. [21]
Separability	Image	Distance between instances and distance between the explanations.	Different instances should not get similar explanations.	Boolean, where truth means separable explanation and false inseparable.	Not specified in the document	Saliency methods (model specific)	Honegger et al. [39]
Sparsity	Graph	Amount of nodes covered by the explanation, total amount of nodes in a graph	Tests how widely over the graph an explanation is scattered. If an explanation is sparse, it is easier to understand the whole system.	Ratio of non-covered nodes in range [0,1]	Not specified in the document	Graph neural networks (model specific)	Pope et al. [22]

Statistical parity	Not specified	μs for different groups	Distance between outcomes of similar instances in different (demographic) groups.	Distance between outcomes of similar instances of different groups	+ Can point out algorithm bias towards groups - Not exhaustive - Not originally intended for evaluating explainability	Model agnostic	Dwork et al. [41]

Table 6: Frameworks for evaluating XAI

Framework	Inputs	How it works	Output	Benefits and challenges	Explainability method	Source
ALTAI	Developers answers to two questions	AI trustworthiness and explainability as subsection of two questions.	Linguistic answers	+ It is a framework introduced by an official committee under European commission thus likely to be considered in future legislation. - Very narrow in evaluating explainability.	Model agnostic	High-Level Expert Group on Artificial Intelligence [43]
Counterfactual fairness preservation	A node, its sensitive perturbation and their explanations and subgraphs	Distance between outcomes of different explanations	A number indicating counterfactual fairness mismatch	Not specified in the document	Graph neural networks (model specific)	Agarwal et al. [23]
MUsE	Explanation and users' answers to interview question	Different questions about usability are answered by users.	Natural language.	+ Based on standards - Laborious	Model agnostic	Dieber & Kirrane [42]

Quantus	Not applicable	Python library for evaluating XAI	Not applicable	+ Contains many XAI evaluation metrics readily implemented.	Model agnostic	Hedström et al. [30]
Z-Inspection	Not applicable	Framework to evaluate trustworthy AI. Explainability is measured with a metric chosen freely.	Not applicable	+ Can be tuned for each case	Model agnostic	Zicari et al. [45]

Table 7: Expert in the loop methods for healthcare XAI evaluation

Framework	Input datatype	Inputs	Logic	Output	Benefits and challenges	Explainability method	Source
Clinical Explainability Failure and clinical explainability failure ratio	Image	Ground truth bounding boxes, explanation bounding boxes, expert's opinion	If the ground truth and explanation do not overlap, a clinician checks the prediction and if there is no clear reason for the difference, the situation is considered to be clinical explainability failure. Finally all CEFs are divided by the total amount of tests.	Ratio between [0,1] where low number is desirable.	+ Evaluates faithfulness but includes an expert to also consider possible "acceptable" mistakes.	Saliency methods (model specific)	Venugopal et al. [47]
DoReMi	Integer	Users' answers to specified Likert scales	Rigorous user experience evaluation is used throughout development of a clinical decision support system to ensure usability and explainability.	Numerical results for the Likert scale.	+ Questions tailored for assessing a clinical decision support system's different aspects.	General (model agnostic)	Schoonderwoerd et al. [4]

Modality Specific Feature Importance	Image	Ground truth maps, modality importance and saliency maps	Shapley value of the share of the saliency map overlapping the ground truth of each modality is weighted by modality importance.	A number in range [0,1]	+ Experts are used to determine importance of different modalities therefore making the output value more relevant.	Saliency methods (model specific)	Jin et al. [46]
Trustworthy Explainability Acceptance	Number, explanation	Explanations by AI and expert and experts' trust ratings.	Checks if the distance between explanations given by the computer and expert is not long and experts' trust in the system is high.	Tuple of trustworthiness of an explanation and experts' confidence in the system.	+ Quantitative method to evaluate trust, suits healthcare applications too. + Designed for healthcare - Requires some work.	Model agnostic	Kaur et al. [3]

Table 8: No-expert metrics for XAI evaluation

Metric	Input datatype	Inputs	Logic	Outputs	Benefits and challenges	Domain	Source
Deletion	Image	Importances for pixels of a saliency map, prediction accuracies when pixels are masked out in relevance	Pixels of input image are masked in decreasing relevance order and drop in prediction accuracy is measured	Similarity score	+ Made for healthcare XAI	Saliency methods (model specific)	Hu et al. [48]
Insertion	Image	Importances for pixels of a saliency map, prediction accuracies when pixels are unmasked in relevance	Pixels of input image are unmasked in decreasing relevance order and rise in prediction accuracy is measured	Similarity score	+ Made for healthcare XAI	Saliency methods (model specific)	Hu et al. [48]

5.2 From a jungle of metrics to valid frameworks

According to the studies reviewed for this thesis, medical domain XAI models mostly rely on general domain metrics without expert in the loop for evaluating them. Even so, it can be argued that particularly in evaluating XAI for healthcare, involvement of experts or even patients is important to achieve trust and satisfactory user experience [45]. Furthermore, to create explanations that support clinical decision making in real life situations and match the user group's conception of a sufficient explanation, having clinicians in the development process and uncovering their mental models is essential [2], [45]. By including interdisciplinary experts in the development and evaluation process of an XAI model, the needs of all stakeholders are more likely to be satisfied, side-effects are more probably revealed and the field specialties considered [4], [7], [45]. Thus, it would be advisable use at least some metrics with experts in the loop to credibly evaluate XAI for healthcare.

Apart from inducing trust in medical decision makers, XAI for healthcare should also be understandable for them, enable them to evaluate accuracy of the system and to provide relevant information about the decision at hand [4]. This should be achieved without causing too heavy a workload that would discourage the experts from using the XAI tools and the clinical requirements should be considered as well [7], [46]. By conforming these user requirements adoption of XAI in healthcare should be more feasible.

Furthermore, XAI needs to conform the many regulations and standards of the field. Probably one of the most important regulations to conform in the EU is the Medical Device regulation under which XAI for healthcare is likely to be classified as a medical device thus forcing the developers to conform the regulations set by it and the GDPR. While explainability is not necessarily required by either, explainability makes it easier to conform at least GDPR when it comes to black-box models [49]. Furthermore, guidelines proposed by AI-HLEG imply that regulation in the EU might be going to the direction of preferring explainability in AI use cases with high risks [50].

One less frequently thought about aspect of explainability is information security. Both Zicari et al. and Suffian et al. point out that XAI is not without the risk of information security breaches and that this is a matter that should be considered and evaluated as well. [45], [51] While no metrics found in this literature review particularly assessed information security of explanations, methods for validating XAI in this sense would be valuable too. Information security and data privacy are essential particularly in medical field as the information collected is typically classified as sensitive and therefore under tighter regulations than insensitive data. Furthermore, for example in the EU the GDPR requires the patient’s consent for using AI in his or her care [49]. Therefore, credibly data-security-wise audited XAI models are essential for feasibility of the application. [45]

Due to this variety of requirements and needs on the XAI and healthcare fields, there is a clear need for officially approved validation processes for XAI. Despite the amount of different evaluation metrics and methods and opinions on what actually should be measured, most researchers seem to agree that external and official validation, even standardisation is needed for the field. [51]–[53] It is clear that interdisciplinary cooperation and determination is needed to create clear and thorough standards or other validation methods to evaluate explainable artificial intelligence used in the complex world of healthcare.

5.3 Limitations of the review

Despite the high effort put in the thesis, the work has limitations. While most of the papers included in the review were clear and expressive, some papers would have required more background knowledge from the reader to be completely understandable.

There is such a variety of metrics and methods for evaluating XAI that certainly some have been accidentally missed out. This means that the tables may not be fully comprehensive. Furthermore, most metrics did not appear as results for the search query but were found from references of other papers and therefore some metrics are almost certainly left undetected.

6. CONCLUSION

In this thesis, different metrics, methods and frameworks for evaluating explainable artificial intelligence in healthcare domain are identified, discussed and finally summarised in tables. Metrics are categorised in the tables according to what they measure and their inputs, logic, outputs and benefits and challenges are also listed in the tables. Metrics are also divided in those that are developed particularly for healthcare domain and those that are not. This was done to achieve clarity, although general domain metrics are often used in evaluating healthcare XAI as well.

Overall, 54 metrics and methods were found for evaluating explainable artificial intelligence. Out of these only 6 were particularly tailored for healthcare applications of XAI but many general domain metrics were found suitable for healthcare as well. The most popular aspect of explainability that has been evaluated is faithfulness of an explanation and 22 metrics were found to assess this.

All in all, there is a wide variety of metrics, particularly those for measuring faithfulness of an explanation. Many metrics are similar to each other and particularly those measuring faithfulness of an explanation often significantly resemble each other. To help developers choose the best metrics to use, a set of metrics could be developed and standardised. This set should include metrics for different types of explainability and different aspects of an explanation ie. faithfulness, robustness and so on.

Involvement of experts in the evaluation loop varies greatly between studies. However, including experts in evaluating healthcare XAI is often encouraged although the available resources must be considered as well as not to conduct too wide user questionnaires. Healthcare is a field where experts have such insight into the use cases that computer scientists should regard it. Furthermore, it would be meaningful to test explainability with the audience that the explainability is implemented for.

All in all, there is a clear need for standardisation of XAI evaluation particularly in healthcare domain. Standardised evaluation processes that include instructions on what to evaluate and with which methods and metrics would ensure sufficiency and credibility of XAI evaluation. Furthermore, official guidelines on when to have experts in the loop would be valuable and harmonise different evaluating paradigms.

REFERENCES

[1] A. Barredo Arrieta *et al.*, ‘Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI’, *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[2] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, ‘Metrics for Explainable AI: Challenges and Prospects’, *ArXiv181204608 Cs*, Feb. 2019, Accessed: Jan. 11, 2021. [Online]. Available: http://arxiv.org/abs/1812.04608

[3] D. Kaur, S. Uslu, A. Durresi, S. Badve, and M. Dundar, ‘Trustworthy Explainability Acceptance: A New Metric to Measure the Trustworthiness of Interpretable AI Medical Diagnostic Systems’, in *Complex, Intelligent and Software Intensive Systems*, vol. 278, L. Barolli, K. Yim, and T. Enokido, Eds. Cham: Springer International Publishing, 2021, pp. 35–46. doi: 10.1007/978-3-030-79725-6_4.

[4] T. A. J. Schoonderwoerd, W. Jorritsma, M. A. Neerincx, and K. van den Bosch, ‘Human-centered XAI: Developing design patterns for explanations of clinical decision support systems’, *Int. J. Hum.-Comput. Stud.*, vol. 154, p. 102684, Oct. 2021, doi: 10.1016/j.ijhcs.2021.102684.

[5] V. Arya *et al.*, ‘One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques’. arXiv, Sep. 14, 2019. doi: 10.48550/arXiv.1909.03012.

[6] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, ‘The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies’, *J. Biomed. Inform.*, vol. 113, p. 103655, Jan. 2021, doi: 10.1016/j.jbi.2020.103655.

[7] L. Sanneman and J. A. Shah, ‘The Situation Awareness Framework for Explainable AI (SAFE-AI) and Human Factors Considerations for XAI Systems’, *Int. J. Human–Computer Interact.*, vol. 0, no. 0, pp. 1–17, Jun. 2022, doi: 10.1080/10447318.2022.2081282.

[8] T. Chakraborti, S. Sreedharan, S. Grover, and S. Kambhampati, ‘Plan Explanations as Model Reconciliation – An Empirical Study’, in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Mar. 2019, pp. 258–266. doi: 10.1109/HRI.2019.8673193.

[9] A. M. Antoniadi, M. Galvin, M. Heverin, L. Wei, O. Hardiman, and C. Mooney, ‘A Clinical Decision Support System for the Prediction of Quality of Life in ALS’, *J. Pers. Med.*, vol. 12, no. 3, Art. no. 3, Mar. 2022, doi: 10.3390/jpm12030435.

[10] A. Holzinger, A. Carrington, and H. Müller, ‘Measuring the Quality of Explanations: The System Causability Scale (SCS)’, *KI - Künstl. Intell.*, vol. 34, no. 2, pp. 193–198, Jun. 2020, doi: 10.1007/s13218-020-00636-z.

[11] M. R. Endsley, ‘Situation awareness global assessment technique (SAGAT)’, in *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*, May 1988, pp. 789–795 vol.3. doi: 10.1109/NAECON.1988.195097.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, ‘Anchors: High-Precision Model-Agnostic Explanations’, *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, Art. no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.11491.

[13] S. Mohseni, N. Zarei, and E. D. Ragan, ‘A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems’, *ArXiv181111839 Cs*, Aug. 2020, Accessed: Jan. 11, 2021. [Online]. Available: http://arxiv.org/abs/1811.11839

[14] F. Sovrano and F. Vitali, ‘How to Quantify the Degree of Explainability: Experiments and Practical Implications’, in *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Jul. 2022, pp. 1–9. doi: 10.1109/FUZZ-IEEE55066.2022.9882574.

[15] D. Alvarez-Melis and T. S. Jaakkola, ‘On the Robustness of Interpretability Methods’, *ArXiv180608049 Cs Stat*, Jun. 2018, Accessed: Feb. 02, 2022. [Online]. Available: http://arxiv.org/abs/1806.08049

[16] U. Bhatt, A. Weller, and J. M. F. Moura, ‘Evaluating and Aggregating Feature-based Model Explanations’, in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, Yokohama, Japan, Jul. 2020, pp. 3016–3022. doi: 10.24963/ijcai.2020/417.

[17] P.-J. Kindermans *et al.*, ‘The (Un)reliability of Saliency Methods’, in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds. Cham: Springer International Publishing, 2019, pp. 267–280. doi: 10.1007/978-3-030-28954-6_14.

[18] S. Dasgupta, N. Frost, and M. Moshkovitz, ‘Framework for Evaluating Faithfulness of Local Explanations’, in *Proceedings of the 39th International Conference on Machine Learning*, Jun. 2022, pp. 4794–4815. Accessed: Jan. 08, 2023. [Online]. Available: https://proceedings.mlr.press/v162/dasgupta22a.html

[19] D. Alvarez Melis and T. Jaakkola, ‘Towards Robust Interpretability with Self-Explaining Neural Networks’, in *Advances in Neural Information Processing Systems*, 2018, vol. 31. Accessed: Feb. 08, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html

[20] H. Yuan, H. Yu, S. Gui, and S. Ji, ‘Explainability in Graph Neural Networks: A Taxonomic Survey’. arXiv, Dec. 30, 2020. doi: 10.48550/arXiv.2012.15445.

[21] E. Amparore, A. Perotti, and P. Bajardi, ‘To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods’, *PeerJ Comput. Sci.*, vol. 7, p. e479, Apr. 2021, doi: 10.7717/peerj-cs.479.

[22] P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann, ‘Explainability Methods for Graph Convolutional Neural Networks’, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 10764–10773. doi: 10.1109/CVPR.2019.01103.

[23] C. Agarwal, M. Zitnik, and H. Lakkaraju, ‘Probing GNN Explainers: A Rigorous Theoretical and Empirical Analysis of GNN Explanation Methods’, in *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, May 2022, pp. 8969–8996. Accessed: Oct. 18, 2022. [Online]. Available: https://proceedings.mlr.press/v151/agarwal22b.html

[24] M. Sundararajan, A. Taly, and Q. Yan, ‘Axiomatic Attribution for Deep Networks’. arXiv, Jun. 12, 2017. doi: 10.48550/arXiv.1703.01365.

[25] A. Shrikumar, P. Greenside, and A. Kundaje, ‘Learning Important Features Through Propagating Activation Differences’, in *Proceedings of the 34th International Conference on Machine Learning*, Jul. 2017, pp. 3145–3153. Accessed: Jan. 12, 2023. [Online]. Available: https://proceedings.mlr.press/v70/shrikumar17a.html

[26] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, ‘Towards better understanding of Gradient-based Attribution Methods for Deep Neural Networks’, Apr. 2018.

[27] G. Montavon, W. Samek, and K.-R. Müller, ‘Methods for interpreting and understanding deep neural networks’, *Digit. Signal Process.*, vol. 73, pp. 1–15, Feb. 2018, doi: 10.1016/j.dsp.2017.10.011.

[28] A. Arias-Duart, F. Parés, D. Garcia-Gasulla, and V. Giménez-Ábalos, ‘Focus! Rating XAI Methods and Finding Biases’, in *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Jul. 2022, pp. 1–8. doi: 10.1109/FUZZ-IEEE55066.2022.9882821.

[29] A. Nguyen and M. R. Martínez, ‘On quantitative aspects of model interpretability’. arXiv, Jul. 15, 2020. Accessed: Jan. 12, 2023. [Online]. Available: http://arxiv.org/abs/2007.07584

[30] A. Hedström *et al.*, ‘Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations’. arXiv, Feb. 14, 2022. doi: 10.48550/arXiv.2202.06861.

[31] L. Rieger and L. K. Hansen, ‘IROF: a low resource evaluation metric for explanation methods’. arXiv, Mar. 09, 2020. doi: 10.48550/arXiv.2003.08747.

[32] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, ‘Evaluating the Visualization of What a Deep Neural Network Has Learned’, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 11, pp. 2660–2673, Nov. 2017, doi: 10.1109/TNNLS.2016.2599820.

[33] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, ‘On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation’, *PLOS ONE*, vol. 10, no. 7, p. e0130140, Jul. 2015, doi: 10.1371/journal.pone.0130140.

[34] L. Arras, A. Osman, and W. Samek, ‘CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations’, *Inf. Fusion*, vol. 81, pp. 14–40, May 2022, doi: 10.1016/j.inffus.2021.11.008.

[35] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, ‘Top-Down Neural Attention by Excitation Backprop’, *Int. J. Comput. Vis.*, vol. 126, no. 10, pp. 1084–1102, Oct. 2018, doi: 10.1007/s11263-017-1059-x.

[36] M. Kohlbrenner, A. Bauer, S. Nakajima, A. Binder, W. Samek, and S. Lapuschkin, ‘Towards Best Practice in Explaining Neural Network Decisions with LRP’, in *2020 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2020, pp. 1–7. doi: 10.1109/IJCNN48605.2020.9206975.

[37] D. Mahajan, C. Tan, and A. Sharma, ‘Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers’. arXiv, Jun. 12, 2020. doi: 10.48550/arXiv.1912.03277.

[38] R. K. Mothilal, A. Sharma, and C. Tan, ‘Explaining machine learning classifiers through diverse counterfactual explanations’, in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, Jan. 2020, pp. 607–617. doi: 10.1145/3351095.3372850.

[39] M. Honegger, ‘Shedding Light on Black Box Machine Learning Algorithms: Development of an Axiomatic Framework to Assess the Quality of Methods that Explain Individual Predictions’. arXiv, Aug. 15, 2018. doi: 10.48550/arXiv.1808.05054.

[40] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, ‘Sanity Checks for Saliency Maps’, in *Advances in Neural Information Processing Systems*, 2018, vol. 31. Accessed: Jan. 30, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html

[41] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, ‘Fairness through awareness’, in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, New York, NY, USA, Jan. 2012, pp. 214–226. doi: 10.1145/2090236.2090255.

[42] J. Dieber and S. Kirrane, ‘A novel model usability evaluation framework (MUsE) for explainable artificial intelligence’, *Inf. Fusion*, vol. 81, pp. 143–153, May 2022, doi: 10.1016/j.inffus.2021.11.017.

[43] C. and T. (European C. Directorate-General for Communications Networks, *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment*. LU: Publications Office of the European Union, 2020. Accessed: Feb. 20, 2023. [Online]. Available: https://data.europa.eu/doi/10.2759/002360

[44] R. V. Zicari *et al.*, ‘Z-Inspection®: A Process to Assess Trustworthy AI’, *IEEE Trans. Technol. Soc.*, vol. 2, no. 2, pp. 83–97, Jun. 2021, doi: 10.1109/TTS.2021.3066209.

[45] R. V. Zicari *et al.*, ‘Co-Design of a Trustworthy AI System in Healthcare: Deep Learning Based Skin Lesion Classifier’, *Front. Hum. Dyn.*, vol. 3, 2021, Accessed: May 24, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fhumd.2021.688152

[46] W. Jin, X. Li, and G. Hamarneh, ‘Evaluating Explainable AI on a Multi-Modal Medical Imaging Task: Can Existing Algorithms Fulfill Clinical Requirements?’, *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 11, pp. 11945–11953, Jun. 2022, doi: 10.1609/aaai.v36i11.21452.

[47] V. K. Venugopal, R. Takhar, S. Gupta, and V. Mahajan, ‘Clinical Explainability Failure (CEF) & Explainability Failure Ratio (EFR) – Changing the Way We Validate Classification Algorithms’, *J. Med. Syst.*, vol. 46, no. 4, p. 20, Mar. 2022, doi: 10.1007/s10916-022-01806-2.

[48] B. Hu, B. Vasu, and A. Hoogs, ‘X-MIR: EXplainable Medical Image Retrieval’, in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2022, pp. 1544–1554. doi: 10.1109/WACV51458.2022.00161.

[49] X. Larrucea, M. Moffie, S. Asaf, and I. Santamaria, ‘Towards a GDPR compliant way to secure European cross border Healthcare Industry 4.0’, *Comput. Stand. Interfaces*, vol. 69, p. 103408, Mar. 2020, doi: 10.1016/j.csi.2019.103408.

[50] ‘Ethics guidelines for trustworthy AI | Shaping Europe’s digital future’, Apr. 08, 2019. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed Mar. 27, 2023).

[51] M. Suffian, P. Graziani, J. M. Alonso, and A. Bogliolo, ‘FCE: Feedback Based Counterfactual Explanations for Explainable AI’, *IEEE Access*, vol. 10, pp. 72363–72372, 2022, doi: 10.1109/ACCESS.2022.3189432.

[52] J. van der Waa, E. Nieuwburg, A. Cremers, and M. Neerincx, ‘Evaluating XAI: A comparison of rule-based and example-based explanations’, *Artif. Intell.*, vol. 291, p. 103404, Feb. 2021, doi: 10.1016/j.artint.2020.103404.

[53] C. Panigutti, A. Perotti, A. Panisson, P. Bajardi, and D. Pedreschi, ‘FairLens: Auditing blackbox clinical decision support systems’, *Inf. Process. Manag.*, vol. 58, no. 5, p. 102657, Sep. 2021, doi: 10.1016/j.ipm.2021.102657.