Explainable Artificial Intelligence (XAI) Techniques - A Review and Case Study

Kaijen Lee
Supervisor(s): Chhagan Lal, Mauro Conti
EEMCS, Delft University of Technology, The Netherlands
20-6-2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

The significant progress of Artificial Intelligence (AI) and Machine Learning (ML) techniques such as Deep Learning (DL) has seen success in their adoption in resolving a variety of problems. However, this success has been accompanied by increasing model complexity resulting in a lack of transparency and trustworthiness. Explainable Artificial Intelligence (XAI) has been proposed as a solution to the need for trustworthy AI/ML systems. A large number of studies about XAI are published in recent years, with a majority discussing the specifics of XAI. Hence this work aims to formalize existing XAI literature from a high-level approach in terms of (1) benefits, (2) requirements, (3) challenges and (4) the underlying building blocks involved. Additionally, this paper presents a case study of XAI within the medical image analysis domain followed by future works and research directions in the field and from a general perspective, all to serve as a foundation and reference point to make the topic more accessible to novices.

1 Introduction

The recent advances in Artificial Intelligence (AI) and Machine learning (ML) techniques have brought great success in their applications within a wide variety of domains, such as healthcare, governance and finance. Taking healthcare for example, Convolutional Neural Networks (CNN) have been widely used within medical imaging to aid diagnosis, such as, skin lesions [1], Alzheimer's [2], lung cancer [3], etc. However, the black-box nature of commonly used traditional ML models, particularly deep learning (DL) and neural networks (NN), lack the transparency and accountability towards its results [4].

With the growing reliance on predictions made through ML models in decision making, the inability of providing a human interpretable explanation to justify the decisions made being free of bias becomes a pressing issue [5]. This may in turn lead to catastrophic consequences for those affected, particularly within high-stakes applications where failure is not an option. In addition to allowing its users to comprehend the key details which lead an ML model to its outcome, the ability to provide an understandable explanation of the decisions from an ML model also plays a role in allowing engineers in their effort of debugging the model and eliminating inherent bias and errors that may be present.

The need for trustworthy AI/ML systems has led to the topic of Explainable Artificial Intelligence (XAI) to be sought-after within the ML research community. As described in [6], XAI systems present themselves as self-explanatory intelligent systems capable of providing human interpretable explanations towards their decision-making processes and logic for end users. Despite the rapidly rising amount of publications involving XAI, there is a lack of literature relating to XAI systems from a general perspective as many tend to be specific to certain fields.

This work aims to (1) bring perspectives to novices towards the importance of XAI as a research topic, and (2) identify the general key benefits of incorporating XAI with different machine learning models, along with its challenges, requirements, and the building blocks in its construction. To achieve this, we survey a wide variety of literature relating to XAI and present concepts that we identify to be common to all domains of XAI applications.

The structure of the paper is as follows: an overview of the literature that we used to conduct our investigation is outlined in section 2. Following that, section 3 provides an analysis of the literature we examined and summarizes our findings towards the importance of XAI as a research topic. Section 4 relates our analysis findings to the application of XAI systems in the field of medical image analysis. Subsequently, we discuss the future work and research direction of XAI from a general perspective and within its application in medical image analysis in Section 5. Section 6 demonstrates the measures we took in our analysis towards responsible research. Lastly, section 7 concludes the findings of the analysis made throughout the paper.

2 Related Work

In this section, background which follows from related work is outlined. The related work we identified can be grouped into three main clusters, namely:

1. XAI Surveys
2. Studies on XAI attributes
3. Use-case Analysis

XAI Surveys
Due to XAI's emergence as topic within the ML research community, there are many surveys of the topic available. However, the surveys we encountered are generally aimed towards the categorization of XAI approaches, and potential research directions that one may take on within this field. Although the surveys may provide great overviews of XAI, the benefits from the adoption of XAI are often presented briefly and, the requirements discussed are usually specific to certain fields of application and lacked uniformity.

Studies on XAI Taxonomy
Related work on taxonomy like [4,7] offer some form of generalised perspective of the application of XAI. These studies have been carried out in effort to standardize commonly used terminologies involved with XAI. Though they may help one to infer some form of generalisation towards the structure which composes an XAI system, they place less emphasis on describing the overall building blocks involved designing an XAI approach.

Use-case Analysis
Literature of this nature, such as [8, 9], investigates the application of XAI on specific use-cases. In this paper, we attempt to generalize the common benefits, requirements, challenges and building blocks one may encounter whilst employing XAI approaches.

This work compiles the concepts brought forth in the various categories of literature mentioned and present them in manner

Table 1: A summary of the literature, grouped into clusters, that were reviewed in detail towards the findings of this paper.

Literature Cluster	Authors	Year	Reference
XAI Surveys	Rawal et al.	2021	[5]
	Saeed and Omlin	2021	[10]
	Adadi and Berrada	2018	[11]
	Longo et al.	2020	[12]
	Doshi-Velez and Kim	2017	[13]
XAI Taxonomy Studies	Schwalbe and Finzel	2021	[7]
	Das et al.	2020	[4]
XAI Use-case Analysis	Tjoa and Guan	2021	[8]
	Singh et al.	2020	[9]
	Lucieri et al.	2020	[14]
	van der Velden et al.	2022	[15]

that is applicable to the general use of XAI. The literature we examined in detail are summarized in Table 1. To this end, the main contributions of this survey include:

- We elaborate the general building blocks involved in designing an XAI approach for different machine models and present the overall key benefits.
- We outline the general requirements involved, including those stipulated by regulatory bodies, towards designing an XAI system.
- We discuss the overall challenge involved with the use of XAI systems.
- We present a case-study in medical image analysis to demonstrate the applicability of our abstraction in practise.

3 Literature Analysis

The different scientific literature clustered in Section 2 will be analyzed and compared in further detail within this section. Firstly, the overall building blocks which constitute an XAI system will be approached in Section 3.1. Secondly, the benefits which follow from the use of XAI will be compiled in Section 3.2. Following that, the general requirements stipulated by different governing bodies will be presented in Section 3.3. Lastly, Section 3.4 outlines the challenges relating to the requirements discussed.

3.1 Building blocks of an XAI system

From a top-level, we examine that the building blocks which constitute an XAI approach consist of the following three main components: (1) the input, (2) the explanator, and lastly (3) the output, as summarized in Figure 1.

Figure 1: The building blocks involved in the construction of an XAI system.

The input
The input of an XAI system specifies the task the explanator is intended to carry out. In this regard, the input can be classified by the task type and the data type involved. Saeed and Omlin [10] proposed that the type of input of an XAI system can be categorized as follows:

- model
- data
- user feedback
- context

The explanator
The explanator refers to the XAI technique that will be implemented to extract the explanations. The choice of explanator is largely influenced by the explanandum which follows from the input of the system. To that extent, the properties of an XAI explanator can be classified as demonstrated in Figure 2.

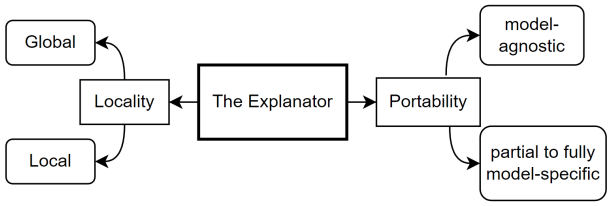

Figure 2: The general properties of the explanator within an XAI system [7].

Generally, the explanator consists of two main properties: portability and locality. The portability of an explanator refers to how far it depends on access to the explanandum model, with model-agnostic explanators only depending on the input and output of the model and model-specific explanators requiring access to the internals of the model [7]. On the other hand, locality refers to the extent of the validity of the explanations given by the explanator. Local explanators only provide explanations valid to one or a group of given input samples, while global explanators provide explanations that are valid to the entire input space [7].

It should be noted that explanators exist as post-hoc explanators or transparent (or self-explaining) models, whereby

transparent models are classified as being model-specific as they involve a special type of model or modification towards existing model architectures [7].

The output
The output of the XAI system relates to the type of information presented to the explainee. The common type of explanations as proposed by [7] includes:

- By example instance
- Contrastive/Counterfactual/Near Miss
- Prototype
- Feature Importance
- Rule Based
- Dimension Reduction
- Dependence Plots
- Explanatory Graphs
- Combination of the above

3.2 Benefits of XAI

With the main purpose of XAI models being to provide human intelligible explanations in their decision-marking processes and logic [6], the explanations provided bring forth benefits to multiple parties that adopted XAI approaches. Here, we dissect the gains from the adoption of XAI approaches we found in literature and outline the ideas that are commonly proposed.

These parties involved in the adoption of XAI approaches were categorized into five *target audiences* in [16], namely: domain experts/users of the model, regulatory entities, managers and executive board members, data scientists, developers, product owners, and users affected by model decisions. On the other hand, [10] further concise these audiences into five perspectives: regulatory perspective, scientific perspective, industrial perspective, model's developmental perspective, and end-user and social perspective. The overall categorization of the multiple parties that may benefit from the use of XAI approaches is summarized in Figure 3.

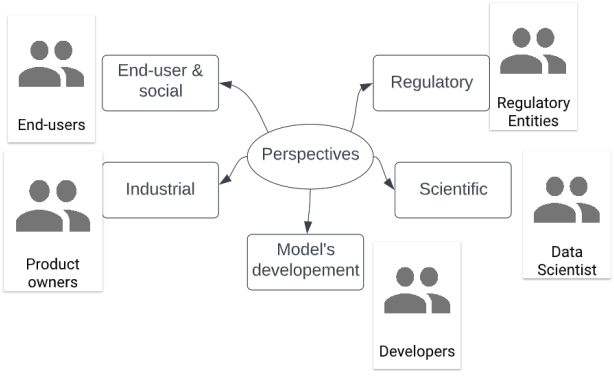

Figure 3: The different perspectives whereby an XAI approach is beneficial [10, 16]

.

In general, the benefits of following an XAI approach can find their source in four main factors [11]: (1) Explain to justify, (2) Explain to control, (3) Explain to improve, and (4) Explain to discover. The four aforementioned factors are related to the proposed categories of target audiences as follows:

- **Explain to justify** relates to the need for the specific reasons or justifications towards a particular outcome of the model, rather than a general description of the underlying model's process or logic of reasoning [11]. Having explanations that fulfil this benefit from the regulatory perspective, industrial perspective and end-user and social perspectives. Regulations such as the European Union General Data Protection Regulation (GDPR) specified the 'right to explanation' which entitles a user to be provided meaningful explanations when considerably affected by a decision made by the algorithm [17]. With XAI techniques, the explanations that entail allow industries to adopt better performing models as opposed to lesser accurate interpretable models whilst complying with necessary regulations [10]. When given an intelligible explanation as evidence, end-users on the other hand would be able to build trust towards such models with the decisions made that it is free of biases and prejudice [10].
- **Explain to control** aims to utilize explanations in the effort of damage control towards erroneous outcomes from the models while **Explain to improve** infers knowledge from explanations to further refine the underlying model [11]. From the model's developmental perspective, the explanations derived from XAI techniques enable engineers to better identify the causes that contribute to inappropriate outcomes of a black-box AI system [10]. The insights disclosed from an XAI approach may be used to further understand the underlying AI model, in the effort to debug and minimize faulty behaviors. This in turn enhances the model's robustness, safety and user trust through the eradication of any bias, unfairness and discrimination that may be present in the model.
- **Explain to discover** looks to uncover the knowledge that is embedded within AI models through the large datasets they are trained upon. From a scientific perspective, the trained black-box AI model represents the basis of knowledge, rather than the data themselves [10]. As a result, XAI approaches serve as a mean to reveal the knowledge inferred by black-box AI models, which may lead to new concepts to be discovered in the field where the model is applied.

Overall, the benefits of XAI stem from two principal factors: (1) regulatory reasons, and (2) knowledge [16]. Explainability is one of the main factors that hindered high-performing state-of-the-art from being implemented in practice. Hence, due to regulatory obligations, strictly regulated sectors such as finance, security and healthcare, are reluctant in deploying modern AI models with low explainability as doing so may put their assets at risk. From the standpoint of the research community, the inference capability of

modern AI techniques through huge amounts of reliable data has only been demonstrated through results and performance metrics such as accuracy, Dice coefficient, or an ROC analysis [5]. Explanations in this regard play the role of aiding researchers in understanding the AI techniques further. This in turn would provide researchers further insight towards improving the techniques and their practical uses.

3.3 Requirements in designing an XAI approach

The requirements in designing an XAI approach derive from the features that the target audience of the system would need. Within this subsection, we examine the requirements proposed, from a high level, across different literature, including surveys, government standards, and regulations. Subsequently, we collate them according to four categorizing features, namely (1) performance, (2) privacy, (3) security and (4) safety.

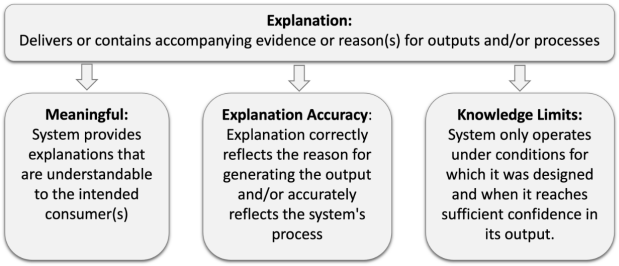

Figure 4: The four principles of XAI as proposed by the National Institute of Standards and Technology (NIST) [18].

Performance
Three out of four principles as proposed by the National Institute of Standards and Technology (NIST) [18] as shown in Figure 4, namely Explanation, Meaningful, and Explanation Accuracy, tie with this requirement. Collectively, they specified that an XAI system must provide or include corresponding evidence or reasons(s) for outcomes and/or its processes which correctly demonstrate the required reasoning and is intelligible to the users of the explanation. This is in line with the *explainability* characteristic of an XAI system which Arrieta et al. [16] defined as "Given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand". Consequently, in the grand scheme, an XAI approach is required to not only be able to provide an explanation of its outcomes(s) and/or its decision-making process, but also to provide an explanation that is sufficient to be understandable by its target audiences.

Privacy
This requirement specifies the need for XAI systems to consider the privacy of information in their design. Generally, AI systems require a large amount of data to be trained upon. When using large datasets from the public domain, XAI systems must be able to protect the consumers privacy. Information provided in the explanations generated is required to be untraceable to the individuals whose data are used during the training phase [19].

Security
As described in Section 3.2, one key benefit of the use of XAI systems is its knowledge extraction capability. However, it could be the case that the knowledge embedded within certain AI models is confidential, thus making it an intellectual property belonging to the corresponding party. This information may be compromised even when only input and output access of the model is available. As a result [20], the XAI system must be resilient to adversary attacks.

Safety
The Knowledge Limits principle as proposed by the NIST is defined as such that an XAI system must operate only under the conditions it is implemented for and only provides and presents its explanation only if it is sufficiently confident [18]. This is important as it concerns the end-user of the system. If the XAI system involved is applied within a domain that it is not intended for or arrives at an outcome with low confidence, the outcome it provides is likely to be unreliable and erroneous. Therefore, those who are affected are at stakes of being met with potentially harmful consequences.

3.4 Challenges of following an XAI approach

Within this subsection, we explore the challenges presented in literature and organise them in the manner which relates them to the four categorizing features proposed in Section 3.3.

Challenges involving XAI Performance
As the performance of an XAI system is measured against its explainability, one may question how can explanations be evaluated in terms of their usefulness to their target audiences? Saeed and Omlin [10] proposed that multidisciplinary research collaboration between fields such as psychology, behavioural and social sciences, human-computer interaction (HCI), physics and neuroscience plays a role in resolving this. Approaching the explanations from a psychological perspective may provide further insights towards how the structure and attributes of explanations may influence the parties on the receiving end [12]. Additionally, HCI studies within XAI are essential. Longo et al. [12] stress the inclusion of humans during the design and deployment phase of the XAI to enhance their explainability. This was further emphasized by the study in [14] where interactive tools enhance the users' comprehension and engagement with AI models.

Challenges involving Privacy
Although XAI systems are required to protect the privacy of the individuals whose data are used in its training, it should also be acknowledged that the right to explanation conflicts with the right to privacy [21]. While XAI systems are capable of providing explanations without disclosing individuals whose data contribute to the outcome, anonymization of data does not guarantee that the data is inherently untraceable [21]. On the other hand, even when the anonymization of data may protect the individuals involved, the party who is entitled to an explanation to the outcome may not be necessary convinced of the explanations provided.

Additionally, the right to be forgotten [22] is also in conflict with the functions of XAI systems. The right to be forgotten stipulates that a user is entitled to have their personal data

erased. In the context of XAI, data preservation is necessary as the whole raw training data is required to be kept to provide the requested justifications [10].

Challenges involving Security
In the context of XAI, security involves the protection of confidential information, whether it is implicitly or explicitly contained within the model itself, against adversarial attacks. Adversarial attacks aim to understand the specific inputs of an ML model which influence its outcome, in the effort of manipulating it.

Arrietta et al [16] proposed that generative models would play an important role in fortifying an XAI system's protection against adversarial attacks. The two ways that generative models could be used in this regard is (1) generating input instances that behaves as a latent representation of the training data to demonstrate its relationship with the output [23], and (2) the creation of counterfactuals to provide the performance boundaries of the model [24].

Challenges involving Safety
The safety of an XAI system concerns the risk taken by processes which depend on the output of the system. It was noted in [16] that erroneous outputs of the system can result in harmful consequences in some domains, compromising the individuals involved. Consequently, regulatory bodies has set out regulations to ensure that decisions cannot be made solely through the manner of data processing [17].

In effort of minimizing the risks and uncertainty of adverse effects caused by the adoption of ML models in the decision-making process, research has been conducted on potential safety measures [25]. By examining the epistemic uncertainty (uncertainty due to lack of knowledge) of the input data and its relationship with the model's output confidence, a user of the system may be informed of the suitability of the model's output for a certain decision-making process [16].

4 Case Study in Medical Image Analysis with Visual Explanations

Within healthcare, AI/ML is commonly used to identify anomalies within medical images in the effort of diagnosing illnesses. In that regard, the most common form of XAI used in this domain is visual explanation, specifically saliency mapping [15], which highlights regions of the image contributing to the outcome of the model as demonstrated in Figure 5.

In this section, the application of XAI approaches within the domain of medical image analysis involving visual explanations is examined closely and related to the concepts presented in Section 3. Firstly, the benefits of the use of XAI approaches in the domain will be outlined in Section 4.1. Subsequently in Section 4.2, the components which constitute an XAI approach for the domain will be analyzed according to the building blocks proposed in Section 3.1. Lastly, the general requirements and challenges of an XAI approach are related to its application in medical image analysis respectively in Sections 4.3 and 4.4.

Figure 5: An example of saliency mapping involving the use of Class Activation Mapping (CAM) technique combined with a modified global average pooling layer of a classification trained Convolutional Neural Network (CNN) to both classify the image and localize class-specific image regions. (e.g.: toothbrush for 'Brushing teeth' and chainsaw for 'Cutting trees') [26]

.

4.1 Benefits of XAI in Medical Image Analysis

As the stakes of medical decision-making are often high, medical experts have voiced their concerns towards the black-box nature of state-of-the-art AI algorithms in medical image analysis [27]. The lack of explainability and complexity in novel AI algorithms makes it difficult for clinicians to verify the results of the model. Hence, XAI plays an essential role in allowing such algorithms to be used in practice by providing the necessary explanations.

Furthermore, the availability of explanations to the outcomes of an AI model could provide further insights to the strengths and weaknesses of the model to improve the model and to uncover new knowledge from the large dataset the model is trained upon [28]. Moreover, XAI approaches could also provide meaningful information for a patient whose diagnosis may result from an AI model's output. This builds trust in the application of state-of-the-art AI techniques among clinicians and patients.

In practice, XAI approaches, particularly, model agnostic post-hoc techniques have high ease of use and can be conveniently integrated into existing applications of AI/ML models to retrieve explanations from their predictions [15].

4.2 Building Blocks of a Visual Explanation XAI approach

According to van der Velden et al. [15], the most common visual explanation XAI technique used in medical image analysis is Class Activation Mapping (CAM) followed by Gradient-weighted Class Activation Mapping (Grad-CAM), in which both follow a backpropagation-based approach, amongst others. The CAM technique provided local, model-specific, post-hoc explanations for Convolutional Neural Networks (CNN) through the means of global average pooling [26]. Grad-CAM behaves similarly but does not require global average pooling, resulting in it being able to work with any type of CNN to produce post-hoc local explanations [29].

With CAM and Grad-CAM, the framework that we proposed in Section 3.1 can be related. The input of the XAI

system involves a classification trained CNN and an image. Following that, the explanator block corresponds to the CAM and Grad-CAM technique themselves, where they are applied post-hoc onto the underlying CNN. As previously mentioned, these explanators are model-specific to CNNs and are only able to produce an accompanying explanation to the image provided as an input. Lastly, the output of the system is a saliency map produced corresponding to the input image, a kind of feature importance which highlights the region of the image that influenced or contributed to the prediction made by the model.

4.3 Requirements of XAI Approaches in Medical Image Analysis

Within the domain of medical image analysis, the four categorizing groups of requirements proposed in Section 3.3 relate as such:

Performance
Where XAI systems are applied for medical image analysis, the predictions made by the model are used within the decision-making process for medical diagnosis. The explanations derived from the predictions are required to provide sufficient information for clinicians upon a diagnosis. As of the time of writing, several modes of explanation evaluation have been proposed by [13], namely: application-grounded, human-grounded, and functionally-grounded evaluations. However, performing either of the aforementioned modes of evaluation within the domain of medical image analysis is currently not a standard practice in studies relating to medical image analysis and can be resource-intensive [15]. Moreover, the explanations provided by the XAI systems have to be suited to the specific users of the system, whereby their fields of expertise may differ. For example, a visual explanation that highlights the location of a disease on an image may suffice as an explanation to a radiologist or medical image analyst, this might not be the case for other clinicians.

Privacy
As required by regulations such as the GDPR [19], patients whose data are used in training the AI models must not be traceable to the identity of the individuals involved, including when the data can be supported by information that can be reasonably accessed from other sources. As an example relating to medical image analysis, Schwarz et al. [30] have studied that anonymous research participants could be identified through the reconstruction of their faces with computed tomography (CT) or magnetic resonance imaging (MRI) data.

With this in mind, explanations derived from a dataset that contains identifiable patient data must be pre-processed and filtered in a manner untraceable to the individuals involved, for instance, the face or skull region of medical images should be removed [31]. Furthermore, a safeguard should be performed by XAI systems applied in this domain where information linking to the identity of individuals must not be uncovered when any type of explanation is provided.

Security
With patient data being highly confidential, their usage in medical image analysis requires protection towards re-identification, dataset reconstruction and tracing attacks [31]. Reports have shown that data-mining companies have adopted a business model performing large-scale re-identification attacks and the sale of re-identified medical records [32]. Individuals compromised will then be susceptible to discrimination. For example, health insurance companies may reduce their financial risk through their consideration of certain groups affected by particular illnesses within their premium pricing model [31].

On the other hand, high stakes are involved in the diagnosis of illnesses within patients. Although clinicians may benefit from the explanations provided by XAI systems in their decision-making process, the explanations may be an area of vulnerability towards adversarial attacks. Adversarial attacks on XAI systems may compromise the explanations and possibly provide misinformation to clinicians, impairing their judgement towards the results.

Safety
Safety concerns the risks that the patients may be exposed to due to the results of the system. In this regard, explanations that XAI systems derive are required to be sufficiently accountable and interpretable when utilized by clinicians while diagnosing illnesses [8]. However, the current state of XAI systems lack a robust way of handling interpretability. Before such is available, Tjoa and Guan [8] emphasizes that human supervision is still necessary where AI models are applied in medical practice, as "interpretable" information derived via XAI systems should only serve as complementary support.

4.4 Challenges with XAI in Medical Image Analysis

The adoption of XAI systems within the medical image analysis domain have been met with challenges in meeting the requirements seen in subsection 4.3. In literature, we observed that the challenges largely involve performance and privacy. In this subsection, we elaborate on the challenges present within the two aforementioned requirements as follows:

Challenges involving Performance
Rudin [33] cautioned the use of black-box models alongside model-specific XAI approaches for high-stakes applications due to several issues. Where a black-box AI model is involved, the explanations provided through the means of an XAI approach may not be completely faithful to the computations of the original model. Additionally, the explanations provided may not provide sufficient detail on the inner workings of the black-box model. In the context of visual explanations, saliency maps of the class may not visually differ much between classes of high probability and low probability [15]. Moreover, Tjoa and Guan [8] stresses that saliency maps are often provided with inadequate evaluations of their utilities within their medical practices. Rudin [33] recommends the use of interpretable model-based XAI approaches instead. However, such approaches typically have low ease of use as the explanation is encapsulated within the design of the neural network [15].

On the other hand, the evaluation of the performance of the explanations from XAI systems is resource-intensive process. The evaluation methods proposed by Doshi-Velez and Kim [13] require either volunteers or clinicians to manually assess the suitability of the explanations. As the field of medical imaging analysis involves a diverse collection of medical experts and given the wide variety of XAIs approaches that can be used, the evaluation process may be time-consuming and be unfeasible in practice.

Challenges involving Privacy
Although pre-processing and deanonymization of patient data may aid in the process of protecting the privacy of the individuals involved, these efforts complicate the data handling process and increase the probability of errors. Where data is manipulated to shield the patients' privacy, any mishandling may present, at worst, an adversarial update to the models involved [31]. Moreover, the GDPR [19] right to be forgotten poses challenges towards this effort. Essentially, a look-up table is necessary to retrieve a patient's data for removal. However, Look-up tables become another area of vulnerability. If not kept safely, patients would then again run the risk of being identified in the event of data theft.

5 Future Work and Research Direction

This section outlines the future work and research direction that we recommend for potential practitioners of the field of XAI upon our examination of XAI-related literature. In this section, we firstly explore the future work and research directions from a general perspective followed by from the context of the medical image analysis domain.

5.1 General XAI

It has been highlighted commonly within literature [8, 10, 11, 13] that research involving XAI systems requires more formalism. Filip et al. [34] have identified that different terminologies have been used by researchers in XAI to describe similar or identical concepts. With that in mind, Adadi and Berrada [11] stress that XAI is a multidisciplinary objective which should be supported by a standalone research community in the effort of engaging formalism. Although works such as [7] have proposed a comprehensive classification of XAI systems, a governing body should provide a clear standardization of terms for common XAI concepts that future research of the field to adhere and extend upon with their contributions.

Additionally, research involving XAI finds its need for more general, quantifiable evaluation metrics and methods [10, 11]. Towards this effort, Doshi-Velez and Kim [13] formulated three evaluation groups, namely: application-grounded, human-grounded, and functionally-grounded evaluations as preliminary works to evaluate the explainability of XAI systems. However, these concepts focus on the qualitative analysis of XAI systems and do not necessarily translate to a fair comparison between different XAI approaches due to their subjective nature. Hence, further work is required towards formulating more generalized and quantifiable metrics to allow for purposeful comparison between different XAI approaches in terms of their explainability.

5.2 XAI in Medical Image Analysis

In medical image analysis, one potential research direction relates to the links between causality and XAI. In the survey performed by van der Velden et al. [15], it was determined that medical image analysis typically consists of correlation instead of causation. While explanations may benefit clinicians regarding how a diagnosis is made by AI models, causality explains the relation between cause and effect, providing further insights into the diagnosis. Furthermore, causal reasoning has been shown to be advantageous in assessing and eliminating bias present in training data [35].

Within visual explanations used in medical image analysis in the form of saliency maps, it was noted by Tjoa and Guan [8] that saliency maps are commonly provided with inadequate consideration towards their utilities with medical practises. Future work in this regard involves further investigation towards explanations and the nature of user experience and expertise to address the domain-specific needs of specific applications and their users [10]. More specifically, considerations must be made towards the inclusion of end-users within the design of XAI systems with clearly stated goals and purposes of the end-users.

6 Responsible Research

Within research, it is important to consider the ethical implications of the work being done. In that regard, this section describes the measures taken in our process to ensure that our findings are transparent and upholds scientific integrity.

Our work takes the form of a literature study to evaluate, from a general perspective, the building blocks, benefits, requirements, and challenges that the adoption of an XAI system entails. In our process, we accompanied our abstractions of the concepts with findings that we have encountered across different categories of research papers outlined in Section 2. Doing so guides our readers to the relevant materials should they require to pursue a certain topic further.

Moreover, in this paper, the ethical implications that XAI systems entail were considered and discussed in the form of requirements and challenges outlined in Sections 3 and 4, respectively from a general perspective and within medical image analysis. Though the adoption of XAI systems has its benefits, we noted several instances along with requirements from governing bodies and regulations to safeguard the safety and privacy of stakeholders involved.

In terms of reproducibility, the works in this paper provided a high-level abstraction involved with the general use of XAI systems. A case study in medical image analysis was provided to indicate how our abstraction relates to an instance of XAI application. With this paper being accessible online via the TU Delft repository and our references available through commonly used scientific databases such as SCOPUS and IEEE Xplore, it is possible for the readers to apply our findings to their relevant investigation. Subsequently, readers are encouraged to critique or extend upon our proposed abstraction where applicable.

7 Conclusion

The overall benefit of adopting an XAI approach stems from regulatory reasons and knowledge extraction. The use of XAI systems facilitates the use of high-performing models in various regulated domains, where the models alone lack the necessary interpretability. Furthermore, the explanations that XAI systems derive allow information that is embedded within models to be uncovered, for the purpose of knowledge extraction or further insights towards improving the inner workings of the model.

Although XAI systems have their benefits, there are notable challenges that follow from their adoption. Namely, the current state of evaluation of the explanations requires human involvement in its processes, making it resource-intensive in practice. Moreover, regulations involving a user's right to explanation conflicts with the right to privacy, as explanations derived from XAI systems may contain the necessary information that is not guaranteed to be untraceable.

Overall in this paper, we examined various state-of-the-art literature on XAI from a high-level approach to compile the building blocks, benefits, requirements and challenges that follow from the general use of XAI systems. Subsequently, the application of XAI approaches in medical image analysis was examined and related to our findings. Lastly, we discussed the future work and research direction within XAI from a general approach and within the domain of medical image analysis.

References

[1] A. Kamalakannan, S. S. Ganesan, and G. Rajamanickam, "Self-learning AI framework for skin lesion image segmentation and classification," *International Journal of Computer Science and Information Technology*, vol. 11, pp. 29–38, Dec. 2019.

[2] A. Farooq, S. Anwar, M. Awais, and S. Rehman, "A deep cnn based multi-class classification of alzheimer's disease using mri," in *2017 IEEE International Conference on Imaging Systems and Techniques (IST)*, pp. 1–6, 2017.

[3] W. Alakwaa, M. Nassef, and A. Badr, "Lung cancer detection and classification with 3d convolutional neural network (3d-cnn)," *Lung Cancer*, vol. 8, no. 8, p. 409, 2017.

[4] S. Das, N. Agarwal, D. Venugopal, F. T. Sheldon, and S. Shiva, "Taxonomy and survey of interpretable machine learning method," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, Dec. 2020.

[5] A. Rawal, J. Mccoy, D. B. Rawat, B. Sadler, and R. Amant, "Recent advances in trustworthy explainable artificial intelligence: Status, challenges and perspectives," *IEEE Transactions on Artificial Intelligence*, pp. 1–1, Dec. 2021.

[6] D. Gunning and D. W. Aha, "Darpa's explainable artificial intelligence program deep learning and security," 2019.

[7] G. Schwalbe and B. Finzel, "A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts," May 2021.

[8] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, pp. 4793–4813, Nov. 2021.

[9] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," June 2020.

[10] W. Saeed and C. Omlin, "Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities," Nov. 2021.

[11] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[12] L. Longo, R. Goebel, F. Lecue, P. Kieseberg, and A. Holzinger, "Explainable artificial intelligence: Concepts, applications, research challenges and visions," vol. 12279 LNCS, pp. 1–16, Springer, 2020.

[13] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," Feb. 2017.

[14] A. Lucieri, M. N. Bajwa, A. Dengel, and S. Ahmed, "Achievements and challenges in explaining deep learning based computer-aided diagnosis systems," Nov. 2020.

[15] B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (xai) in deep learning-based medical image analysis," *Medical Image Analysis*, vol. 79, p. 102470, July 2022.

[16] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[17] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision making and a "right to explanation"," 2017.

[18] P. J. Phillips, C. A. Hahn, P. C. Fontana, A. N. Yates, K. Greene, D. A. Broniatowski, and M. A. Przybocki, "Four principles of explainable artificial intelligence," Sep. 2021.

[19] The European Parliament and The Council of the European Union, "General data protection regulation (gdpr) – official legal text," *Official Journal of the European Union*, Apr. 2016.

[20] T. Orekondy, B. Schiele, and M. Fritz, "Knockoff nets: Stealing functionality of black-box models," 2018.

[21] T. D. Grant and D. J. Wischik, "Show us the data: Privacy, explainability, and why the law can't have both," *Geo. Wash. L. Rev.*, vol. 88, p. 1350, 2020.

[22] E. F. Villaronga, P. Kieseberg, and T. Li, "Humans forget, machines remember: Artificial intelligence and the

right to be forgotten," *Computer Law and Security Review*, vol. 34, pp. 304–313, Apr. 2018.

[23] C. F. Baumgartner, L. M. Koch, K. C. Tezcan, J. X. Ang, and E. Konukoglu, "Visual feature attribution using wasserstein gans," Nov. 2017.

[24] S. Liu, B. Kailkhura, D. Loveland, and Y. Han, "Generative counterfactual introspection for explainable deep learning," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1–5, 2019.

[25] K. R. Varshney and H. Alemzadeh, "On the safety of machine learning: Cyber-physical systems, decision sciences, and data products," Oct. 2016.

[26] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization."

[27] X. Jia, L. Ren, and J. Cai, "Clinical implementation of ai technologies will require interpretable ai models," *Medical Physics*, vol. 47, pp. 1–4, Jan. 2020.

[28] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang, "Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation," Sep. 2020.

[29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," Oct. 2016.

[30] C. G. Schwarz, W. K. Kremers, T. M. Therneau, R. R. Sharp, J. L. Gunter, P. Vemuri, A. Arani, A. J. Spychalla, K. Kantarci, D. S. Knopman, R. C. Petersen, and C. R. Jack, "Identification of anonymous mri research participants with face-recognition software," *New England Journal of Medicine*, vol. 381, pp. 1684–1686, Oct. 2019.

[31] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, vol. February, pp. 305–311, June 2020.

[32] A. Tanner, *Our Bodies, Our Data: How Companies Make Billions Selling Our Medical Records*. Beacon, January 2017.

[33] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," May 2019.

[34] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 0210–0215, May 2018.

[35] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *Nature Communications*, vol. 11, Dec. 2020.