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

Why is eXplainable Artificial Intelligence (XAI) an important research topic?

TU Delft Delft University of Technology

Author: Kaijen Lee - LeeKaijen@student.tudelft.nl
Supervisor: Dr. Chhagan Lal
Responsible Professor: Prof. Mauro Conti

(1) BACKGROUND

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. XAI systems are self-explanatory intelligent systems capable of **providing human interpretable explanations** within their decision-making processes and logic for end-users [1].

A large number of studies about XAI are published each year, with a majority discussing the specifics of XAI. Hence it is imperative to formalize existing XAI literature from a high-level approach to serve as a foundation and reference point to make the topic more accessible to novices.

(2) RESEARCH QUESTION

Main question: Why is eXplainable Artificial Intelligence (XAI) an important research topic?

Research sub-questions:

1. What are the **key benefits, requirements, building blocks, and challenges** involved with the use of XAI for different machine learning models?
2. How do the identified factors **relate** to a specific use case of XAI?

(3) BENEFITS

The overall goal of XAI is to provide human interpretable reasoning behind a black-box AI/ML model's outcome. Its main benefits stem from regulatory purposes and knowledge extraction. These can be viewed from 5 perspectives, as proposed by [2]:

Table 1: The five main perspectives for the need for XAI and their accompanying goals [2].

Perspectives	Goals of XAI
Regulatory	To allow stakeholders influenced by an algoritm's decision to be provided with explanation(s)
Scientific	To access the scientific knowledge embedded within the black-box AI models
Industrial	To access better performing models while complying with regulations relating to model's explanability
Model's Development	To improve the model with insights into its inner working
End-user and Social	To improve trust such that the model is free of inherent biases and prejudice

(4) REQUIREMENTS

The National Insitute of Standards and Technology (NIST) presented 4 principles that XAI systems should adhere to overall [3]:

- Explanation
- Meaningful
- Explanation Accuracy
- Knowledge Limits

Additionally, requirements also take the form of:

- Performance
- Privacy
- Security
- Safety

(5) CHALLENGES

The challenges related to the requirements that were identified are as follows:

Performance: The evaluation of explanations provided by XAI systems.

Privacy: Conflicts between GDPR "Right to explanation", "Right to privacy", and "Right to be forgotten".

Security: Protection against adversarial attacks to safeguard confidential information.

Safety: Minimizing the risk and uncertainty of adverse effects from the use of XAI.

(6) BUILDING BLOCKS

Figure 1: The building blocks of a general XAI system.

REFERENCES

[1] D. Gunning and D. Aha, "DARPA's Explainable Artificial Intelligence (XAI) Program," *AIMag*, vol. 40, no. 2, pp. 44–58, Jun. 2019, doi: 10.1609/aimag.v40i2.2850.
[2] W. Saeed and C. Omlin, "Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities," arXiv:2111.06420 [cs], Nov. 2021, Accessed: May 12, 2022. [Online]. Available: http://arxiv.org/abs/2111.06420
[3] P. J. Phillips et al., "Four Principles of Explainable Artificial Intelligence," National Institute of Standards and Technology, Sep. 2021. doi: 10.6028/NIST.IR.8312.
[4] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," Medical Image Analysis, vol. 79, p. 102470, Jul. 2022, doi: 10.1016/j.media.2022.102470.

(7) CASE STUDY

A case study has been performed regarding the use of XAI in the form of **visual explanations within medical image analysis** to investigate the relevance of the factors identified.

Benefits:

- Aids clinicians in verifying model results.
- Provide insights to improve the model.
- Helps researchers uncover new knowledge from the model.

Requirements:

- **Performance:** Provide sufficient information about the predictions made by AI models suited to the field of expertise of the clinicians involved.
- **Privacy:** Patients' data used in training the models must be protected and untraceable.
- **Security**: The system must be resilient against adversarial attacks.
- **Safety:** Explanations must be accountable and interpretable when involved in decision-making.

Challenges:

- **Evaluation of XAI explanations -** requires the involvement of clinicians which can be resource-intensive.
- **Area of vulnerability -** the existence of look-up tables for anonymized patients' data in compliance with GDPR's "Right to be forgotten".

Building Blocks: With Class Activation Mapping (CAM) followed by Gradient-weighted Class Activation Mapping (Grad-CAM) are the most commonly used visual explanation XAI technique in medical image analysis [4], we relate them to our building blocks abstraction as shown in Figure 2.

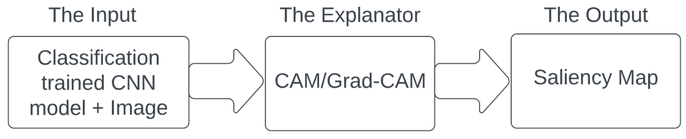

Figure 2: The building blocks of an XAI system following the CAM and Grad-CAM techniques.

(8) FUTURE WORK

General XAI :

- The formalism of XAI concepts and terminologies.
- More quantifiable and general evaluation metrics and methods for meaningful comparison between different XAI approaches.

XAI in medical imaging analysis:

- Investigation towards the links between causality and XAI
- Consideration of explanations and their utilities by clinicians of different areas of expertise.