Making ML Algorithms auditable

Presentation of a 30-question ML Auditing Criteria Catalog as Guide

Markus Schwarz[1] Ludwig Christian Hinske[2] Ulrich Mansmann[3]

Fady Albashiti[1]

[1]Medical Data Integration Center (MeDIC$^{\text{LMU}}$), LMU University Hospital Munich, Planegg
[2]Institute for Digital Medicine, University Hospital Augsburg, Neusäß
[3]Institute for Medical Information Processing, Biometry and Epidemiology (IBE), LMU Munich, München

MIRACUM-DIFUTURE-Kolloquium, Cisco Webex, 25.02.2025

Motivation

- Low adaption of ML applications (especially in the healthcare sector)
- Lack of trust and missing transparency as important reasons
- Need of commonly agreed criteria, processes and tools for ML auditing

⇒ Motivation: Help making ML model predictions more **transparent**, so they can be **used more often**, effectively and safely.

Definitions (1/2)

Auditable AI (AAI)

- AI systems that “answer questions asked by humans and interact with them in a human understandable way.” Dengel et al., 2021, p. 91
- “[A]uditability ... ensur[es] that the AI model behaves as expected.” Benchekroun et al., 2020, p. 2

Explainable AI (XAI)

- “Humans are able to derive a sufficiently causal understanding of the model’s inner workings.” Kiseleva et al., 2022, p. 7

Definitions (2/2)

- Focus on Black Box eXplanation[s] (BBX) . . . to allow "humans ... [to] debug, interpret, control, and reason about [deep neural networks]." Chatila et al., 2021, p. 23, Dengel et al., 2021, p. 91

Interpretable AI (IAI)

- "Refers to the observation and representation of cause and effect within a system." Dengel et al., 2021, p. 94
- Has XAI as necessary, but not always sufficient prerequisite. Kiseleva et al., 2022, pp. 7-8

Objective of the Catalog

A **tool** that helps creating a better ML algorithm/AI product for a given **use case**.

Whereas a use case is constituted by:

- Environment (classical IT systems, human actors or local restrictions)
- Data (training, testing/validation and live input/population)
- Model (design, assumptions, hyperparameters)

Catalog Contents

Conceptual Basics:

- AI Opportunities vs. AI Risks (2 questions)
- Risk Management (2 questions)
- Methodology (5 questions)
- Audit Process (3 questions)
- Quality Assurance (5 questions)

Catalog Contents - Example Questions (1/2)

AI Opportunities vs. AI Risks

?⃝ “Is the expected $\mathit{benefit} \ast \mathit{benefitProbability}$ of a successful ML use case implementation greater than the $\mathit{damage} \ast \mathit{damageProbability}$ in case of failure?”

Risk Management

?⃝ “Do you have a proactive, reactive and/or non-reactive risk management strategy in place? For example, have you planned to implement a 'kill switch' with measures to (temporarily) go back to the old process?”

Catalog Contents

Data & Algorithm Design:

- Data Properties (4 questions)
- Algorithm Design (3 questions)

Assessment Metrics:

- Qualitative Assessment (3 questions)
- Quantitative Assessment (3 questions)

Catalog Contents - Example Questions (2/2)

Data Properties

- ⓘ "Is the data generation process (DGP) of the training, testing and validation dataset sufficiently known? Could there be unknown confounders or mediator variables influencing the observed data?"

Algorithm Design

- ⓘ "Did you establish a correct ML use case hypothesis with concrete problem description and expected behavior (acceptance criteria, metrics, statistical testing results)?"

Catalog Contents

Thus, the catalog consists of **30 questions**:

- Conceptual Basics (17)
- Data & Algorithm Design (7)
- Assessment Metrics (6)

⇒ The questions are thought to **guide** the **ML** development/implementation **team**.

Catalog Contents

The ML Auditing Core Criteria Catalog has been published in IEEE Access: https://doi.org/10.1109/ACCESS.2024.3375763 (Open Access)

IEEE Access

Received 7 February 2024, accepted 1 March 2024, date of publication 11 March 2024, date of current version 20 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3375763

RESEARCH ARTICLE

Designing an ML Auditing Criteria Catalog as Starting Point for the Development of a Framework

MARKUS SCHWARZ[1], LUDWIG CHRISTIAN HINSKE[2], ULRICH MANSMANN[3], AND FADY ALBASHITI[1]

[1]Medical Data Integration Center (MeDIC LMU), LMU University Hospital Munich, 82152 Planegg, Germany
[2]Institute for Digital Medicine, University Hospital Augsburg, 86356 Neusäß, Germany
[3]Institute for Medical Information Processing, Biometry and Epidemiology (IBE), Ludwig-Maximilians-University Munich, 81377 München, Germany

Corresponding author: Markus Schwarz (markus.schwarz@campus.lmu.de)

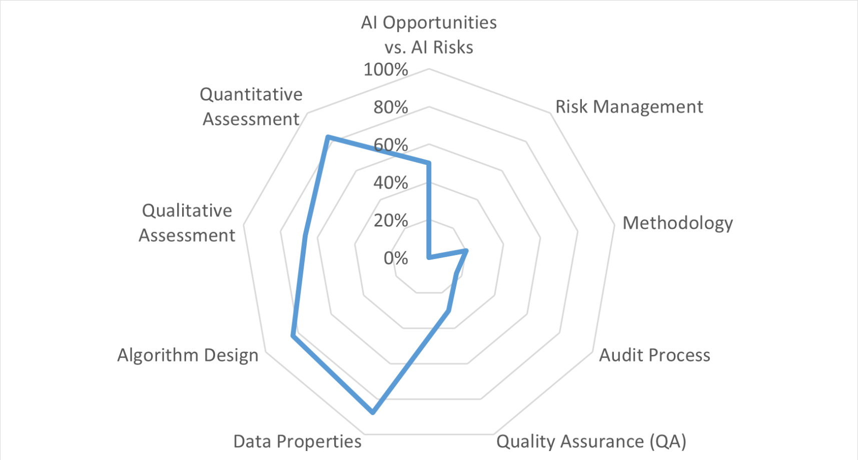
ABSTRACT Although AI algorithms and applications become more and popular in the healthcare sector, only few institutions have an operational AI strategy. Identifying the best suited processes for ML algorithm implementation and adoption is a big challenge. Also, raising human confidence in AI systems is elementary to building trustworthy, socially beneficial and responsible AI. A commonly agreed AI auditing framework that provides best practices and tools could help speeding up the adoption process. In this paper, we first highlight important concepts in the field of AI auditing and then restructure and subsume them into an ML auditing core criteria catalog. We conducted a scoping study where we analyzed sources being associated with the term "Auditable AI" in a qualitative way. We utilized best practices from Mayring (2000), Miles and Huberman (1994), and Bortz and Döring (2006). Based on referrals, additional relevant white papers and sources in the field of AI auditing were also included. The literature base was compared using inductively constructed categories. Afterwards, the findings were reflected on and synthesized into a resulting ML auditing core criteria catalog. The catalog is grouped into the categories: Conceptual Basics, Data & Algorithm Design and Assessment Metrics. As a practical guide, it consists of 30 questions developed to cover the mentioned categories and to guide ML implementation teams. Our consensus-based ML auditing criteria catalog is intended as a starting point for the development of evaluation strategies by specific stakeholders. We believe it will be beneficial to healthcare organizations that have been or will start implementing ML algorithms. Not only to help them being prepared for any upcoming legally required audit activities, but also to create better, well-perceived and accepted products. Potential limitations could be overcome by utilizing the proposed catalog in practice on real use cases to expose gaps and to further improve the catalog. Thus, this paper is seen as a starting point towards the development of a framework, where essential technical components can be specified.

INDEX TERMS AAI, AI auditing, auditable AI, AI governance, ML auditing core criteria catalog, AI auditing framework.

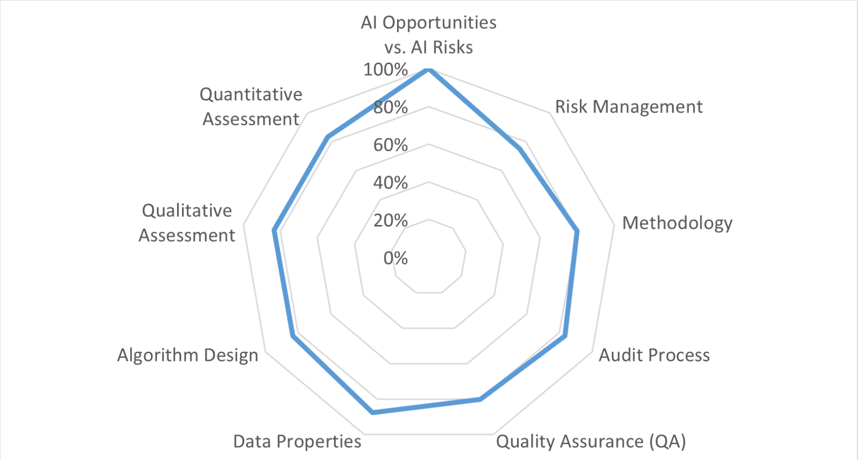
I. BACKGROUND

Artificial Intelligence (AI), especially Machine Learning (ML) algorithms, become more and more popular in the healthcare market. In the U.S., only 7% of the hospitals have a fully operational AI and automation strategy, even though 90% started a draft [4, p. 4]. Identifying the best suited processes for ML algorithm implementation is one of the biggest challenges according to Sage Growth Partners and Olive [4, p. 4]'s study. It is not easy to answer this ex-ante.

Hypothetical Result for ML Development Project



Hypothetical Result for ML Implementation Project



Current Activities

- All 30 questions of the auditing catalog have been applied to a prediction project in the medical area
- Goal was to assess the publication, all supplementary files and the code base retrospectively
- Replication of the Python data processing pipeline and results as major activity

⇒ Our manuscript is currently under peer review.

Thank you!

Feel free to contact

markus.schwarz@campus.lmu.de

or

fady.albashiti@med.uni-muenchen.de

Q&A