Aus der

Klinik für Anaesthesiologie
Klinikum der Ludwig-Maximilians-Universität München

Auditing Framework of ML Models Applied in Medicine

Dissertation
zum Erwerb des Doctor of Philosophy (Ph.D.)
an der Medizinischen Fakultät
der Ludwig-Maximilians-Universität München

vorgelegt von

Markus Schwarz

aus

Lahr / Deutschland

Jahr

2025

Mit Genehmigung der Medizinischen Fakultät der
Ludwig-Maximilians-Universität München

Erstes Gutachten:	Prof. Dr. Ludwig Christian Hinske
Zweites Gutachten:	Dr. Fady Albashiti
Drittes Gutachten:	Prof. Dr. Kristian Unger
Viertes Gutachten:	Prof. Dr. Frederick Klauschen

Dekan:	Prof. Dr. med. Thomas Gudermann

Tag der mündlichen Prüfung:	30.10.2025

Affidavit

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Promotionsbüro Medizinische Fakultät

MMRS

Affidavit

Schwarz, Markus

Surname, First Name

I hereby declare that the submitted thesis entitled

"Auditing Framework of ML Models Applied in Medicine"

is my own work. I have only used the sources indicated and have not made unauthorised use of services of a third party. Where the work of others has been quoted or reproduced, the source is always given.

I further declare that the dissertation presented here has not been submitted in the same or similar form to any other institution for the purpose of obtaining an academic degree.

Munich, 9th November 2025

Place, Date

Markus Schwarz

Signature Doctoral Candidate

Confirmation of Congruency

LMU

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Promotionsbüro
Medizinische Fakultät

MMRS

Confirmation of congruency between printed and electronic version of the doctoral thesis

Schwarz, Markus

Surname, First Name

I hereby declare that the electronic version of the submitted thesis entitled

“Auditing Framework of ML Models Applied in Medicine”

is congruent with the printed version both in content and format.

Munich, 9th November 2025

Place, Date

Markus Schwarz

Signature Doctoral Candidate

Table of Contents

List of Abbreviations

AAI	Auditable AI
ABx	Administration of Antibiotics
ACP	Algorithm Change Protocol
AGI	Artificial General Intelligence
AI	Artificial Intelligence
AI4H	Artificial Intelligence for Health
ALTAI	Assessment List of Trustworthy Artificial Intelligence
APC	Article Processing Charge
ATDD	Acceptance Test-Driven Development
attn	Deep Self-Attention Model
AUC	Area Under the Curve
AUMC	Amsterdam University Medical Center (Intensive Care Database)
B	Billion
BBX	Black Box Explanation
BPT	Breusch-Pagan Test
cAI	Connectionist Artificial Intelligence
CEN	Comité Européen de Normalisation [European Committee for Standardization]
CENELEC	Comité Européen de Normalisation Électrotechnique [European Committee for Electrotechnical Standardization]
cIT	Classical Information Technology
CITI	Collaborative Institutional Review Board Training Initiative

COBIT	Control Objectives for Information and Related Technologies
COSO	Committee of Sponsoring Organizations
CPU	Central Processing Unit
CRediT	Contributor Roles Taxonomy
CRISP-DM	Cross Industry Standard Process for Data Mining
CSVP	Cosine Similarity Vector Pairs
CT	Computed Tomography
CUDA	Compute Unified Device Architecture
DGP	Data Generation Process
DI	Disparate Impact
DL	Deep Learning
DNN	Deep Neural Network
DSS	Decision Support System
e.g.	Exempli Gratia [For Example]
E/E/PE	Electrical, Electronic and Programmable Electronic Devices
EC	European Commission
eICU	Electronic Intensive Care Unit (Collaborative Research Database)
EKNZ	Ethikkommission Nordwest- und Zentralschweiz [Ethics Committee for Northwestern and Central Switzerland]
ERM	Enterprise Risk Management
ESM	Epic Sepsis Model
ETSI	European Telecommunications Standards Institute
EU	European Union
FDA	Food and Drug Administration (of the United States)
FG	Focus Group
FN	False Negative
FP	False Positive

GAFAI	Generalized Audit Framework for Artificial Intelligence
GDPR	General Data Protection Regulation
GLM	Generalized Linear Model
GMLP	Good Machine Learning Practice
GPU	Graphics Processing Unit
gru	Recurrent Neural Networks Employing Gated Recurrent Units
HiRID	High time-Resolution ICU Dataset
HLEG	High Level Expert Group on Artificial Intelligence
i.i.d.	Independent and Identically Distributed Random Variables
ICU	Intensive Care Unit
IEC	International Electrotechnical Commission
IRR	Inter-Rater Reliability
ISACA	Information Systems Audit and Control Association
ISO	International Organization for Standardization
ITU	International Telecommunication Union
lgbm	Light Gradient-Boosting Machine
LIME	Local Interpretable Model-Agnostic Explanations
LLM	Large Language Model
LMU	Ludwig-Maximilians-University Munich
lr	Logistic Regression
LTS	Long Time Support
M	Million
MCC	Matthews Correlation Coefficient
MeDIC$^{\text{LMU}}$	Medical Data Integration Center of LMU University Hospital
MEWS	Modified Early Warning Score
MIMIC-III	Medical Information Mart for Intensive Care III
ML	Machine Learning

MLM	Multilevel Models
MLOps	Machine Learning Operations
MPIB	Max Planck Institute of Biochemistry
MRI	Magnetic Resonance Imaging
NLP	Natural Language Processing
PPV	Positive Predictive Value
QA	Quality Assurance
qSOFA	Quick Sequential Organ Failure Assessment
R&D	Research and Development
RACI	Responsible Accountable Consulted Informed
RAM	Random-Access Memory
ricu	Intensive Care Unit Data with R
ROC	Receiver Operating Characteristics
RT	Radiation Therapy
sAI	Symbolic Artificial Intelligence
SaMD	Software as Medical Device
SD	Standard Deviation
SHAP	Shapley Additive Explanations
SI	Suspected Infection
SMACTR	Scoping Mapping Artifact Collection Reflection
SOFA	Sequential Organ Failure Assessment
SP	Statistical Parity
SPS	Software as Medical Device Pre-Specification
SWT	Shapiro-Wilk Test
TAC	Thesis Advisory Committee
TN	True Negative
TP	True Positive

TPR	True Positive Rate
TSViz	Time Series Visualization Framework
TSVR	Total Sobol's Variance Ratio
TSXplain	Time-Series Explanation
U.S.	United States of America
UNHR	United Nations Human Rights (Law)
VIF	Variance Inflation Factor
VM	Virtual Machine
XAI	Explainable AI
XbD	Explanation by Design

List of Publications

My Ph.D. project consists of two publications:

1) M. Schwarz, L. C. Hinske, U. Mansmann, F. Albashiti, “Designing an ML Auditing Criteria Catalog as Starting Point for the Development of a Framework”. In: IEEE Access 12 (2024), pp. 39953-–39967. DOI: 10.1109/ACCESS.2024.3375763.

2) M. Schwarz, L. C. Hinske, U. Mansmann, F. Albashiti, “ML Auditing and Reproducibility: Applying a Core Criteria Catalog to an Early Sepsis Onset Detection System”. In: IEEE Access 13 (2025), pp. 104899–104915. DOI: 10.1109/ACCESS.2025.3579631.

Chapter 1

Contribution to the Publications

For both publications of this Ph.D. project, CRediT (Contributor Roles Taxonomy) is used to describe my contribution and interaction with the co-authors (NISO, 2025). This is depicted in a compressed tabular representation using bullet points. The following author abbreviations are used:

FA	Fady Albashiti
LH	Ludwig Christian Hinske
MS	Markus Schwarz
UM	Ulrich Mansmann

In both publications, MS was the first author and served as the corresponding author.

1.1 Contribution to Paper I

CRediT Term	Description of Contributions
Conceptualization	• MS followed the idea of FA to focus on the auditing criteria in the first paper • MS followed the advice of UM to use the term “core criteria catalog” • MS created the paper outline and discussed it with FA

Methodology

- MS followed the advice from UM to perform a qualitative content analysis after Mayring (2000)
- MS followed the advice from UM to register a scoping study on OSF.io[1]
- MS utilized best practices from Mayring (2000), Miles and Huberman (1994), and Bortz and Döring (2006) to create a three-stage qualitative content analysis
- MS discussed the process steps of the method with FA and integrated FA's feedback

Software

- MS set up a LaTeX[2] writing environment using Citavi[3] as a literature database
- MS set up auxiliary tools (e.g., FreeMind[4]) necessary for the investigation

Analysis & Investigation

- MS defined the literature search strategy and condensed knowledge in source summaries
- MS performed a qualitative content analysis, leading to a mind map with 12 categories and 800 artifacts[5]
- MS selected, reflected on, and synthesized relevant artifacts to identify 5 categories with 34 artifacts
- MS restructured, subsumed, and connected the artifacts to form the 30-question ML auditing core criteria catalog
- MS discussed the status of the analysis and the design process regularly with FA and integrated FA's feedback

Resources

- MS integrated additional sources from FA, LH and UM into the literature search strategy
- MS used a workplace at the Medical Data Integration Center of LMU University Hospital (MeDIC$^{\text{LMU}}$) provided by FA

[1] "OSF is a free, open platform to support ... research and enable collaboration" (COS, 2025).

[2] "LaTeX, ... is a document preparation system for high-quality typesetting. It is most often used for medium-to-large technical or scientific documents" (The LaTeX Project, 2025).

[3] "Citavi is [an] all-in-one reference management tool with knowledge organization" (Lumivero, 2025).

[4] "FreeMind is a ... free mind-mapping software written in Java" (Foltin, 2023).

[5] "Here 'artifact' is used to describe a virtual product from an author of a scientific article. Such a product is generated content of specific type (e.g., idea description, conclusion, figure, table or code). It is often the result of a thought process steming from synthesis or analysis of a topic" (Schwarz et al., 2024a, p. 39955).

Writing - Original Draft	• MS wrote the original draft of the manuscript
Writing - Review & Editing	• MS regularly consulted FA with the current draft version and integrated FA's annotations and change suggestions • MS provided LH and UM with the final draft and addressed LH's and UM's remarks and change requirements • MS submitted the manuscript to the IEEE Access[6] journal • MS corresponded with the journal during the whole review-process until the paper was accepted and the final page-proof version was released
Visualization	• MS presented the contents of the publication at the 69th Annual GMDS Conference[7] (Schwarz et al., 2024b) and the Miracum-DIFUTURE Colloquium on 25th February 2025[8] (Schwarz & Albashiti, 2025)
Supervision	• MS received operational supervision from FA within the Thesis Advisory Committee (TAC) • MS received formal supervision from LH within the TAC • MS received additional supervision from UM within the TAC
Project administration	• MS was responsible for the project planning and execution, regularly receiving feedback from FA • MS organized the Ph.D. topic into actionable items on a timeline with guidance from FA, LH and UM
Funding acquisition	• MS sought for partial reimbursement of the article processing fee (APC) at the LMU Open Access Fund in collaboration with FA

[6] "IEEE Access is a multidisciplinary, online-only, gold fully open access journal, continuously presenting the results of original research or development across all IEEE fields of interest" (IEEE, 2024).

[7] At this conference, "leading professional societies from Germany in the fields of biomedical informatics, biometrics, epidemiology, social medicine, prevention, medical sociology and public health have come together" (GMDS, 2024).

[8] Here, experts from the medical informatics fields regularly present new projects, results or developments and discuss them with their peers (Mannheim University of Applied Sciences, 2025).

1.2 Contribution to Paper II

CRediT Term	Description of Contributions
Conceptualization	• MS proposed to "test" the 30-question ML auditing core criteria catalog on a practical example • MS applied the catalog to a publication from Moor et al. (2023a) according to advice of FA • MS followed the recommendation of FA that the goal of the 2nd paper should be to employ an "external auditor's point of view" and not to "challenge the medical soundness" (Schwarz et al., 2025, p. 2)
Methodology	• MS proposed using the first three steps of SAI of Finland et al. (2020, p. 16), which are also mentioned under the key word "Audit Process" in the first paper for answering the catalog questions, as well as for undertaking the reproduction study • MS had the idea to use a 3-point ordinal scale for codifying the raters' responses and justified it with literature • MS applied inter-rater reliability (IRR) techniques including utilizing a second rater from MeDIC^{LMU} to reduce the subjectivity of the catalog application • MS conducted data perturbation tests to investigate the robustness of the deep self-attention model (attn) being confronted with minor input data changes

Software

- MS set up a LaTeX writing environment using Citavi as a literature database
- MS set up auxiliary tools (e.g., Visual Studio Code[9]) necessary for the investigation
- MS followed strictly the instructions provided by Moor et al. (2023b) when reproducing the complete data preprocessing, ML training and ML evaluation pipeline
- MS wrote a detailed protocol about every step conducted while doing the reproduction study (incl. when and why it was necessary to deviate from the given instructions)
- MS configured the reproduction software environment on his virtual machine (VM)
- MS did all code changes necessary to successfully execute the reproduction
- MS received help from FA's team members who configured the reproduction software environment on the GPU server
- MS created docker images[10] of the reproduction environments with the help of FA's team
- MS created Python code necessary for the perturbation experiment
- MS uploaded the docker images and all relevant reproduction files on OSF.io to ensure the highest level of transparency

[9] "Visual Studio Code combines ... a code editor with what developers need for their core edit-build-debug cycle. It provides ... code editing, navigation, and understanding support along with lightweight debugging" (Microsoft, 2025).

[10] "Docker is an open platform for developing, shipping, and running applications. Docker enables ... to separate ... applications from ... infrastructure" (Docker Inc, 2024).

Analysis & Investigation

- MS conducted an in-depth study of Moor et al. (2023a)'s paper, their paper supplement, and their GitHub repository
- MS identified relevant information necessary to answer the 30 questions of the ML auditing core criteria catalog
- MS answered all questions of the catalog with detailed reference to the acquired body of knowledge
- MS integrated the answers from the second rater and calculated the weighted Cohen's kappa agreement coefficient[11]
- MS chose a radar diagram to display the aggregated results of each rater
- MS executed all computing steps on his VM as part of the reproduction study
- MS supervised the execution of all computing steps on the GPU server as part of the reproduction study
- MS conducted a root cause analysis on the reproduction differences and elaborated on the findings
- MS provided regular updates on answering the catalog questions and performing the reproduction study to FA, who gave feedback on the next steps

Resources

- MS used a workplace and computing infrastructure provided by FA at MeDIC$^{\text{LMU}}$ during the reproduction
- MS was in regular exchange with a member of Karsten Borgwardt's work group[12] at the Max Planck Institute of Biochemistry (MPIB) after an established connection by FA
- MS sought contact with Moor et al. (2023a) to acquire original results to identify the root cause of reproduction discrepancies

Data Curation

- MS acquired access to the four ICU datasets necessary for the reproduction study
- MS performed data cleansing, harmonization, and transformation as part of the reproduction study pipeline

Writing - Original Draft

- MS wrote the original draft of the manuscript

[11] This coefficient can be utilized to quantify the extent of agreement between two raters (Gwet, 2014, p. 102).

[12] Karsten Borgwardt is the corresponding author of the sepsis prediction project publication (Moor et al., 2023a).

Writing - Review & Editing	• MS regularly consulted FA with the current draft version and integrated FA's annotations and change suggestions • MS provided LH and UM with the final draft/revision and addressed LH's and UM's remarks and change requirements • MS submitted the manuscript to the IEEE Access journal • MS replied in detail to all issues from the reviewers and revised the manuscript accordingly • MS corresponded with the journal during the whole review-process until the paper was accepted and the final page-proof version was released
Visualization	• MS presented the results of the catalog application and reproduction study to Karsten Borgwardt and his team on 12th February 2025
Supervision	• MS received operational supervision from FA within the Thesis Advisory Committee (TAC) • MS received formal supervision from LH within the TAC • MS received additional supervision from UM within the TAC
Project administration	• MS was responsible for the project planning and execution, regularly receiving feedback from FA • MS organized the Ph.D. topic into actionable items on a timeline with guidance from FA, LH and UM
Funding acquisition	• MS sought for partial reimbursement of the article processing fee (APC) at the LMU Open Access Fund in collaboration with FA

Chapter 2

Introductory Summary

2.1 Motivation

Artificial Intelligence plays an increasingly significant role in our daily lives. This applies to both private and professional domains. Hardly a day passes without an AI-related news headline describing another breakthrough or new development. OpenAI recently released a research preview of *GPT-4.5* that was “designed to be more general-purpose” and would provide a “more natural [interaction]” than previous model versions (OpenAI, 2025). On their road towards Artificial General Intelligence (AGI), OpenAI (2023)’s mission is to “empower humanity to maximally flourish in the universe.” There is also a recent uprising of new large language models (LLM) like *DeepSeek-R1*[13] or *Mistral Large 2*[14], which will certainly further boost the competition. “AI Prompt Engineer” is now a fully fledged job role that larger companies advertise (Manatal, 2025).

When you are an office worker whose corporation adopts *Microsoft 365 Copilot*[15] technology, you might already use the integrated copilot to help generating meeting minutes by transcribing what was said in the previous meeting on *Teams*[16]. It can also be employed to extract key insights of existing word documents, or to collect facts about a new topic your manager wants you to make a short presentation about. Also,

[13] This model was developed by Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co., Ltd. (Sina Corporation, 2025). They claim that “DeepSeek-R1 achieves [a] performance comparable to OpenAI-o1 across math, code, and reasoning tasks” (DeepSeek, 2025).

[14] “Mistral AI, [is] a ... French artificial intelligence startup founded in April 2023” (Mistral AI, 2025). “Mistral Large 2 is designed for single-node inference with long-context applications in mind” (Mistral AI, 2024).

[15] The main platform was formerly called “Microsoft 365” and contains many other applications or services e.g., like Word, Excel or Sharepoint (Bott, 2025). “Copilot can access [the] organization’s data ... to generate a response [to a prompt] that is contextually relevant to the user’s task” (Ohlinger & Carter, 2025).

[16] “Teams, ... [is a] hub for teamwork, ... where people ... can actively connect and collaborate in real time to get things done” (Smith et al., 2025).

in the business world, LLMs are quite often used for creating outlines of presentations or for brainstorming.

However, when you are a doctor working in a hospital or have your own practice, there is not an abundance of AI tools that you can choose from to increase the quality of care for your patients or to make the overall interaction with them more efficient.

You will most likely get in contact with AI that is already integrated into workflows of medical devices that are specialized in different medical departments. For example, in the radiology department, the software that is used to analyze structures and artifacts within the human anatomy coming from computed tomography (CT) scans or magnetic resonance imaging (MRI), may already contain image recognition algorithms. Those can e.g., automatically generate contours of internal organs to avoid damage caused by radiation therapy (RT) of tumor volumes (MIM Software Inc., 2025).

In the U.S., medical devices are subject to regulation by the FDA. Starting with their initial white paper (FDA, 2019), the FDA has put a lot of emphasis on guiding manufacturers and developers on the approval of (self-learning) AI/ML components (FDA, 2025). For the contouring algorithm example above, there are five entries in the FDA's "AI/ML-Enabled Medical Devices List," indicating that they have "met the FDA's applicable premarket requirements" (FDA, 2024).

The research and development (R&D) of new algorithms in the medical sector is flourishing. Many of them have the potential to increase patient's quality of life, or to reduce the burden of labor-intensive manual tasks when following medical guidelines for diagnostics. They can also help patients to prevent diseases in the first place or assist doctors with the treatment of illnesses.

Back in 2020, my operational supervisor Dr. Fady Albashiti[17] and I asked ourselves the questions:

- "What are the reasons of underutilization of promising algorithms in clinical practice?"
- "What can be done to promote their safe and effective adoption?"

Over many discussion rounds, we identified that *trust* plays a key role and that there seems to be a general tendency to distrust recommendations or decisions made by ML algorithms, especially among doctors (Kuan, 2019).

[17] He is the CEO of the Medical Data Integration Center of LMU University Hospital (MeDIC$^{\text{LMU}}$). Its purpose is to collect and harmonize hospital's routine data, make it available for research and assist researchers in data questions.

This led us to ask the question: “What can be done to increase the trust in (good) ML algorithms for medical professionals?”

After having done an initial literature search, it became evident that we oriented ourselves towards the field of *Auditable AI*. Auditing in the business world usually refers to external companies that use standardized and internationally recognized methods and procedures to check whether internal activities are done correctly and in the best interest of the organization in scope and its stakeholders.

It became clear that there is a research gap in exactly those methods, tools, and procedures necessary for auditing an organization or individuals that develop or implement ML algorithms in the medical sector.

Consequently, my Ph.D. project with the title “Auditing Framework of ML Models Applied in Medicine” was formed. In the first two years of this Ph.D. project, the focus was set on completing required coursework, further specification of the research questions and AAI knowledge acquisition. The actual research activities took place in the last three years of this Ph.D. project.

2.2 Background

Market research organizations quantify the value that AI brought to the healthcare sector to $18.7B in 2023, with a potential to grow to over $300B by 2032 (Global Market Insights Inc, 2024). This is largely because algorithmic ML models promise process improvements, efficiency gains or enable new medical procedures in the first place (Healthcare Tech, 2019). For example, in the use case of assisting radiologists to classify cell structures, DNN ML models are already superior to humans (Bizzego et al., 2019, p. 15).

The highly estimated market growth potential of more than 37% annually can only be achieved, if medical practitioners are convinced that the outcome of the ML algorithm or product; may it be for prevention, diagnosis, or treatment of diseases, can be trusted. In the medical industry, even more than in the automotive or aviation industry, there is not much room for error. Severe consequences of malfunctioning AI products can lead to patient harm and huge liability claims could follow.

The term “Auditable AI” (AAI) is new in academic literature. In Schwarz et al. (2024a, p. 39954) it is distinguished from the related terms “Explainable AI” (xAI) and “Interpretable AI”. In March 2023, the search of AAI on Google Scholar led to only 40 hits.

In March 2025, exactly after two years, the same search achieved 163 hits[18]. This is a considerable growth of research activity, but publications dealing with AAI are by far only a tiny fraction of the overall body of 4.74M hits existing with the term "Artificial Intelligence."[19]

This indicates that back in 2019, when this Ph.D. project's proposal was written, there existed almost no academic literature about AAI and its implications. This was another reason to ground the work of this Ph.D. project in the field of AAI.

2.3 Objectives

This Ph.D. project focuses on solving a "practical problem, ... that imposes ... a tangible cost that [society doesn't] want to pay" (Booth et al., 2008, p. 55). "Society" in this context refers primarily to ML developers, healthcare professionals and patients. The main aspects that describe the practical problem are:

- Distrust of clinicians on ML model's/AI product's decision
- Fear of manipulation of AI behavior by patients and regulators
- Ineffective patient data usage for individual prevention and treatment decisions
- Unleveraged efficiency gains
- Low ML model/AI product adoption in healthcare compared to other industries

The research problem is described by:

1. Lack of transparency in ML model predictions
2. Not existing auditing criteria of ML models applied in medicine
3. Missing framework how to audit ML models effectively and efficiently in a scientific way

The overall goal is to contribute to identifying relevant assessment determinants for ML models/AI products. Those would promote the safe and effective ML model/AI product adoption by health scientists and practitioners.

The first publication, *Designing an ML Auditing Criteria Catalog as Starting Point for the Development of a Framework*, "[sets] the focus ... on carving out relevant auditing aspects in form of a core criteria catalog" (Schwarz et al., 2024a, p. 39954). This is especially

[18] The search was conducted on 19.03.2025.

[19] The search was also conducted on 19.03.2025.

relevant, since in literature it is “generally agreed [that a] framework for auditing AI systems is required” and stated that “’[there] is no agreed framework for assessing or reporting the results of health AI models.’”

The second publication, *ML Auditing and Reproducibility: Applying a Core Criteria Catalog to an Early Sepsis Onset Detection System*, where this catalog is applied to an actual ML development project, is used “to gauge the catalog’s usefulness” (Schwarz et al., 2025, p. 2). This goal also entails obtaining practical insights related to the 3rd research problem. Lastly, by conducting a comprehensive reproducibility study, which is also part of the second publication, important lessons about the “ease of reproducing ... [an] algorithm development pipeline and ... acquired results” can be learned. The reproduction process and results also aim at the 1st research problem, namely, to contribute to making ML model predictions more transparent.

2.4 Methods

We first conducted a “scoping study” for “identifying commonalities and differences among themes, artifacts and patterns” (Schwarz et al., 2024a, p. 39954). Therefore a “qualitative content analysis” was carried out “[utilizing] best practices from Mayring (2000, pp. 3–6), Miles and Huberman (1994, pp. 245–287), and Bortz and Döring (2006, pp. 149–154).”

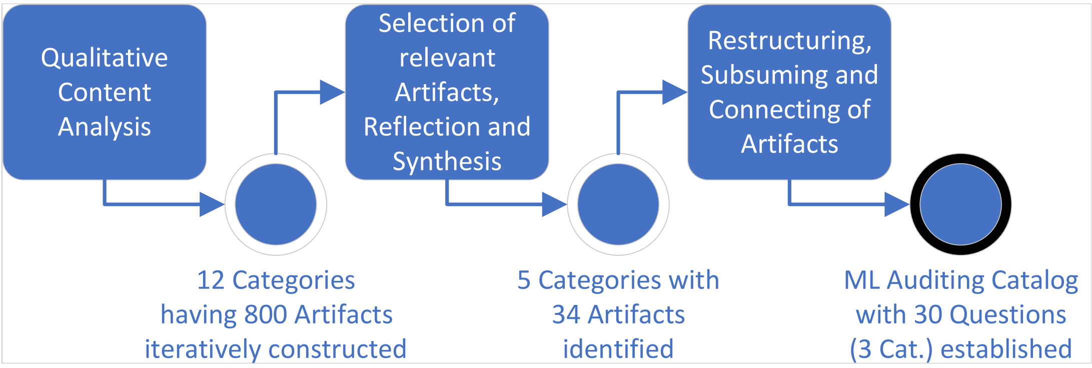

Qualitative Content Analysis
Selection of relevant Artifacts, Reflection and Synthesis
Restructuring, Subsuming and Connecting of Artifacts
12 Categories having 800 Artifacts iteratively constructed
5 Categories with 34 Artifacts identified
ML Auditing Catalog with 30 Questions (3 Cat.) established

FIGURE 2.1: Process Steps of the 30-Question ML Auditing Core Criteria Catalog Synthesis

Source: Schwarz et al. (2024a, p. 39956)

In Schwarz et al. (2024a, pp. 39954–39955) we started the literature assessment with the “exact term ’Auditable AI’” being “executed on 02.03.2023” in Google Scholar. Also, “[AAI] white papers” and additional “sources based on expert recommendation” were added. A selection process led to a total of 41 relevant publications being used in the qualitative content analysis, where we could “[reveal] concepts and ideas behind

AAI." This was done iteratively growing an exhaustive mind map[20], capturing and ordering "12 main categories having a total number of 800 artifacts." After two selection, reflection and synthesis repetitions, we established the final "ML auditing core criteria catalog consisting of 30 questions." The described process steps are outlined in figure 2.1.

As described in Schwarz et al. (2025, pp. 2–3), in the catalog application and reproducibility study, "we considered the first three steps described by SAI of Finland et al. (2020, p. 16)," which are also given below:

1. "'Reviewing the documentation'"
2. "'Close inspection of the data and a review of the code'"
3. "'Reproduction of ... the model training, testing, scoring and performance measures'"

First, a "thorough study of the existing paper, paper supplement and GitHub repository from Moor et al. (2023a, 2023b)," is necessary before the 30 catalog questions can be answered. "To reduce the amount of subjectivity, ... we utilize inter-rater reliability (IRR) techniques," where "a second, independent rater ... also answered each of the 30 questions retrospectively," using "a 3-point ordinal scale." Then a "suitable agreement coefficient" is calculated.

For "the reproduction of the sepsis project, [which refers to] steps two and three of the audit process," we first conducted an "in-depth inspection of the utilized datasets, as well as the complete data processing pipeline" (Schwarz et al., 2025, pp. 2–3). Once a "sufficient understanding of the code base's functioning" was acquired, we started "working on the reproducibility of the results."

2.5 Results

The "ML auditing core criteria catalog," as being presented in Schwarz et al. (2024a, p. 39955), consists of 30 questions that are grouped around the categories "Conceptual Basics" (17 questions), "Data & Algorithm Design" (7 questions) and "Assessment Metrics" (6 questions).

[20] For this activity the program *FreeMind* was used. "FreeMind is a ... free mind-mapping software written in Java" (Foltin, 2023).

The category “Conceptual Basics” consists of the following subcategories:

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

“Data & Algorithm Design” contains:

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

And the last category “Assessment Metrics” includes:

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

This 30-question catalog is intended to be applicable for various industry areas and different ML model types. “The questions are balanced in terms of breadth and [depth] to provide an operationalizable starting point for diverse stakeholders” (Schwarz et al., 2024b). The questions are thought to be “beneficial to ... organizations that have been or will start implementing ML algorithms” (Schwarz et al., 2024a, p. 39963). They would also help them “being prepared for any upcoming legally required audit activities.” All 30 questions of the catalog are provided in appendix A. Figure 2.2 presents the contents of the catalog in a word cloud.

The application of the ML auditing core criteria catalog to an “early sepsis onset detection system use case” is done in Schwarz et al. (2025). The 30 questions of the catalog were successfully answered by the corresponding author, as well as by a second rater from MeDIC$^{\text{LMU}}$, using “the existing [sepsis] paper, paper supplement and GitHub repository” (Schwarz et al., 2025, p. 2). Both rating results, aggregated per subcategory, are depicted as a radar diagram in figure 2.3.

There are two important take aways from the diagram. First, there is evident agreement between both auditors, who applied the catalog questions independently and without any interaction or instruction. This is also confirmed by a “weighted Cohen’s kappa coefficient of $\kappa = 0.51$, ... constituting a ’fair to good’ agreement” (Schwarz et al., 2025, p. 15). Second, “the focus of ... [the sepsis prediction project] is rather on the left side of the diagram, including algorithm design, data properties, assessment

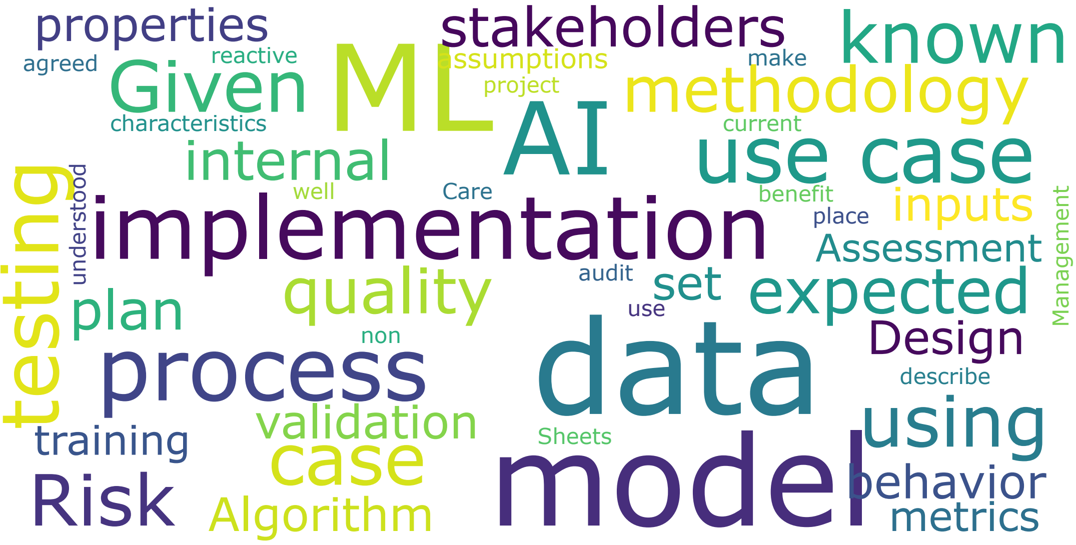


FIGURE 2.2: Word Cloud of the 30-Question ML Auditing Core Criteria Catalog

Each sentence is split into words and then displayed according to the calculated rank.
Words with less than two characters are ignored.
Source: Schwarz et al. (2024a, p. 39964)

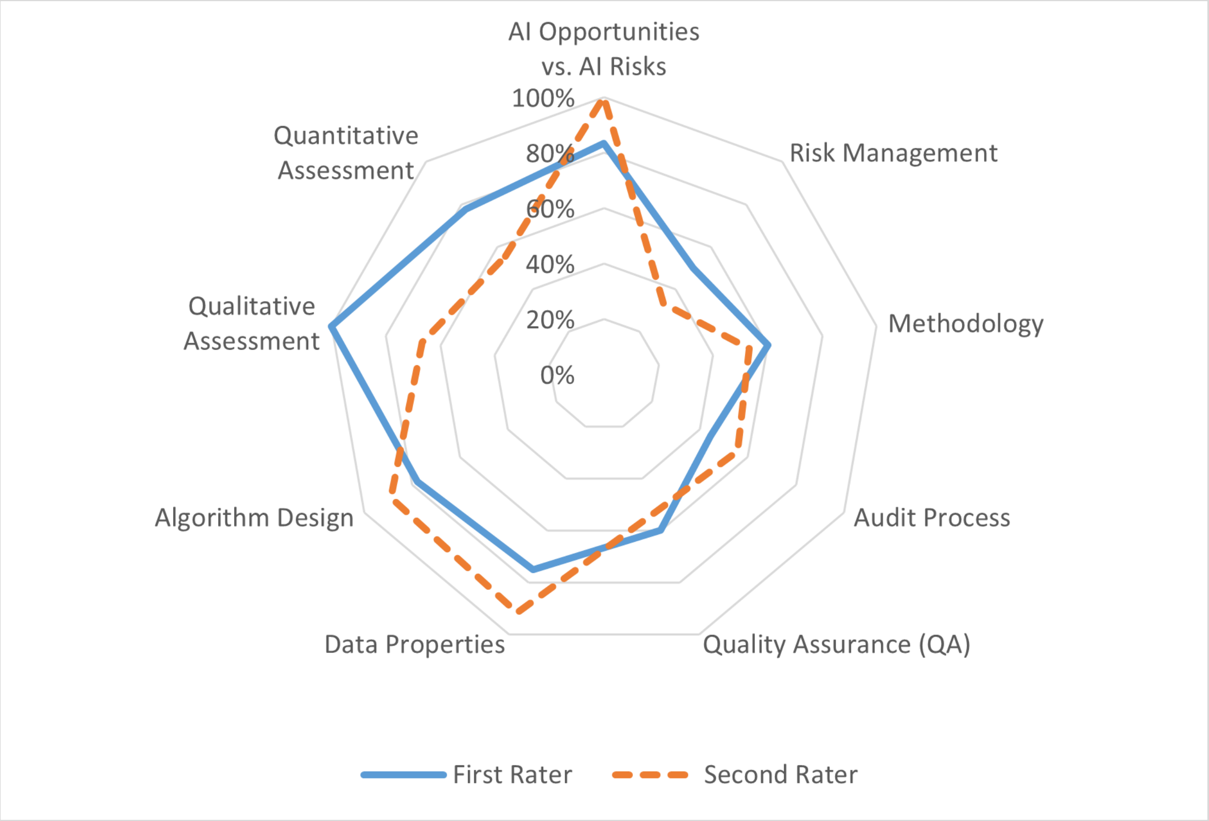


FIGURE 2.3: Sepsis Project Audit Catalog Subcategory Coverage

Questions were summed up per subcategory using their codified numeric value.
Afterwards the sum of each subcategory was divided by the maximum possible sum per subcategory.
Source: Schwarz et al. (2025, p. 9)

metrics, as well as presenting opportunities of such a sepsis early warning system" (Schwarz et al., 2025, p. 9). This is typical for a development project and in contrast to an implementation project, where the scope is extended to include "areas like risk management, quality assurance or audit process."

Looking at the reproducibility part of the second publication, and focusing on the deep self-attention model being externally validated, Schwarz et al. (2025, pp. 13–15) acquired an AUC of 0.717 (-5.83%) compared to Moor et al. (2023a, pp. 5–6) and a PPV of 28.3 (-11.03%). The lead time to sepsis onset metric did not yield to "meaningful values." Taking the AUC and the PPV deviations into account, together with the fact that a similar ranking of the best performing models was achieved, it can be concluded that "the magnitude of ... [the] reported performance metrics [can be reproduced]."

2.6 Discussion

Considering the results of both publications, we were able to contribute to the Auditable AI community by two main artifacts. First, by creating a 30-question ML auditing core criteria catalog, constituting a tool that is easy to operationalize for a wide variety of use cases. The practical test of the catalog, being done retrospectively using an ML algorithm development project in the medical sector, indicates that the questions have validity and tend to be applied reliably among a team of auditors.

Second, the process of the catalog application, which triggered a sophisticated reproduction study that consumed a lot of resources of this Ph.D. project, also provided important practical implications related to auditing existing ML models. For example, it became evident that dependencies on soft- and hardware environments as well as code versions being worked on by different people play a significant role for the success of a reproducibility study.

We also must mention a few limitations of this Ph.D. project. The decision to structure the ML auditing core criteria catalog in three categories containing 30 questions in total was made to ensure actionability for a broad spectrum of use cases and users (Schwarz et al., 2024a, p. 39963). Consequently, it is plausible that certain aspects of AAI have not been considered with the necessary depth.

Also, when performing the application of the catalog questions to the sepsis prediction project, I as the first rater had quite differently answered three questions compared to my colleague at MeDIC$^{\text{LMU}}$, who acted as the second rater (Schwarz et al., 2025, p. 15). The reported Cohen's kappa agreement coefficient is also "still [showing] room for improvement." This "can be mitigated if more groups (multiple auditors) would apply our catalog to various ML development or implementation projects (ideally first within the healthcare sector). Because then a kind of 'catalog application guideline' with practical recommendations could be established."

Lastly, the "probable existence of an algorithmic error," causing us to not acquire meaningful results of the third metric of the sepsis prediction project, should be looked at in more detail (Schwarz et al., 2025, p. 15). However, this would require additional resources within a team of external auditors as well as the willingness of Moor et al. (2023a) to explicitly dedicate resources for a collaborative in-depth code inspection.

Those limitations hopefully can be worked on by a potential successor at MeDIC$^{\text{LMU}}$, respectively by the broader auditable AI community in general.

Ideally, for future ML development or implementation projects, our ML auditing core criteria catalog should be used in parallel with the development or implementation process. Then, external auditors who are familiar with the catalog itself and its application, can become sparring partners within the whole project, ensuring transparency, as well as reproducibility that is independent of soft- or hardware environments.

Chapter 3

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

Title:	Designing an ML Auditing Criteria Catalog as Starting Point for the Development of a Framework
Authors:	Markus Schwarz*, Ludwig Christian Hinske, Ulrich Mansmann, Fady Albashiti
Journal:	IEEE Access
Submitted:	7th February 2024
Accepted:	1st March 2024
Published:	11th March 2024
Year:	2024
Volume:	12
Pages:	39953–39967
DOI:	10.1109/ACCESS.2024.3375763
OSF Project ID:	tdr3p
JIF Percentile:	65.4 (2023, Computer Science, Information Systems)
JIF Rank:	87/250 (2023, Computer Science, Information Systems)
Rights:	Creative Commons Attribution 4.0 United States (CCBY 4.0)

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

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. To do so, the algorithm needs to be successfully audited, having robust evidence available that proofs the algorithm

The associate editor coordinating the review of this manuscript and approving it for publication was Thomas Canhao Xu.

is able to perform the task within defined quality control metrics.

Von Twickel et al. [5, p. 2] highlight that a generally agreed upon framework for auditing AI systems is required. This is in line with Wiegand et al. [6, p. 10] who state that "[t]here is no agreed framework for assessing or reporting the results of health AI models." Brundage et al. [7, p. 11] mention that AI systems are intransparent and often closed source. This is crucial as raising human confidence in AI systems is elementary to building trustworthy, socially beneficial and responsible AI [8, p. 83]. According to SAI of Finland [9, p. 11], for "a well-functioning public sector", personal data protection, decision explanation and bias are few of the main challenges of ML algorithm implementation. If not tackled well, those challenges could lead to "obscured inefficiency …[and] damaged trust."

The goal of this paper is to contribute to designing an approach how to audit machine learning algorithms. To do so, the focus is set on carving out relevant auditing aspects in form of a core criteria catalog.

A commonly agreed AI auditing framework that provides best practices and tools could help speeding up the ML adoption process in the healthcare market.

Before describing ways *how* to perform audits and *what* pillars a supporting criteria catalog may be made of, it is important to first get a good understanding of the term *Auditable AI (AAI)*. In this context, Dengel et al. [10, p. 91] mention AI systems that "should be able to answer questions asked by humans and interact with them in an understandable way." To operationalize this definition, Benchekroun et al. [11, p. 2] state that "[a]uditability …ensur[es] that the AI model behaves as expected." Next to this definition, Dengel et al. [10, p. 103] also provide an idea how this could be achieved: By "[querying] AI systems …externally with hypothetical cases", whereas those cases can be based on real world data or being artificially created.

Explainability, on the other hand, which is often abbreviated by the term *Explainable AI (XAI)*, makes sure that humans are able to derive a sufficient understanding of the model's inner workings [12, p. 7]. Ideally, the explanations are exhaustive and adjusted to the individual [13, p. 1048]. According to Chatila et al. [14, p. 23], XAI's main focus lays on generating *Black Box eXplanation[s] (BBX)*, in contrast to inherently understandable *white box* models.[1] BBX should allow "humans …[to] debug, interpret, control, and reason about [deep neural networks]" [10, p. 91].

Lastly, "*Interpretability* refers to the observation and representation of cause and effect within a system" [10, p. 94]. Russell and Norvig [13, p. 729] accentuate that the focus is on comprehending the input/output behavior of an ML model, without necessarily opening the black box. Many authors argue that XAI is a necessary, but not always sufficient, prerequisite to achieve human comprehension of an AI system [12, pp. 7–8].

In the first part of this paper, we explore important artifacts in the field of AAI using inductively constructed categories. In the second part, we further group and synthesize the findings, having an ML auditing core criteria catalog as a result.

TABLE 1. Overview of analyzed literature.

Year	No. of Publications	Most Frequent Publication Type
2018	3	Internet Document
2019	7	Report or Gray Literature, Journal Article
2020	7	Report or Gray Literature
2021	11	Contribution in Conference, etc.
2022	10	Report or Gray Literature
2023	3	Journal Article
	41	

Source: Authors

II. METHODS

In order to achieve the aforementioned goal, it is necessary to get a good understanding of ongoing research in the field of AAI. Thus, we first conducted a scoping study, where we focus on identifying commonalities and differences among themes, artifacts and patterns [15, p. 408]. Our method of choice was a qualitative content analysis. Here we utilized best practices from Mayring [1, pp. 3–6], Miles and Huberman [2, pp. 245–287] as well as Bortz and Döring [3, pp. 149–154].

In our search we used the exact term "Auditable AI" in Google Scholar, leading to 40 hits that translated into 41 sources.[2]

Out of the initial 41 sources, 24 sources were assessed being relevant and 13 sources as not being relevant. Four sources were not accessible.

In addition to the 24 relevant sources of the initial literature research, we added seven relevant white papers in the field of AI auditing to the literature base as well. Furthermore, we included 10 additional sources based on expert recommendation.

We performed a detailed text analysis on the overall body of summarized literature, revealing concepts and ideas behind AAI. Those 41 sources, which can be seen in table 1, aim to provide a concept, methodology, framework or use case how to audit ML algorithms. The table also shows that the field of AAI is very new and rapidly changing, as the oldest publications are from 2018.

For the text analysis, we applied the step model process from Mayring [1, p. 4]. The outcome of the process

[1] Those *white box* models already contain *eXplanation by Design (XbD)*, meaning the model itself is inherently explainable [14, p. 23].

[2] The search was executed on 02.03.2023. A hit is not the same as a source, since sometimes different text passages of the same source are referred as multiple hits or one hit contains several articles of a series. No recursive search based on the source bibliographies was conducted.

are artifacts[3] that are grouped around inductively created categories.

The results structured as an AAI mind map consisted of 12 main categories having a total number of 800 artifacts. The next, main step was to identify relevant artifacts and reflect using own experience on how they could be synthesized into a resulting criteria catalog. We first created five categories containing 34 artifacts. Finally, we restructured the artifacts and put them into a logical connection, establishing an ML auditing core criteria catalog consisting of 30 questions. Those questions are grouped around the final categories *Conceptual Basics*, *Data & Algorithm Design* and *Assessment Metrics*. The final categories and their artifacts are presented in detail in the results section. A summary of the explained method can be seen in figure 1.

III. RESULTS

A. CONCEPTUAL BASICS

AI might provide many *Opportunities*, for example, to increase the economic productivity or to allow the development of new technologies/products. According to Clarke [16, pp. 429–430], "AI's purpose is to extend human capabilities." In that sense, improving products that require repetitive tasks with very high precision and replicability, where humans usually do poor, could gain leverage. Application areas like decision support (DSS) are on the rise, where ML algorithms provide briefings to a human necessary for a decision.

On the other hand, as with every new technology, there are *Risks* involved. Lack of incentives for industry to evaluate and steer threats, political manipulation and bias/no human control in decision (support) systems are often among mentioned risks [17, pp. 3–4]. AI algorithms may even have purposely embedded bias or disinformation so that certain stakeholders achieve their goals [18]. Often, the benefit resulting from a decision and the accountability/liability in case of harm, are not within the same natural or legal person [17, pp. 3–4]. Last but not least, the famous phrase "garbage in, garbage out" also applies for AI [19, p. 130].

Thus, it is very important to establish a good *Risk Management*. Clarke [20, pp. 411–414] categorizes risk management strategies in "proactive", "reactive" and "non-reactive". Tamboli [21, pp. 91–92, 98–100] suggests the use of "red teams" (ethical hacker teams) who challenge AI products and produce problematic inputs of natural or malicious type. For any "residual risks", after all proactive measures were taken, the accountability among stakeholders should be clarified and options like utilizing "AI insurances" taken into account [21, pp. 105–106, 108–111]. Von Twickel et al. [5, p. 5] propose to "transfer the concepts and methods from the classical IT domain to the AI domain". In the context of risk management, they point to the "IEC[4] 61508 functional safety standard [life cycle]". It is a technical framework for "electrical, electronic and programmable electronic devices (E/E/PE)", which "sets requirements for the avoidance and [damage] control of systematic faults" [22, pp. 7–9].

Before working on the ML algorithm itself, as with any well managed project, it is recommended to align on *Methodology*. Groza and Marian [23, pp. 4–5] suggest adapting the "CRISP-DM" concept. It was drafted in 1999 for the area of data mining and contains the following six process steps [24, p. 7]:

1) Business understanding
2) Data understanding
3) Data preparation
4) Modeling
5) Evaluation
6) Deployment

In the business understanding phase, the focus is on identifying the correct variables/features representing the business model. In step two the correlation among features as well as the data distribution is examined. Then the focus is set on standardization and harmonization of the data. In step four modeling starts, always having Occam's razor principle, as well as GDPR's "right to explanation" in mind (details see later here and in section III-B). In the evaluation step, the ML algorithm is checked against acceptance criteria metrics as well as for occurrence of bias. During the deployment, questions of retraining and updating the model need to be discussed, to comply with the "learning" aspect of the ML algorithm.

Another concept is the "five-step-cAI"[5] life cycle from von Twickel, Samek and Fliehe [5, p. 20]. It consists of five steps (phases):

1) Planning
2) Data
3) Training
4) Evaluation
5) Operation

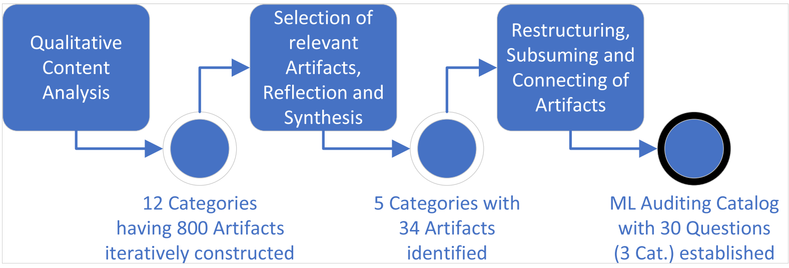
It is possible to repeatedly jump back- and forward between phases. In the planning phase, the developer sets the boundary conditions.[6] This leads to the AI model characteristics like model family, dataset, algorithm and (hyper-)parameters. In the next three phases, the mostly experience driven data processing and design work takes place. The developer needs to closely supervise the training, test intermediate results and adjust (hyper-)parameters. Once the agreed development

[3] Here "artifact" is used to describe a virtual product from an author of a scientific article. Such a product is generated content of specific type (e.g., idea description, conclusion, figure, table or code). It is often the result of a thought process steming from synthesis or analysis of a topic.

[4] International Electrotechnical Commission.

[5] The term "connectionist" refers to the connected logical units of a neural network. Those try to emulate the network of neurons by processing activation signals from many different "input wires" to one output, being organized in many different layers [13, p. 42]. Another school of thinking is "sAI", which works with "internal symbol[s] . . . that represent . . . an external reality through association, convention or resemblance" [25].

[6] The local context of an ML use case determines requirements, for example, in terms of IT-security, stability and interpretability. Given the available ML technology at hand, "boundary conditions" are defined [5, p. 22].


Qualitative Content Analysis
Selection of relevant Artifacts, Reflection and Synthesis
Restructuring, Subsuming and Connecting of Artifacts
12 Categories having 800 Artifacts iteratively constructed
5 Categories with 34 Artifacts identified
ML Auditing Catalog with 30 Questions (3 Cat.) established

Source: Authors

FIGURE 1. **Main process steps of the method.**

criteria is met (e.g., performance), the model is moved into operation.

The last methodological concept presented here comes from FDA [26], [27] dealing with "AI/ML as Software as Medical Device (SaMD)". They bring "retraining" and "learning" of an ML algorithm into focus. Questions like "how developers can automatically deploy an updated version" are discussed. The FDA's goal is to establish good machine learning practices (GMLP). They do pre-market reviews, safety and performance monitoring and publish reports. There are two core elements given: "SaMD pre-specifications (SPS)" and "algorithm change protocol (ACP)." The first addresses "what" an algorithm becomes as it learns (with new data or updates). The ACP deals with methods to control risks to SaMD's modifications (include GMLP, good data governance, retraining description or performance evaluation update procedure).

Before going into concrete project activities, it necessary to be aware of relevant *Legal Acts/Policies and Standards*. From the ones identified in literature, which can be seen in table 2, two selected legal acts/policies are briefly mentioned.

The EU AI Act aims at specifying an audit process including conformity assessments in areas like "data and data governance, documentation and recording keeping, transparency and provision of information to users, human oversight, robustness, accuracy and security" [38]. The legislative draft mandates for "unacceptable" and "high risk" applications to describe the type of information AI producers have to provide, the form of the information and the addressees of the ML algorithm in order to allow a pre-market assessment [12, p. 4]. In June 2023 the European Parliament has found consensus in a negotiating position that is being presented to the EU member countries in the EU council.

The EU General Data Protection Regulation (GDPR) has transparency as the overarching principle, which shall be achieved by transparent data processing, explaining what data is being processed and providing access to it. It also demands explanations of decisions and requires an informed consent before processing any data [12, p. 5]. According to Chatila et al. [14, p. 22], the most relevant point regarding AI auditing is that it dictates the right to explanation in Recital 71 in combination with Article 22: "obtain an explanation of the decision." Analogously, an existence of automated decision making, as well as logic and consequences of data processing, needs to be made clear to the user [14, p. 22].

There are tools and methods of *ML Documentation* that can help fulfilling those legal requirements. First, there are "data sheets" that provide information about data collection and data properties in a tabular summarized overview [39, p. 283]. They strongly remind of descriptive statistics patient cohort tables often used in life sciences. Then there are "AI model cards" that provide characteristics capturing the model's effectiveness [14, p. 21]. "Model fact labels" are similar, but include additional metrics [39, p. 283]. Furthermore exist "AI fact sheets" that are supposed to standardize capturing and representation of facts about the whole AI model [14, p. 21]. Lastly, Chatila et al. [14, p. 21] introduced the term "Care Labels" providing guidance how to use and "treat" the algorithm.

Of course, similar as with reviewing a company's bookkeeping on a yearly basis, it is appropriate do also implement an *Audit Process* for the ML project implementation/operation. How does such a process typically look like? SAI of Finland et al. [9, p. 6] suggest those steps:

1) Documentation review
2) Code review
3) Reproduction of model training, testing, performance measures and perturbation tests
4) Alternative models development

The "SMACTR auditing method" from Google stands for [23, pp. 3–4]:

- *S*coping
- *M*apping
- *A*rtifact *C*ollection
- *T*esting
- *R*eflection

According to Groza and Marian [23, pp. 3–4], scoping looks at the motives and intended impact, as well as concepts for development. Mapping examines the data and how decision are made. Artifact collection refers to documentation (e.g., in the form of model cards and data sheets). In the testing step, the developer engages thoroughly with the artifact in order to test for contradictions, biases or breakdowns. Lastly, during reflection, a risk analysis and mitigation strategy is set up and questions discussed (e.g., how to deal with inconsistencies among subgroups).

TABLE 2. **AI auditing related legal acts/policies or standards.**

Legal Acts/Policies	
Name/Description	**Source**
UN Human rights law (UNHR)	Chatila et al. [14, p. 28]
EU AI Act	Kiseleva, Kotzinos and de Hert [12] and Becker et al. [28]
EU General Data Protection Regulation (GDPR)	Kiseleva, Kotzinos and de Hert [12] and Chatila et al. [14]
U.S. Medical Device Safety Framework	FDA [26, 27]
(Technical) Standards	
Name/Description	**Source**
International: ISO, IEC, ITU	ISO [29], IEC [30] and ITU [31]
European: CEN, CENELEC, ETSI	CEN-CENELEC [32] and ETSI [33]
German Standardization Roadmap for AI	Wahlster and Winterhalter [34]
Assessment Standard for AI (German Public Auditors)	IDW [35]
COBIT	ISACA [36]
COSE ERM	COSO [37]

Source: Authors

It is important, following Chatila et al. [14, p. 30], to "[include] external oversight by independent, competent and properly resourced regulatory authorities with appropriate powers of investigation and enforcement." One aspect that has gotten more and more attention recently is the consideration of *Ethical AI Design*. A synonym is "ethics by design", highlighting that ethical concepts are not designed around an ML model, rather that the ML model is designed around ethical concepts [40, p. 2]. Making sure that human rights are not violated is the essence of those concepts.

Transparency, especially for ML algorithms applied in medicine, health care or life science is of great importance when advocating the use of ML products. Kiseleva et al. [12, p. 8] mention three central functions: accountability, ensuring safety/quality and allowing to make informed decisions. The first is geared towards the vendor of the ML algorithm, making sure the physician acquires all necessary information about a "black box model" (see section III-B). The second implies continuous testing, debugging and auditing due to the high stakes in medicine. Only if the vendor provides enough information, physician and patient are able to make a joint decision about the use of the ML algorithm in the current medical case, considering the patient's state, risk and potential outcome.

Lo [41] argues in his theory about the *Paradoxical Transparency of XAI* that with the increasing usage of deep neural networks, the asymmetry of knowledge between developer and user diminishes [41, pp. 6–9]. The "code around the black box" would be relatively easy to write and to audit [41, pp. 9–12]. The focus should be set on increasing the transparency in the pre-modelling stage by auditing the data collection, cleansing and governance of ML algorithm producers, as well as in the post-modelling phase by documented deployment and monitoring [41, pp. 9–12]. Lo [41, pp. 9–12] mentions that big companies would be "OK" with external audits of that kind (Facebook, Apple, Amazon, Netflix, Google).

For AAI as with any classical IT software program, the concept of *Verification* and *Validation* is important. The first asks if "specifications of systems [are satisfied]" and "internal structural correctness of systems" is given, and the latter "compares the system to the needs of stakeholders" [42, p. 18]. Verification and validation have a long history in classical software engineering. However, according to von Twickel et al. [5, p. 8], when designing AI algorithms, only for a controlled hypothesis space and adequately understood structures, a formal verification is possible.

As mentioned in section I, especially when human lives depend upon AI (supported) decisions, *Trust* plays a crucial role. In this context, Dengel et al. [10, p. 100] argues that "trustworthiness …is more important than accuracy." For human agents, Rempel et al. [43, pp. 96–97]'s model of trust consists of "predictability", "dependability" and "faith". Larsen [44] tries to extend this model to AI agents by adding "consistency", "utility" and "understanding" and bringing those six components in a circular relationship.

Trusted AI, as being described by Knowles and Richards [45, p. 2], contains the idea of establishing "AI-as-an-institution." They explain that by using elements of signification (symbols of trust), legitimation (norms and values) and domination (allocative and authoritative resources) a society of systemic trust, which properly uses AI technology, can be established. An example of an "element of signification" is a public repository containing good cases of AI applications that successfully underwent an *AI Certification* [45, p. 7]. Such a certification process would make use of the three pillars of an *AI Assurance Case*, which consists of "objectives & constraints", "argument" and "evidence" [28, pp'. 33–34].

Trust is also captured by the "Assessment List of Trustworthy AI (ALTAI)", which was developed by the High Level Expert Group on AI (HLEG) of the EU as part of their "ethics guidelines" [46, p. 3]. There, trustworthy AI is described by seven requirements, for example, "1. human agency and oversight", "4. transparency" or "7. accountability." For each requirement, the European Commission and Directorate-General for Communications Networks, Content and Technology [46, pp. 7–22] provides a series of questions for self-assessment of developers, users or other third-party persons involved with the AI product (implementation).

When a credit application was rejected by an AI agent, often questions about *Fairness* and *Bias* arise. The first can be characterized by "distributive", "procedural" and "interactional justice", as well as "personal principles of fairness" [47, pp. 39–40]. For the second, light is being shed on differential prediction (e.g., improper treatment of minority groups) or intentional discrimination (e.g., no proper representation of minority groups) [47, pp. 40–42].

B. DATA & ALGORITHM DESIGN

Switching to technical aspects, it is vital to first look on the "driver of growth and change [of this century]", *Data* itself [48]. The *Assumptions* made when capturing data that is used to train, test or validate an ML model are very important. The encoded occurences of values between feature and target variables, their correlation and other statistical properties are created by an underlying data generation process (DGP) that should be well understood. Individual sample records should always come from that same DGP, are independent of each other and have equally randomly distributed errors.[7] If the DGP is not well understood, there is a chance that "unknown unknowns" in the form of mediators or confounders are left out in the ML model design phase. Following Breiman [49, p. 204], the idea is to include all those "unmeasured variables" in the "model box" (the ML algorithm) in order to "emulate[] nature's box" (the DGP under scrutiny). This exercise is not a one way path, meaning that after having acquired the model's accuracy (see section III-C), and thus a measure of how well (or not) the DGP is captured by the model, it is common to work again on the data assumptions. Otherwise, when not doing so, the model's performance will be poor or biased.

Speaking about bias, it is also very critical that data taken for training, test or validation does represent relevant characteristics of the population in scope for the model. Clarke [16, p. 428] speaks here from the "inferencing process's applicability to the particular problem."

A prerequisite for any successful ML product design is well structured data in a high quality. Following Loshin [50, pp. 89–93], relevant *Data Quality Dimensions* in this paper's context are: "accuracy", "consistency", "completeness" and "currency". Accuracy describes if the information adequately represents the underlying real world object and consistency indicates if the piece of information is not contradicted by another source. Completeness deals with missing and implausible values and currency asks if the data collection time represents the latest possible state.

The "art of data preparation" that makes use of the four data quality dimensions, is very important before starting with the model design itself. It is dangerous to have uncleansed and not consolidated data, especially with missing values and coming from multiple sources [16, p. 428]. Each source has to be well understood, so that it is clear what data type is contained in a field, what the scale (nominal, ordinal, interval or ratio) and the reference range is.

After the origin and assumptions of the data have been clarified, it can be used to train an ML algorithm. Before doing so, it is imperative to take a look on the most important *ML Algorithm Properties*. The terms "causality" and "correlation" need to be clearly distinguished. The first describes the relationship of a cause to an effect, which ideally can be naturally derived. Imagine a person standing on rollerblades pushing against a wall. The person exerts a force to the direction of the wall ("actio") and as a result experiences an oppositely directed force of equal strength ("reactio"). This is a typical example of a causal relationship that can be easily determined by a third-party observer.

Now, when dealing with a complex machine learning model that predicts the creditworthiness of a customer, the concept of causality is difficult to proof. There is no formula following the laws of physics stating that people of a certain age, income and education will not be able to pay a credit back. Instead, within the domain of stochastic statistics, the concept of "correlation" can be used. It describes how strongly one variable is related to another variable (influenced by or dependent on). It can be among features (X) or between features and outcome (Y). A typical measure is the "Pearson's Correlation Coefficient".

A value of 1 indicates a perfect relation between X and Y, 0 indicates no relation and -1 a perfect inverse relation. The normalization allows a comparison of the relationship among different variables (features and target) with different scales. It is important to note that a causal relationship always has perfect correlation, but having a correlation coefficient of 1 does not automatically proof a causal relationship.

Looking at ML design options, there are two different *Types of ML Algorithms*: "black box" models and "white box" models. The first describes an entity that conceals its inner workings to the user or even the developer. Data goes in and a result comes out, without giving the possibility to the user to comprehend the mechanics or rules of decision making. The most prominent example of this ML model type is a neural network, which is often used for natural language processing (NLP) or image recognition.

Traditionally, statisticians have been working with parametric models to directly design a formal relationship between input and output [49, p. 202]. Those "white box models" allow the user, within reasonable effort, to get an understanding of the inner workings. The user would be able to break down the decision points that determine the result. Typical examples of this type are generalized linear models (GLM) like linear or logistic regression, as well as decision trees.

In his well cited "The Two Cultures", Breiman [49, pp. 209–214] argues that black box models are superior to white box models in terms of predictive power, as well as in terms of understanding the DGP (the internal mechanics). He uses examples to show that a random forest can achieve a better prediction (lower misclassification rate) than logistic

[7] This is also known as "independent and identically distributed random variables (i.i.d.)" in literature.

regression while correctly identifying the variables majorly influencing the decision (feature importance). His main point is that statisticians should work together with computer scientists to solve a problem using both black box and white box models rather than just assuming being able to directly derive a DGP by establishing a parametric model.

However, according to "Occam's razor principle", given a black box and a white box model have the same accuracy (performance) and both are able to sufficiently explain a hypothesis, the less complex white box model should be preferred [11, pp. 1–2].

At the stage of ML algorithm development, before going into feature selection/engineering and model type selection, the *Specifications* need to be clear among the stakeholders. This includes a correctly formulated hypothesis (narrowly described problem situation with expected algorithm behavior). Once the algorithm starts generating results, statistical testing should be done. Only then the question can be answered, if the right solution was built and if it was properly designed. Additionally, when writing the specifications, acceptance criteria and metrics of success have to be defined. In terms of product safety, a "kill switch" should be integrated, allowing a human operator to shut down the ML algorithm in case of misbehavior (e.g., autopilot of a plane). The implementation of the ML algorithm is usually a composition of classical IT (cIT) components and AI components.

C. ASSESSMENT METRICS

The next aspect that plays an important role for an AI auditing criteria catalog are assessment metrics. In principle, the nature of those is either *Qualitative* or *Quantitative*. The first utilizes descriptions or narratives, which are often given from a rather subjective perspective. That is in contrast to the latter, where an objective measurement is possible, often with subsequent mathematical calculations.

For the qualitative area, adequate model assumptions are very important. Those assumptions are constantly being made during the modeling process and might state how to deal with missing data, the engineering of model features, the model type itself and the chosen hyperparameters. Those are all made choices that should be justified. Methods like sensitivity analysis or model validation can be used to "assess the appropriateness of a particular model specification and to appreciate the strength of the conclusions being drawn from such a model" [51, p. 263]. In literature, the dimension "auditability" of an AI product is closely related with "audit trials", which is a "traceable log" that "cover[s] all steps of the AI development process" [7, p. 24]. Only if such a log exists, an appropriate level of human oversight could be executed, which is also required in the proposed EU AI Act in article 14 [38]. Important aspects of transparency and trust that were already described in section III-A, can be used to perform a textual assessment of a given AI model and use case. Lastly for fairness, Becker et al. [28, p. 36] suggest to use Acceptance Test-Driven Development (ATDD) to formulate fairness in a way that is understandable to business and the developers.

TABLE 3. Confusion matrix for binary credit worthiness classifier.

		Clerk's decision (Truth)		
		Accepted (1)	Rejected (0)	Total
ML Prediction	Accepted (1)	TP	FP	$TP + FP$
	Rejected (0)	FN	TN	$FN + TN$
	Total	$TP + FN$	$FP + TN$	N

Source: Authors

Looking at the quantitative area, usually first the *Statistical Properties* of the training data set are assessed. Bhaumik et al. [52, p. 4] provide three measures: Variance Inflation Factor (VIF) for the degree of collinearity, Shapiro-Wilk Test (SWT) to examine normality and Breusch-Pagan Test (BPT) to check for heteroskedasticity.

Coming to *Performance Indicators*, the "Area Under the Curve (AUC)" or "F1-Score" using "Receiver Operating Characteristics (ROC)" is being commonly utilized. Using the example of assessing the credit worthiness of a bank customer, the first step in measuring the performance of the used ML product is to create a "Confusion Matrix", as can be seen in table 3.

The ML product is used on N credit applications and for every run the manual decision of the clerk is compared to the automated prediction of the algorithm. In case the algorithm accepts an application that the clerk also accepted, it is a "true positive (TP)." In case only the algorithm rejects the application it is a "false negative (FN)." If both the clerk and the algorithm reject the application, it is a "true negative (TN)" and in case only the algorithm accepts the application it is a "false positive (FP)." Afterwards confusion metrics like sensitivity or fallout can be calculated.

A "ROC graph" is then generated by plotting the sensitivity on the Y-axis and the fallout on the X-axis. A perfect classifier would be on the top left corner (0,1), having no false negatives (FN) or false positives (FP) and a random classifier would be on a diagonal from origin (0,0) to top right (1,1) [53, pp. 862-863]. The AUC, measuring the area (integral) underneath the plotted line, would be maximal in the first case and minimal in the second case.

A widely used metric for accuracy is the "F1-Score", where a value of < 0.5 is considered a bad classifier. Alternatively the "Matthews Correlation Coefficient (MCC)" can be used, where values > 0.7 indicate a strong correlation and thus a good classifier.

Robustness against naturally occurring perturbation is another useful indicator for ML model assessment. Here Bhaumik and Dey [54, pp. 5–6] suggest "Total Sobol's Variance Ratio (TSVR)" and "Cosine Similarity Vector Pairs (CSVP)".

When looking at metrics to evaluate potential discrimination, according to Bhaumik et al. [52, pp. 5–6], "Statistical Parity (SP)" or "Disparate Impact (DI)" can be used. As an illustration, the DI is used in the U.S. labor law to investigate potential discrimination of women towards

TABLE 4. Overview of AI auditing related tools or methods.

Conceptual/Theoretic	
Name/Description	**Source**
"AI Guardians"	Dengel et al. [10] and Etzioni and Etzioni [56]
"50 principles for responsible AI application"	Clarke [20, pp. 414–417]
"Regulatory framework dealing with threats"	Clarke [57, pp. 406–407]
"Audit framework to technically assess [MLM and binary classifiers]"	Bhaumik, Dey and Kayal [52] and Bhaumik and Dey [54]
"Framework to evaluate fairness and bias"	Landers and Behrend [47]
7 suggested dimensions for AI audit	Groza and Marian [23, p. 3]
Google SMACTR auditing method	Groza and Marian [23, pp. 3–4]
"Generalized Audit Framework for AI (GAFAI)"	Markert, Langer and Danos [58]
Hands-On/Practical	
Name/Description	**Source**
Amazon's face recognition tool audit	Dengel et al. [10, p. 104]
Allan Institute "WhyLabs"	Dengel et al. [10, p. 104]
Checklist to audit your AI systems	Yakobovitch [59, p. 104]
"FG-AI4H Online Benchmarking Platform"	Wiegand et al. [60, p. 7]
AI4H Developer Questionnaire	Oala et al. [39, p. 283]
"Aequitas" Toolkit	Oala et al. [39] and Saleiro et al. [61]
Auditability Checklist (Excel Helper Tool)	SAI of Finland et al. [9, p. 6]
"Credo AI Lens"	Rawat [62]

Source: Authors

men in employment settings [47, p. 40]. There, an industry convention has formed saying it should be $\geq$ 0.8, after including the disturbance term, otherwise one might suspect discrimination [55].

D. TOOLS & METHODS

The assessed literature provides an abundance of tools or methods related to AI auditing or XAI. Those tools could be used complementary when working with the ML auditing core criteria catalog presented in section III-E. *Audit related Tools or Methods* can be further subgrouped into being more conceptual/theoretic and more hands-on/practical as can be seen in table 4. In the next paragraphs, a few exemplary tools and methods are chosen and further explained.

The framework from Bhaumik et al. [52], and Bhaumik and Dey [54] was already presented in section III-C when assessing the statistical properties. Clarke [20, pp. 414–417] groups his "50 principles for responsible AI application" around ten themes (e.g., "2. complement humans", "7. embed quality assurance" or "9. ensure accountability for obligations").

Looking at the hands-on tools, there are audit checklists like the one from Yakobovitch [59] or the one from SAI of Finland [9], which look more into the auditability of the solution itself. The first takes care of aspects like "AI business outcome", "data source" or "privacy of data".

The second comprises an audit catalog consisting of the six "CRISP-DM" steps, which were already presented in section III-A, but adding the seventh step "operation of the model and performance in production" [9, p. 6].

Additionally, there exist tool kits of different scope. "Credo AI Lens" looks at responsible AI and is supposed to help companies manage risks to achieve fair, compliant and auditable ML products [62]. It provides a technical evaluation of the data set and ML model, as well as guides developers with critical questions during the development process.

XAI Tools or Methods are usually grouped by the ML algorithm development stage when they are applied. An overview is provided in table 5.

There are also "meta frameworks" that contain collections of XAI tools for all development stages. One example of this is the "Censius Platform" or the "Tag Tool & Deep Learning Sandbox". The first contains 23 tools for MLOps,[8] as well as AI observability, monitoring and explainability tools and contains an enterprise AI audit checklist [66]. The latter provides a web interface for testing purposes, mainly focusing on capabilities and performance metrics [10, p. 97].

Next to those company products, also not-for-profit projects like "OpenML" have been established, where ML algorithms and datasets can be shared between researchers, developers and users in order to "to work more effectively and collaborate on a global scale" [67, p. 4].

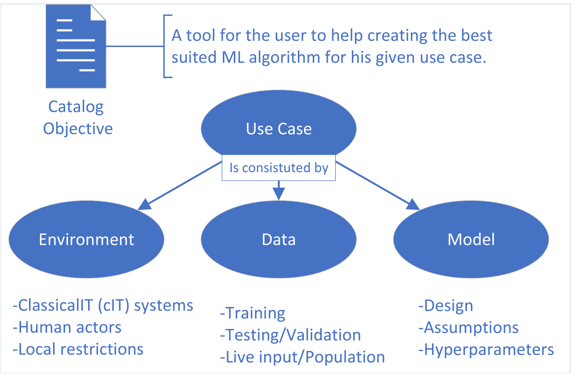
It becomes clear that before modeling starts, the "method" is rather to have a good strategy how to achieve a representative, high quality training data set. The "tools" are mainly descriptive statistics that look at correlation and data distributions. During modeling, the focus is on the already described "Occam's razor" principle (see section III-B), meaning to rather prefer a "white box" model over a "black box" model, given the relevant metrics are identical. Most tools look at the post-modeling stage. Well cited are "LIME" and "SHAP". The first, model agnostic method, fits linear regressions to "simulate" local behavior for concrete data input, as well as providing global explanation [11]. The second uses Shapley additive explanations to figure out the average marginal contribution of features [11].

[8] "ML Ops is a set of practices that combines Machine Learning, DevOps and Data Engineering, which aims to deploy and maintain ML systems in production reliably and efficiently" [65].

TABLE 5. Overview of explainable AI (XAI) related tools or methods.

Pre-modeling	
Name/Description	**Source**
Data cleansing: Improve quality of incoming data stream	Liu et al. [63, pp. 11–12]
Data description, analysis, exploration and feature engineering	Benchekroun et al. [11, pp. 5–6]
DeepLift: data pertubation to understand inner workings	Benchekroun et al. [11, p. 5]
ProtoDash: select protoype of underlying data distribution	Liu et al. [63, p. 17]
Explainability quantification: amount/interaction of features vs. explanatory value	Liu et al. [63, pp. 23–25]
During modeling	
Name/Description	**Source**
"Make reasoning of model more understandable to user"	Liu et al. [63, p. 12]
Inherently explainable models (white boxes), like linear regression or decision trees	Benchekroun et al. [11, p. 6]
Post-modeling	
Name/Description	**Source**
"How result and or prediction of model came to be"	Liu et al. [63, pp. 12–13]
TSViz: "How decision was made?"	Dengel et al. [10, pp. 100–101]
TSXplain: "Why decision as made?"	Dengel et al. [10, pp. 100–101]
LIME and SHAP: Approximation methods building a proxy	Benchekroun et al. [11, p. 6]
AIX360, Skater, "What-If-Tool", EL15, Activation Atlases: Popular XAI frameworks	Chinu and Bansal [64, p. 20]

Source: Authors



Source: Authors

FIGURE 2. Objective of the ML auditing core criteria catalog.

E. ML AUDITING CRITERIA CATALOG

After having identified and described artifacts being relevant for auditing machine learning algorithms, the next step is to subsume and connect the artifacts into an ML auditing core criteria catalog. As a methodology, the authors follow the suggestion from Gawande [68, p. 26] who lists requirements and procedures of creating a good checklist. The objective of the catalog is given in figure 2.

Having this in mind, the resulting ML auditing core criteria catalog consists of 30 questions that are grouped around the categories: "Conceptual Basics", "Data & Algorithm Design" and "Assessment Metrics". The questions are thought to guide the ML development team.

1) CONCEPTUAL BASICS

a: AI OPPORTUNITIES VS. AI RISKS

- ✓ Is the expected *benefit* · *benefitProbability* of a successful ML use case implementation greater than the *damage* · *damageProbability* in case of failure?
- ✓ Do you expect a productivity gain, improved quality or a new functionality compared to the current manual/non-ML process?

b: RISK MANAGEMENT

- ✓ Are the roles and responsibilities (RACI[9]) and liabilities before, during and after the implementation clearly defined?
- ✓ 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?

c: METHODOLOGY

- ✓ Have you aligned and agreed on the methodology with all project stakeholders (e.g., for implementation CRISP-DM and internal audit SMACTR)?
- ✓ Are the implications in case the ML use case falls in the "high risk" category of the EU AI Act understood?
- ✓ Do you plan to make use of Data Sheets to describe the data collection process as well as the data properties?
- ✓ Do you plan to create AI Model Cards/AI Fact Sheets to describe the model characteristics?
- ✓ Do you plan to prepare AI Care Labels to instruct internal stakeholders how to use and "treat" the algorithm?

d: AUDIT PROCESS

- ✓ Have you established an internal advisory committee consisting of senior IT governance specialists and business/medical specialists who critically accompany

[9] RACI says that when working in teams it needs to be clear who is **R**esponsible for a given task, who is **A**ccountable especially if something goes wrong, who needs to be **C**onsulted for advice and who must be **I**nformed about the progress.

the implementation (e.g., watch for sufficient documentation and methodology adherence)?

- ✓ Do you ensure the ML implementation is not violating ethical concepts (“ethics by design” is considered)?
- ✓ Do you have protocols in place that allow independent, external auditors to critically review the ML use case implementation?

e: QUALITY ASSURANCE (QA)

- ✓ Did you perform a verification of the ML output behavior using a set of expected, representative inputs of the productive usage?
- ✓ Did you perform a validation whether the project’s specification and stakeholders’ needs are met?
- ✓ Do you think the ML model would pass an external AI Certification/AI Assurance case fulfilling the six components of trust: predictability, dependability, faith, consistency, utility and understanding?
- ✓ Given inputs from different test users, does the ML model adhere to the principles of distributive, procedural and interactional justice?
- ✓ Given inputs from different test users, does the ML model avoid differential prediction and intentional discrimination?

2) DATA & ALGORITHM DESIGN

a: DATA PROPERTIES

- ✓ Is the data generation process (DGP) of the training, testing and validation data set sufficiently known? Could there be unknown confounders or mediator variables influencing the observed data?
- ✓ Does the training data capture relevant characteristics of the population in scope for the ML use case?
- ✓ Are the required data quality dimensions (e.g., accuracy, consistency, completeness and currency) well understood and taken care of?
- ✓ Are the procedures necessary for data cleansing and consolidation known, and is the understanding of data scales and references ranges given?

b: ALGORITHM DESIGN

- ✓ Is the difference between causality and correlation known? In the absence of known counterfactuals for each individual, population samples can only give associations with a certain strength (e.g., given by the Pearson’s correlation coefficient).
- ✓ Did you apply Occam’s Razor principle for the model type selection? Meaning in case a black box model (e.g., DNN, NLP) is to be used, does it provide substantial benefit (e.g., accuracy) over a white box model (e.g., logistic regression, decision tree)?
- ✓ Did you establish a correct ML use case hypothesis with concrete problem description and expected behavior (acceptance criteria, metrics, statistical testing results)?

3) ASSESSMENT METRICS

a: QUALITATIVE ASSESSMENT

- ✓ Are the model assumptions (e.g., how to deal with missing data, model type, hyperparameters) transparently described?
- ✓ Did you establish a traceable log of those model assumptions/testing results being used during the whole development process?
- ✓ Did you discuss with all stakeholders the strength of conclusions that can be drawn with the current model assumptions (and make sure the conclusions are appropriate)?

b: QUANTITATIVE ASSESSMENT

- ✓ Did you determine the statistical properties of the training, testing and validation data set? For example, by using Variance Inflation Factor (VIF), Shapiro-Wilk Test (SWT) and Breusch-Pagan Test (BPT)?
- ✓ Did you conduct extensive performance testing according to the agreed metrics? For example, using Receiver Operating Characteristics (ROC): creating the confusion matrix and calculating the F1-Score, Matthews Correlation Coefficient (MCC) or Area Under the Curve (AUC)?
- ✓ Did you assess the resistance of the ML model’s output behavior to natural perturbation, for example, using Total Sobol’s Variance Ratio (TSVR) or Cosine Similarity Vector Pairs (CSVP)?

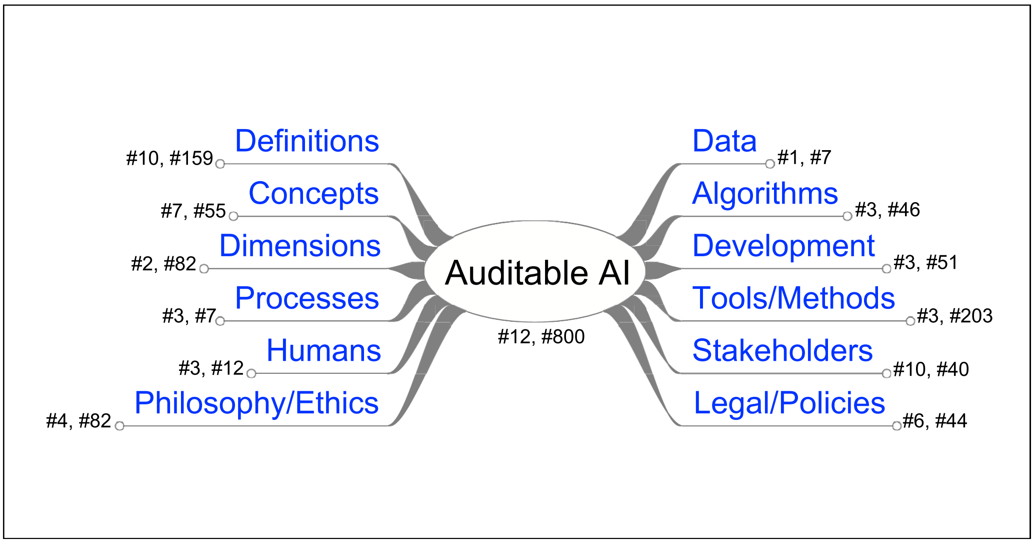
IV. DISCUSSION

The goal of our paper was twofold, first we wanted to explore and highlight the most important auditable AI related artifacts existing in literature. Second, we wanted to provide valuable input for auditing ML algorithms in form of a core criteria catalog.

Our main outcome is a 30-question, consensus-based ML auditing core criteria catalog that is intended as a starting point for the development of evaluation strategies by specific stakeholders.

The catalog is organized in the categories *Conceptual Basics*, *Data & Algorithm Design* and *Assessment Metrics*. Section III (results) presents referred artifacts in the same logical sequence as they appear in the catalog, to simplify its usage.

We grouped artifacts dealing with AI opportunities/risks, risk management, methodology, audit process and quality assurance (QA) under the category *Conceptual Basics* (section III-A). This term was chosen, because our recommendations mostly deal with business, process and people related topics. Those need to be taken care of before going into the *Data & Algorithm Design* (section III-B). There, the main artifacts describe data properties and algorithm design. In *Assessment Metrics* (section III-C), we distinguished qualitative assessments from quantitative assessments and provided detailed instructions how latter can be calculated.



Source: Authors

FIGURE 3. Iteratively created categories of the "Auditable AI" qualitative content analysis (top level view). The first # behind each node gives the number of children, and the second # the sum of descendants. The total number of artifacts is 800.

The contents of the catalog are the result of the "auditable AI" qualitative content analysis (see section II). We utilized many iterations to sort out irrelevant artifacts and synthesize relevant ones into new ideas and categories. We believe that this catalog 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 a better, well-perceived and accepted product for patients/ customers.

With section III-D (tools & methods), even though not contained in the catalog, our intention is to provide a summary of programs and little helpers that could be used aside of the ML auditing criteria catalog. We discovered those during the literature study.

We needed to decide on a trade-off between the breadth and the depth for our ML auditing criteria catalog. Also, for practical reasons, it was demanded to limit the scope and time period of the qualitative content analysis in the first place. Therefore, it is possible that there are AAI aspects missing and/or have not been dealt with in the necessary detail.

Those limitations can be overcome by utilizing the proposed catalog in practice on real use cases or hypothetical use cases with artificially generated data. Doing so, potential gaps could be exposed and the catalog further improved. Thus, this paper is seen as a starting point towards the development of a framework, where essential technical components can be specified.

V. CONCLUSION

Our motivation for this paper was the existing lack of commonly agreed procedures and content for auditing ML algorithms, which we point out in different sources. We opened with a Gartner study showing a lack of AI adoption in the healthcare sector. In order to help speeding up this adoption, we provided a core criteria catalog as a starting point for further developments.

As an additional valuable input for the new and rapidly evolving field of "Auditable AI (AAI)", we included a definition of the term and contrasted it to "eXplainable AI" and "Interpretability".

The 30 questions of the core criteria catalog are contained within the categories *Conceptual Basics*, *Data & Algorithm Design* and *Assessment Metrics*. The level of detail is chosen purposely rather compact, to focus on the most relevant artifacts, being valid for most stakeholders who want to implement an ML algorithm.

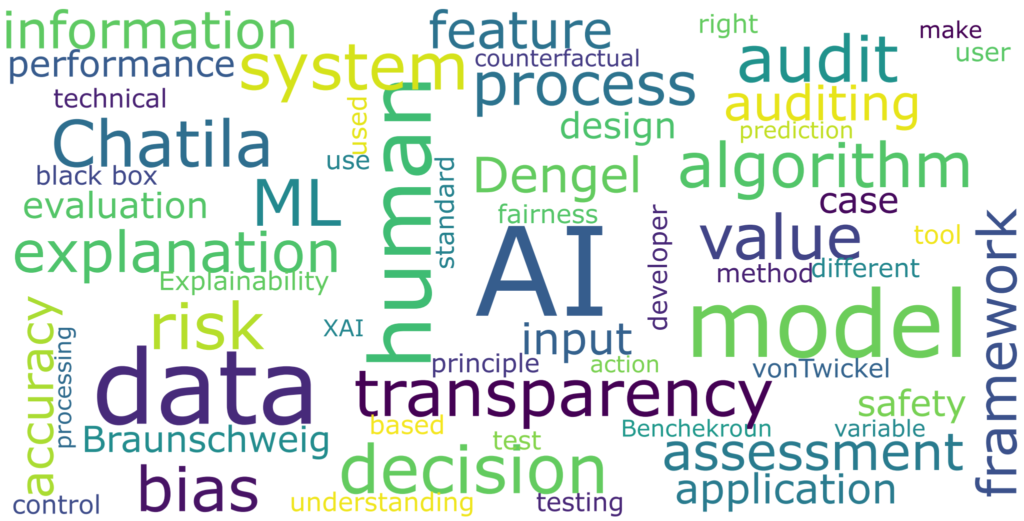
APPENDIX A
DIFFERENT REPRESENTATIONS OF AAI MINDMAP AND CRITERIA CATALOG

The most relevant artifacts are explored in an AAI mind map that is shown in figure 3 (top level view). It consists of 12 main categories having a total number of 800 artifacts. The category *Tools/Methods* contains most artifacts (203), followed by *Definitions* (159) and *Dimensions* (82) as well as *Philosophy/Ethics* (82).

Figure 4 is a representation of the AAI mind map as a word cloud.

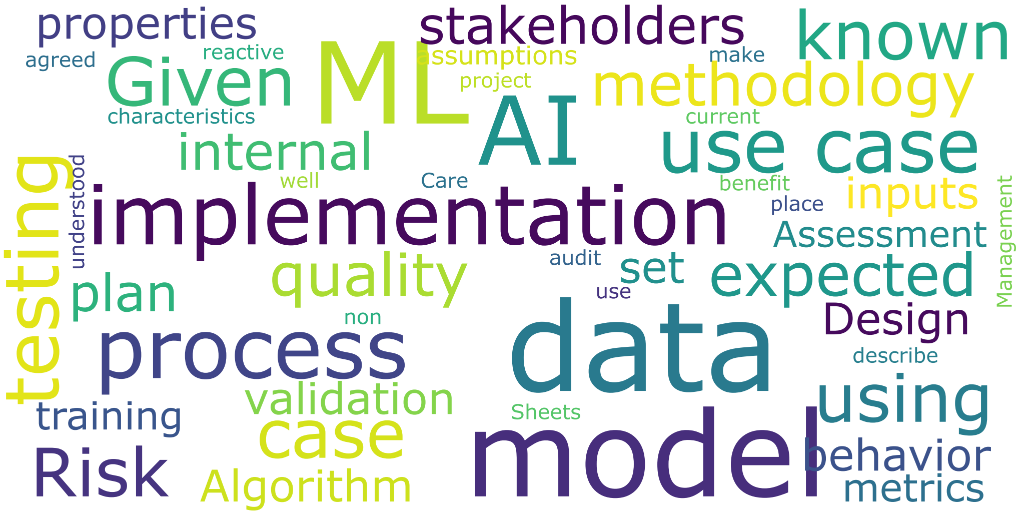
In figure 5 the ML auditing core criteria catalog is displayed as well as a world cloud with the same parameters as the AAI mind map.

The complete mind map file, as well as auxiliary files of the core criteria auditing catalog can be accessed via OSF at: https://osf.io/tdr3p.



Source: Authors

FIGURE 4. Word cloud of the "Auditable AI" mind map file. Each text attribute of the 800 artifacts (nodes) is split into words and then displayed according to the calculated rank. Words with less than two characters are ignored.



Source: Authors

FIGURE 5. Word cloud of the 30-question ML auditing core criteria catalog. Each sentence is split into words and then displayed according to the calculated rank. Words with less than two characters are ignored.

APPENDIX B
FORMULAE

The "Pearson's Correlation Coefficient" is given in equation (1).

$$\rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{\mathrm{E}\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sigma_X\sigma_Y} \qquad \sigma_X\sigma_Y > 0 \qquad (1)$$

The μ_Y and μ_X are the variable means and E calculates the expected value (weighted average of $(x_i - \bar{x})(y_i - \bar{y})$). σ_X and σ_Y are the respective standard deviations. The value and direction ($-\infty \leq \mathrm{cov} \leq +\infty$) of the relationship is already determined by the covariance and then normalized by the standard deviation, leading to values $\in \mathbb{R}$ in $[-1, 1]$.

$$\mathrm{VIF}_i = \frac{1}{1-R_i^2} = -\frac{\sum_i (y_i - \bar{y})^2}{\sum_i \left(y_i - \hat{y}_i\right)^2} \qquad (2)$$

The Variance Inflation Factor (VIF) measures the degree of collinearity (linear correlation between features). It is given in equation (2), y_i is the i^{th} observed outcome value of a data set, $\hat{y}_i$ the i^{th} predicted outcome value and $\bar{y}$ the mean of all observed outcome values.

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{3}$$

The Shapiro-Wilk Test (SWT) examines normality (normally distributed values of each feature). As shown in equation (3), for the SWT of a sample with size n of a given variable x, $x_{(i)}$ is the i^{th} order statistic (smallest to largest value) and $\bar{x}$ the sample mean. $(a_1, \ldots, a_n) = \frac{m^{\mathsf{T}} V^{-1}}{C}$, whereas $m = (m_1, \ldots, m_n)^{\mathsf{T}}$ is created from the expected values of the order statistics, V is the $n \times n$ covariance matrix of the order statistics and $C = \left(m^{\mathsf{T}} V^{-1} V^{-1} m\right)^{\frac{1}{2}}$.

```
Algorithm 1 Breusch-Pagan Test (BPT)
Assumption:   Null hypothesis $H_0$: homoscedasticity is
     present (equally distributed variance of residuals)
          Alternative hypothesis $H_1$: heteroscedasticity is
     present (unequally distributed variance of residuals)
Step 1:   Fit a regression model so that $\hat{y}_i = \mathrm{f}\left(X_i, \hat{\beta}\right)$
Step 2:   Calculate the squared residuals $\hat{e}_i^2 = \left(y_i - \hat{y}_i\right)^2$
Step 3:   Fit a new regression model on those residuals $\hat{y'}_i^2 =$
     $\mathrm{f}\left(X_i, \hat{\beta}'\right)$
Step 4:   Calculate the "new" $R'^2 = 1 - \frac{\sum_i \left(\hat{y}_i - \hat{y'}_i\right)^2}{\sum_i \left(\hat{y}_i - \bar{\hat{y}}\right)^2}$
Step 5:   Perform the Chi-Square test $\mathcal{X}^2 = nR'^2$
Step 6:   Derive the $p$-value using $\mathcal{X}^2$ and DF according to $X$
Result:   if $p$-value $\leq 0.05$ then
              $H_0$ can be rejected and $H_1$ applies: heteroscedasticity is present
          else
              $H_0$ cannot be rejected: homoscedasticity is present
          end if
```

The Breusch-Pagan Test (BPT) checks for heteroskedasticity (increase in variance). The sequence of the BPT with explanation is given in algorithm 1.

Typical confusion metrics are Sensitivity $= \frac{\text{TP}}{\text{TP+FN}}$, Specificity $= \frac{\text{TN}}{\text{TN+FP}}$, Precision $= \frac{\text{TP}}{\text{TP+FP}}$, Negative Predictive Value $= \frac{\text{TN}}{\text{TN+FN}}$ or Fallout $= \frac{\text{FP}}{\text{FP+TN}}$. Each of those metrics result in values $\in \mathbb{R}$ in $[0, 1]$.

The "F1-Score" is given in equation (4). It is the harmonic mean of sensitivity and precision, being as well $\in \mathbb{R}$ in $[0, 1]$, and not of symmetric nature. That means if the rejected and accepted group members are interchanged (swapping of class labels), having everything else constant, the F1-Score will be different.

$$\begin{aligned} F_1 &= \frac{2(\text{Sensitivity} \cdot \text{Precision})}{\text{Sensitivity} + \text{Precision}} \\ &= \frac{2\text{TP}}{\text{FN} + \text{FP} + 2\text{TP}} \end{aligned} \tag{4}$$

Due to the symmetric nature of the "Matthews Correlation Coefficient (MCC)", class label swap does not have any impact. It is given in equation (5) and its values $\in \mathbb{R}$ are in $[-1, 1]$. A MCC of 0 means there is a weak correlation between the features e.g., (customer's characteristics) and the outcome e.g., (granting of credit).

$$\text{MCC} = \frac{\text{TN} \cdot \text{TP} - \text{FN} \cdot \text{FP}}{\sqrt{(\text{FN+TN})(\text{FN+TP})(\text{FP} + \text{TN})(\text{FP} + \text{TP})}} \tag{5}$$

The "Total Sobol's Variance Ratio (TSVR)" explains the total effect (incl. interactions) of an input variable X_i on the variance of the output Y and is given in equation (6).

$$S_{\text{Ti}} = 1 - \frac{\text{Var}_{X_{\sim i}}\left(\text{E}_{\text{Xi}}\left(Y \mid X_{\sim i}\right)\right)}{\text{Var}\,(Y)} \tag{6}$$

$X_{\sim i}$ denotes the set of all input variables except X_i. The larger S_{Ti}, the larger the contribution of X_i on Y's variance.

The "Cosine Similarity Vector Pairs (CSVP)" is used in the context of binary classification models in order to "find …the number of vector pairs that are very similar in values but have been predicted to be in different classes."

$$\begin{aligned} S_C\left(\vec{A}, \vec{B}\right) := \cos(\theta) &= \frac{\vec{A} \cdot \vec{B}}{\left\|\vec{A}\right\| \left\|\vec{B}\right\|} \\ &= \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \end{aligned} \tag{7}$$

The formula is given in equation (7). $\vec{A}$ might contain the input variables (features) of a customer whose credit request got approved and $\vec{B}$ might contain the input variables from a customer who got rejected. An $S_C\left(\vec{A}, \vec{B}\right)$ value close to 1 indicates that the customers are very similar and a value close to -1 indicates the customers are very different.

The Statistical Parity (SP) is shown in equation (8) and the Disparate Impact (DI) is shown in equation (9). They describe if an ML model tends to predict the positive result more often for members of a privileged group than for members of a minority group. The ML model would do so, even though the attributes or characteristics of the privileged group are not supposed to act as features in the model.

$$\begin{aligned} \text{SP} = \Pr&\left(\hat{Y} = 1 | S = 1\right) \\ &- \Pr\left(\hat{Y} = 1 | S = 0\right) \\ &\text{SP} \leq \epsilon \end{aligned} \tag{8}$$

$$\text{DI} = \frac{\Pr\left(\hat{Y} = 1 | S = 0\right)}{\Pr\left(\hat{Y} = 1 | S = 1\right)} \quad \text{DI} \geq 1 - \epsilon \tag{9}$$

$\hat{Y} = 1$ is a positive predicted outcome and $S = 1$ indicates membership in the privileged group respective $S = 0$ indicates membership in the non-privileged group.

Equation (8) can be extended to include the true outcome Y so it becomes the Equalized Odds (EO) term shown in equation (10).

$$\begin{aligned} \mathrm{EO} &= \Pr\left(\hat{Y}=1|S=1,Y=y\right) \\ &\quad - \Pr\left(\hat{Y}=1|S=0,Y=y\right) \\ &\quad y \in \{0,1\}, \mathrm{EO} \leq \epsilon \end{aligned} \tag{10}$$

This means that members of the privileged group have equal Sensitivity and Fallout as members of the non-privileged group. Ideally, the observed differences should only result in the disturbance term (being close to 0).

REFERENCES

[1] P. Mayring, "Qualitative content analysis," *FQS*, vol. 1, no. 2, Jun. 2000.

[2] M. B. Miles and A. M. Huberman, *Qualitative Data Analysis. An Expanded Sourcebook / Matthew*, 2nd ed., B. Miles and A. M. Huberman, Eds. Thousand Oaks, CA, USA: SAGE, 1994.

[3] J. Bortz and N. Döring, *Forschungsmethoden Und Evaluation Für Human- und Sozialwissenschaftler. Mit 87 Tabellen*, 4th ed. Cham, Switzerland: Springer, 2006.

[4] Sage Growth Partners and Olive, *The State of Healthcare Automation*, Baltimore, MD, USA, 2021. [Online]. Available: https://sage-growth.com/wp-content/uploads/2022/05/State_of_Healthcare_Automation_022221.pdf

[5] A. von Twickel, W. Samek, and M. Fliehe, "Towards auditable AI systems," in *Bundesamt Für Sicherheit in der Informationstechnik, Fraunhofer-Institut Für Nachrichtentechnik, Verband der TÜV*, 2021. [Online]. Available: https://www.hhi.fraunhofer.de/fileadmin/News/2021/White_Paper/20210504_Whitepaper__Towards_Auditable_AI_Systems_-_Current_status_and_future_directions__final.pdf

[6] T. Wiegand, R. Krishnamurthy, M. Kuglitsch, N. Lee, S. Pujari, M. Salathé, M. Wenzel, and S. Xu, "WHO and ITU establish benchmarking process for artificial intelligence in health," *Lancet*, vol. 394, no. 10192, pp. 9–11, Jul. 2019.

[7] M. Brundage, "Toward trustworthy AI development: Mechanisms for supporting verifiable claims," 2020, *arXiv:2004.07213*.

[8] R. V. Zicari, J. Brodersen, J. Brusseau, B. Dudder, T. Eichhorn, T. Ivanov, G. Kararigas, P. Kringen, M. McCullough, F. Moslein, N. Mushtaq, G. Roig, N. Sturtz, K. Tolle, J. J. Tithi, I. van Halem, and M. Westerlund, "Z-inspection: A process to assess trustworthy AI," *IEEE Trans. Technol. Soc.*, vol. 2, no. 2, pp. 83–97, Jun. 2021. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9380498

[9] Supreme Audit Institutions of Finland, Germany, The Netherlands, Norway and the U.K. (2020). *Auditing Machine Learning Algorithms*. [Online]. Available: https://www.auditingalgorithms.net/auditing-ml.pdf

[10] A. Dengel, O. Etzioni, N. DeCario, H. Hoos, F.-F. Li, J. Tsujii, and P. Traverso, "Next Big Challenges in Core AI Technology," in *Reflections on Artificial Intelligence for Humanity* (Lecture Notes in Computer Science), vol. 12600, B. Braunschweig and M. Ghallab, Eds. Cham, Switzerland: Springer, 2021, pp. 90–115.

[11] O. Benchekroun, A. Rahimi, Q. Zhang, and T. Kodliuk, "The need for standardized explainability," 2020, *arXiv:2010.11273*.

[12] A. Kiseleva, D. Kotzinos, and P. De Hert, "Transparency of AI in healthcare as a multilayered system of accountabilities: Between legal requirements and technical limitations," *Frontiers Artif. Intell.*, vol. 5, May 2022, Art. no. 879603.

[13] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach, Global Edition*, 4th ed. London, U.K.: Pearson, 2021.

[14] R. Chatila, V. Dignum, M. Fisher, F. Giannotti, K. Morik, S. Russell, and K. Yeung, "Trustworthy AI," in *Reflections on Artificial Intelligence for Humanity* (Lecture Notes in Computer Science), vol. 12600, B. Braunschweig and M. Ghallab, Eds. Cham, Switzerland: Springer, 2021, pp. 13–39.

[15] M. Peters, C. Godfrey, P. McInerney, Z. Munn, A. Tricco, and H. Khalil, "Chapter 11: Scoping reviews," in *JBI Manual for Evidence Synthesis*, E. Aromataris and Z. Munn, Eds. Denizli, Turkey: JBI, 2020. [Online]. Available: https://jbi-global-wiki.refined.site/space/MANUAL/4685874/Downloadable+PDF+-+current+version

[16] R. Clarke, "Why the world wants controls over artificial intelligence," *Comput. Law Secur. Rev.*, vol. 35, no. 4, pp. 423–433, Aug. 2019.

[17] B. Braunschweig and M. Ghallab, "Reflections on AI for humanity: Introduction," in *Reflections on Artificial Intelligence for humanity* (Lecture Notes in Computer Science), vol. 12600, B. Braunschweig and M. Ghallab, Eds. Cham, Switzerland: Springer, 2021, pp. 1–12.

[18] J. Guszcza, I. Rahwan, W. Bible, M. Cebrian, and V. Katyal. (2018). *Why We Need to Audit Algorithms*. [Online]. Available: https://hbr.org/2018/11/why-we-need-to-audit-algorithms

[19] L. A. Celi, M. S. Majumder, P. Ordóñez, J. S. Osorio, K. E. Paik, and M. Somai, *Leveraging Data Science for Global Health*. Cham, Switzerland: Springer, 2020.

[20] R. Clarke, "Principles and business processes for responsible AI," *Comput. Law Secur. Rev.*, vol. 35, no. 4, pp. 410–422, Aug. 2019.

[21] A. Tamboli, *Keeping Your AI Under Control*. New York, NY, USA: Apress, 2019.

[22] International Electrotechnical Commission. (2010). *Functional Safety of Electrical, Electronic, Programmable Electronic Safety Related Systems*. [Online]. Available: https://webstore.iec.ch/preview/info_iec61508-1%7Bed2.0%7Db.pdf

[23] A. Groza and I. Marian. (2022). *Towards Assuring Conformance of AI Systems*. [Online]. Available: https://www.researchgate.net/profile/Adrian-Groza/publication/360709226_Towards_assuring_conformance_of_AI_systems/links/628641d96e41e5002d312c11/Towards-assuring-conformance-of-AI-systems.pdf

[24] P. Chapman. (2000). *The CRISP-DM User Guide*. [Online]. Available: https://s2.smu.edu/~mhd/8331f03/crisp.pdf

[25] S. N. P. Nam. (2018). *Symbolic and Connectionist Artificial Intelligence: Comparing Paradigms Using Marvin Minsky's Views on Natural Intelligence*. [Online]. Available: https://foliojournal.wordpress.com/2018/12/10/symbolic-and-connectionist-artificial-intelligence-comparing-paradigms-using-marvin-minskys-views-on-natural-intelligence-by-sean-ng/

[26] U.S. Food and Drug Administration. (2019). *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)*. [Online]. Available: https://www.fda.gov/media/122535/download

[27] U.S. Food and Drug Administration. (2021). *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan*. [Online]. Available: https://www.fda.gov/media/145022/download

[28] N. Becker, R. Adler, G. Borges, M. Hauer, J. Heidrich, S. Hilpisch, R. Hoffmann, P. Junginger, L. Jöckel, M. Kläs, D. Krupka, L. Martinez, A. Sesing, and K. Zweig. (2021). *Abschlussbericht ExamAI—KI Testing Und Auditing*. [Online]. Available: https://testing-ai.gi.de/fileadmin/PR/Testing-AI/Abschlussbericht_ExamAI_-_KI_Testing_und_Auditing.pdf

[29] International Organization for Standardization. (2023). *Standards*. [Online]. Available: https://www.iso.org/standards.html

[30] International Electrotechnical Commission. (2023). *Technical Committees and Subcommittees*. [Online]. Available: https://www.iec.ch/technical-committees-and-subcommittees#tclist

[31] International Telecommunication Union. (2023). *ITU-T Recommendations and Other Publications*. [Online]. Available: https://www.itu.int/en/ITU-T/publications/Pages/default.aspx

[32] European Committee for Standardization and the European Committee for Electrotechnical Standardization. (2023). *Search Standards*. [Online]. Available: https://standards.cencenelec.eu/dyn/www/f?p=CEN:105:RESET::

[33] European Telecommunications Standards Institute. (2023). *Search & Browse Standards*. [Online]. Available: https://www.etsi.org/standards#Pre-defined%20Collections

[34] W. Wahlster and C. Winterhalter. (2020). *German Standardization Roadmap on Artificial Intelligence*. [Online]. Available: https://www.din.de/resource/blob/772610/e96c34dd6b12900ea75b460538805349/normungsroadmap-en-data.pdf

[35] *IDW PS: Prüfung von KI-Systemen*, Institut der Wirtschaftsprüfer in Deutschland, Berlin, Germany, 2023.

[36] Information System Audit and Control Association. (2023). *COBIT | Control Objectives for Information Technologies*. [Online]. Available: https://www.isaca.org/resources/cobit

[37] Committee of Sponsoring Organizations. (2023). *Guidance*. [Online]. Available: https://www.coso.org/guidance-erm

[38] European Commission. (2021). *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[39] L. Oala, J. Fehr, L. Gilli, A. Leite, S. C. Ramírez, D. Xie-Li, G. Nobis, E. Alejandro, É. M. Alvarado, G. Jaramillo-Gutierrez, A. Sprl, C. Spain, F. Kherif, B. Sanguinetti, T. Wiegand, E. Alsentzer, M. Mcdermott, F. Falck, and S. Hyland, "ML4H auditing: From paper to practice," in *Proc. 34th Conf. Neural Inf. Process. Syst.*, H. Larochelle, Ed., 2020, pp. 280–317. [Online]. Available: http://proceedings.mlr.press/v136/oala20a/oala20a.pdf

[40] A. P. Martín, C. G. Mateo, L. D. Fernández, and M. C. D. L. Pérez, "Ethics guidelines for the development of virtual assistants for e-health," in *Proc. IberSPEECH*, A. M. Peinando, A. M. Gomez, J. L. Perez-Cordoba, and J. A. Gonzalez-Lopez, Eds. Incheon, South Korea: International Speech Communication Association, 2022, pp. 121–125. [Online]. Available: https://www.academia.edu/94773752/Ethics_Guidelines_for_the_Development_of_Virtual_Assistants_for_e_Health

[41] F. T. H. Lo, "The paradoxical transparency of opaque machine learning," *AI Soc.*, 2022, doi: 10.1007/s00146-022-01616-7.

[42] A. Engel, *Verification, Validation, and Testing of Engineered Systems* (Wiley Series in Systems Engineering and Management). Hoboken, NJ, USA: Wiley, 2010. [Online]. Available: https://studylib.net/doc/25812563/verification–validation-and-testing-of-engineered-system.

[43] J. K. Rempel, J. G. Holmes, and M. P. Zanna, "Trust in close relationships," *J. Pers. Soc. Psychol.*, vol. 49, no. 1, pp. 95–112, 1985. [Online]. Available: https://www.researchgate.net/publication/232554295_Trust_in_close_relationships_J_Pers_Soc_Psychol

[44] K. K. Larsen. (2018). *Trust Thou AI? AI Strategy & Policy Blog*. [Online]. Available: https://aistrategyblog.com/2018/12/03/trust-thou-ai/

[45] B. Knowles and J. T. Richards, "The sanction of authority," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, Mar. 2021, pp. 262–271.

[46] European Commission and Directorate-General for Communications Networks, Content and Technology. (2020). *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self Assessment*. [Online]. Available: https://op.europa.eu/en/publication-detail/-/publication/73552fcd-f7c2-11ea-991b-01aa75ed71a1

[47] R. N. Landers and T. S. Behrend, "Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models," *Amer. Psycholog.*, vol. 78, no. 1, pp. 36–49, Jan. 2023.

[48] The Economist. (2017). *Data is Giving Rise to a New Economy*. [Online]. Available: https://www.economist.com/briefing/2017/05/06/data-is-giving-rise-to-a-new-economy

[49] L. Breiman, "Statistical modeling: The two cultures," *Stat. Sci.*, vol. 16, no. 3, pp. 199–215, 2001.

[50] D. Loshin, *Master Data Management*. San Mateo, CA, USA: Morgan Kaufmann, 2009.

[51] *Secondary Analysis of Electronic Health Records*. MIT Critical Data, Cambridge, MA, USA, 2016.

[52] D. Bhaumik, D. Dey, and S. Kayal, "A framework for auditing multilevel models using explainability methods," 2022, *arXiv:2207.01611*.

[53] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[54] D. Bhaumik and D. Dey, "An audit framework for technical assessment of binary classifiers," 2022, *arXiv:2211.09500*.

[55] S. Ronaghan. (2022). *AI Fairness—Explanation of Disparate Impact Remover*. [Online]. Available: https://towardsdatascience.com/ai-fairness-explanation-of-disparate-impact-remover-ce0da59451f1

[56] A. Etzioni and O. Etzioni, "Designing AI systems that obey our laws and values," *Commun. ACM*, vol. 59, no. 9, pp. 29–31, Aug. 2016.

[57] R. Clarke, "Regulatory alternatives for AI," *Comput. Law Secur. Rev.*, vol. 35, no. 4, pp. 398–409, Aug. 2019.

[58] T. Markert, F. Langer, and V. Danos, "GAFAI: Proposal of a generalized audit framework for AI," in *Proc. INFORMATIK*, in Lecture Notes in Informatics (LNI). Bonn, Germany: Gesellschaft für Informatik e.V. (GI), 2022.

[59] D. Yakobovitch. (2021). *A Checklist to Audit Your AI Systems*. [Online]. Available: https://www.linkedin.com/pulse/checklist-audit-your-ai-systems-david-yakobovitch/

[60] T. Wiegand, N. Lee, S. Pujari, M. Singh, S. Xu, M. Kuglitsch, M. Lecoultre, A. Riviere-Cinnamond, E. Weicken, M. Wenzel, A. Werneck Leite, S. Campos, and B. Quast. (2020). *Whitepaper for the ITU/WHO Focus Group on Artificial Intelligence for Health*. [Online]. Available: https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/FG-AI4H_Whitepaper.pdf

[61] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani, "Aequitas: A bias and fairness audit toolkit," 2018, *arXiv:1811.05577*.

[62] M. Rawat. *This California-based Startup, 'Credo AI' is Bringing a Responsible AI Platform to Build a Fair, Compliant, and Auditable AI Model*. [Online]. Available: https://www.marktechpost.com/2022/06/29/this-california-based-startup-credo-ai-is-bringing-a-responsible-ai-platform-to-build-a-fair-compliant-and-auditable-ai-model/

[63] H. Liu, C. Zhong, A. Alnusair, and S. R. Islam, "FAIXID: A framework for enhancing AI explainability of intrusion detection results using data cleaning techniques," *J. Netw. Syst. Manage.*, vol. 29, no. 4, p. 40, Oct. 2021.

[64] U. Bansal, "Explainable AI: To reveal the logic of black-box models," *New Gener. Comput.*, pp. 1–35, Feb. 2023, doi: 10.1007/s00354-022-00201-2.

[65] C. Breuel. (2020). *ML Ops: Machine Learning as an Engineering Discipline*. [Online]. Available: https://towardsdatascience.com/ml-ops-machine-learning-as-an-engineering-discipline-b86ca4874a3f

[66] Census. (2023). *AI Audit*. [Online]. Available: https://censius.ai/wiki/ai-audit

[67] J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo, "OpenML: Networked science in machine learning," *ACM SIGKDD Explor. Newslett.*, vol. 15, no. 2, pp. 49–60, 2013. [Online]. Available: https://dl.acm.org/doi/10.1145/2641190.2641198, doi: 10.1145/2641190.2641198.

[68] A. Gawande, *The Checklist Manifesto*. New Delhi, India: Metropolitan Books, 2010.

[69] B. Braunschweig and M. Ghallab, Eds., *Reflections on Artificial Intelligence for Humanity* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence), vol. 12600. Cham, Switzerland: Springer, 2021, doi: 10.1007/978-3-030-69128-8.

Chapter 4

Paper II: ML Auditing and Reproducibility: Applying a Core Criteria Catalog to an Early Sepsis Onset Detection System

Title:	ML Auditing and Reproducibility: Applying a Core Criteria Catalog to an Early Sepsis Onset Detection System
Authors:	Markus Schwarz*, Ludwig Christian Hinske, Ulrich Mansmann, Fady Albashiti
Journal:	IEEE Access
Submitted:	22nd April 2025
Accepted:	11th June 2025
Published:	13th June 2025
Year:	2025
Volume:	13
Pages:	104899–104915
DOI:	10.1109/ACCESS.2025.3579631
OSF Project ID:	bca7n
JIF Percentile:	65.4 (2023, Computer Science, Information Systems)
JIF Rank:	87/250 (2023, Computer Science, Information Systems)
Rights:	Creative Commons Attribution 4.0 United States (CCBY 4.0)

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 April 2025, accepted 11 June 2025, date of publication 13 June 2025, date of current version 23 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3579631

RESEARCH ARTICLE

ML Auditing and Reproducibility: Applying a Core Criteria Catalog to an Early Sepsis Onset Detection System

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

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

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

ABSTRACT **Background:** On the way towards a commonly agreed framework for auditing ML algorithms, in our previous paper we proposed a 30-question core criteria catalog. In this paper, we apply our catalog to an early sepsis onset detection system use case. **Methods:** The assessment of the ML algorithm behind the sepsis prediction system takes place in a kind of external audit. We apply the questions of our catalog with described context to the available sepsis project resources made publicly available. For the audit process we considered three steps proposed by the Supreme Audit Institutions of Finland et al. and utilized inter-rater reliability techniques. We also conducted an extensive reproduction study, as being encouraged by our catalog, including data perturbation experiments. **Results:** We were able to successfully apply our 30-question catalog to the sepsis ML algorithm development project. 37% of the questions were rated as *fully addressed*, 33% of the questions as *partially addressed* and 30% of the questions as *not addressed*, based on the first auditor. The weighted Cohen's kappa agreement coefficient results in $\kappa = 0.51$. The focus of the sepsis project is on algorithm design, data properties and assessment metrics. In our reproduction study, using externally validated pooled prediction on the self-attention deep learning model, we achieved an AUC of 0.717 (95% CI, 0.693-0.740) and a PPV of 28.3 (95% CI, 24.5-32.0) at 80% TPR and 18.8% sepsis-case prevalence harmonization. For the lead time to sepsis onset, we could not reproduce meaningful values. In the perturbation experiment, the model showed an AUC of 0.799 (95% CI, 0.756-0.843) with modified input data in contrast to an AUC of 0.788 (95% CI, 0.743-0.833) with original input data, when trained on the AUMC dataset and validated externally. **Discussion:** The catalog application results are visualized in a radar diagram, allowing an auditor to quickly assess and compare strengths and weaknesses of ML algorithm development or implementation projects. In general, we were able to reproduce the magnitude of the sepsis project's reported performance metrics. However, certain steps of the reproduction study proved to be challenging due to necessary code changes and dependencies on package versions and the runtime environment. The extent of the deviation in the result metrics was −5.83% for the AUC and −11.03% for the PPV, presumably explained by our absence of tuning. The AUC change of 1.45% indicates resilience of the self-attention deep learning model to input data manipulation. An algorithmic error is most likely responsible for the missing lead time to sepsis onset metric. Even though the acquired weighted Cohen's kapa coefficient is interpreted as having a ''fair to good'' agreement between both auditors, there exists potential subjectivity showing room for improvement. This could be mitigated if more groups (multiple auditors) would apply our catalog to existing ML development and implementation projects. A subsequent ''catalog application guideline'' could be established this way. Our activities might also help development or implementation teams to prepare themselves for future, legally required audits of their newly created ML algorithms/AI products.

The associate editor coordinating the review of this manuscript and approving it for publication was Rajeeb Dey.

INDEX TERMS AI auditing framework, auditable AI, AI reproducibility, ML auditing core criteria catalog application, reproduction study, sepsis prediction audit.

I. BACKGROUND

In 2020 the European Commission started working towards a legislative act that is supposed to "ensure that AI systems … are safe and respect existing law on fundamental rights," as well as to "enhance … effective enforcement of existing law" [1]. The European Commission [1] also wants to "facilitate the development of a single market for … trustworthy AI applications." This initiative is known under the "EU AI Act," which is effective since 1st August 2024 and can be enforced in all EU member states by 2nd August 2026 [2]. A major part of this initiative is about "[ensuring] … protection for … fundamental rights" [1]. These fundamental rights encompass e.g., in Article 7 the "respect for private and family life" and in Article 8 the "protection of personal data" [3].

The lack of trust in AI/ML-based decisions that impact human lives, as well as the lack of transparency behind the details of an ML algorithm/AI product development often lead to low acceptance. For example, in the healthcare sector, when compared to the e-commerce or the automotive sector, the utilization of applications or tools incorporating an ML core component remains limited. This implies that good, value-providing applications that, e.g., could improve patient outcomes or clinical processes are not introduced at a rate at which they could be.

In Schwarz et al. [4, p. 44] we presented a 30-question "ML Auditing Core Criteria Catalog" that is supposed to guide interdisciplinary ML development teams through the design and deployment of an ML application, with the aim of enhancing its acceptance. The targeted consequence is a speed up in the adoption of such products in general practitioner's offices, (outpatient) clinics or other actors in the healthcare sector.

In this paper we apply our catalog to an early sepsis onset detection system use case from Moor et al. [5]. Our first goal is to gauge the catalog's usefulness by answering each of our 30 questions retrospectively using the sepsis project resources. Doing so would also allow us to judge how our learnings can be generalized, e.g., for similar use cases. Our second goal, as encouraged by our catalog, is to assess the ease of reproducing Moor et al. [5]'s algorithm development pipeline and their acquired results.

Moor et al. [5]'s project was chosen because it is a well-documented, openly accessible, and newly developed ML classification proposal within the medical application area. Since our catalog paper was published after the sepsis prediction paper, Moor et al. [5] could not have considered any of the 30 catalog questions in their work.

It is important to emphasize that the objective of our study is not to challenge the medical soundness or rationale behind Moor et al. [5]'s sepsis project. Instead, we adopt an external auditor's point of view, employing our ML auditing core criteria catalog presented in Schwarz et al. [4]. We pursue an academic perspective and do not have any affiliation with existing professional auditing companies or regulatory organizations.

This paper is structured as follows: First, we give an overview of Moor et al. [5]'s project. Then we describe the existing resources of their project including its code repository on GitHub [6]. The main part of this paper consists of applying each of the 30 questions within the three catalog categories *Conceptual Basics*, *Data & Algorithm Design* and *Assessment Metrics* to the sepsis project. Afterwards we describe the setup and the results of our sepsis project reproduction undertaking. Finally, we contrast our findings to our initial goals and conclude with an outlook of the next steps.

II. METHODS

For the audit process utilizing our catalog questions given in Schwarz et al. [4, pp. 9-11], we considered the first three steps described by SAI of Finland et al. [7, p. 16][1]:

1) "Reviewing the documentation"
2) "Close inspection of the data and a review of the code"
3) "Reproduction of … the model training, testing, scoring and performance measures"

Those steps were performed by the corresponding author. Specifically, the first step entails a thorough study of the existing paper, paper supplement and GitHub repository from Moor et al. [5], [6]. Once a comprehensive understanding of the project's documentation is achieved, we can work through the 17 conceptual basics, seven data & algorithm design and six assessment metrics questions from our ML auditing core criteria catalog. The information deemed relevant for answering each question is always supported by a reference to the respective source document.

Thus, in section III-C to section III-E, each of the 30 questions of the ML auditing core criteria catalog is applied to the sepsis paper resources (described in section III-B). We utilized a 3-point ordinal scale consisting of the categories:

⊗ The question is not addressed (codified with integer value 1)

? The question is partially addressed (codified with integer value 2)

✓ The question is fully addressed (codified with integer value 3)

The decision to use trichotomous item levels instead of the more common 5- or 7-point Likert-type scales was made because of two reasons. First, we are primarily interested whether the respective question of our auditing catalog is covered by Moor et al. [5] or not. However, simple binary response categories are not sufficient to capture the state

[1] We also briefly mentioned those steps under the key word *Audit Process* in our catalog paper.

when partial aspects of a question were answered. Second, in a tradeoff between reducing complexity for the rater and ensuring validity, there is "evidence … that … validity … [is] independent of the number of scale points used for Likert-type items" [8, p. 666]. In a more recent study, Jae Jeong [9, p. 133] also found that "simpler measurement scales like 3-point … scales provided … estimates that are highly equivalent to those from the original 5-point scale."

To reduce the amount of subjectivity, before presenting the catalog coverage results, in section III-F we utilize inter-rater reliability (IRR)[2] techniques. A second, independent rater from the Medical Data Integration Center of LMU University Hospital (MeDIC$^{\text{LMU}}$) also answered each of the 30 questions retrospectively. Afterwards we present a suitable agreement coefficient and highlight questions where there is disagreement.

In the second part of our paper, where we focus on the reproduction of the sepsis project, steps two and three of the audit process were conducted. This includes an in-depth inspection of the utilized datasets, as well as the complete data processing pipeline. A sufficient understanding of the code base's functioning needs to be established, before working on the reproducibility of the results. The Turing Way Community [11] defines *reproduction* as "the same analysis steps performed on the same [datasets]." We also included data perturbation tests as part of our reproduction study.

III. RESULTS

A. SEPSIS ONSET PREDICTION PROJECT

The paper from Moor et al. [5] has the title "Predicting sepsis using deep learning across international sites: a retrospective development and validation study" and was published in eClinicalMedicine (Lancet) in 2023. The objective of their paper is to develop and validate an ML-based early sepsis onset detection system [5, p. 2]. As motivation for their paper, Moor et al. [5, p. 2] state the "lack of … annotated, multi-centre data and … external validations of predictive models for sepsis." Additionally, they refer to the finding of Wong et al. [12] about the mediocre performance of the proprietary Epic Sepsis Model (ESM). In an external validation setting, "it identifies only 7% of patients with sepsis who were missed by a clinician," while "not [identifying] 67% of patients with sepsis" and "generating alerts on 18% of all hospitalized patients" [12].

Moor et al. [5] claim to help solve those issues by multiple contributions. First, their work would "[unify] ICU data from multiple sources [building] an open-access platform for developing and externally validating sepsis prediction approaches" [5, p. 2]. Second, Moor et al. [5, p. 2] would "[implement] sepsis annotations … and [develop a] sepsis early warning system using state-of-the-art machine learning (ML) algorithms." And third, they would create "an evaluation strategy … for the … trade-off between *accurate* and *early* alarms … [vs.] false alarms" [5, p. 2].

Their target variable is a sepsis label indicating that sepsis developed within a patient after being admitted to the ICU. To create the sepsis label ("sepsis label annotation"), they use the Sepsis-3 definition from Singer et al. [13, p. 20] pointing to a "total SOFA change $\geq$ 2 points consequent to the infection." Singer et al. [13, p. 24 and footnote b] "[defines a suspected infection] as the … administration of … antibiotics and sampling of body fluid cultures." Moor et al. [5, Suppl. p. 2] "[follow] the approach … as closely as possible" and determine a suspected infection (SI) by administration of antibiotics (ABx) with subsequent fluid sampling within 24h or by fluid sampling and ABx within 72h. SOFA stands for *Sequential Organ Failure Assessment* and is a standard ICU patient condition evaluation score consisting of six subscores (e.g., respiratory or cardiovascular). Each subscore can add 0 to 4 points to form the total SOFA score $\in [0, 1, \ldots, 24]$. Moor et al. [5, Suppl. p. 2] label a patient-ICU stay record as having developed sepsis when the total SOFA scores change by two or more within a SI window of $[\text{SI time} - 48h; \text{SI time} + 24h]$.

The SOFA subscores are calculated from raw data; in case no appropriate raw data is available, the subscore is set to 0 [5, Suppl. p. 2]. In case the dataset lacks information about body fluid samplings, an "alternative definition of suspected infection, which was defined as co-occurrence of multiple antibiotic administrations" is used [5, Suppl. p. 3].

The four utilized ICU datasets in Moor et al. [5, p. 2]'s paper are MIMIC-III, eICU, HiRID and AUMC, with collected patient-ICU stay information ranging from 2001 to 2016.

Their exclusion criteria for patient-ICU stay records are [5, Suppl. p. 3]:

- Age $<$ 14 years
- Hospital centers in eICU with Sepsis-3 prev. $<$ 15%
- ICU stay $<$ 6h
- Measurements done at less than 4 different points in time
- Missing data window $>$ 12h
- Onset of sepsis outside (before) ICU stay
- Onset of sepsis before 4h or after 168h of ICU stay

Across the four datasets, Moor et al. [5, Suppl. p. 4] determined "59 temporal variables (vital signs and lab tests)" that serve as independent features for the ML models. For each hour of a patient-ICU stay, the respective value was calculated using the median of all samplings taken within the hourly interval [5, Suppl. p. 4]. If no measurement took place, the respective value was considered missing [5, Suppl. p. 4].

Moor et al. [5, Suppl. pp. 3-4] constructed four ML models:

1) Light gradient-boosting machine (lgbm) after Ke et al. [14]
2) Least absolute shrinkage and selection operator regularized logistic regression (lr) after Robert Tibshirani [15]
3) Deep (neural network) self-attention model (attn) after Vaswani et al. [16]

[2] "Inter-rater reliability refers to the portion of data reliability that is affected by the specific components of the data production system that you call raters" [10, p. 4].

4) Recurrent neural networks employing gated recurrent units (gru) after Cho et al. [17]

Thus, for the deep learning models (3 and 4), next to the 59 temporal values (observed measurements), 59 missingness indictors were created, as well as 59 measurement counts [5, Suppl. pp. 3-4]. Together with nine derived domain knowledge features (e.g., the SOFA subscores) and four static variables (demographics) this results in 190 features [5, Suppl. pp. 3-4].

For the non-deep learning (DL) models (1 and 2) "multi-scale look back statistics" were created resulting in 1269 features [5, Suppl. p. 4]. Those include the maximum, minimum, median, mean and variance value for each variable calculated over periods of 4, 8 and 16 hours [5, Suppl. p. 4].

For each feature, a standardization was performed by computing the mean and SD across all patient-ICU stay values $\frac{x-\bar{x}}{\sigma_{\bar{x}}}$ [5, Suppl. p. 4].

The experimental setup and data preprocessing is conducted as follows [5, Suppl. pp. 5-6]:

1) Splitting data into development and test sets containing each five splits with two stratified, randomized folds (training and validation, respectively fine tuning, and test boosting)
2) Optimization on first split of the training fold (identifying the best hyperparameters)
3) Tuning of best hyperparameters on validation fold
4) Training with the best hyperparameters on all splits of the training fold
5) Repeated subsampling to harmonize sepsis-case prevalence across the four datasets
6) Evaluate performance internally (on test boosting fold of the same dataset) and externally (on test boosting fold of all other datasets)
7) Fine tuning done on respective fold of the test data

As outcome, Moor et al. [5, p. 5] report the best area under the curve (AUC) value achieved with internal validation with model 3 averaged across all datasets of 0.846. The average positive predictive value (PPV) is 42.0% with 3.7h lead time to sepsis onset at a fixed 80% sensitivity (TPR) and sepsis-case prevalence harmonization of 18.8% [5, p. 5]. This implies that for 1 true sepsis prediction (TP) there are $\approx$ 1.4 false sepsis predictions (FP).[3]

For external validation, using prediction pooling averaged across all datasets, Moor et al. [5, p. 6] acquire with model 3 a best AUC of 0.761. The average PPV is 31.8% with 1.71h lead time to sepsis onset also at a fixed 80% TPR and sepsis-case prevalence of 18.8% [5, p. 6]. Analogously, this leads to 1 TP among $\approx$ 2.1 FP.[4]

B. AUDIT: ANALYZED RESOURCES

The sepsis paper itself comprises 13 pages and the provided supplementary document consists of 32 pages. The project also contains a public GitHub repository that has 1,556 commits done between 21st February 2020 and 9th October 2023 from five authors [6]. The repository's *multicenter-sepsis-master* branch consists of 353 files in 49 folders. There is a *README.md* file that describes the "Data setup," "Python pipeline," "Preprocessing," "Training," "Evaluation pipeline," "Results and plots" and "Pooled predictions" [6]. There are also two files called *requirements_full.txt* and *requirements_minimal.txt* that contain the necessary Python packages incl. version Moor et al. [6] used.

The repository's structure looks like:

1) . (7 files)
2) config (122 files, 4 folders)
3) datasets (3 files, 2 folders)
4) img (1 file)
5) r (33 files, 1 folder)
6) results (2 files, 3 folders)
7) revisions (49 files, 10 folders)
8) scripts (69 files, 6 folders)
9) src (55 files, 11 folders)
10) tests (12 files, 3 folders)

Folders 2, 5, 8 and 9 are important for the reproduction study explained later in section III-G.

C. AUDIT: CONCEPTUAL BASICS

a: QUESTIONS 1-2: AI OPPORTUNITIES VS. AI RISKS

? Q1. "Is the expected *benefit* $*$ *benefitProbability* of a successful ML use case implementation greater than the *damage* $*$ *damageProbability* in case of failure?"

This question aims at quantifying concrete benefits and risks with attributed probabilities for an ICU onsite implementation of the algorithm. There is no direct reference to this in the sepsis paper, however, potential benefits and drawbacks are theoretically discussed.

Moor et al. [5, p. 11] mention several benefits of their sepsis prediction algorithm: "making predictions in hourly intervals, … [being] closer to a bed-side monitoring scenario. … [addressing] alarm fatigue by … [using] an upper bound of at most one *single* false alarm … [per] ICU stay" and making "harmonised and annotated datasets [available] … for other researchers … to evaluate their prediction models on external hospital sites."

As limitations, the "[assessment of] clinical utility [at] bed-side," the "[introduction of] selection bias" and the "alternative definition of suspected infection" are mentioned [5, 11-12]. The selection bias might "affect the … performance for certain subgroups" [5, p. 12].

✓ Q2. "Do you expect a productivity gain, improved quality or a new functionality compared to the current manual/non-ML process?"

In addition to the previously outlined benefits, there is a claimed "clinically useful" improvement of sepsis prediction performance compared to the existing proprietary ESM model [5, p. 11]. Additionally, the automatic calculation of a sepsis label using Sepsis-3 suspected infection and

[3] $\text{PPV} = \frac{\text{TP}}{\text{TP}+\text{FP}} \Rightarrow \text{FP} = \frac{\text{TP}-\text{PPV}*\text{TP}}{\text{PPV}} \Rightarrow \text{FP} = \frac{1-0.42*1}{0.42} = \frac{0.58}{0.42} \approx 1.4.$

[4] $\text{FP} = \frac{1-0.318*1}{0.318} = \frac{0.682}{0.318} \approx 2.1.$

SOFA-subscores derivations from raw data could be considered as a process improvement over manually calculated clinical guidelines (like qSOFA) that currently help clinicians to identify sepsis onset.

b: QUESTIONS 3-4: RISK MANAGEMENT

?⃝ Q3. “Are the roles and responsibilities (RACI[5]) and liabilities before, during and after the implementation clearly defined?”

A clinical implementation of the model is not part of Moor et al. [5]'s paper. There are no details provided on the location or way a potential prototype ICU deployment of the suggested ML model has been or is planned. Moor et al. [5, 12] suggest that before deployment, “international clinical validation studies” need to take place. However, since the question also aims at the RACI of the development stage (before implementation), there the roles are clearly defined among the sepsis paper authors in the *Contributors* section [5, p. 12].

⊗ Q4. “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?”

The discussion of a kill switch or when to fall back to the “established” guidelines or clinician's experience for determining a sepsis onset is not part of Moor et al. [5]'s paper.

c: QUESTIONS 5-9: METHODOLOGY

?⃝ Q5. “Have you aligned and agreed on the methodology with all project stakeholders (e.g., for implementation CRISP-DM and internal audit SMACTR)?”

Within the given resources, there is no explicit method mentioned that dictates the planning and the necessary steps of the sepsis onset prediction project. However, upon examination of all the steps that were conducted during the study, which are described in greater detail in the paper's supplement [5, Suppl. pp. 2-6], it becomes evident that CRISP-DM has been implicitly applied. All activities described in between step 1 *Business understanding* and step 5 *Evaluation* were conducted and can be found in the resources. Step 6 *Deployment* of the suggested ML model at a specific site is not discussed, but Moor et al. [5, Suppl. p. 6] point towards “[indispensable] clinical trials” that should follow.

⊗ Q6. “Are the implications in case the ML use case falls in the ‘high risk’ category of the EU AI Act understood?”

Any legal aspects or requirements that would come with the on-site introduction of such a sepsis early warning system are not considered.

✓ Q7. “Do you plan to make use of Data Sheets to describe the data collection process as well as the data properties?”

Moor et al. [5, p. 3,fig. 1] describe the data curation pipeline, including standardization and how to deal with missing data, as well as the feature engineering process. For each of the four data sources, an extensive patient characteristics table is provided, highlighting details of the respective development and test sets [5, Suppl. pp. 25,28-31].

?⃝ Q8. “Do you plan to create AI Model Cards/AI Fact Sheets to describe the model characteristics?”

The two non-DL models (1 and 2), as well as the two DL models (3 and 4), are well described in Moor et al. [5, Suppl. pp. 3-4]'s paper. Especially for model 3, the *attn* model, which is found best performing, the sepsis project authors provide sufficient details [5, Suppl. p. 4]. Additionally, for the DL models, there is an overview of the utilized, possible hyperparameters combination given [5, Suppl. p. 24,tab. S3]. The GitHub repository's *README.md* provides information about training details and performance evaluation of the models [6]. Indirectly, by looking at the dependencies and source packages used by Moor et al. [6]'s custom functions, an experienced scientist or developer can infer the underlying details of the algorithm and evaluation steps. However, a tabular representation as an AI Model Card or AI Fact Sheet, listing the model details, intended use etc., which would allow contrasting those four models, is not contained in Moor et al. [5]'s resources.

⊗ Q9. “Do you plan to prepare AI Care Labels to instruct internal stakeholders how to use and ‘treat’ the algorithm?”

The *README.md* file existing at the GitHub repository allows other scientists to reproduce the sepsis onset project's algorithm and developers to review the whole code base [6]. However, there is an absence of “care” labels for the end user, thereby leaving clinicians without instructions on how to utilize and operate the algorithm.

d: QUESTIONS 10-12: AUDIT PROCESS

⊗ Q10. “Have you established an internal advisory committee consisting of senior IT governance specialists and business/medical specialists who critically accompany the implementation (e.g., watch for sufficient documentation and methodology adherence)?”

Moor et al. [5] do not include thought processes about how to establish a project management or advisory committee of an onsite implementation.

⊗ Q11. “Do you ensure the ML implementation is not violating ethical concepts (‘ethics by design’ is considered)?”

Even though Moor et al. [5, p. 5] mention under “Ethics approval” their acquired permission from “Ethikkommission Nordwest- und Zentralschweiz EKNZ” for their sepsis paper study, any ethical implications concerning the introduction of

[5] “RACI says that when working in teams it needs to be clear who is **R**esponsible for a given task, who is **A**ccountable especially if something goes wrong, who needs to be **C**onsulted for advice and who must be **I**nformed about the progress” [4, p. 9 and footnote 9].

such a sepsis early warning system at an ICU hospital site are not described.

⍰ Q12. "Do you have protocols in place that allow independent, external auditors to critically review the ML use case implementation?"

There are no suggestions included in Moor et al. [5]'s paper, how the proposed prediction system might be audited (and thus also no suggested protocols). However, since Moor et al. [6]'s GitHub repository is freely accessible and the code base appears sufficiently commented, it seems possible with some effort to reverse engineer the model's internal mechanics.

e: QUESTIONS 13-17: QUALITY ASSURANCE (QA)

✓ Q13. "Did you perform a verification of the ML output behavior using a set of expected, representative inputs of the productive usage?"

The experimental setup, how to proceed with the harmonized and cleansed data of the four datasets in the model training and evaluation step is comprehensible [5, Suppl. p. 5]. For each dataset, a five times randomly sampled test partition sized 10% of the total dataset's size was always kept separate from the training data [5, Suppl. p. 5]. Moor et al. [5, p. 5] combined the predictions on the test data of three datasets not used for training in a "pooled prediction," allowing to test how well the trained algorithm can be transferred from one ICU site (location) to another.[6]

Secondly, Moor et al. [5, Suppl. pp. 9-12,figs. S4-S7] always compare the performance of their four ML models against existing clinical baselines, such as the "Modified Early Warning Score" (MEWS) or the SOFA.

✓ Q14. "Did you perform a validation whether the project's specification and stakeholders' needs are met?"

Moor et al. [5, p. 2] wanted to contribute to closing the gap of the availability of publicly available and harmonized ICU datasets that also contain externally validated sepsis prediction algorithms. Compared to the mentioned baselines, in the "pooled prediction" scenario, the *attn* DL model [5, Suppl. pp. 9-12,figs. S4-S7] performs in three out of four cases better in terms of AUC. Also, Moor et al. [6]'s resource contribution includes a GitHub repository, which is subject to an open-source license.

⊗ Q15. "Do you think the ML model would pass an external AI Certification/AI Assurance case fulfilling the six components of trust: predictability, dependability, faith, consistency, utility and understanding?"

Moor et al. [5] do not comment on or discuss the process of any (theoretical) external AI certification/assurance case. Also, there is no section performing a fit-gap analysis of the sepsis onset detection algorithm against the requirements of the six components of trust. Thus, no conclusion can be drawn regarding the perception of the sepsis prediction model being trustworthy by an onsite clinician.

⊗ Q16. "Given inputs from different test users, does the ML model adhere to the principles of distributive, procedural and interactional justice?"

Ethical aspects and investigations of any already mentioned bias towards patients occurring in the prediction of a sepsis onset are not part of Moor et al. [5]'s paper. As the onsite deployment of the algorithm is not discussed, no evaluation with new patient input from potential representative test users has been conducted.

⊗ Q17. "Given inputs from different test users, does the ML model avoid differential prediction and intentional discrimination?"

The race is not included as a variable in the algorithm, but the sex, age, height, and weight are, which could also lead to misconceptions. Additionally, the location sites of the four datasets are all based in Western Europe and the United States, potentially leading to an underrepresentation of other patient populations. But, as already mentioned in the previous question, ethical aspects and investigations of any bias are not part of the sepsis paper.

D. AUDIT: DATA & ALGORITHM DESIGN

a: QUESTIONS 18-21: DATA PROPERTIES

⍰ Q18. "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?"

According to Moor et al. [5, p. 2], the "four … databases [contain] clinical and laboratory ICU data that was routinely collected." They also mention for each of the four data sources the respective document indicating that pseudonymization took place [5, Suppl. p. 3]. In an auxiliary analysis, Moor et al. [5, Suppl. p. 15,fig. S10] point out that "AUMC … [has] a large proportion (80%) of surgical patients." Besides this, the sepsis paper does not provide any further details on the initial data collection process and the potential influence of unknown factors on the observed data.

⍰ Q19. "Does the training data capture relevant characteristics of the population in scope for the ML use case?"

The focus of the sepsis paper is on the time series deep learning self-attention model (*attn*), where the general maxim states the more data, the better. It is not discussed why and how the 59 clinical and laboratory time-dependent variables including derivatives were chosen. The datasets used contain much more potential variables that could have been selected or engineered. Moor et al. [5, p. 3] state that "clinically valid ranges" were "determined by an experienced ICU clinician," which could imply that this clinician also helped in the variable selection in the first place.

⍰ Q20. "Are the required data quality dimensions (e.g., accuracy, consistency, completeness and currency) well understood and taken care of?"

In the experimental setup and the feature engineering part of Moor et al. [5, Suppl. pp. 4-5]'s paper, it is well described how the data was preprocessed, and missing values are dealt

[6] The sepsis paper authors talk here about an *external validation.*

with. However, it has not been explicitly discussed if the data quality is similar between the four datasets. It was also not discussed if there might be crucial ICU related information that is not captured in the datasets or if the captured data could be outdated (e.g., the first MIMIC-3 patient-ICU stay records are from 2001).

✓ Q21. “Are the procedures necessary for data cleansing and consolidation known, and is the understanding of data scales, and references ranges given?”

The data processing steps necessary to acquire a usable database with the needed features and target for the study are well described [5, Suppl. pp. 2-5]. Moor et al. [5, p. 3] discussed the reference ranges with an “experienced ICU clinician”. The sepsis paper supplement explains details of the features used [5, Suppl. pp. 20-24]. Furthermore, Moor et al. [6]’s GitHub repository enables an experienced developer to derive comprehensive details concerning data types, scales, and reference ranges, as well as standardization protocols.

b: QUESTIONS 22-24: ALGORITHM DESIGN

? Q22. “Is the difference between causality and correlation known? In the absence of known counterfactuals for each individual, population samples can only give associations with a certain strength (e.g., given by the Pearson’s correlation coefficient).”

The sepsis prediction model does not intend to find or explain the causes of a sepsis onset. Thus, there is no intention to establish a causal relationship between ICU time series variables and a sepsis onset event. Moor et al. [5]’s paper focusses on prediction, namely the creation of a model that identifies correlations between time series ICU data and sepsis onset, with the subsequent generation of a warning for clinicians that a patient is at risk of developing sepsis. However, the model does investigate which variables have a greater influence on a sepsis onset event than others using variable importance methods [5, Suppl. pp. 4-5].

? Q23. “Did you apply Occam’s Razor principle for the model type selection? Meaning in case a black box model (e.g., DNN, NLP) is to be used, does it provide substantial benefit (e.g., accuracy) over a white box model (e.g., logistic regression, decision tree)?”

Although models 1 (lgbm) and 2 (lr) generally demonstrate inferior performance in terms of AUC compared to DL model 3 (attn), the average margin is -1.5% (vs. lgbm) with internal validation, respective -8.0% with external validation [5, Suppl. pp. 9-12,figs. S4-S7]. On the HiRID dataset, when internal validation is considered, model 1 exhibits a marginal superiority over model 3, with an increase in AUC of $+0.6\%$ [5, Suppl. p. 11,fig. S6]. Additionally, on the same dataset, when doing an external pooled validation, the clinical baseline SOFA used as a predictor achieves an AUC of $+4.4\%$ compared to model 3 [5, Suppl. p. 11,fig. S6].

However, when examining the other metrics, PPV and median earliness in hours before sepsis onset, the margin between model 3 (attn) and model 1 (lgbm) becomes more pronounced. Is the average margin of PPV -5.7% with internal validation (vs. lgbm),[7] it already becomes -17.5% with external validation [5, Suppl. pp. 9-12,figs. S4-S7]. The difference in median earliness in hours between model 3 (attn) and model 1 (lgbm) is even -38.0% with internal validation (vs. lgbm) respective -50.4% with external validation [5, Suppl. pp. 9-12,figs. S4-S7].

Taking both findings into account, Moor et al. [5] could have considered discussing the implications for a clinician who must rely on black box sepsis predictions, explained only by feature importance methods, in contrast to e.g., already established, transparent 6-dimensional SOFA assessment scores. It is also noteworthy that the deep learning models 3 (attn) and 4 (gru) are able to “automatically [learn] feature representations from sequential data,” in contrast to models 1 (lgbm) and 2 (lr), and are thus better suited to work with 1h sampled real time clinical and laboratory data [5, Suppl. p. 4].

✓ Q24. “Did you establish a correct ML use case hypothesis with concrete problem description and expected behavior (acceptance criteria, metrics, statistical testing results)?”

Moor et al. [5, p. 2] describe their use case, problem, and algorithm’s goal well (see section III-A). The three primary metrics (AUC, median earliness and PPV at 80% TPR) and statistical testing methods employed are provided in the necessary detail [5, Suppl. pp. 5-6]. Moor et al. [5, p. 12] see their “harmonised dataset … and performed analyses …help pave the way …to deploy sepsis prediction models.” However, there is no debate around acceptance criteria of a theoretical end user (e.g., ICU clinician) for such deployment scenarios.

E. AUDIT: ASSESSMENT METRICS

a: QUESTIONS 25-27: QUALITATIVE ASSESSMENT

✓ Q25. “Are the model assumptions (e.g., how to deal with missing data, model type, hyperparameters) transparently described?”

The four ML models and the strategy on how to select suitable hyperparameters are well described [5, Suppl. pp. 3-4]. Moor et al. [5, Suppl. p. 4] also describe how they standardize the variables and proceed with missing values.

✓ Q26. “Did you establish a traceable log of those model assumptions/testing results being used during the whole development process?“

Moor et al. [5, p. 12] mention in their “data sharing statement” that the “raw data used … is publicly available for accredited researchers.” They continue that “all code … will be made available … , ensuring end-to-end reproducibility” [5, 12]. Thus, their GitHub repository with the complete R and Python code files can be seen as a “traceable log”

[7] The comparison of PPV and median earliness in hours of sepsis onset given in this paragraph are approximations, because Moor et al. [5, Suppl. pp. 9-12,figs. S4-S7] only provide graphics without numerical values in their paper supplement.

because of three points. First, the *README.md* instructions file is a sufficient starting point to comprehend the data processing and modeling steps, allowing users to delve into Moor et al. [6]'s code. Second, the scripts existing under ∼/scripts as well as the functions existing under ∼/src are quite well commented. And third, Moor et al. [6] provide temporary files of their original calculations, e.g., under ∼/config/splits four **.json* split files exist indicating the allocation of patient-ICU stay IDs across the development and test sets.

✓ Q27. "Did you discuss with all stakeholders the strength of conclusions that can be drawn with the current model assumptions (and make sure the conclusions are appropriate)?"

The discussion part of Moor et al. [5, pp. 8-12]'s paper contains three main strengths. First the "size and heterogeneity of the cohort," then the "depth of the external validation" and third the "simulated … real-time prediction scenario" [5, p. 11]. The latter would be "closely aligned with a possible clinical implementation of an early warning system" [5, p. 11]. Moor et al. [5, pp. 11-12] also provide limitations that were already presented in questions 1 and 3. Those facts, in combination with the statement that "all [eight] authors contributed to the interpretation of the findings" implies that the strengths, weaknesses and conclusions were sufficiently discussed among all stakeholders [5, p. 12].

b: QUESTIONS 28-30: QUANTITATIVE ASSESSMENT

✓ Q28. "Did you determine the statistical properties of the training, testing and validation dataset? For example, by using Variance Inflation Factor (VIF), Shapiro-Wilk Test (SWT) and Breusch-Pagan Test (BPT)?"

Although the example metrics (VIF, SWT, BPT) have not been applied, Moor et al. [5, Suppl. pp. 28-31,tabs. S6-S7] provide statistical properties of the development and test sets of the four data sources used during modeling. They also investigate the underlying data generation process among the four datasets using density plots on the harmonized features [5, Suppl. p. 8,fig. S2A]. Because those plots show variances, they mention that the features of each dataset cannot be considered indiscernible, but that "would not even be desirable for a credible simulation" [5, p. 3]. Additionally, they "harmonised the prevalence of sepsis cases to the across-dataset average of 17%" [5, Suppl. p. 5].

✓ Q29. "Did you conduct extensive performance testing according to the agreed metrics? For example, using Receiver Operating Characteristics (ROC): creating the confusion matrix and calculating the F1-Score, Matthews Correlation Coefficient (MCC) or Area Under the Curve (AUC)?"

The sepsis paper itself contains the combined comparisons of the four ML models with the respective four datasets using confusion matrices and AUC curves, as well as a reported PPV at fixed 80% Sensitivity (TPR) [5, p. 7]. The paper supplement provides details for each model and dataset as well as the pooled cohort; the performance is assessed using many different area under the ROC curve (AUC) graphs, as well as PPV to median earliness scatterplots [5, Suppl. pp. 9-12,15-16].

⊗ Q30. "Did you assess the resistance of the ML model's output behavior to natural perturbation, for example, using Total Sobol's Variance Ratio (TSVR) or Cosine Similarity Vector Pairs (CSVP)?"

The assessment of how robust the model behaves when it is fed with naturally, unintentionally, or intentionally manipulated input data is not assessed in Moor et al. [5]'s paper.

F. AUDIT: INTER-RATER RELIABILITY CHECK AND SCORING

As mentioned in section II, to decrease subjectivity and increase the reliability of the ML auditing core criteria catalog assessment, a second rater also answered each of the 30 questions. The same sepsis project resources and codification strategy were utilized. The rating results of both auditors are given in table 1.

TABLE 1. Sepsis project audit rating results.

Catalog Question	First Rater	Second Rater
Q1	2	3
Q2	3	3
Q3	2	1
Q4	1	1
Q5	2	1
Q6	1	1
Q7	3	3
Q8	2	2
Q9	1	1
Q10	1	1
Q11	1	2
Q12	2	2
Q13	3	3
Q14	3	1
Q15	1	2
Q16	1	1
Q17	1	1
Q18	2	2
Q19	2	3
Q20	2	3
Q21	3	3
Q22	2	2
Q23	2	3
Q24	3	3
Q25	3	3
Q26	3	2
Q27	3	1
Q28	3	1
Q29	3	3
Q30	1	1

The rating values are based on a 3-point ordinal scale.
1 $\triangleq$ "not addressed," 2 $\triangleq$ "partially addressed" and 3 $\triangleq$ "fully addressed."
Source: Authors

Since the response categories both raters use constitute a 3-point ordinal scale, even if the first rater declares a question as "fully addressed" (3) and the second rater only as "partially addressed" (2), it becomes evident that there is a notion of agreement among both. The weighted Cohen's Kappa coefficient, given in equation 1 in the appendix, can be used to quantify the level of agreement between two raters [10, p. 102]. Following Gwet [10, pp. 107-108], quadratic weights were used for calculating the weights matrix w_{ij}. The acquired agreement coefficient between the first and the second rater results in $\kappa = 0.51$.

In summary, the first rater determined 11 questions (37%) of the ML auditing core criteria catalog as fully addressed, 10 questions (33%) as partially addressed and 9 questions (30%) as not addressed. The second rater codified 11 questions (37%) with "3," seven questions (23%) with "2" and 12 questions (40%) with "1".

Questions 14, 27 and 28 show the biggest disagreement between the first rater and the second rater.[8] For question 14, the second rater states "while the study conducts extensive technical validation … the paper does not describe any formal process for verifying that specific stakeholder needs or project specifications were met." He also makes a similar argument in question 27: "The paper does not describe any formal stakeholder alignment process … where the model's assumptions and conclusions were explicitly discussed." In question 28 the second rater concludes that "the authors … do not mention running formal statistical tests" and "their approach seems more pragmatic."

The results acquired by both raters are used to visualize the extent by which the sepsis project addresses the ML auditing catalog questions. Here, Harpe [18, p. 840]'s "implication … that the intervals between numbers may actually be equal" is followed, assuming equal numeric distances between the response categories 1, 2 and 3. This in combination with having the catalog subcategory coverage as the "phenomenon of interest," allows the aggregated evaluation of items [18, p. 840].

Thus, the subcategory ratings are built by summing the numeric values of the contained questions divided by the respective maximum possible sum.[9] The coverage of the catalog's nine subcategories based on the sepsis project resources is given in figure 1.

As is evident in the figure, the focus of Moor et al. [5]'s paper is rather on the left side of the diagram, including algorithm design, data properties, assessment metrics, as well as presenting opportunities of such a sepsis early warning system. The not covered questions belong to areas like risk management, quality assurance or audit process, usually related to a concrete implementation at a hospital site.

G. SEPSIS PROJECT REPRODUCTION STUDY

Question 12 of the catalog aims at "[allowing] … external auditors to critically review the ML use case implementation" [4, p. 10]. This implies, in combination with question 26 ("traceable log"), a reproduction of all the important steps done for the development and evaluation of the sepsis prediction model [4, p. 10].

The hard- and software utilized at MeDIC$^{\text{LMU}}$ premises for the reproduction study is given below:

- VM within internal network (models 1 & 2, non-DL)
  - 24 Core Intel Xeon Gold 6152 CPU @2.10GHz
  - 128GiB RAM, 256GiB Swap
  - x86_64-pc-linux-gnu platform with Ubuntu 20.04.6 LTS
  - R version 4.4.1[10]
  - Anaconda conda-build version: 24.5.1
  - New conda environment *p37* with Python 3.7.16[11]
- GPU server within internal network (models 3 & 4, DL)
  - 2 × 64 Core AMD EPYC 7742 CPU @2.25GHz
  - 1024GiB RAM, 512GiB Swap
  - x86_64-pc-linux-gnu platform with Ubuntu 22.04.5 LTS
  - NVIDIA H100 Tensor Core GPU 80GiB PCIe 5.0
  - Anaconda conda-build version: 24.9.2
  - New conda environment *p39* with Python 3.9.20[12]

The *README.md* file of the GitHub repository, previously referenced in section III-B, serves as a valuable source of information [6]. The authors of this paper understand that, as already mentioned in catalog question 26, Moor et al. [6] created this file to make their development process transparent, "ensuring end-to-end reproducibility of all results" [5, p. 12].

According to Moor et al. [6, Suppl. p. 2] the four data sources used are the following:

- MIMIC-III 1.4, 2001-2012, [20]
- eICU 2.0, 2014-2015, [21]
- HiRID 1.1.1, 2008-2016, [22]
- AUMC 1.0.2, 2003-2016, [23]

The first three are accessible on *Physionet* with regular login (user and password), the last is accessible on *life-sciences.datastations.nl* where the login is facilitated by an *eduID* that sends a magic link with a token via email. To get access to MIMIC-III and eICU, it is necessary to obtain a CITI "Data or Specimens Only Research" certificate [24]. Furthermore, a separate data access justification is required for HiRID and AUMC, with an additional reference from a practicing intensivist being necessary for AUMC.

H. REPRODUCTION: DATA PREPARATION IN R

The data preparation is done in R, Moor et al. [6] call this step "data setup." They make use of the *ricu* package, which "allows … to load various clinical concepts through a unified interface, … shielding the user from the data-cleaning and data-mining process" [25, p. 6]. Moor et al. [6] do not mention directly (in the *README.md*) or indirectly (in any file of the *multicenter-sepsis-master* repository) the

[8] Please refer to section A in the appendix for the complete audit protocol of the second rater.

[9] When all questions in a category would be answered with "fully addressed" (3).

[10] Moor et al. [5, p. 5] used R version 4.1 in their study. However, it was determined that this should not have any influence on the reproduction outcome, if the same *ricu* package version is used.

[11] Moor et al. [5, p. 5] indicate in their paper that they used "Python … 3.7.4". The preferred version 3.7.16 used here is identical besides additionally applied security bugfixes [19].

[12] The H100 GPU is set up with CUDA runtime 12.1.105, which is only supported starting from PyTorch 2.4. This is why it was decided to directly use the latest version PyTorch 2.5, which in turn requires at least Python 3.9, differing from Moor et al. [6]'s utilized version.

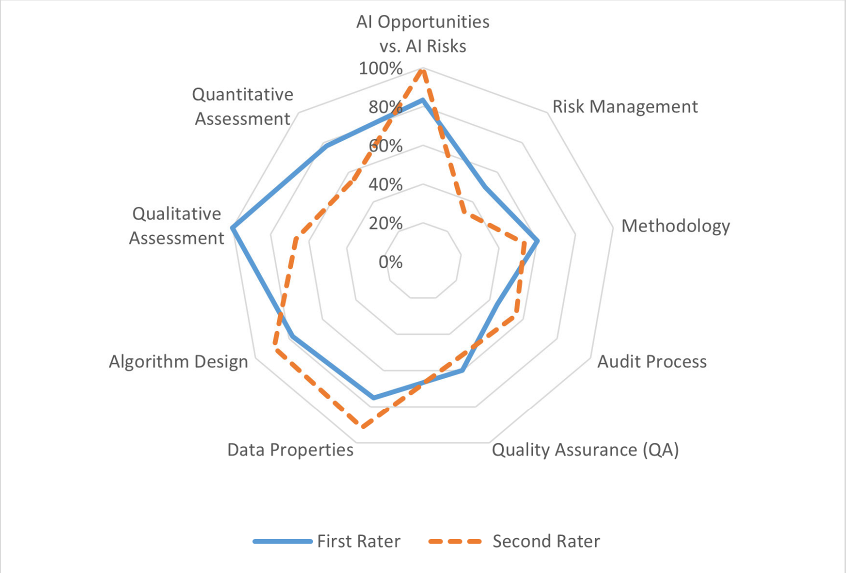


FIGURE 1. Sepsis project audit catalog subcategory coverage. Questions were summed up per subcategory using their codified numeric value. Afterwards the sum of each subcategory was divided by the maximum possible sum per subcategory. Source: Authors.

utilized package version of ricu. When using the currently existing version *0.5.6* an error occurs in ricu-load.R:249:5. After an informal discussion, it became apparent that *ricu 0.4.0* was the version Moor et al. [6] utilized in their data preparation step.

After the successful installation of *ricu 0.4.0*, the following three steps can be executed within R:

1) Downloading: `download_src()` to fetch and store the raw files e.g., (compressed) *.csv from the web
2) Importing: `import_src()` to preprocess the data (convert the raw files to the *.fst fast storage format[13])
3) Attaching: `attach_src()` to make the dataset available for the ricu package

Steps 1, 2 and 3 are combined in a function called `setup_src_data()`. It is necessary to create the data-export folder before execution having the default location at ~/data-export. The exported four dataset files have the **.parquet* extension.[14]

The `export_data()` command will first load concepts, which represent a variable (“feature” or “target”) that Moor et al. [6] use in their model. The outcome variable *sep3* (Sepsis-3 prevalence) is also defined there. However, by looking at line 371 of ~/R/utils/ricu-load.R it becomes clear that the outcome variable present in the **.parquet* files will be *is_case*.

Utilizing the defined properties for each variable, the concepts are then applied to each of the four datasets. The transformations include deletions of records due to the already mentioned exclusion criteria or those that do not correspond to defined column properties (e.g., value ranges).

I. REPRODUCTION: PYTHON PIPELINE

For the ML model training, the scikit-learn [28] and the PyTorch [29] toolkits are used. Moor et al. [6] provide a file *requirements_full.txt* that lists 119 Python packages in the form *[package-name]==[version]*. On the VM within the Python 3.7 environment, precisely those package versions were installed. However, on the GPU server with the Python 3.9 environment, many packages needed to be brought to a later version due to incompatibilities with PyTorch 2.5.

Functions that use or inherit from scikit-learn or PyTorch are located in the src folder of the GitHub repository, which was copied to the Ubuntu home directory ~ [6]. The scripts utilized by Moor et al. [6] for the pipeline can be found in the similarly named folder of the GitHub repository, which was copied to ~ as well.

For the creation of patient splits according to the earlier described experimental setup (see section III-A), as well as for the feature extraction and standardization, the files to be examined are:

- ~/scripts/run_preprocessing.sh script that was separated in run_preprocessing_splits.sh and run_preprocessing_other.sh
- ~/src/datasets/utils.py referenced Python source file
- ~/src/variables/mapping.py referenced Python source file
- ~/src/splits/create_splits.py referenced Python source file

[13] “Fst … provides a fast, easy and flexible way to serialize data frames.… to unlock the potential of high speed solid state disks… [allowing] full random access, both in column and rows” [26].

[14] “Apache Parquet is an open source, column-oriented data file format designed for efficient data storage and retrieval” [27].

Several adjustments were necessary to successfully execute the scripts:

- *Corrected version = "0-4-0" to version = "0.4.0" to match parquet filenames (stored in ~/data-export) coming from R data preprocessing*
- *Corrected -version to _version to match parquet filenames (stored in ~/data-export) coming from R data preprocessing*
- *Corrected the default output_col from 'sep3' to 'is_case' to match column name coming from the R data preparation step*
- *Adjusted ~/config/... to ~/datasets/... for the splits, normalizer, and lambdas folder (new folder for the purpose of reproduction)*

The four prepared dataset **.parquet* files are copied from ~/data-export to the ~/datasets/downloads folder and have the following dimensions (see table 2).

TABLE 2. Overview of datasets after data preprocessing in R.

Dataset	No. of rows	No. of columns
mimic_0.4.0.parquet	2,371,614	993
eicu_0.4.0.parquet	3,515,168	993
hirid_0.4.0.parquet	904,156	993
aumc_0.4.0.parquet	974,116	993

Source: Authors

The column names are all aligned between the datasets. This was accomplished in the preceding R data preparation step.

Upon completion of the operation, the ~/datasets/splits folder contains four **.json* files that indicate the distribution of the stay_id (ICU stay identifier) across the splits.

A sklearn pipeline is used to create two different sets of *.parquet files: *small* and *middle*. The first contains 205 columns and the latter contains 1284 columns. At the end, for each dataset, a *_metadata* file was written that e.g., contains the total number of rows. The files are stored in ~/datasets/{mimic, eicu, hirid, aumc}/data/parquet/{features_middle, features_small}. The number of rows and row groups can be seen in table 3. The number of row groups determine the number of resulting **.parquet* files.

TABLE 3. Overview of datasets after feature transformation in python.

Dataset	No. of rows	No. of row groups
mimic	1,813,796	1,214
eicu	2,968,722	1,801
hirid	878,258	459
aumc	538,093	497

Source: Authors

The next step consists of standardization,[15] which is done on each of the five train splits existing in the development dataset. For every split, on the middle version of the *.parquet files, for 1209 features the mean values and standard deviations are calculated. The results are saved as *.json files in ~/datasets/normalizer with the name syntax normalizer_{mimic, eicu, hirid, aumc}_rep_{0, 1, 2, 3, 4}.json

There were 25 encounters of *RuntimeWarning* when running the normalization. However, when comparing the normalization files to the "reference" files provided in the GitHub repository, there are no meaningful differences in the structure and content [6]. The calculated mean and variance values per variable only differ at the 10-12th digit after the decimal separator.

Moor et al. [6] first focus on training the non-DL model 1 (lgbm) and non-DL model 2 (lr) with the help of the *sklearn* toolkit. The files in scope are:

- ~/scripts/run_lgbm.sh and ~/scripts/run_lgbm_rep.sh scripts that were merged in run_lgbm_together.sh
- ~/scripts/run_lr.sh and ~/scripts/run_lr_rep.sh scripts that were merged in run_lr_together.sh
- ~/scripts/run_baselines.sh script
- ~/src/sklearn/main.py referenced Python source file
- ~/src/sklearn/baseline.py referenced Python source file

Also, adjustments were necessary to successfully execute the scripts:

- *Adjusted –cv_n_jobs = 1 to –cv_n_jobs = {1, 2, 8},*[16] *to achieve a better core utilization*
- *Corrected code for n_jobs: −1 to n_jobs: args.cv_n_jobs to use the cv_n_jobs parameter for the clf_params n_jobs*
- *Adjusted n_iter_search to 4000 for model 2 (lr) to avoid non-convergence*[17]

TABLE 4. Variable set and feature set configuration.

Variable set	Feature set	Key	No. of Elements
full	small	groups	7
full	small	columns	192
full	middle	groups	13
full	middle	columns	1,271
full	raw	groups	6
full	raw	columns	64
physionet	small	groups	7
physionet	small	columns	112
physionet	middle	groups	13
physionet	middle	columns	640
physionet	raw	groups	5
physionet	raw	columns	45

Source: Authors

As can be seen in table 4, Moor et al. [6] allow reductions in the number of features used for training (key = "columns").

[15] In their *README.md* file, as well as in the code itself, Moor et al. [6] use the term "normalization" for calculating mean values and standard deviations. In the paper, as well as in the paper supplement, they use the term "standardisation" [5, p. 4] and [5, Suppl. p. 4].

[16] This parallelization option was adjusted according to the resource utilization of the model and dataset. If set too high, the amount of memory reserved for each job could be too little, resulting in "UserWarning: A worker stopped while some jobs were given to the executor." If set too low, the runtime per model and dataset would be excessive.

[17] This arbitrary large number was reduced to 100 by sklearn during execution because "the total space of parameters 100 is smaller than n_iter = 4000."

For their sepsis project they used the combination of variable set = "full" and feature set = "middle" to achieve 1269 features[18] for models 1 and 2 (non-DL) and variable set = "full" and feature set = "small" to achieve 190 features for the DL models 3 and 4 [5, Suppl. pp. 2-3].

Initially, before performing a (randomized) hyperparameter search for the best parameter combination, a series of auxiliary functions are invoked that e.g., perform data transformations like standardization and replacing NaN and invalid numbers with 0. Then, `RandomizedSearchCV()` is set up, including the search grid and the target scorers to be optimized for. The parameters depend on model technology and thus differ between model 1 (lgbm) and model 2 (lr).

The search grid for model 1 (lgbm) is:

- n_estimators ∈ {100, 300, 500, 1000}
- boosting_type ∈ {gbdt, dart}
- learning_rate ∈ {0.0001, 0.01, 0.1, 0.5}
- num_leaves ∈ {30, 50, 100}
- reg_alpha ∈ {0, 0.1, 0.5, 1, 3, 5}
- scale_pos_weight: 7.71[19]

The search grid for model 2 (lr) is:

- penalty: l1
- C: `np.logspace(-3, 2, 50)`
- solver ∈ {liblinear, saga}

Afterwards, the models are trained with the best identified hyperparameters (best estimator) on the train batch of each of the five development splits (split_0 to split_4). The results are outputted under ~/output whereas *cv_results.csv* contains each of the 50 CV run results, *results.json* contains the predictions of the best estimator using the defined scorers and *best_estimator.pkl* is the Python object for the best estimator. Additionally, *model_repetition_{0-4}.pkl* is outputted as well, representing the Python objects trained on each of the five development splits.

In addition to the aforementioned corrections in the *run_lgbm_together.sh* and *run_lr_together.sh* scripts, the authors of this paper encountered several issues while attempting to reproduce the results for the latter. The hyperparameter search for model 1 (lgbm) was successfully conducted on all the four datasets. However, for model 2 (lr) only the execution on the HiRID dataset was successful. For all other datasets, besides runtimes of over 328h respectively 13 days, the models did not converge.

For training the DL models, the attn model 3 and the gru model 4, Moor et al. [6] utilize the *PyTorch* toolkit in combination with the *Weights & Biases (W&B)* platform.[20] Here the files in scope are:

- ~/scripts/wandb/submit_job.sh script that was rewritten as ~/scripts/wandb/submit_sweep.sh
- ~/scripts/run_pytorch_rep.sh new script file
- ~/src/torch/train_model.py referenced Python source file
- ~/src/torch/models/__init__.py referenced Python source file
- ~/src/torch/models/base_model.py referenced Python source file
- ~/src/torch/models/attention_model.py referenced Python source file
- ~/src/torch/models/recurrent_model.py referenced Python source file
- ~/src/torch/datasets.py referenced Python source file
- ~/src/sklearn/utils/validation.py referenced Python source file

As was necessary for the non-DL models, several adjustments were required for a successful execution of the code for the DL models:

- *Removed LSTMModel and RNNModel since those classes are not defined*
- *Commenting of the … features_small_cache … line and adding the line below …features_small …for all four datasets*
- *Added imputation and normalization for each dataset using the prepared Normalize() and Impute() classes*
- *Manual specification of monitoring outputs in PyTorch Lightning due to version 2.5 requirement*
- *Changed 'lengths' to 'lengths.cpu()' to avoid error due to GPU usage*
- *Adjusted pl.Trainer() due to PyTorch Lightning version 2.5 requirement*
- *Created script to acquire the necessary five repetitions*

To ascertain the optimal hyperparameters for each DL model and dataset combination, eight Sweeps were configured in W&B, as can be seen in table 5.

TABLE 5. Weights & biases sweep configuration.

Sweep Name	Run Count	Compute Time in Days
attn_AUMC	502	9
attn_EICU	147	11
attn_Hirid	273	1
attn_MIMIC	238	13
gru_AUMC	512	8
gru_EICU	140	8
gru_Hirid	288	8
gru_MIMIC	187	8

Source: Authors

For each Sweep, the platform randomly selects a configuration from the following hyperparameter space:

- batch_size ∈ {16, 32, 64, 128}
- d_model ∈ {32, 64, 128, 256}
- dropout ∈ {0.3, 0.4, 0.5, 0.6, 0.7}
- learning_rate ∼ $U(ln(0.0001), ln(0.001))$
- weight_decay ∈ {30, 50, 100}
- reg_alpha ∈ {0.1, 0.01, 0.001, 0.0001}

[18] The index features "stay_id" and "stay_time" are not used for ML training and hence not reported by Moor et al. [5, Suppl. pp. 2-3] in their feature count.

[19] This parameter is not part of the set that is varied to find the best estimator. It is a constant determined by the dataset coming from the calculation of $\frac{1-\text{prevalence}}{\text{prevalence}}$.

[20] Weights & Biases [30]' "mission is to build the best tools for AI developers." Specifically relevant for this reproduction study is the platform's *Sweep* capability that allows "[visualizing] all … hyperparameter … [searches] in one unified place" [31].

Such a configuration constitutes a *Run* that is executed in PyTorch on the first training split (rep = 0) of the development dataset. After the successful training, the metric *online_val/loss* is calculated on the first validation split (rep = 0) of the development dataset. The Run with the lowest value of the metric for each Sweep is chosen as the ideal hyperparameter combination for each model and dataset. Next, using those best hyperparameters, each model and dataset was trained on the remaining four trainings splits (rep $\in [1, 4]$) to acquire all five repetitive runs.

For the evaluation of all models, the following files are used:

- ∼/scripts/eval_sklearn_create_subsamples.sh new script file
- ∼/scripts/eval_sklearn.sh script file
- ∼/scripts/wandb/submit_evals.sh script file
- ∼/scripts/eval_torch.sh script file
- ∼/scripts/plots/gather_data.py referenced Python source file
- ∼/scripts/plots/plot_roc.py referenced Python source file
- ∼/scripts/plots/plot_scatterplots.py referenced Python source file
- ∼/scripts/map_model_to_result_files.py script file
- ∼/scripts/pool_predictions.sh script file
- ∼/scripts/eval_pooled.sh script file
- ∼/scripts/make_heatmap.py referenced Python source file

There were also adjustments necessary to get a successful evaluation:

- *Created script to acquire subsamples with target sepsis prevalence of 0.188 of the split files*
- *Commented the if 'wandb_run' ...section because 'model_params' does not exist in the preds_max_pooled prediction json files*
- *In case the dictionary key 'dataset_train' is an array (list), the first item is taken from the array (to work with a dictionary)*
- *Commented the assert np.allclose() line because of assertion errors when running on GPU due to less precise CUDA calculations*

J. REPRODUCTION: RESULTS

The results of this reproduction study are given in table 6 and table 7. Focusing on the self-attention model 3, when validating on the same dataset (internal scenario), the authors of this paper achieved an AUC of 0.797 (95% CI, 0.789-0.805) and a PPV of 39.4 (95% CI, 33.2-45.7) at the same fixed 80% TPR and sepsis-case prevalence harmonization of 18.8%. This corresponds to $\approx$ 1.5 (95% CI, 1.2-2.0) FP per 1 TP.

When using the prediction pooling scenario that is validated externally across all datasets on model 3, the authors of this paper achieved an AUC of 0.717 (95% CI, 0.693-0.740) and a PPV of 28.3 (95% CI, 24.5-32.0) at 80% TPR and 18.8% sepsis-case prevalence harmonization. This corresponds to $\approx$ 2.5 (95% CI, 2.1-3.1) FP per 1 TP.

For the third reported metric from Moor et al. [5, p. 5], the lead time to sepsis onset, the authors of this paper were unable to reproduce meaningful values (for any of the four models and four datasets). When plotting the "proportion of TPs raised before X hours," as part of the described evaluation pipeline, for any decision threshold, the proportion stayed constantly at 0, leading to an earliness median of 0. This occurred with all models and datasets.

K. REPRODUCTION: PERTURBATION EXPERIMENT

To examine the robustness of the self-attention model 3, which is also suggested in question 30 of our catalog, minor perturbations in the input data were introduced as described in the following experiment.

The first parquet file (*part.0.parquet*) of each dataset existing in ∼/datasets/{mimic, eicu, hirid, aumc}/data/parquet/{features_small} was taken under scrutiny. The file contains 29, 37, 52 and 31 patient-ICU stays (for MIMIC-III, eICU, HiRID and AUMC). It includes all the feature values, as well as the sepsis onset target value per stay_id and stay_time. The values of the 59 raw clinical and lab observation features were modified. Random numbers from a Gaussian distribution were drawn having $\mu = 0$ and σ as the calculated standard deviation of the feature from the respective parquet file. Next the result was added to the original feature value.

Then the respective split files existing under ∼/datasets/splits/splits_{mimic, eicu, hirid, aumc}.json were adjusted so that only the previously modified stay_ids were contained in the respective test sets. Afterwards model 3 (attn), the best identified model, trained on the AUMC dataset was tasked to predict the sepsis onset for the selected patient-ICU stays using the modified input data records, as well as using the original, unmodified input data records. This was done five times with all the model repetitions (trained on all five splits of the development data set).

As is visible in table 8, using the perturbated AUMC parquet file with the AUMC attn model, an AUC of 0.934 (95% CI, 0.920-0.949) is achieved, respectively an AUC of 0.799 (95% CI, 0.756-0.843) when averaging over all other datasets except AUMC with the AUMC attn model.

Comparing this to the unmodified AUMC parquet file, an AUC of 0.933 (95% CI, 0.923-0.943) is achieved when testing internally and an AUC of 0.788 (95% CI, 0.743-0.833) in the external test case.

To be fully transparent and allow a repetition of our evaluation steps, relevant files and documents that were created as part of this reproduction study are made public on OSF[21] (see section B in the appendix).

[21] "OSF is a free, open platform to support ... research and enable collaboration" [32].

TABLE 6. Reproduction study area under the curve (AUC) results.

Model Name	AUC Int.	Std.	AUC Ext.	Std.	AUC Pooled	Std.
Model 1 (lgbm)	0.798	0.002	0.635	0.011	0.663	0.010
Model 2 (lr)*	0.792	0	0.697	0.004	-	-
Model 3 (attn)	0.797	0.004	0.682	0.021	0.717	0.012
Model 4 (gru)	0.808	0.003	0.683	0.014	0.716	0.012

*Model 2 (lr) did only converge on the HiRID dataset and thus the results are not comparable with the other models.
Source: Authors

TABLE 7. Reproduction study positive predictive value (PPV) at 80% sensitivity (TPR) results.

Model Name	PPV Int.	Std.	PPV Ext.	Std.	PPV Pooled	Std.
Model 1 (lgbm)	36.0	1.1	22.4	0.7	24.0	0.6
Model 2 (lr)*	34.5	0.2	23.1	0.3	-	-
Model 3 (attn)	39.4	3.1	26.2	2.2	28.3	1.9
Model 4 (gru)	37.2	1.2	24.1	0.9	27.1	1.2

*Model 2 (lr) did only converge on the HiRID dataset and thus the results are not comparable with the other models.
Source: Authors

TABLE 8. Perturbation experiment area under the curve (AUC) results.

Setting	AUC Int.	Std.	AUC Ext.	Std.
Model 3 (attn) with Perturbated Input Data Trained on AUMC	0.934	0.017	0.799	0.086
Model 3 (attn) with Original Input Data Trained on AUMC	0.933	0.012	0.788	0.089

Source: Authors

IV. DISCUSSION

In this paper we pursued two goals: First, we sought to assess the feasibility of applying our ML auditing core criteria catalog [4, p. 9] to an existing ML development project. This also includes a meta layer to evaluate the usefulness of doing so. Second, we wanted to ascertain the ease by which we can reproduce the steps and results described by Moor et al. [6] in their sepsis project.

Applying the 30 questions contained in the categories *Conceptual Basics*, *Data & Algorithm Design* and *Assessment Metrics* was straightforward. By using Moor et al. [5]'s paper, their paper supplement, and their GitHub repository [6], we could quickly determine if a question was addressed or not. It proved to be more difficult to differentiate between the evaluation categories *fully addressed* and *partially addressed*. Here our criteria was to use the first category when all aspects of a question were covered and the latter, if only some aspects were covered.

Our ML auditing catalog was designed as a "core criteria catalog" that is seen as a "starting point for the development of evaluation strategies" [4, p. 39962]. As such, the questions were selected and formulated in a rather generic way, allowing them to be applied to a broad spectrum of ML scenarios. For example, question 29 asks "did you conduct extensive performance testing according to the agreed metrics?" [4, 39962]. Metrics like the F1-Score or the Area Under the Curve (AUC) are mentioned as common examples but not dictated to be used. It depends on the concrete model type and application context, which metrics make sense and were discussed with the relevant stakeholders.

The proposed radar diagram visualization given in figure 1 allows an auditor to quickly identify strengths and weaknesses of an ML algorithm development or implementation project. He could also construct an auditing database with various ML algorithm projects organized according to different dimensions (e.g., algorithm maturity/product development stage, medical application group, medical application area, ML model family).

The specific rating results reveal that questions concerning risk management or ethical concepts (as part of the *Audit Process* catalog subcategory) are not covered in the sepsis project. We acknowledge that the scope of Moor et al. [5]'s paper is on the conceptual level, representing rather the profile of a development project than that of an implementation project. However, we would like to highlight the importance of both topics when seeking to implement the sepsis prediction algorithm. Using the categorization of risk management strategies of question 4 from our catalog, a "proactive" strategy can be to utilize "ethical hacker teams" who "produce problematic inputs of natural or malicious type" [4, p. 39955]. A "reactive" strategy e.g., constitutes a "'kill switch' … allowing a human operator to shut down the ML algorithm in case of misbehavior" [4, p. 39959]. An important ethical concept that could be followed is "'ethics by design,'" where the "ML model is designed around ethical concepts …. making sure that human rights are not violated" [4, p. 39957]. Those are also mandated in policies like the EU AI Act that e.g., requires "'transparency and provision of information to users, [and] human oversight,'" as well as the EU General Data Protection Regulation (GDPR) that "demands explanations of decisions [of ML algorithms/AI products]" [4, p. 39956].

It was immensely helpful that Moor et al. [6] provided the entire code in their GitHub branch *multicenter-sepsis-master* as well as an informative *README.md* file. Nevertheless, the reproduction of the ML algorithm of the sepsis project and its evaluation proved to be challenging due to several reasons. The effort required to achieve the postulated "end-to-end reproducibility of all results" far exceeded our initial expectation [5, p. 12].

Ascertaining the correct *ricu* version in R or adjusting the file/folder references in the Python script files was easy. However, we also encountered errors that forced us to inquire the logic and dependencies of Moor et al. [6]'s Python code located in the *src* folder. Examples here were the change of the outcome variable to *is_case*, the removal of the irrelevant *LSTMModel* and *RNNModel* and the uncommenting of code dealing with missing or NaN values. We utilized the Python debugger and many breakpoints to reverse engineer what is happening internally within Moor et al. [6]'s code to find solutions to the described errors. Another type of error encountered during the reproduction process was related to the run time itself. For instance, on our 24-core VM, besides various modifications in the parameter setup, we were unable to achieve convergence for three out of four datasets with model 2 (lr), even after multiple runs, with each requiring nearly 14 days of computing time.

Given the available infrastructure of our H100 GPU server, we needed to upgrade to a later PyTorch version, which in turn necessitated upgrades in dependent packages. Additionally, it was necessary to rewrite how the *LightningModule* captures monitoring metrics like *online_val/loss* [33].

Comparing our obtained results for model 3 (attn) with the postulated target metrics from Moor et al. [5, pp. 5-6], there was a deviation of -5.87% for the AUC and a deviation of -6.10% for the PPV at 80% TPR on average in the internal validation scenario. When executing the prediction pooling scenario with external validation averaged across datasets, we achieved a deviation of -5.83% for the AUC and a deviation of -11.03% for the PPV at 80% TPR.

This outcome indicates that, in general, we were able to reproduce the magnitude of Moor et al. [5, pp. 5-6]'s reported performance metrics. Also in our case, model 3 (attn) emerged as best performing, closely followed by deep learning model 4 (gru) and non-deep learning model 1 (lgbm). Model 3 (attn) also proved to be quite resilient to input data manipulation, as our perturbation experiment showed. The AUC only changed by 0.14% in the internal setting, and by 1.45% in the external setting.

We suspect that the divergent AUC and PPV values can be explained by our decision to not perform any tuning or fine tuning (step 3 and 7 of Moor et al. [5, Suppl. pp. 5-6]'s experimental setup) after the best hyperparameters were identified on the first split of the training fold.

Model 3 (attn) is a deep neural network (DNN) model, which is inherently opaque, and often, "clinician trust [is built] through transparency and accessibility" [34, p. 6]. Therefore, it might be advisable to gain a deeper understanding of the model's internal mechanics by employing explainable AI (XAI) related tools and methods suggested in Schwarz et al. [4, pp. 39960-39961], to conduct a more detailed analysis of the performance differences.

The invalid results for the third metric, the lead time to sepsis onset, suggest a probable existence of an algorithmic error in the published *multicenter-sepsis-master* code base [6]. We analysed the evaluation pipeline and traced the issue back to the shifting of sepsis labels. For example, when looking at patient-ICU stay record 5616 from the AUMC dataset, which exists in the file …/prediction_output/lgbm_aumc_aumc_classification_middle_cost_5_50_iter_rep_0_full_set.json, the list of labels (per stay_time) looks like: [1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, …]. However, the (unpublished) file we received from Moor et al. [6] after a personal correspondence looks like: [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 1.0, …]. Since the algorithm identified a sepsis case at the 7th stay_time entry, the expected behavior is that all sepsis labels of previous stay_time entries are shifted to 0, which does not take place in our reproduction study. The identification of the respective Python module and function that causes this behavior would require an in-depth analysis and additional reverse engineering. Since the reproduction study already consumed much more time than anticipated, and we see our role as an external auditor primarily in the application of our ML auditing catalog, we did not pursue any further steps. However, we encourage other groups to do so[22] and would recommend Moor et al. [6] to verify that the latest code revision has been published to the *multicenter-sepsis-master* branch.

As limitations we must note that if our ML auditing core criteria catalog is used to assess different ML project use cases, only similar profiles should be compared. For example, whether the intention is at the exploratory, development level or already constitutes a concrete implementation study at a hospital site.

Our acquired weighted Cohen's kappa coefficient of $\kappa = 0.51$, even though after Fleiss [35, p. 218]' interpretation constituting a "fair to good" agreement, still shows room for improvement. This is also indicated by questions 14, 27 and 28 that were assessed quite differently by both auditors. When we compare the response justifications, it becomes evident that the first auditor interpreted the question more loosely, considering also other, rather informal and not explicitly mentioned examples, whereas the second rater interpreted the question more strictly, insisting on the existence of the exactly mentioned examples.

[22] We will also put the original result files received from Moor et al. [6] in our OSF storage (see section B in the appendix).

These limitations can be mitigated if more groups (multiple auditors) would apply our catalog to various ML development or implementation projects (ideally first within the healthcare sector). Because then a kind of "catalog application guideline" with practical recommendations could be established. This guideline could also provide more specific instructions for the auditors on how rather openly formulated questions should be interpreted or can be operationalized when assessing an ML algorithm project.

V. CONCLUSION

We see our ML auditing core criteria catalog as a tool that is practical and lean, requiring little effort to use. As explained in Schwarz et al. [4, pp. 39962-39963], we tried to condense the most important aspects related to AI auditing in three categories containing together only 30 questions, which are easy to apply. This differentiates our catalog to other frameworks like the "AI Risk Management Framework (AI RMF)" or the "ISO/IEC 42001" [36], [37].

The first aims at "[helping] manage the many risks of AI and promote trustworthy and responsible development and use of AI systems" [36, p. 2]. The second goes in a similar direction, asking to create an "AI management system within the context of an organization" [37, p. v]. "Addressing risks related to the design and operation of AI systems" constitutes such a system's main purpose [37, p. 17]. This means that in contrast to our two questions contained in the subcategory "Risk Management," the AI RMF and the ISO/IEC 42001 delve much deeper into this area. Thus, especially for major corporations involving many stakeholders or high-risk application use cases, both frameworks can be used complementary to our catalog.

With this paper we wanted to conduct a practical test of our 30-question ML auditing core criteria catalog on a real-world ML algorithm project example. We were motivated by identifying strengths and weaknesses of our catalog and reveal potential caveats when performing a reproduction study. Our activities may also assist development or implementation teams in preparing for future, legally mandated audits of their newly created ML algorithms/AI products.

APPENDIX A
SECOND RATER AUDIT PROTOCOL

The audit protocol containing a justification of the chosen response category for each question by the second rater can be accessed via OSF at: https://osf.io/bca7n.

APPENDIX B
SEPSIS REPRODUCTION FILES

Our internal protocol of setting up and executing the sepsis project reproduction study, as well as our code file modifications and the generated result files can also be accessed via OSF at: https://osf.io/bca7n. There are two Docker[23] images available containing the described reproduction environment.

APPENDIX C
FORMULAE

The "weighted Cohen's Kappa" coefficient is given in equation (1).

$$\kappa = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} x_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} m_{ij}} \tag{1}$$

k denotes the number of response categories, w, x and m are $k \times k$ element matrices of weights, observed frequencies and expected frequencies.

ACKNOWLEDGMENT

The authors would like to thank *Lucas Miranda* and *Karsten Borgwardt* from the Machine Learning and Systems Biology research group at the Max Planck Institute of Biochemistry in Munich. Lucas provided firsthand insights from Moor et al. [5] and offered support for their sepsis project reproduction study. They also thank *Michael Moor* from the Department of Biosystems Science and Engineering at the Federal Institute of Technology Zurich. He shared some of his original results that were helpful in investigating reproduction differences.

Lastly, they would like to thank *Paul Elbers* from the Department of Intensive Care Medicine and Center for Critical Care Computational Intelligence at the Amsterdam University Medical Center. He is a practicing intensivist who granted them access to the AUMC v1.0.2 dataset.

REFERENCES

[1] European Commission. (2021). *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[2] European Commission. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text With EEA Relevance). PE/24/2024/REV/1*. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L202401689

[3] European Commission. (2012). *Charter of Fundamental Rights of the European Union*. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT

[4] M. Schwarz, L. C. Hinske, U. Mansmann, and F. Albashiti, "Designing an ML auditing criteria catalog as starting point for the development of a framework," *IEEE Access*, vol. 12, pp. 39953–39967, 2024.

[5] M. Moor, N. Bennett, D. Plečko, M. Horn, B. Rieck, N. Meinshausen, P. Bühlmann, and K. Borgwardt, "Predicting sepsis using deep learning across international sites: A retrospective development and validation study," *eClinicalMedicine*, vol. 62, Aug. 2023, Art. no. 102124.

[6] M. Moor, N. Bennett, D. Plečko, M. Horn, and B. Rieck. (2023). *GitHub—BorgwardtLab/Multicenter-Sepsis*. [Online]. Available: https://github.com/borgwardtlab/multicenter-sepsis

[7] Supreme Audit Institutions of Finland, Germany, The Netherlands, Norway and the U.K. (2020). *Auditing Machine Learning Algorithms*. [Online]. Available: https://www.auditingalgorithms.net/auditing-ml.pdf

[8] M. S. Matell and J. Jacoby, "Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity," *Educ. Psychol. Meas.*, vol. 31, no. 3, pp. 657–674, Oct. 1971.

[9] H. Jae Jeong, "'The level of collapse we are allowed: Comparison of different response scales in safety attitudes questionnaire,'" *Biometrics Biostatistics Int. J.*, vol. 4, no. 4, pp. 128–134, Sep. 2016.

[23] "Docker is an open platform for developing, shipping, and running applications. Docker enables … to separate …applications from … infrastructure" [38].

[10] K. L. Gwet, *Handbook of Inter-Rater Reliability*, 4th ed. Gaithersburg, MD, USA: Advances Analytics, 2014.

[11] The Turing Way Community. (2022). *Table of Definitions for Reproducibility*. [Online]. Available: https://book.the-turing-way.org/reproducible-research/overview/overview-definitions#reproducible-matrix

[12] A. Wong, E. Otles, J. P. Donnelly, A. Krumm, J. McCullough, O. Detroyer-Cooley, J. Pestrue, M. Phillips, J. Konye, C. Penoza, M. Ghous, and K. Singh, "External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients," *JAMA Internal Med.*, vol. 181, no. 8, pp. 1065–1070, Jun. 2021.

[13] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith, R. S. Hotchkiss, M. M. Levy, J. C. Marshall, G. S. Martin, S. M. Opal, G. D. Rubenfeld, T. van der Poll, J.-L. Vincent, and D. C. Angus, "The third international consensus definitions for sepsis and septic shock (Sepsis-3)," *J. Amer. Med. Assoc.*, vol. 315, no. 8, pp. 801–810, Feb. 2016.

[14] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Dec. 2017, pp. 3149–3157. [Online]. Available: https://dl.acm.org/doi/10.5555/3294996.3295074

[15] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B, Stat. Methodology*, vol. 58, no. 1, pp. 267–288, Jan. 1996. [Online]. Available: http://www.jstor.org/stable/2346178

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 5998–6008, Jun. 2017. [Online]. Available: https://proceedings.neurips.cc/paperfiles/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[17] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proc. 8th Workshop Syntax, Semantics Struct. Stat. Transl.*, 2014, pp. 103–111.

[18] S. E. Harpe, "How to analyze Likert and other rating scale data," *Currents Pharmacy Teaching Learn.*, vol. 7, no. 6, pp. 836–850, Nov. 2015.

[19] Python Software Foundation. (2022). *Python Release Python 3.7.16*. [Online]. Available: https://www.python.org/downloads/release/python-3716/

[20] A. Johnson, T. Pollard, and R. Mark, "MIMIC-III clinical database (version 1.4)," PhysioNet, RRID:SCR_007345, MIT Lab. Comput. Physiol., Cambridge, MA, USA, Tech. Rep. C2XW26, 2016. [Online]. Available: https://doi.org/10.13026/C2XW26

[21] T. J. Pollard, A. E. W. Johnson, J. Raffa, and O. Badawi, "The eICU collaborative research database (version 2.0)," PhysioNet, RRID:SCR_007345, MIT Lab. Comput. Physiol., Cambridge, MA, USA, Tech. Rep. C2WM1R, 2019. [Online]. Available: https://doi.org/10.13026/C2WM1R

[22] M. Faltys, M. Zimmermann, X. Lyu, M. Hüser, S. Hyland, G. Rätsch, and T. Merz, "HiRID, a high time-resolution ICU dataset (version 1.1.1)," PhysioNet, RRID:SCR_007345, MIT Lab. Comput. Physiol., Cambridge, MA, USA, Tech. Rep. nkwc-js72, 2021. [Online]. Available: https://doi.org/10.13026/nkwc-js72

[23] P. W. G. Elbers, "AmsterdamUMCdb v1.0.2," DANS Data Station Life Sci., Amsterdam Univ. Med. Center Amsterdam, The Netherlands, Tech. Rep. dans-22u-f8vd_2019, 2019. [Online]. Available: https://doi.org/10.17026/dans-22u-f8vd

[24] MIT Laboratory for Computational Physiology. (2014). *CITI Course Instructions*. [Online]. Available: https://physionet.org/about/citi-course/

[25] N. Bennett, D. Plečko, I.-F. Ukor, N. Meinshausen, and P. Bühlmann, "Ricu: R's interface to intensive care data," *GigaScience*, vol. 12, Jun. 2023, Art. no. giad041.

[26] M. Klik. (2024). *FST: Lightning Fast Serialization of Data Frames*. [Online]. Available: https://www.fstpackage.org/

[27] Apache Parquet. (2024). *Overview*. [Online]. Available: https://parquet.apache.org/docs/overview/

[28] scikit Learn. (2024). *User Guide*. [Online]. Available: https://scikit-learn.org/stable/userguide.html

[29] Linux Foundation. (2024). *PyTorch Documentation—PyTorch 2.4 Documentation*. [Online]. Available: https://pytorch.org/docs/stable/index.html

[30] Weights & Biases. (2024). *About Us*. [Online]. Available: https://wandb.ai/site/company/about-us/

[31] Weights & Biases. (2024). *Sweeps*. [Online]. Available: https://wandb.ai/site/sweeps/

[32] Center for Open Science. (2025). *OSF*. [Online]. Available: https://osf.io/

[33] A. Wälchli. (2023). *Lightning 2.0: Fast, Flexible, Stable*. [Online]. Available: https://github.com/Lightning-AI/pytorch-lightning/releases/tag/2.0.0#bc-changes-pytorch

[34] C. Ding, T. Yao, C. Wu, and J. Ni, "Advances in deep learning for personalized ECG diagnostics: A systematic review addressing inter-patient variability and generalization constraints," *Comput. Biol. Med.*, vol. 271, Jun. 2025, Art. no. 117073. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0956566324010807

[35] J. L. Fleiss, *Statistical Methods for Rates Proportions* (Wiley Series in Probability and Mathematical Statistics), 2nd ed., Hoboken, NJ, USA: Wiley, 1981.

[36] E. Tabassi, "Artificial intelligence risk management framework (AI RMF 1.0)," U.S. Dept. Commerce, Nat. Inst. Standards Technol., Washington, DC, USA, Tech. Rep. NIST AI 100-1, 2023.

[37] International Organization for Standardization. (2023). *ISO/IEC 42001: 2023*. [Online]. Available: https://www.iso.org/standard/81230.html

[38] Docker Inc. (2024). *What is Docker?* [Online]. Available: https://docs.docker.com/get-started/docker-overview/

MARKUS SCHWARZ received the B.Eng. degree in media technology and economics, the M.Sc. degree in media economics, and the M.Sc. degree in business information systems engineering. He is currently pursuing the Ph.D. degree with the Medical Data Integration Center (MeDIC LMU), University Hospital LMU Munich, Germany. He is the Team Lead Global Master Data Management of SYNLAB HQ, Munich, Germany.

LUDWIG CHRISTIAN HINSKE received the M.D. degree from LMU Munich, and the Master of Science degree in biomedical informatics from MIT. He is currently the Director of the Institute for Digital Medicine, University Hospital Augsburg, Germany. He is also with the Department of Anesthesiology, University Hospital LMU Munich, Germany. His research interest includes federated analytics to AI applications in medicine.

ULRICH MANSMANN received the Diploma (Univ.) and Dr.rer.nat. degrees in mathematics. He is currently the Director of the Institute for Medical Information Processing, Biometry and Epidemiology (IBE), LMU Munich, Germany. His research interests include clinical trials, clinical epidemiology, prognostic and predictive models, computational biology, survival and event data, complex multivariate models in clinical epidemiology, Bayesian modeling, computer intensive Bayesian methods, and statistical methods in molecular medicine.

FADY ALBASHITI received the M.Sc. and Dr.sc.hum. degrees in medical informatics. He is currently the CEO of the Medical Data Integration Center (MeDIC LMU), University Hospital LMU Munich, Germany. His research interests include data integration, data platforms that provide real-world data (RWD) to generate real-world evidence, and on offering Digital Health Intelligence products around data and AI for research and health care.

References

Bizzego, A., Bussola, N., Chierici, M., Maggio, V., Francescatto, M., Cima, L., Cristoforetti, M., Jurman, G., & Furlanello, C. (2019). Evaluating reproducibility of AI algorithms in digital pathology with DAPPER [Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't The authors have declared that no competing interests exist.]. *PLoS computational biology*, *15*(3), e1006269. https://doi.org/10.1371/journal.pcbi.1006269

Booth, W. C., Colomb, G. G., & Williams, J. M. (2008). *The craft of research* (3rd ed.). University of Chicago Press.

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler: Mit 87 Tabellen* (4., überarb. Aufl.) [Bortz, Jürgen (Verfasser) Döring, Nicola (Verfasser)]. Springer.

Bott, E. (2025). *The Microsoft 365 Copilot launch was a total disaster* (ZDNET, Ed.). Retrieved March 20, 2025, from https://www.zdnet.com/home-and-office/work-life/the-microsoft-365-copilot-launch-was-a-total-disaster/

Center for Open Science. (2025). *OSF*. Retrieved April 11, 2025, from https://osf.io/

Docker Inc. (2024). *What is Docker?* Retrieved April 11, 2025, from https://docs.docker.com/get-started/docker-overview/

Foltin, C. (2023). *FreeMind*. Retrieved March 12, 2025, from https://freemind.sourceforge.io/wiki/index.php/Main_Page

German Society for Medical Informatics, Biometry and Epidemiology e. V. (2024). *Gesundheit-gemeinsam.de - cooperation conference 2024*. Retrieved March 12, 2025, from https://gesundheit-gemeinsam.de/en/

Global Market Insights Inc. (2024). *Artificial Intelligence in Healthcare Market: By Offering, By Application, By End Use - Global Forecast, 2024 to 2032*. Global Market Insights Inc. Retrieved March 19, 2025, from https://www.gminsights.com/industry-analysis/healthcare-artificial-intelligence-market

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (Fourth edition). Advances Analytics LLC.

Hangzhou DeepSeek Artificial Intelligence Basic Technology Research. (2025). *GitHub - deepseek-ai/DeepSeek-R1*. Retrieved March 20, 2025, from https://github.com/deepseek-ai/DeepSeek-R1

Healthcare Tech. (2019). *AI's impact on healthcare industry*. https://www.healthcaretechoutlook.com/news/ai-s-impact-on-healthcare-industry-nid-925.html

Institute of Electrical and Electronics Engineers. (2024). *Learn More About IEEE Access*. Retrieved March 12, 2025, from https://ieeeaccess.ieee.org/about-ieee-access/learn-more-about-ieee-access/

Kuan, R. (2019). *Adopting AI in Health Care Will Be Slow and Difficult* (Harvard Business Review, Ed.). Retrieved April 30, 2025, from https://hbr.org/2019/10/adopting-ai-in-health-care-will-be-slow-and-difficult

Lumivero. (2025). *Citavi - The Only All-in-One Writing and Referencing Solution*. Retrieved March 12, 2025, from https://lumivero.com/products/citavi/

Manatal. (2025). *AI Prompt Engineer Job Description for Recruiters*. Retrieved March 20, 2025, from https://www.manatal.com/job-description/ai-prompt-engineer-job-description

Mannheim University of Applied Sciences. (2025). *MIRACUM-DIFUTURE-Kolloquium*. Retrieved March 12, 2025, from https://sites.google.com/master-bids.de/miracum-difuture-kolloquium/home

Mayring, P. (2000). Qualitative Content Analysis. *Forum: Qualitative Social Research*, *1*(2). https://doi.org/10.17169/fqs-1.2.1089

Microsoft. (2025). *GitHub - microsoft/vscode: Visual Studio Code*. Retrieved May 7, 2025, from https://github.com/microsoft/vscode

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook / Matthew B. Miles, A. Michael Huberman* (2nd ed.). SAGE.

MIM Software Inc. (2025). *Zero-Click Auto-Contouring: Contour ProtégéAI+™*. Retrieved March 20, 2025, from https://www.mimsoftware.com/radiation-oncology/contour-protegeai-plus

Mistral AI. (2024). *Large Enough: Mistral Large 2*. Retrieved March 20, 2025, from https://mistral.ai/news/mistral-large-2407

Mistral AI. (2025). *About us*. Retrieved March 20, 2025, from https://mistral.ai/about

Moor, M., Bennett, N., Plečko, D., Horn, M., Rieck, B., Meinshausen, N., Bühlmann, P., & Borgwardt, K. (2023a). Predicting sepsis using deep learning across international sites: a retrospective development and validation study. *EClinicalMedicine*, *62*, 102124. https://doi.org/10.1016/j.eclinm.2023.102124

Moor, M., Bennett, N., Plečko, D., Horn, M., & Rieck, B. (2023b). *GitHub - BorgwardtLab/multicenter-sepsis*. Retrieved September 25, 2024, from https://github.com/borgwardtlab/multicenter-sepsis

National Information Standards Organization. (2025). *CRediT taxonomy – JATS4R*. Retrieved March 11, 2025, from https://jats4r.niso.org/credit-taxonomy

Ohlinger, M., & Carter, B. (2025). *How does Microsoft 365 Copilot work?* (Microsoft, Ed.). Retrieved March 20, 2025, from https://learn.microsoft.com/en-us/copilot/microsoft-365/microsoft-365-copilot-architecture

OpenAI. (2023). *Planning for AGI and beyond*. Retrieved March 20, 2025, from https://openai.com/index/planning-for-agi-and-beyond/

OpenAI. (2025). *OpenAI GPT-4.5 System Card*. Retrieved March 20, 2025, from https://cdn.openai.com/gpt-4-5-system-card-2272025.pdf

Schwarz, M., & Albashiti, F. (2025). *Making ML Algorithms Auditable: Presentation of a 30-Question ML Auditing Criteria Catalog as Guide: MIDI-Kolloquium 2025-02-25*. Retrieved May 6, 2025, from https://sites.google.com/master-bids.de/miracum-difuture-kolloquium/archiv/2025/2025-02-25_making-ml-algorithms-auditable

Schwarz, M., Hinske, L. C., Mansmann, U., & Albashiti, F. (2024a). Designing an ML Auditing Criteria Catalog as Starting Point for the Development of a Framework. *IEEE Access*, *12*, 39953–39967. https://doi.org/10.1109/ACCESS.2024.3375763

Schwarz, M., Hinske, L. C., Mansmann, U., & Albashiti, F. (2024b). Designing an ML Auditing Criteria Catalog as Starting Point for the Development of a Framework [German Medical Science GMS Publishing House]. *Gesundheit – gemeinsam. Kooperationstagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS), Deutschen Gesellschaft für Sozialmedizin und Prävention (DGSMP), Deutschen Gesellschaft für Epidemiologie (DGEpi), Deutschen Gesellschaft für Medizinische Soziologie (DGMS) und der Deutschen Gesellschaft für Public Health (DGPH)*. https://doi.org/10.3205/24GMDS105

Schwarz, M., Hinske, L. C., Mansmann, U., & Albashiti, F. (2025). ML Auditing and Reproducibility: Applying a Core Criteria Catalog to an Early Sepsis Onset Detection System. *IEEE Access*, *13*, 104899–104915. https://doi.org/10.1109/ACCESS.2025.3579631

Sina Corporation. (2025). *DeepSeek breaks news!* Retrieved March 20, 2025, from https://finance.sina.com.cn/jjxw/2025-02-01/doc-inehyqcx9694053.shtml

Smith, D., Jupudi, A., Mandalika, S., Payne, H., Hu, R., Buck, A., Lobo, A., Gupta, A., Saldanha, L., Coulter, D., Soysal, S., Jacobsen, L., Burden, T., Penna, M., Woitasen, D., Borys, A., Baumgartner, P., Peacock, R., Chin, L., … Gatimu, K. (2025). *Introduction to Microsoft Teams for admins - Microsoft Teams*. Retrieved March 20, 2025, from https://learn.microsoft.com/en-us/microsoftteams/teams-overview

Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK. (2020). *Auditing machine learning algorithms* [Whitepaper]. Retrieved October 13, 2021, from https://www.auditingalgorithms.net/auditing-ml.pdf

The LaTeX Project. (2025). *Introduction to LaTeX*. Retrieved March 12, 2025, from https://www.latex-project.org/about/

U.S. Food and Drug Administration. (2019). *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD): Discussion Paper and Request for Feedback*. Retrieved January 18, 2021, from https://www.fda.gov/media/122535/download

U.S. Food and Drug Administration. (2024). *Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices*. Retrieved March 20, 2025, from https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

U.S. Food and Drug Administration. (2025). *Artificial Intelligence and Machine Learning in Software*. Retrieved March 20, 2025, from https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

Appendix A

ML Auditing Core Criteria Catalog Questions

The questions of the ML auditing core criteria catalog printed below are taken from Schwarz et al. (2024a, pp. 39961–39962).

Conceptual Basics

AI Opportunities vs. AI Risks

1. "Is the expected $benefit * benefitProbability$ of a successful ML use case implementation greater than the $damage * damageProbability$ in case of failure?"
2. "Do you expect a productivity gain, improved quality or a new functionality compared to the current manual/non-ML process?"

Risk Management

3. "Are the roles and responsibilities (RACI[21]) and liabilities before, during and after the implementation clearly defined?"
4. "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?"

Methodology

5. "Have you aligned and agreed on the methodology with all project stakeholders (e.g., for implementation CRISP-DM and internal audit SMACTR)?"

[21] "RACI says that when working in teams it needs to be clear who is **R**esponsible for a given task, who is **A**ccountable especially if something goes wrong, who needs to be **C**onsulted for advice and who must be **I**nformed about the progress" (Schwarz et al., 2024a, p. 39961).

6. “Are the implications in case the ML use case falls in the ’high risk’ category of the EU AI Act understood?”
7. “Do you plan to make use of Data Sheets to describe the data collection process as well as the data properties?”
8. “Do you plan to create AI Model Cards/AI Fact Sheets to describe the model characteristics?”
9. “Do you plan to prepare AI Care Labels to instruct internal stakeholders how to use and ’treat’ the algorithm?”

Audit Process

10. “Have you established an internal advisory committee consisting of senior IT governance specialists and business/medical specialists who critically accompany the implementation (e.g., watch for sufficient documentation and methodology adherence)?”
11. “Do you ensure the ML implementation is not violating ethical concepts (’ethics by design’ is considered)?”
12. “Do you have protocols in place that allow independent, external auditors to critically review the ML use case implementation?”

Quality Assurance

13. “Did you perform a verification of the ML output behavior using a set of expected, representative inputs of the productive usage?”
14. “Did you perform a validation whether the project’s specification and stakeholders’ needs are met?”
15. “Do you think the ML model would pass an external AI Certification/AI Assurance case fulfilling the six components of trust: predictability, dependability, faith, consistency, utility and understanding?”
16. “Given inputs from different test users, does the ML model adhere to the principles of distributive, procedural and interactional justice?”
17. “Given inputs from different test users, does the ML model avoid differential prediction and intentional discrimination?”

Data & Algorithm Design

Data Properties

18. “Is the data generation process (DGP) of the training, testing and validation data set sufficiently known? Could there be unknown confounders or mediator variables influencing the observed data?”

19. “Does the training data capture relevant characteristics of the population in scope for the ML use case?”

20. “Are the required data quality dimensions (e.g., accuracy, consistency, completeness and currency) well understood and taken care of?”

21. “Are the procedures necessary for data cleansing and consolidation known, and is the understanding of data scales and references ranges given?”

Algorithm Design

22. “Is the difference between causality and correlation known? In the absence of known counterfactuals for each individual, population samples can only give associations with a certain strength (e.g., given by the Pearson’s correlation coefficient).”

23. “Did you apply Occam’s Razor principle for the model type selection? Meaning in case a black box model (e.g., DNN, NLP) is to be used, does it provide substantial benefit (e.g., accuracy) over a white box model (e.g., logistic regression, decision tree)?”

24. “Did you establish a correct ML use case hypothesis with concrete problem description and expected behavior (acceptance criteria, metrics, statistical testing results)?”

Assessment Metrics

Qualitative Assessment

25. “Are the model assumptions (e.g., how to deal with missing data, model type, hyperparameters) transparently described?”

26. “Did you establish a traceable log of those model assumptions/testing results being used during the whole development process?”

27. “Did you discuss with all stakeholders the strength of conclusions that can be drawn with the current model assumptions (and make sure the conclusions are appropriate)?”

Quantitative Assessment

28. “Did you determine the statistical properties of the training, testing and validation data set? For example, by using Variance Inflation Factor (VIF), Shapiro-Wilk Test (SWT) and Breusch-Pagan Test (BPT)?”
29. “Did you conduct extensive performance testing according to the agreed metrics? For example, using Receiver Operating Characteristics (ROC): creating the confusion matrix and calculating the F1-Score, Matthews Correlation Coefficient (MCC) or Area Under the Curve (AUC)?”
30. “Did you assess the resistance of the ML model’s output behavior to natural perturbation, for example, using Total Sobol’s Variance Ratio (TSVR) or Cosine Similarity Vector Pairs (CSVP)?”

Acknowledgements

Over the course of this five-year Ph.D. journey, I have met many people who were supportive, asked critical questions and provided useful contacts and resources.

It all started with Assoc. Prof. Dr. Jan Stratil who nurtured my idea of pursuing a Ph.D. and brought my initial attention to this Ph.D. program of LMU in Munich. A dialogue with Prof. Dr. Michael Laxy narrowed down the list of potential topics and guided me towards the necessary institutional structure of such an endeavor.

Prof. Dr. Ulrich Mansmann, who gave important academic advice throughout the whole project, was very affirmative of my idea and directed me to Dr. Fady Albashiti, who is heading the MeDIC$^{\text{LMU}}$ where my Ph.D. topic was eventually rooted. Not only did Dr. Albashiti help me in finding and formulating the exact topic, as my operational supervisor he was a close mentor on every concern of my journey over uncountable meetings and discussions. Prof. Dr. Ludwig Christian Hinske was encouraging about the topic, agreed to being my formal supervisor and also served as a mentor providing his unique perspective.

My colleagues at MeDIC$^{\text{LMU}}$ showed genuine interest in my Ph.D. topic, leading to many fruitful conversations. They also supported me in terms of administration, software, hardware, or conceptual concerns. To name a few: Elena Druidi, Martin Knauer, Nenad Stefanac, Reinhard Thasler, Konstantin Böll and Ardit Selfo.

I had many insightful and encouraging conversations with my fellow Ph.D. students, helping to overcome difficulties being part of such an undertaking. To also name a few: Stefan Buchka, Dr. Halimuniyazi Haliduola and Andrea Gutmann. The annual Ph.D. retreat provided a special opportunity for those conversations and was a highlight of the program itself.

The Ph.D. program would not be what it is today without the work of the program coordinators Dr. Annette Hartmann and Monika Darchinger who have been providing superb support for all student concerns. The same is true for Dr. Antje Hentrich from the PhD office for all regulatory matters and concerns around the submission process.

Finally, I would like to mention my line manager Alexander Knechtges at SYNLAB International GmbH in Munich, who gave me the necessary flexibility by reducing my working hours and allowing me to change the working days according to the needs of my Ph.D. project.

To all of you, and people I did not explicitly mention, I would like to express my sincere gratitude for accompanying me during my Ph.D. journey.