IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, October 2026
Impact Factor: 8.2

Developing a Practical ML Auditing Criteria Catalog: A Literature-Grounded Starting Point for Framework Design

Dhanashri Karale[1], Prof. P. V. Gaikwad[2], Dr. H. B. Jadhav[3]
[1] Third Year Computer Engineering Student, Adsul's Technical Campus, Chas, Maharashtra, India
[2] Project Guide, Department of Computer Engineering, Adsul's Technical Campus, Chas, Maharashtra, India
[3] Head of Department, Department of Computer Engineering, Adsul's Technical Campus, Chas, Maharashtra, India

Abstract: *Machine learning (ML) systems are increasingly used to support decisions in healthcare and other high-impact domains. However, a model's predictive performance alone does not establish that it is safe, fair, explainable, reliable, or appropriate for its intended use. Auditing therefore needs to examine the full lifecycle: the purpose and risks of a use case, the quality and representativeness of data, model design, testing evidence, and governance. This paper presents a structured, literature-grounded approach for organizing ML auditing criteria into a practical catalog that can serve as the starting point for a broader auditing framework. The structure follows three core groups identified in the source study: Conceptual Basics, Data & Algorithm Design, and Assessment Metrics. A qualitative literature-oriented approach is described, and an adapted checklist is provided to help teams identify evidence that should be documented before deployment and during monitoring. The paper does not claim new experimental results; it synthesizes the supplied research article into an implementation-oriented student paper. The proposed framework emphasizes traceability, stakeholder responsibility, data quality, fairness, validation, robustness, and continuous improvement.*

Keywords: Machine learning auditing, auditable AI, AI governance, model validation, data quality, fairness, accountability.

I. INTRODUCTION

Artificial intelligence (AI) and machine learning are being introduced into systems that influence decisions, services, and human wellbeing. In healthcare, ML may assist with classification, prediction, prioritization, or analysis of clinical data. These potential benefits also create responsibilities: teams must understand what the model is intended to do, what could go wrong, which people may be affected, and how its behavior will be evaluated. A model that performs well on a test dataset may still fail when data distributions change, when a subgroup is under-represented, or when users interpret its output incorrectly.

An ML audit is a systematic examination of a model and the processes surrounding it. It can include governance, documentation, data provenance, design decisions, quantitative testing, fairness, and the suitability of the model for its intended use. The research by Schwarz et al. proposes a 30-question core criteria catalog and organizes it into three broad categories. This paper adapts that structure into a concise, practical framework for planning an audit and identifying evidence required for responsible deployment [1].

II. BACKGROUND AND PROBLEM STATEMENT

A. Background

Auditable AI is concerned with whether an AI system can be examined using meaningful evidence. It is related to explainable AI, but the two concepts are not identical. Explainability focuses on making model behavior or outputs understandable. Auditing is broader: it may inspect governance, data, assumptions, testing, documentation, human

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, October 2026
Impact Factor: 8.2

oversight, and operational controls in addition to explanations. Interpretability refers to how readily a person can understand a model or its behavior. These concepts can support one another, but a clear explanation alone is not proof that a system is fair or dependable.

The source study reports a qualitative analysis of literature associated with "Auditable AI." The authors iteratively grouped relevant artifacts into categories and describe a mind map containing 12 main categories and approximately 800 artifacts. They identify Tools/Methods (203 artifacts), Definitions (159), Dimensions (82), and Philosophy/Ethics (82) among the largest categories. These are findings from the source article, not results generated independently for the present paper [1].

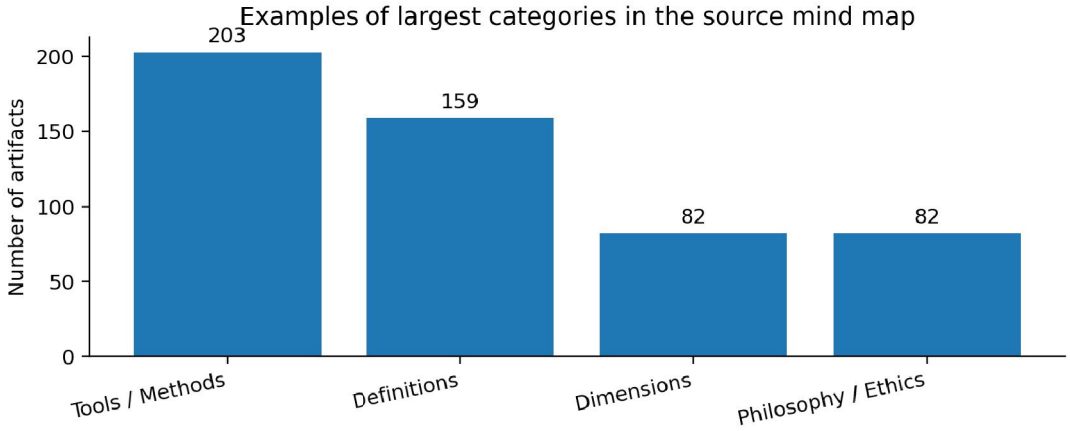


Fig. 1. Selected artifact-category counts reported by Schwarz et al. [1].

B. Problem Statement

Many ML projects are evaluated mainly through model accuracy or other performance scores. Such a narrow view can miss critical questions: whether the project addresses a valid need, whether training data represent the intended population, whether assumptions are documented, whether different groups experience unequal errors, and whether performance remains stable after deployment. Audit information may also be spread across technical notes, project plans, datasets, and test reports, making independent review difficult.

C. Need for the Proposed Work

A compact catalog can help teams convert broad principles into reviewable questions. It does not replace domain expertise, law, standards, or formal certification. Instead, it offers a common starting point for planning evidence, assigning responsibilities, and deciding when additional testing or expert review is required. The catalog must remain adaptable to the domain, risk level, intended users, and consequences of failure.

D. Aim and Objectives

Aim: To present a structured ML auditing criteria catalog as a foundation for developing a practical auditing framework.

- Organize audit considerations into conceptual, technical, and assessment categories.
- Identify documentation and evidence expected across the ML lifecycle.
- Highlight data quality, model choice, fairness, validation, and robustness checks.
- Provide an implementation workflow that supports traceable review and continuous improvement.

III. RELATED WORK AND RESEARCH METHODOLOGY

A. Related Work

The source article draws on research and guidance related to AI governance, fairness, quality assurance, explainability, testing, and audit practices. It refers to the need for a shared approach to evaluating AI systems and notes that available

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, October 2026
Impact Factor: 8.2

tools and methods are diverse. The catalog approach responds to this fragmentation by grouping individual artifacts into a manageable set of questions. Relevant related concepts include trustworthy AI self-assessment, ML auditing in healthcare, explainable AI, and technical assessment of classifiers [1].

B. Research Approach

This paper uses a literature-grounded synthesis of the supplied source article rather than an independent systematic review. The original study describes a qualitative scoping/content-analysis process, informed by methods associated with Mayring, Miles and Huberman, and Bortz and Döring. Relevant sources were compared, categorized iteratively, and consolidated into a core criteria catalog. The present paper retains that overall logic and reframes the output for student-level implementation planning [1].

C. Proposed Steps

Step	Activity	Expected output
1	Define use case	Record the intended purpose, users, expected benefit, and potential harms.
2	Collect audit artifacts	Gather project documents, data descriptions, model information, test reports, and policies.
3	Categorize evidence	Map each artifact to governance, data/model design, or assessment.
4	Review gaps	Mark each criterion as complete, partial, missing, or not applicable.
5	Resolve and monitor	Assign owners, address high-risk gaps, document decisions, and repeat checks after changes.

D. Scope and Assumptions

The proposed catalog is a starting point for ML systems and is particularly relevant to high-impact applications such as healthcare. It is not a claim that every criterion applies identically to every model. The depth of review should be proportional to risk, intended use, data sensitivity, model complexity, and the possible consequences of an incorrect prediction. Domain experts and affected stakeholders should participate in deciding what constitutes adequate evidence.

E. Research Questions

- Which major dimensions should be considered when auditing an ML system?
- How can broad AI governance principles be translated into reviewable criteria?
- What evidence can support assessment of data quality, model performance, fairness, and robustness?

IV. PROPOSED SYSTEM DESIGN

A. Framework Overview

The proposed design connects three audit layers. Conceptual Basics establishes the purpose, risk, responsibilities, and review process. Data & Algorithm Design examines how the data and model were constructed. Assessment Metrics evaluates whether the system meets agreed requirements and behaves acceptably across relevant test conditions. Evidence from all three layers is recorded in an audit log and used to make a deployment, remediation, or monitoring decision.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, October 2026
Impact Factor: 8.2

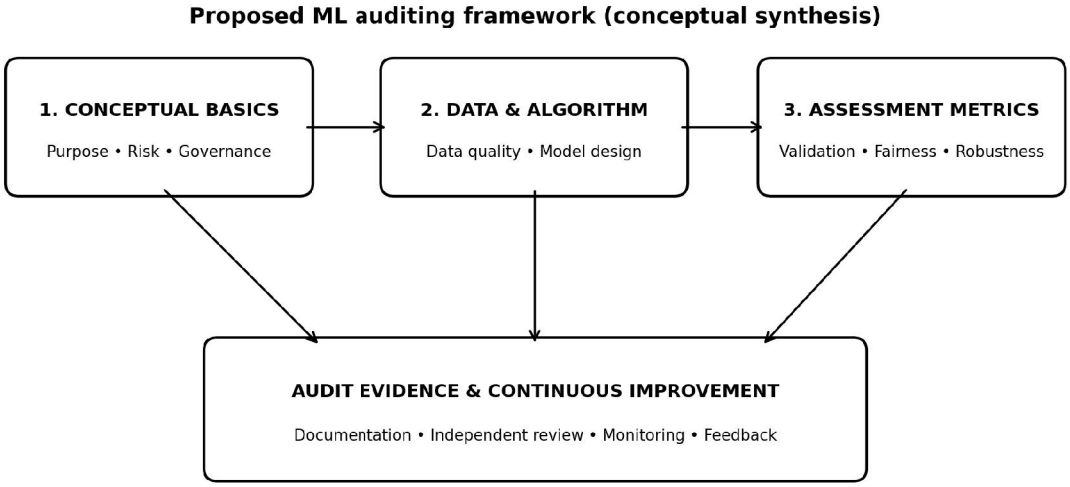


Fig. 2. Conceptual ML auditing framework synthesized from the three-category catalog in [1].

B. Main Modules

Module	Core focus	Example evidence
Conceptual Basics	Use-case value and risk; governance roles; methodology; ethics; audit access; quality assurance.	Risk register, responsibility matrix, use-case specification, review records.
Data & Algorithm Design	Data generation and representativeness; data quality; preprocessing; model selection; assumptions.	Dataset documentation, preprocessing log, model card, design rationale.
Assessment Metrics	Qualitative review; quantitative performance; fairness; robustness; validation.	Test plan, confusion matrix, metric results, subgroup analysis, limitations.
Monitoring & Improvement	Post-deployment drift checks, incident handling, change control, re-audit.	Monitoring reports, incident log, updated approvals and documentation.

C. Inputs and Outputs

Inputs include the use-case description, stakeholder requirements, training/validation/test data descriptions, model configuration, performance reports, and applicable policies. Outputs include a completed audit checklist, a list of unresolved gaps, evidence references, risk-based recommendations, and a monitoring plan. An auditor should be able to trace each conclusion to a documented source rather than relying only on verbal assurances.

D. Roles and Accountability

The project team prepares technical evidence; the domain owner confirms that the use case is appropriate; data stewards explain data provenance and limitations; governance or compliance specialists review obligations; and independent reviewers challenge assumptions. A responsibility matrix can clarify who is responsible, accountable, consulted, and informed for each audit activity. The exact roles depend on the organization and use case.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, October 2026
Impact Factor: 8.2

V. ML AUDITING CRITERIA CATALOG

The following checklist is an adapted, implementation-oriented set of questions based on the themes and examples in the source article [1]. It is not presented as a verbatim reproduction of all 30 questions in the original publication. Teams can record each item as Yes / Partly / No / Not applicable and link the answer to evidence.

A. Conceptual Basics

ID	Audit question	Example evidence
C1	Is the expected benefit of the use case greater than the plausible harm if the model fails?	Use-case assessment and risk register
C2	Are the intended users, affected groups, and operational setting clearly described?	Use-case specification
C3	Have stakeholders agreed on the development and audit methodology?	Project plan and approvals
C4	Have applicable legal, regulatory, and organizational requirements been identified?	Compliance mapping
C5	Are model risks and mitigation owners recorded and reviewed?	Risk register
C6	Are responsibilities and escalation routes clear?	RACI/responsibility matrix
C7	Are data and model documentation planned and maintained?	Dataset documentation and model card
C8	Is ethical review included in design and decision-making?	Ethics review notes
C9	Can an independent reviewer access sufficient evidence to assess the system?	Audit trail and access procedure
C10	Has verification checked expected outputs for representative inputs?	Verification report
C11	Has validation checked whether requirements and stakeholder needs are met?	Validation report

B. Data & Algorithm Design

ID	Audit question	Example evidence
D1	Is the data-generation process and data provenance sufficiently understood?	Data source/provenance record
D2	Does the dataset represent the population and setting where the model will be used?	Coverage and subgroup analysis
D3	Are accuracy, completeness, consistency, and currency of data assessed?	Data quality report
D4	Are missing values, outliers, duplicates, scales, and reference ranges handled transparently?	Preprocessing log
D5	Are possible confounders, leakage, and sampling bias considered?	Data/model review
D6	Is the selected model complexity justified against simpler alternatives?	Model selection rationale
D7	Are the problem hypothesis, target variable, and acceptance criteria explicit?	Specification and test plan

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, October 2026
Impact Factor: 8.2

V. ML AUDITING CRITERIA CATALOG (CONTINUED)

C. Assessment Metrics

ID	Audit question	Example evidence
A1	Are model assumptions, missing-data decisions, hyperparameters, and limitations documented?	Assumption log
A2	Can design decisions and test results be traced across development?	Versioned experiment log
A3	Have domain experts reviewed whether conclusions are appropriate for the evidence?	Expert review record
A4	Are training, validation, and test sets kept distinct and their statistical properties examined?	Dataset split and diagnostics
A5	Are suitable performance measures defined before final testing?	Evaluation protocol
A6	Are confusion matrix, precision, recall, F1, ROC-AUC, or other suitable metrics reported where appropriate?	Metric report
A7	Are results examined for relevant subgroups and possible unfair error differences?	Fairness/subgroup report
A8	Is robustness tested against plausible perturbations and distribution changes?	Stress-test results
A9	Are uncertainty and failure modes communicated to intended users?	User guidance and limitations
A10	Is there a plan to monitor drift, incidents, and performance after deployment?	Monitoring and incident plan
A11	Are changes to data, code, or model followed by appropriate re-evaluation?	Change-control records
A12	Are decisions to deploy, restrict, or reject the model documented with rationale?	Audit decision record

D. Interpreting Assessment Metrics

Metric selection must follow the task and the consequences of errors. Accuracy can be misleading when classes are imbalanced. Precision describes how often positive predictions are correct, while recall describes how many actual positive cases are identified. The F1-score combines precision and recall. ROC-AUC summarizes ranking discrimination across thresholds, but it does not by itself establish calibration, fairness, or clinical usefulness. A confusion matrix helps show false positives and false negatives, which may carry different costs. For regression tasks, measures such as MAE or RMSE may be more suitable. No single metric is sufficient for every use case.

Fairness assessment should be planned with domain experts and relevant stakeholders because the appropriate comparison depends on context, labels, populations, and the harms under consideration. Robustness tests should reflect plausible changes in the operating environment. Statistical tests such as VIF, Shapiro–Wilk, or Breusch–Pagan may be relevant to specific assumptions, but they should not be applied mechanically to every ML model. The audit report should explain why each test was selected and what its result does—and does not—show [1].

E. Audit Status and Prioritization

Status	Meaning	Suggested action
Complete	Evidence is available and meets the agreed requirement.	Retain evidence and review periodically.
Partial	Some evidence exists, but a gap remains.	Assign owner and due date.
Missing	No adequate evidence is available.	Assess risk and create remediation action.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, October 2026
Impact Factor: 8.2

N/A	Criterion does not apply to this use case.	Record justification and approver.

VI. DISCUSSION

A. Expected Benefits

A structured catalog makes auditing more consistent by translating broad ideas into questions that can be answered with evidence. It can encourage teams to consider risk and governance before model development becomes difficult to change. It also creates a common vocabulary for technical teams, domain experts, managers, and reviewers. When used throughout the lifecycle, the catalog can reveal missing documentation, weak data coverage, untested assumptions, and gaps in monitoring plans before these issues lead to harm.

B. Practical Application Scenario

Consider a hypothetical healthcare model designed to help prioritize cases for additional clinical review. Before deployment, the team would document the intended use and explicitly state that the model supports rather than replaces professional judgment. The data review would examine provenance, missing values, representation of patient groups, and potential leakage. Evaluation would report task-appropriate metrics and inspect errors across relevant subgroups. Domain specialists would review whether the outputs are useful and safe in the workflow. Finally, the organization would define human oversight, escalation, incident reporting, and periodic revalidation. This scenario illustrates how the catalog can organize evidence; it is not an experiment or a claim of measured clinical benefit.

C. Limitations

The present paper is a synthesis of one supplied source article and does not claim to conduct a new exhaustive literature review, implement the framework in an organization, or validate its effectiveness experimentally. The adapted checklist is a practical interpretation of the source themes rather than a verbatim reproduction of the original catalog. The adequacy of any audit depends on the application domain, quality of available evidence, expertise of reviewers, and the risk posed by the system. A checklist can also become a box-ticking exercise if answers are accepted without critical review.

D. Future Scope

- Pilot the checklist on real or carefully simulated ML use cases and record where criteria are unclear or missing.
- Map criteria to relevant laws, standards, organizational policies, and domain-specific guidance.
- Develop a scoring and evidence-tracking tool with ownership, deadlines, and audit history.
- Add lifecycle-specific checks for deployment, monitoring, model updates, and incident response.
- Evaluate the framework with independent auditors and domain experts before claiming effectiveness.

E. Implementation Recommendations

Organizations should tailor the checklist to the risk level and purpose of each model. Each response should include evidence, a responsible owner, a date, and an explanation for any not-applicable decision. High-risk unresolved issues should be escalated rather than averaged into a score that hides their importance. Auditing should recur when data sources, model versions, intended users, or operating conditions change.

VII. CONCLUSION

This paper presents a structured approach for designing an ML auditing criteria catalog as the starting point for a broader framework. Following the three core groups described in the source research—Conceptual Basics, Data & Algorithm Design, and Assessment Metrics—the proposed structure connects governance and risk review with technical evidence and measurable evaluation. The adapted checklist gives project teams a practical way to identify documentation, assign responsibility, examine data and model choices, evaluate performance and fairness, and plan post-deployment monitoring.

The catalog should be treated as a living review aid, not as proof of compliance or a guarantee that a model is safe. Its value depends on honest evidence, critical interpretation, stakeholder participation, and follow-up action. Future work should test the checklist in practical settings, identify gaps, align it with relevant standards and domain needs, and

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, October 2026
Impact Factor: 8.2

evaluate whether it improves audit completeness and decision quality. The central contribution of this paper is an organized, evidence-oriented starting point for that development process.

ACKNOWLEDGMENT

The author expresses sincere thanks to the project guide and the Head of Department for their guidance and academic support. The author also acknowledges the original researchers whose published work provided the foundation for this literature-grounded synthesis.

REFERENCES

[1] 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, doi: 10.1109/ACCESS.2024.3375763.
[2] P. Mayring, "Qualitative Content Analysis," Forum: Qualitative Social Research, vol. 1, no. 2, 2000.
[3] M. B. Miles and A. M. Huberman, Qualitative Data Analysis: An Expanded Sourcebook, 2nd ed. Thousand Oaks, CA, USA: Sage, 1994.
[4] European Commission, "The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment," 2020.
[5] L. Oala et al., "ML4H Auditing: From Paper to Practice," in Advances in Neural Information Processing Systems, 2020, pp. 280–317.
[6] T. Fawcett, "An Introduction to ROC Analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.
Note: This student paper is a paraphrased synthesis based primarily on reference [1]. The reported mind-map counts are attributed to the source article; the proposed diagram and checklist are conceptual adaptations and have not been experimentally validated.