The Review of Socionetwork Strategies (2023) 17:255–273
https://doi.org/10.1007/s12626-023-00148-w

ARTICLE

Towards a Risk-Based Continuous Auditing-Based Certification for Machine Learning

Dorian Knoblauch[1] · Jürgen Großmann[1]

Received: 4 March 2023 / Accepted: 1 August 2023 / Published online: 29 August 2023

Abstract

Machine learning systems have gained widespread adoption across various industries. This includes highly regulated ones that need to match certain quality requirements based on a given risk exposure. The MLOps paradigm, following a similar approach to DevOps, promises major improvements in quality and speed, with a focus on deploying ML models at a fast pace with high quality on an automated basis. However, traditional point-in-time certifications with manual audits are inadequate for MLOps setups due to frequent changes to the ML system. To overcome this challenge, we propose Continuous Audit-Based Certification (CABC), which uses automated audits to issue or revoke certificates based on an automated assessment of artifacts from the MLOps lifecycle. Our approach utilizes artifacts from the MLOps lifecycle for quality measurements based on standards such as ISO 25012. We propose a risk-based measurement selection, an audit API for standardized retrieval of data for measurement, a tamper-proof data collection process, and an architecture for separation of duties in the certification process. CABC aims to improve efficiency, enhance trust in the ML system, and support highly regulated industries in achieving their quality goals.

Keywords MLOps · Machine learning quality dimensions · Certification

1 Introduction

Nowadays, machine learning (ML) systems are being used across various industries, including highly regulated ones that demand systematic evidence for achieving quality goals or certification. Also, the industry has adopted MLOps, which follows a similar approach to DevOps, promise major improvements in quality and speed, but

✉ Dorian Knoblauch
dorian.knoblauch@fokus.fraunhofer.de

Jürgen Großmann
juergen.grossmann@fokus.fraunhofer.de

[1] Fraunhofer Fokus, Berlin, Germany

with a focus on deploying ML models at a fast pace with high quality on an automated basis [2]. MLOps enables faster deployment of ML systems and allows for frequent changes, leading to new models and may lead to different predictions that can impact the entire system's behavior.

Traditional point-in-time certifications with manual audits are inadequate for MLOps setups because the frequent changes to the ML system are unaudited until the recertification which occurs at fixed long term intervals. This results in in a lack of trust and may not be feasible for highly regulated industries that require a continuous assessment of quality. Recently several bodies and institutions, such as ISO and NIST, have released AI Risk Management Frameworks that help identify AI-based risks and mitigate them with suitable quality requirements. These requirements need to be measured to ensure their effectiveness, and this can partly be automated based on artifacts from the MLOps lifecycle.

To overcome the challenges of traditional certification processes, we propose Continuous Audit-Based Certification (CABC). This approach uses automated audits to issue or revoke certificates based on an automated assessment of artifacts from the MLOps lifecycle. By automating the certification process, CABC improves efficiency, enhances trust in the ML system, and supports highly regulated industries in achieving their quality goals.

Furthermore, our proposed CABC approach is technology agnostic. This means it can seamlessly operate with various machine learning methodologies, whether they are Supervised, Unsupervised, Semi-supervised, or Reinforcement Learning, as long as there are automated measurements.

However, it's important to note that while CABC is generally applicable to a wide range of methodologies, there are significant limitations when it comes to Federated Learning. This is because Federated Learning involves distributing the measurements, which can result in a heavy overhead for the user. Nonetheless, we are actively researching ways to mitigate this limitation, in order to make CABC even more universally applicable in the ML domain.

Traditional certifications are carried out under a "certification scheme," which includes a methodology that the auditing party uses to conduct the assessment. The requirement are defined in a "standard" against which a system's conformity is evaluated.[1] AI Risk Management Frameworks are not standalone industry standards, but rather they provide requirements that might include or even demand the mitigation of risks from ML systems.

In this paper, we propose the CABC approach for ML in a sport that leverages a risk management approach to initialize the framework and continuously audits the derived requirements for ensuring high-quality and trustworthy outcomes. The proposed approach includes roles and processes for continuous audit, a methodology to operationalize quality requirements, and a trustworthy infrastructure for continuous auditing. The core of CABC is the automated assessment of the ML system with a set of given requirements for various quality dimensions. These requirements come from AI Risk Management Frameworks, such as those from ISO, NIST, or other

[1] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/.

organizations and bodies, as well as catalogs like the AI Cloud Service Compliance Criteria Catalogue (AIC4) from the BSI.

However, the frameworks and catalogs are ambiguous about the measurement of requirements, which depends on the specific ML system. While existing standards for data quality, such as ISO25012, provide a comprehensive set of quality measurements, and ISO25024 defines measurements and metrics, the level of abstraction still requires the implementation of these measurements based on the system being assessed.

To address this challenge, we propose a methodology that makes it easy to implement measurements from scratch or use existing tools to provide information on quality. We introduce an audit API for standardized retrieval of data for measurement, which functions as a clear separation between the test subject being audited and the party performing the audit. The measurement results are evaluated off-premise to achieve an independent assessment of compliance with quality goals. Depending on whether the goals are met, a certificate is issued.

A significant challenge in the measurement process is the large part of information gathering done on the auditee's premise, which makes the information provided by the auditee less trustworthy since it can be altered. To address this challenge, we propose collecting data in a tamper-proof manner using a trusted execution environment included in the MLOps Execution.

Although the feasibility of CABC has been piloted in the domain of cloud security certification in previous works, it still requires human assessment and operationalization. Some aspects of quality assessment are also reduced for the benefit of automation. In this paper, we transfer CABC to the ML domain and aim to mitigate these downsides.

2 Related Work

2.1 MLOps/DevOps

DevOps, as an established field in software development, champions the integration of development and operations to streamline the process of software deployment. By advocating for frequent and high-quality software releases, DevOps has significantly transformed traditional software development practices [7].

However, as machine learning has become integral to many modern software systems, the need for a specialized DevOps approach arose. This led to the evolution of MLOps, a discipline that adapts and extends DevOps principles to machine learning workflows. MLOps not only seeks to improve the speed and quality of machine learning system deployment but also addresses the unique challenges of deploying such systems [6].

One of the pivotal studies that shaped MLOps was by D. Sculley et al. [25]. They identified several sources of technical debt unique to machine learning systems and proposed best practices for managing them, emphasizing the importance of considering the entire machine learning lifecycle.

Sergio Moreschi et al.'s comprehensive review of 84 MLOps tools, "Toward End-to-End MLOps Tools Map," is another survey in this field. The study provides valuable insights into the use of different MLOps tools across the various DevOps phases and identifies potential incompatibilities [20]. A challenge that also occur when performing quality measurements based on the output of these different tools.

As the field of MLOps continues to evolve, various trends and advancements are emerging. A key trend is the development of continuous certification frameworks for MLOps, borrowing from the DevOps concept but extending it to cater to the unique needs of machine learning workflows [5].

Automated deployment of machine learning models is another current focus in MLOps. It involves the intricate process of implementing CI/CD pipelines in applications with machine learning components, a subject of ongoing research due to its inherent challenges [6].

Additionally, a comprehensive understanding of the machine learning lifecycle is being pursued. The emphasis is on providing quality assessment at every stage of the ML lifecycle and establishing trust through continuous certification.

In our work, we establish a continuous certification framework for MLOps that shares some aspects with evaluating artifacts along the pipeline, but extends it in areas such as actually mapping the artifacts to quality measurement inputs.

2.2 Risk Management

The KI-Prüfkatalog [1], developed by Fraunhofer IAIS, provides a comprehensive set of quality criteria for AI systems, covering reliability, safety, and robustness, as well as criteria such as explainability, fairness, and data protection. Similarly, the Kriterienkatalog für KI-Cloud-Dienste - AIC4 [12], developed by the Federal Office for Information Security (BSI), provides criteria that cloud service providers offering AI services should fulfill, covering security, data protection, and transparency. Both catalogs provide detailed descriptions of each criterion and how they should be implemented. By operationalizing these criteria, requirements for a ML system can be derived and measured through CABC to ensure compliance with the quality goals set by these catalogs.

ISO/IEC 23894:2023 also known as "Information technology — Guidance on risk management," is a document that provides guidance on managing risks related to AI systems [13]. This guidance can be used in conjunction with ISO 31000:2018 to assist organizations in integrating risk management into their AI-related activities and functions. The document provides an overview of the underlying principles of risk management, as well as a framework and processes for managing risks related to AI systems. Furthermore, the document provides common AI-related objectives and risk sources in annexes, which can be used to derive quality requirements for AI systems. Those requirements are assessed automatically during the CABC process.

Similarly, the AI Risk Management Framework (AI RMF) provides practical guidance for organizations and individuals involved in the AI system lifecycle, to increase the trustworthiness of AI systems and promote responsible design, development, deployment, and use of AI systems.[2] This is essential for meeting the requirements for responsible and ethical AI systems.

2.3 Certification

Granlund et al. [8] define and evaluate an MLOps process that produces regulatory-compliant models. The process includes running a continuous deployment that selects the best model and packages it. This package is then considered as "locked" and becomes part of a product. The entire product is then verified. This approach was chosen due to regulatory requirements that do not consider the special character of ML systems. However, the certification process is still manual. In contrast, our proposed CABC approach does not have the limitation of locking the ML system. The MLOps cycle can still be rerun independently of the rest of the system without rendering the product certification obsolete, but it requires the certification body to adopt continuous audit-based certification.

In our previous works [16, 18], we developed and evaluated the approach of CABC for security certification of cloud services. While security standards for cloud services are well-established, the emergence of quality standards for machine learning (ML) poses a new challenge. However, we observed that the high-level requirements and the need for operationalization in both domains share similarities. This similarity led us to define a new assessment process for ML systems that builds on the CABC approach.

As part of the EU-Sec project,[3] CABC underwent an evaluation pilot to demonstrate its applicability in different IT infrastructures. The pilot focused on two different services with the same use case, provided by a partner from the banking industry. The use case involved the exchange of sensitive personal data between banks and regulators, subject to national, industry, and international requirements. CABC was used to continuously ensure compliance with these requirements. In the pilot, a subset of requirements relevant to the partner, defined during the preparations, were implemented. The controls utilized in the assessment were derived from the Cloud Control Matrix.[4] The first implementation of the use case involved an open-source cloud storage solution, along with a custom plug-in for encrypted message and document exchange on the client side. The second service, provided by another partner from the banking industry, enabled the same use case. Both services offered an Audit API, which was leveraged by Clouditor[5] to gather evidence for assessing the fulfillment of objectives. However, due to architectural differences, the evidence collection process varied between the two services. In the case of the first service, deployed on AWS EC2 with connected EBS and RDS, Clouditor also verified the associated infrastructure services. This step was not necessary for the second service. The assessment results were subsequently shared with the certifying body for

[2] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[3] https://www.sec-cert.eu.

[4] https://cloudsecurityalliance.org/research/cloud-controls-matrix/.

[5] https://www.aisec.fraunhofer.de/de/forschungsabteilungen/SAS/Clouditor.html.

the pilot, which issued a certificate upon successful evaluation. This confirmed the technical applicability of the CABC approach.

In a previous publication titled “Towards Continuous Audit-based Certification for MLOps” [17] we presented a continuous audit-based approach to certify ML systems, which elaborates on CABC for MLOps like the present work. However, our previous work proposed an incomplete Quality Attribute Catalogue, which we have repurposed to serve as a basis for automated measurements. Our current approach employs a risk-based strategy to derive quality requirements, which we view as a way to mitigate risks. In other words, instead of relying solely on the Quality Attribute Catalogue, we look to manage risks as a means of shaping quality requirements.

2.4 Data Quality Measurement

The ISO has released standards such as ISO/IEC 25012 [14] and ISO/IEC 25024 [15] to address the quality of data. ISO/IEC 25012 defines a set of characteristics for data quality, including Accuracy, Completeness, Consistency, Currentness, and Credibility. ISO/IEC 25024 provides means to evaluate certain features of the data by assigning quality properties to the quality characteristics introduced in ISO/IEC 25012. For example, accuracy can be evaluated via the properties of syntactic accuracy, semantic accuracy, and accuracy range. While measurement formulas are defined for quality assessment, suitable quality characteristics and requirements must be implemented. Our approach builds upon those data quality measurements.

The ISO/ICE 25000 series does not provide guidance on how to obtain the values in the measurement formulas, leaving it to the judgment and experience of the assessing party. In a related work [9], the authors present a data quality evaluation process carried out from establishing quality requirements to executing the quality evaluation. They obtained the information needed for the measurement by querying the auditee’s database. As they pointed out, the actual specification of data quality and the measurements for their assessment depend on the actual system as well as the business case. Our approach builds upon data quality measurements, as defined in the ISO/IEC 25012 and ISO/IEC 25024 standards, to ensure the effectiveness of quality requirements for ML systems. By using artifacts from the MLOps lifecycle, we can automate the assessment and don’t require manual integration work.

Quality dimensions are often on a level of abstraction that requires further refinement of the measurements required in the MLOps process. For automation, tools are required that can perform measurements at the necessary frequency and thus ensure scalability. In a related work [19], the authors provide a survey of data quality measurement and monitoring tools, where they evaluated 13 tools, including the used metrics for data quality measurement and their capabilities for continuous data quality monitoring. To enable continuous assessment of data quality based on rules, DaQL 2.0 has been introduced, a tool that allows continuous data quality measurement once rules are implemented for complex data objects [19]. Similar schema-based approaches have been implemented by Schelter et al. [24] and Brek et al. [4]. In our proposed CABC approach, we embed these tools as a way of retrieving information on the quality of the MLOps lifecycle artifacts. Our methodology provides a standardized way of implementing and selecting the measurements for quality requirements.

3 Deriving Quality Requirements

As ML systems become more complex, their quality becomes a crucial factor for their adoption and success. Quality requirements are necessary to mitigate risks associated with ML systems related to fairness, data protection, reliability and more. This chapter discusses how to derive quality requirements suitable for a particular ML system based on AI risk frameworks and catalogs.

The first step towards creating quality requirements involves recognizing the risks associated with the ML system. AI Risk Frameworks and catalogs provide valuable insights into risk management and the identification of quality requirements for ML systems. These resources encompass various aspects of quality dimensions, often referred to as quality objectives or criteria, depending on the source. The following list features dimensions commonly found in two or all three of these frameworks: ISO/IEC 23,894 [13], IAIS' KI-Prüfkatalog [1], and AIC4 [12], thereby highlighting the importance of these dimensions:

- Fairness: The ML system should be designed and trained to be fair and unbiased. This can be achieved by ensuring that the training data is diverse and representative of the population and by using algorithms and techniques that minimize bias. (ISO, IAIS, AIC4 associates bias with fairness)
- Explainability: The ML system should be transparent and explainable so that decisions made by the system can be understood and audited.(ISO, AIC4)
- Security: Security in AI involves ensuring that AI systems are protected from attacks and unauthorized access. It includes safeguarding data used by and generated from the AI system, and ensuring that the AI system cannot be exploited to perform malicious actions. (ISO, IAIS, AIC4 mentioned security explicitly in Security & Robustness)
- Reliability: The ML system should be reliable and accurate in its predictions and outcomes.(ISO, IAIS, AIC4)
- Robustness: The ML system should be able to handle unexpected or adversarial inputs and continue to operate effectively. (ISO, AIC4 mentioned security explicitly in Security & Robustness)
- Data protection: The ML system should protect sensitive or personal information and prevent unauthorized access or use. (ISO listed as Privacy, IAIS)
- Transparency: The ML system should be transparent about how it uses and processes data, and how it makes decisions. (ISO, IAIS)
- Data Quality and Management: This involves ensuring the availability of sufficient, representative, and high-quality data for training and testing the AI system, and managing this data responsibly. Good data management practices include appropriate collection, storage, access, and processing of data, while data quality involves ensuring the data is accurate, complete, reliable, relevant, and timely. High-quality data and effective data management are crucial for the performance,

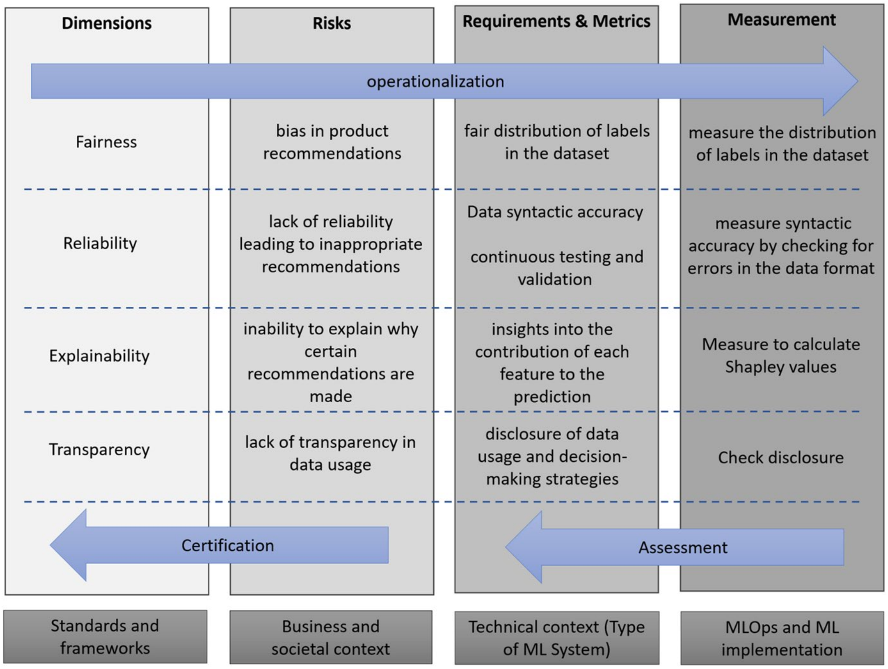
Dimensions
Risks
Requirements & Metrics
Measurement
operationalization
Fairness
bias in product recommendations
fair distribution of labels in the dataset
measure the distribution of labels in the dataset
Reliability
lack of reliability leading to inappropriate recommendations
Data syntactic accuracy
continuous testing and validation
measure syntactic accuracy by checking for errors in the data format
Explainability
inability to explain why certain recommendations are made
insights into the contribution of each feature to the prediction
Measure to calculate Shapley values
Transparency
lack of transparency in data usage
disclosure of data usage and decision-making strategies
Check disclosure
Certification
Assessment
Standards and frameworks
Business and societal context
Technical context (Type of ML System)
MLOps and ML implementation

Fig. 1 Applying Quality Requirements to an example ML System. This figure illustrates the process of identifying risks and establishing quality requirements for an ML system in an e-commerce company. The example showcases how fairness, reliability, explainability, and transparency can be implemented in a personalized product recommendation system

fairness, and reliability of AI systems. (ISO listed as Availability and quality of training and test data, AIC4 listed as Data Quality and Data Management)

Those and other dimensions mention solely in one of the frameworks can guide the derivation and measurement of requirements through CABC to ensure responsible and ethical use of AI technologies. Once the risks associated with the ML system have been identified, the next step is to establish quality requirements that mitigate these risks. Quality requirements are defined as a means to mitigate risk and are assessed via measurements.

Imagine the following example, depicted in Fig. 1: An e-commerce company utilizes a ML system to personalize product recommendations for its customers. However, it understands the importance of ensuring its ML system doesn't inadvertently create an unfair or opaque experience for its users. In order to maintain customer trust and abide by ethical considerations, the company translates four crucial dimensions into specific risks: fairness, reliability, explainability, and transparency. For each identified risk, quality requirements are established, which are then operationalized into measurable entities.

For instance, a requirement derived from fulfillment might be measured by checking the distribution of labels in the dataset. A requirement derived from reliability can be determined by verifying the accuracy of the system's predictions. A requirement derived from explainability may be validated by the interpretability of the models used. Lastly, a requirement derived from transparency might be evaluated based on how well the system's data usage and decision-making processes are documented and communicated.

The assessment of these measurements is a crucial step towards fulfilling the quality requirements. If the ML system meets these requirements, it mitigates the risks associated with its use, thereby ensuring compliance with relevant standards and regulations. Through a robust certification process, the e-commerce company can demonstrate that it has effectively mitigated these risks, leading to greater trust in its ML system from users and regulators alike.

- Fairness: Could be achieved by establishing requirements for bias mitigation strategies, such as ensuring a fair distribution of labels in the dataset.
- Reliability: This can be achieved by establishing requirements for data quality, such as syntactic accuracy, and by ensuring that the system is tested and validated throughout the development process.
- Explainability: Could be achieved by establishing requirements for interpretability strategies, such as using inherently interpretable models like decision trees or logistic regression, or by applying post-hoc explainability methods like LIME or SHAP to more complex models.
- Transparency: Could be promoted by disclosing the system's data usage, processing methods, and decision-making strategies.

Assessing the fulfillment of quality requirements is vital for ensuring that the ML system meets desired quality standards. A minimal set of measurements for assessing ML system quality, collected from ISO 25012 and current standards, has been compiled. However, ML systems exhibit significant diversity in terms of the data they process, the architecture used, the learning paradigm applied, among other factors. For example, some ML systems are designed for natural language processing, others for image recognition or time series forecasting, and some are built on deep learning architectures, while others use decision trees or support vector machines.

Due to this diversity, assessing the quality of ML systems requires different measurements that are tailored to each specific system. For instance, image recognition models may be evaluated using metrics such as precision, recall, and F1-score, while natural language processing models may be assessed using metrics such as accuracy, perplexity, and BLEU score. Similarly, reinforcement learning models may require different metrics such as reward maximization and exploration, while clustering models may be evaluated using metrics such as silhouette score or Dunn index.

In summary, the wide variety of ML systems available requires different measurements to assess their quality. These measurements should be tailored to the specific type of ML system and the problem it is trying to solve. For our previous example, the explainability of a model trained on tabular data might be measured differently than a model trained on images. For tabular data, explainability could be measured by the feature importance rankings, while for images, explainability could be measured by methods that generate visual explanations, like saliency maps or activation maximization.

Table 1 Initial set of quality metrics from the Data domain

Metric	Description
Accuracy	"Data accuracy is the degree to which data has attributes that represent the actually value of a concept" (ISO 25012)
Completeness	"The degree to which subject data associated with an entity has values for all expected attributes" (ISO 25012)
Consistency	"The degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use" (ISO 25012)
Timeliness	"The degree to which data has attributes that are of the right age in a specific context of use" (ISO 25012)

Table 2 Initial set of quality metrics from the Model domain

Metric	Description
Accuracy	Accuracy is the ratio of predictions that a ML-model predicts correctly [11, 23]
Generalization	Generalization means the capability of a model to function correctly with unseen types of data [10, 21]
Fairness	Fairness means the capability of the model to correct biased tendencies [3]
Robustness	The capability of the model to deal with intentionally or unintentionally wrong input [22]

For our previous example, syntactic accuracy for tabular data needs to be measured differently than for images. For tabular data, syntactic accuracy can be measured by checking for errors in the format of the data, such as missing values or incorrect data types. In contrast, for images, syntactic accuracy can be measured by checking for errors in the annotation or image resolution. We are addressing currently three MLOps domains with the current minimal proof of concept set, each having distinct characteristics:

- Data quality is the core part of every ML system, and quality assurance in this domain has the main focus. The measurements for data quality are taken from ISO 25012 [14], see Table 1, and the measurements are performed on static artifacts produced in the early steps of each cycle. ISO 25024 introduces generic measurements for the listed quality measurements, which can be used to assess the quality of data. For example, to measure the completeness of data, the percentage of missing data can be calculated. Similarly, to measure the accuracy of data, the percentage of correct data can be calculated.
- Model quality measurements are compiled from different state-of-the-art contributions and the measurements for assessing model quality are highly dependent on the kind of ML system, see Table 2. Models are deployed, so they need to be monitored in production, which leads to different kinds of artifacts that are indicators of quality. Usually, those are logs of the prediction tasks containing information on aspects like accuracy. Information on factors like robustness might even only be obtained by specific tests.

Table 3 Initial set of quality metrics from the MLOps Domain

Metric	Description
Deployment Frequency	The frequency in which a new model gets deployed after a MLOps cycle
Lead Time for Changes	The time depends on “explorative phase in Data Science, Duration of the ML model training and duration of manual steps during the deployment process” (ml-ops.org)
Mean Time To Restore	“Mean Time To Restore refers to the duration of the rollback of the ML model to the previous version” (ml-ops.org)
Change Failure Rate	“ML Model Change Failure Rate can be expressed in the difference of the currently deployed ML model performance metrics to the previous model.” (ml-ops.org)

– To evaluate the quality of the MLOps process itself, the metrics in Table 3 from the ml-ops.org site[6] are used. The quality of a process is a valuable indicator of product quality.

The set of measurements we have compiled serves as a feasibility study for CABC for MLOps, and we expect that as the field continues to evolve, the set of measurements will need to be updated and refined.

4 Continuous Audit Based Certification

Continuous Auditing Based Certification for MLOps can be divided into two phases: the initialization phase and the continuous phase. The initialization phase involves the initial manual setup, while the continuous phase involves automated execution, as shown in Fig. 2.

During the initialization phase, the proper operationalization of the selected set of quality requirements takes place. The key actions in this phase include defining the scope, identifying the measurements associated with each quality requirement, determining the frequencies at which each quality goal should be checked, and implementing the mapping of evidence and quality measurement input. Mapping involves leveraging the raw data consumed or produced in the MLOps process as so called artifacts to usable measurement input via parsing, transforming, or executing testing tools on them.

Figure 2 also shows the three roles involved in CABC and their main activities during the two phases. By following the proper operationalization and mapping and leveraging MLOps artifacts to evidence, the auditee can ensure that the CABC process can effectively monitor and assess the ML systems compliance posture.

The initialization phase of CABC for MLOps involves several key activities that ensure the quality and compliance of the ML system. These activities are:

[6] https://ml-ops.org/content/mlops-principles#ml-based-software-delivery-metrics-4-metrics-from-accelerate.

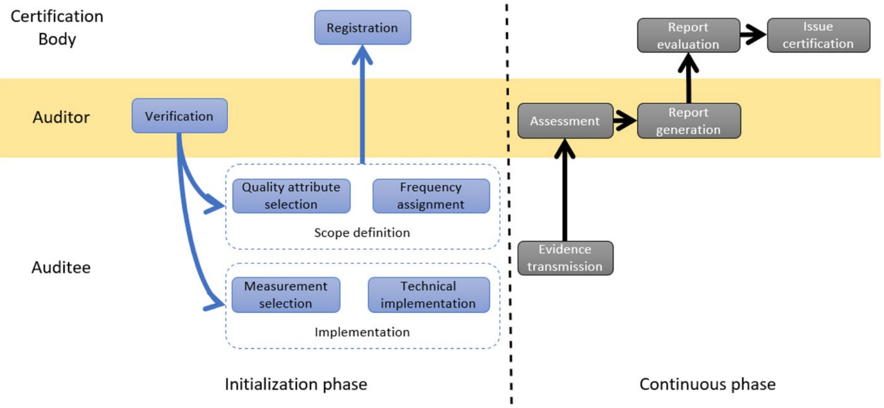

Fig. 2 This image depicts the roles and processes involved in Continuous Auditing Based Certification during both the initialization phase and the continuous phase. The blue processes, such as defining the scope and selecting the requirements, require a one-time execution. The black processes, such as evidence collection, assessment, and reporting, are executed continuously to ensure the quality and compliance of the ML system

- Defining the scope: As a prerequisite the auditee needs to establish a risk management plan that identifies potential risks and their impacts on the ML system. Based on this plan, they can determine the quality requirements that will mitigate the identified risks. Once the requirements that are subject to the assessment are defined suitable frequencies with which the assessment for each requirement should take place need to be assigned. The scope definition is checked by the auditor for its suitability as part of the verification process. Once verified, the auditee submits the verified scope to the certification body to initialize the continuous phase.
- Selecting corresponding quality measurements: Depending on the requirements, corresponding quality measurements need to be selected. For each measurement, the desired value boundaries need to be assigned according to the quality goals. Since the industry is still moving towards the first set of standards for ML, the mapping of requirements and measurements has to be solved by expert knowledge from the auditee and auditor. One is selecting and the other one verifying. There might also be a scenario where the auditor executes the initialization phase on the auditees' site.
- Defining the mapping between artifacts and measurement input values: Since each MLOps process is implemented with a different set of tools and libraries that differ in usage and produced artifacts, a mapping between the artifacts and the measurement function's input values needs to be defined as part of the technical implementation. Sometimes the artifacts of the MLOps process are not sufficient for a proper measurement. In that case, specialized measurement means need to be installed to fill this gap.
- Verification of the initial setup and changes: The initial setup, as well as changes for CABC for MLOps, need to be verified by an auditor to establish trust.

As an example, let's consider an auditee that runs a business that requires accurate customer data for its operations. Inaccurate data could lead to customer dissatisfaction, loss of business, and potential legal issues. The auditee can mitigate this risk by making high-quality data a requirement. For instance, they can set up a quality requirement for data accuracy and select syntactic accuracy as the corresponding measurement. In our use case, our data frame is a set of customer information, such as names, addresses, and contact details. Syntactic accuracy in this case would be the percentage of customer data that is complete and correctly formatted, divided by all customer data in the system. To ensure high data quality, the auditee can set a desired value boundary of 95% for syntactic accuracy. By setting up these quality requirements and measurements, the auditee can ensure that their ML system is compliant with relevant standards and regulations, and the CABC process can effectively monitor and assess the organization's compliance posture.

The continuous phase of CABC for MLOps involves the following key activities:

- Artifacts production and usage: Artifacts are produced or used during an MLOps cycle, and they are used as input for measurements. Some artifacts reveal the measurement result through parsing, while others require fully-fledged test suites to obtain accurate results.
- Mapping artifacts to measurement inputs: The result of the measurements is mapped to a standardized interface that allows the auditor to consume them and perform the assessment. This mapping takes place before the actual transmission of evidence, which is triggered after every cycle and for the monitoring data based on the frequency of the reflecting quality requirement.
- Evaluation of evidence: As part of the assessment, the auditor evaluates the received evidence and matches the result against the predefined values, which reflect the quality goals.
- Report generation: Based on the assessment, a report is generated, mentioning each measurement result and the final verdict on the fulfillment of the quality goals.
- Evaluation by certification body: This report then gets evaluated by the certification body, which assesses whether the ML system is compliant with the predefined promises.
- Update registry: Based on the evaluation of the report, the registry for the specific continuous certification gets updated with either the new certificate or the revocation of the old one.

In our example either in the Data Extraction and Analysis or in the Data preparation phase of MLOps, some check on the consistency will happen and the results get logged. For instance, if warnings occur during the loading of the data frame indicating missing information for certain customer information, those warnings would be counted as syntactically incorrect. After subtracting the number of incorrect data from the overall data count both values as well as the raw logs containing the warnings get submitted to the auditor for assessment. The auditor then divides both numbers and matches to result to the boundary defined by the expert. This then gets done for all measurements corresponding to the quality requirement and for all

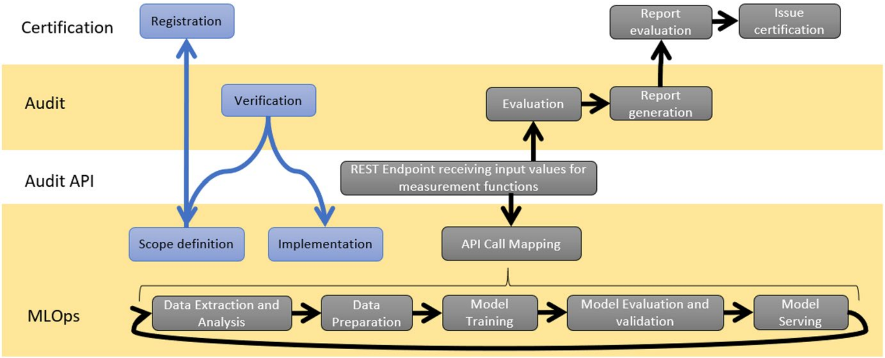

Fig. 3 Layered architecture to facilitate separation of roles. Blue processes require a one time execution, black are executed continuously. The white and yellow areas are symbolizing the affiliation to the particular entity. E.g scope definition and the implementation of the measurement are part of the auditees MLOps setup

requirements in the assessment process and finally leads to a audit report. Based on this assessment report, the certification body would confirm whether the auditee's ML system is compliant with the predefined quality requirements and update the registry for the specific continuous certification with either a new certificate or the revocation of the old one. The auditee can ensure that their ML system remains compliant with relevant standards and regulations, and the CABC process can effectively monitor and assess the organization's compliance posture, mitigating the risk of inaccurate customer data.

5 CABC Layered Architecture

To establish trust in the quality of ML systems, a third party certification is often required. To facilitate this, our proposed framework consists of four distinct layers: certification, auditing, evidence provisioning, and mapping of MLOps artifacts to information for measurement. This layered architecture ensures clear separation of duties in the certification process and promotes technical encapsulation to prevent tampering. As illustrated in Fig. 3, the blue phase represents the initialization phase while the black phase represents the continuous phase. At the bottom of the figure, a generic MLOps process is depicted. This approach enables separation of implementations and promotes encapsulation, which are critical components in maintaining the integrity and trustworthiness of the certification process.

- Certification layer: This layer facilitates the registration of the ongoing CABC process. It maintains a machine-readable log of the scope and frequencies of the assessment process to ensure that all assessments are submitted in a timely manner. Stakeholders can retrieve information on the certification status, scope, and frequencies. This layer evaluates the audit report and issues the certificate, which

is stored and published in the registry. The auditee initially submits a scope definition that has been verified by an auditor to the registry.
- Audit layer: This layer verifies the scope and implementation of the ML system and creates a digital fingerprint to ensure traceability and tamper resistance. The measurement functions are evaluated on this layer, similar to traditional auditing where provided evidence is evaluated.
- Audit API layer: This layer handles the handover of requested evidence from the audited party to the auditor. It defines the input parameters for the measurements and requires the audited system to provide the corresponding values.
- MLOps layer: This layer involves the implementation of the MLOps process, which runs on the premises of the audited party. It produces evidence that is required for the quality assessment, as not all evidence can be retrieved from existing artifacts. In that case the measurement takes place on the auditees premisses but in a sealed environment. The results is then submitted to the corresponding endpoint of the Audit API.

5.1 Information Retrieval Inside MLOps

During each step of the MLOps process artifacts like logs, configurations, checkpoints, metadata, models, etc. get created. The proposed framework requires the extraction of input parameters for measurements from various artifacts generated during each step of the MLOps process, such as logs, configurations, checkpoints, metadata, and models. This process involves both passive aggregation of information through techniques like log parsing and active counting, querying, testing, and monitoring, which must be implemented specifically to the software stack in use. To make the framework more technology-agnostic, the Audit API layer is introduced. This REST-Endpoint receives data on the parameters required for the metrics, including the parameter value and raw data or hashes to store as evidence. These values enable measurements, which allow for the assessment of individual quality requirements and the overall quality of the ML system. The result is an automated audit report suitable for certification by an accepting certification body. Initial experimental results have been realized in Kubeflow and MLflow through components and plugins for information retrieval.

5.2 Mapping of Artifacts to Parameters

The proposed framework includes a mapping component that facilitates the transformation of raw data generated during various steps of the MLOps process into values and evidence transmitted to the Audit API, as illustrated in Fig. 4. The MLOps process produces raw data, which is either input to active testing or directly to the extractor. The extracted information is then mapped to the API. In industry, the MLOps process is often supported by frameworks like Mlflow or Kubeflow, which can be extended to manage artifacts. Raw data such as logs, metadata, or trained models are extracted from these frameworks, and depending on the type of data, it is either parsed to extract information or used as input

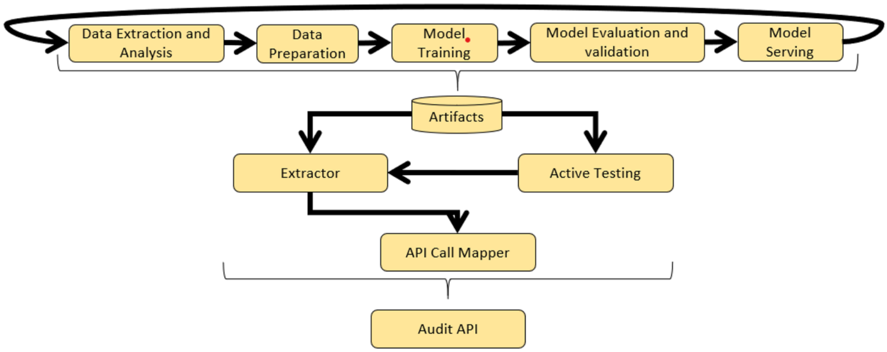

Fig. 4 Mapping of the raw data to the input of the API calls

for other testing tools. The extractor must be implemented for each artifact type, and some artifacts may require specific active testing tools. Finally, the extracted information is mapped to a specific API call parameter, resulting in a specific measurement value. The handling of each artifact type depends on the tool that generated it, and some artifacts require dedicated parsers or extractors.

5.3 Trusted Execution Environment

As auditees have the theoretical ability to alter artifacts in the CABC process, this can lead to a lack of trust in the framework. To address this issue, the implementation is verified by the auditor. During verification, the setup is cryptography fingerprinted by hashing configurations, code, and other elements that could potentially alter the MLOps process or information retrieval. Ideally, the MLOps process should run in a container, virtual machine, or on cloud premises to allow for the fingerprinting of the entire setup. For example, the auditee can implement information retrieval and mapping and package it as a container. In this scenario, the auditor simply needs to seal the container and ensure that it is the same one executed. During each evidence transmission, the auditor's signature is transmitted to the Audit API and checked for validity before evaluating the information. If the scope or implementation changes, the auditor must re-verify the scope and implementation. After re-verification, the implementation must be fingerprinted again to ensure trust in the CABC process.

The separation of layers in the CABC framework enables the audit layer and certification layer logic to run in a trusted execution environment on the auditee's premises. In this scenario, the auditor and certification body provide a sealed execution environment that the auditee can run, enhancing trust in the CABC process.

6 Roles and Processes

Compared to traditional point-in-time certification, Continuous Auditing Based Certification requires adjustments to the roles and processes involved. In a traditional certification, trustworthy third parties, usually organizations and humans, are introduced to establish trust. However, in CABC, the challenge is to achieve the same level of trustworthiness primarily through automated technical means.

Certification is typically used to demonstrate compliance to customers or authorities. However, a self-proclamation from an auditee does not generate the same level of trust as a third-party audit due to a conflict of interest. Therefore, certification schemes often require two or even three parties. For our CABC for MLOps, we foresee the following parties:

- The Certification body defines the rules for the certification process. It lays out the criteria under which an audit is conducted and defines the form of the audit report. According to the audit report, the certification body issues or suspends a certification. In the case of CABC, the certification body provides a registry of the ongoing certification process, which serves as a trusted resource for the defined scope and current certification status for potential stakeholders.
- The Auditing party conducts the audit under the rules of the certification body. It verifies the scope provided by the auditee for its suitability and adherence to given requirements. The auditing party also verifies the initial setup of the continuous auditing and facilitates the automated measurements and assessments during operation. Additionally, the auditing party provides the means to receive evidence from the auditee.
- The auditee is responsible for establishing CABC by defining the scope. The scope reflects the quality necessary for the ML system. To do this, the auditee defines quality requirements and suitable measurements, which come from AI Risk Management Frameworks and Criteria Catalogues. The Auditee then carries out the technical implementations of those measurements to provide measurement results in an automated manner in the form of evidence to the auditor. The actual technical implementation is part of the MLOps process. By fulfilling these responsibilities, the auditee ensures that the ML system is compliant with relevant standards and regulations.

By involving these three parties, CABC ensures the independence of the audit process and increases objectivity. Additionally, to ensure data security and privacy, secure protocols are used for data transfer and storage, and access controls are implemented to limit who can view or modify the data.

7 Conclusion and Future Work

In this paper, we have proposed a risk-based continuous audit-based certification (CABC) approach for MLOps that includes a methodology, initial quality measurements derived from risk mitigation efforts, and a trustworthy infrastructure for continuous auditing. We have presented details on information extraction, mapping, and audits based on automated assessment, building on a certification schema that was already successfully piloted in cloud security certification.

We have also set up an initial technical demo for CABC for MLOps.[7] Moving forward, we plan to pilot this approach with a real use case in the area of image recognition for energy infrastructure maintenance scenarios.

Future work will focus on refining the methodology, expanding the measurements, and evaluating the effectiveness and scalability of the CABC approach. We also plan to investigate the use of machine learning techniques to enhance the automation of the certification process and the accuracy of the assessments.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was partially funded by the German BMBF under grant agreement ITEA- 2021-20219-IML4E as well as the German BMWI project KI-LOK.

Data Availability Not applicable.

Declarations

Conflict of Interest The authors of this paper, Dorian Knoblauch and Jürgen Großmann, declare the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: - The work presented in this paper was partially funded by the German Federal Ministry of Education and Research (BMBF) under grant agreement ITEA-2021-20219-IML4E as well as the German Federal Ministry for Economic Affairs and Energy (BMWI) project KI-LOK. The authors confirm that this work is free from any other potential conflicts of interest.

References

1. Ki-pruefkatalog. (2021). https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkatalog/202107_KI-Pruefkatalog.pdf. Accessed 27 Feb 2023
2. Alla, S., & Adari, S. K. (2021). *What Is MLOps?* (pp. 79–124). Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4842-6549-9_3
3. Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. *Nips Tutorial, 1*, 2.
4. Breck, E., Polyzotis, N., Roy, S., Whang, S., & Zinkevich, M. (2019). Data validation for machine learning. In: MLSys
5. Calefato, F., Lanubile, F., & Quaranta, L. (2022). A preliminary investigation of mlops practices in github. In: Proceedings of the 16th ACM / IEEE International Symposium on Empirical Software Engineering and Measurement. p. 283-288. ESEM '22, Association for Computing Machinery, New York, NY, USA https://doi.org/10.1145/3544902.3546636
6. Garg, S., Pundir, P., Rathee, G., Gupta, P., Garg, S., & Ahlawat, S. (2021). On continuous integration/continuous delivery for automated deployment of machine learning models using mlops. In: 2021 IEEE

[7] https://iml4e.org/en/iml4e/cabc.

Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE). pp. 25–28. https://doi.org/10.1109/AIKE52691.2021.00010

7. Gokarna, M., & Singh, R. (2021). Devops: a historical review and future works. In: 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS). pp. 366–371. https://doi.org/10.1109/ICCCIS51004.2021.9397235
8. Granlund, T., Stirbu, V., & Mikkonen, T. (2021). Towards regulatory-compliant mlops: Oravizio's journey from a machine learning experiment to a deployed certified medical product. *SN Computer Science, 2*(5), 1–14.
9. Gualo, F., Rodriguez, M., Verdugo, J., Caballero, I., & Piattini, M. (2021). Data quality certification using ISO/IEC 25012: industrial experiences. *Journal of Systems and Software, 176*, 110938. https://doi.org/10.1016/j.jss.2021.110938
10. Hardt, M., Recht, B., & Singer, Y. (2016). Train faster, generalize better: stability of stochastic gradient descent. In: International conference on machine learning. pp. 1225–1234. PMLR
11. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284.
12. For Information Security (BSI). (2021). F.O.: Kriterienkatalog für ki-cloud-dienste - aic4. https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/CloudComputing/AIC4/AI-Cloud-Service-Compliance-Criteria-Catalogue_AIC4.pdf?__blob=publicationFile &v=4. Accessed 27 Feb 2023
13. International Organization for Standardization: Information technology - artificial intelligence - guidance on risk management. (2023). https://www.iso.org/standard/77304.html
14. ISO/IEC: Systems and software engineering – systems and software quality requirements and evaluation (SQuaRE) – Data quality model. (2008). ISO/IEC 25012, International Organization for Standardization, Geneva, Switzerland
15. ISO/IEC: ISO/IEC 25024, Systems and Software Engineering - Systems and Software Quality Requirements and Evaluation (SQuaRE) - Measurement of Data Quality. No. 25024, ISO, Geneva, Switzerland. (2015). https://books.google.de/books?id=wYXKtAEACAAJ
16. Knoblauch, D., & Banse, C. (2019). Reducing implementation efforts in continuous auditing certification via an audit API. In: 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE). pp. 88–92. IEEE
17. Knoblauch, D., & Großmann, J. (2022). Towards continuous audit-based certification for mlops. In: Proceedings of the International Workshop on AI Compliance Mechanism. WAICOM 2022 Co-Chairs
18. Knoblauch, D., Großmann, J., Strick, L., & Pannetrat, A. (2019). Europäisches rahmenwerk für continuous auditing based certification. In: Tagungsband zum 16. IT-Sicherheitskongress des BSI. pp. 495–504. SecuMedia
19. Lettner, C., Stumptner, R., Fragner, W., Rauchenzauner, F., & Ehrlinger, L. (2021). Daql 2.0: Measure data quality based on entity models. *Procedia Computer Science, 180*, 772–777. https://doi.org/10.1016/j.procs.2021.01.327
20. Moreschi, S., Recupito, G., Lenarduzzi, V., Palomba, F., Hastbacka, D., & Taibi, D. (2023). Toward end-to-end mlops tools map: a preliminary study based on a multivocal literature review
21. Mukherjee, S., Niyogi, P., Poggio, T., & Rifkin, R. (2006). Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics, 25*(1), 161–193.
22. Rauber, J., Brendel, W., & Bethge, M. (2017). Foolbox v0.8.0: A python toolbox to benchmark the robustness of machine learning models. CoRR **abs/1707.04131**, http://arxiv.org/abs/1707.04131
23. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS One, 10*(3), e0118432.
24. Schelter, S., Lange, D., Schmidt, P., Celikel, M., Biessmann, F., & Grafberger, A. (2018). Automating large-scale data quality verification. *Proceedings of the VLDB Endowment, 11*(12), 1781–1794. https://doi.org/10.14778/3229863.3229867
25. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., & Crespo, J. (2015). Hidden technical debt in machine learning systems. In: Advances in neural information processing systems. pp. 2503–2511