BULLETIN of Cherkasy State Technological University

Journal homepage: https://bulletin-chstu.com.ua/en

Vol. 31 No. 3. 2026

UDC 004.056:006.91:004.8
DOI: 10.62660/bcstu/3.2026.77

Metrological principles for the application of artificial intelligence in software certification

Oleh Liubuskyi*
Postgraduate Student
Lviv Polytechnic National University
79000, 12 Stepan Bandera Str., Lviv, Ukraine
https://orcid.org/0009-0009-8796-4753
Tetiana Bubela
Doctor of Technical Sciences, Professor
Lviv Polytechnic National University
79000, 12 Stepan Bandera Str., Lviv, Ukraine
https://orcid.org/0000-0002-2525-9735

Abstract. The relevance of this research stems from the need for metrological assurance for the application of artificial intelligence methods in software certification procedures, which require quantitatively defined quality indicators, traceability of results, and the assessment of uncertainty in certification conclusions. The aim of this work was to develop a metrology-oriented conceptual model for the use of intelligent methods in software conformity assessment processes. The study employed systems analysis, the method of decomposing certification processes, a classification-based approach to generalising artificial intelligence methods, and a methodology for establishing metrological characteristics for measuring the quality indicators of models. As a result, a representative sample of contemporary scientific sources was compiled, on the basis of which the main areas of application of intelligent algorithms in certification procedures were identified. A structural and functional analysis of the software assessment life cycle was carried out, and key points for integrating intelligent methods into the processes of evidence base formation, artefact tracing, quality metric assessment and decision support were identified. A system of metrological characteristics has been developed for the subprocesses of data preparation, model training, testing and generalisation of results, which includes quality metrics, sources of uncertainty, methods for its assessment, traceability mechanisms and reproducibility criteria. An approach has been proposed for establishing the uncertainty budget of a certification decision as an aggregated characteristic, taking into account the contribution of all stages of data processing and model operation. Artificial intelligence methods have been classified according to their role in certification procedures, taking into account the level of metrological certainty, the risk of application and the need for uncertainty assessment. A structural-logical model for the integration of intelligent methods into software certification has been developed, combining the processes of data processing, metrological evaluation and conformity decision-making. The practical value of the results obtained lies in the possibility of using the proposed model to develop standardised procedures for evaluating artificial intelligence models, implementing reproducibility protocols, ensuring the traceability of certification artefacts, and providing a quantitative justification for conclusions regarding software conformity

Keywords: metrological assurance; conformity assessment; measurement uncertainty; artefact traceability; reproducibility of results

Article's History: Received: 22.05.2026; Revised: 31.08.2026; Accepted: 21.09.2026; Published: 07.10.2026

Suggested Citation:
Liubuskyi, O., & Bubela, T. (2026). Metrological principles for the application of artificial intelligence in software certification. *Bulletin of Cherkasy State Technological University*, 31(3), 77-91. doi: 10.62660/bcstu/3.2026.77.

*Corresponding author (Liubuskyi120@proton.me)

Copyright © The Author(s). This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/)

INTRODUCTION

The rapid integration of artificial intelligence (AI) methods into software development and testing processes necessitates a rethinking of approaches to software evaluation and certification within the modern metrological paradigm. Intelligent algorithms are increasingly being used for automated test generation, defect prediction, code quality analysis and decision support throughout the software lifecycle, which significantly improves the efficiency of engineering processes, bAut at the same time raises new challenges regarding trust in the results obtained, their reproducibility and traceability. As noted by M. Ogrizović *et al.* (2024), the application of machine learning methods in modern software systems operating at the intersection of software engineering, AI, and data analysis is accompanied by increasing requirements for explainability, reliability, safety, and the proper assessment of the quality of such solutions. Furthermore, according to A. Escalante-Viteri & D. Mauricio (2025), the application of AI is transforming classical testing procedures, shifting the focus from traditional metrics to more complex multidimensional quality indicators that include semantic, dynamic and behavioural characteristics of software systems.

At the same time, the proliferation of deep learning models in software engineering has revealed a critical problem regarding the reproducibility of experimental results, caused by the stochastic nature of the learning process, sensitivity to the data sample and the incompleteness of reproducibility artefacts. In particular, according to the findings of a study by C. Liu *et al.* (2021), only a small proportion of research studies allow for the full replication of the results obtained, which undermines their scientific validity and complicates their use in certification procedures. For metrology, this means a lack of the necessary traceability of measurement procedures and uncertainty regarding the stability of the assessed software quality indicators.

An additional factor is that intelligent methods are increasingly being used to analyse anomalies and behavioural deviations in software systems, which broadens the range of measurable characteristics but requires a formalised definition of the metrological parameters for such assessments. In particular, as noted by A.B. Nassif *et al.* (2021), when analysing anomalies, it is essential to take into account confidence intervals, errors and acceptance criteria. In this context, according to the study by V. Polužanski *et al.* (2022), the issue of assessing the uncertainty of AI results takes on key importance, as the data used to train models may be incomplete, noisy or heterogeneous, which directly affects the reliability of predictions. As demonstrated in the study by T. Dorst *et al.* (2023), failing to account for measurement uncertainty leads to significant distortions in machine learning results, whereas applying GUM principles and propagating uncertainty through the data processing stages makes it possible to increase the robustness of models and ensure the validity of the estimates obtained. This confirms the need to integrate metrological approaches into the validation and certification processes for intelligent software systems.

At the level of digital quality infrastructure, a new field is emerging – AI metrology. In particular, as defined by H. Rabus *et al.* (2025), its aim is to develop standard metrics for assessing explainability, robustness, data quality and model uncertainty, as well as to ensure traceability to reference datasets and reference procedures. In the regulatory sphere, this is consistent with approaches to assessing the conformity of AI systems, where, as noted by L. Floridi *et al.* (2022), conformity procedures are to be carried out on the basis of formalised criteria of transparency, non-discrimination and reproducibility. Despite the significant body of research into the application of AI in software engineering, there is no comprehensive, metrology-oriented model that combines the requirements of reproducibility, traceability and uncertainty assessment in software certification procedures. As noted by A. Escalante-Viteri & D. Mauricio (2025), existing approaches largely focus on individual aspects – algorithmic efficiency, prediction accuracy or test automation – without formalising them as measurable metrological characteristics suitable for use in conformity assessment systems.

Thus, the relevance of this research stems from the need to establish a metrological framework for the application of AI in software certification, which will ensure greater reliability in assessing its quality and compliance with regulatory requirements. The aim of this article was to systematise contemporary scientific approaches to the use of AI methods in software assessment and to develop a classification-based, metrology-oriented conceptual model for their application in certification procedures, taking into account the requirements of reproducibility, traceability and uncertainty assessment. To achieve this aim, three interrelated objectives were identified in the study. The first objective involved systematising modern approaches to the application of AI methods in software testing, quality assessment and certification. The second objective involved formalising the metrological characteristics of the use of intelligent models, in particular indicators of reproducibility, traceability and the quantitative assessment of uncertainty in results. The third objective was to develop a conceptual, metrology-oriented model for integrating AI methods into software certification procedures, taking into account the criteria of acceptability, risk and the evidential basis of the certification decision.

LITERATURE REVIEW

Current scientific research into the application of AI methods for assessing software quality forms a multidimensional scientific discourse covering the tasks of

defect prediction, code structure analysis, ensuring the reproducibility of results, the traceability of artefacts and the development of conformity procedures. One of the key areas is the prediction of software defects based on complexity metrics and structural characteristics of the code. According to M. Ali *et al*. (2024), the effectiveness of defect prediction can be enhanced by combining the selection of informative features with the use of several heterogeneous classifiers and an ensemble approach, which yields better results compared to individual baseline models. At the same time, the authors emphasise the importance of the integrity and quality of the input data, as well as the need to apply a set of formalised evaluation metrics, which is directly linked to metrological requirements for the reproducibility of measurements.

Further development in this area is presented in the work by A. Khalid *et al*. (2023), which carries out a comparative analysis of classical and optimised machine learning models for defect prediction using clustering and meta-heuristic optimisation. The results achieved an accuracy of up to 99.8%; however, the authors emphasise the need to standardise sets of metrics and evaluation procedures to ensure the validity of model comparisons. Probabilistic approaches to defect prediction are discussed by M.J. Hernández-Molinos *et al*. (2023), who used Bayesian networks to construct classification models for software modules. It is shown that such models are characterised by lower variability in results during cross-validation, which is an important prerequisite for the metrological stability of software quality measurements.

An empirical study by W. Albattah & M. Alzahrani (2024) demonstrated the superiority of deep recurrent neural networks over traditional algorithms in defect prediction tasks based on a set of internal class metrics, specifically cohesion, connectivity and complexity. The authors emphasise the need for a multi-criteria evaluation of models based on the F-measure, accuracy and completeness, which is consistent with the principles of a metrological approach to multi-parameter measurements. In an industrial context, H. Kang & S. Do (2024) investigated the application of machine learning models for defect prediction in embedded telecommunications software, demonstrating an improvement in the F-measure through the introduction of new metric features. At the same time, the authors highlight the problem of limited generalisability of models across different domains, which points to the need for standardised metrological procedures for model calibration.

A separate area of research concerns methods for the automated detection of structural code defects. I. Ali *et al*. (2025) proposed a transformer model for detecting code smell using relational embeddings, which achieves high accuracy but is characterised by significant computational costs and reduced completeness. This highlights the need to assess the uncertainty of results and to take resource constraints into account in certification procedures. A systematic approach to software quality assurance using AI was proposed by J.R. Laracy *et al*. (2025), who view software systems as dynamic socio-technical ecosystems. Within this approach, quality is defined in terms of operational stability, feedback mechanisms and control of changes in the system's state, which is consistent with the metrological principles of managing measurement processes. A bibliometric analysis by P. Kokol (2022) revealed an exponential growth in the number of studies in the field of software quality and identified key thematic areas, amongst which machine learning methods for defect prediction and automated testing occupy a leading position. The author emphasises the absence of a comprehensive system for knowledge synthesis, which justifies the need to develop a generalised metrological evaluation model.

The issue of the reproducibility of machine learning results is addressed by M. Hartley & T.S.G. Olsson (2020), who proposed an approach to automatically recording the provenance of models and data based on the FAIR principles (Findable, Accessible, Interoperable, Reusable). This ensures the traceability of experiments and lays the groundwork for the metrological validation of AI models. A generalisation of approaches to ensuring reproducibility in software engineering is presented by A. Cordeiro & E. Oliveira Jr. (2025), who classified artefacts, environments and methods for assessing reproducibility, identifying containerisation, data repositories and standardised workflows as key tools. This classification can be interpreted as a basis for metrological assurance of the repeatability of quality measurements. An important aspect of software certification is the traceability of requirements and artefacts. E.R. Batot *et al*. (2022) proposed a feature model for systematising approaches to traceability, which allows the formalisation of relationships between requirements, code and tests. In turn, J. Lin *et al*. (2021) applied transformer-based language models to automatically establish such links, thereby improving the accuracy of traceability matrix construction. This creates the prerequisites for formalising metrological traceability in certification processes.

The issues of conformity and auditing of AI systems are addressed by M. von Zahn *et al*. (2025), who proposed a design framework for evaluating the fairness, explainability and performance of models. The proposed approach can be interpreted as a multi-criteria evaluation system that complies with metrological principles for constructing comprehensive quality indicators. In turn, J. Mökander *et al*. (2021) consider pre-market conformity assessment and post-market monitoring throughout a system's life cycle as complementary audit mechanisms for high-risk AI systems. At the same time, D. Hartmann *et al*. (2024) identify a regulatory gap linked to insufficient access by independent

researchers and civil society organisations to data and models, and justify the need for their involvement in shaping the European AI audit ecosystem.

A separate area of research is devoted to the formalisation and quantitative assessment of uncertainty in machine learning systems. The study by T. Siddique *et al.* (2022) presents a comprehensive review of approaches to assessing epistemic and aleatory uncertainty, which allows the results of models to be interpreted not merely as point estimates, but as distributions with confidence intervals, which is critically important in the context of metrological assurance of the reliability of predictions. Further development of uncertainty assessment methods is demonstrated in the study by C.E. Montanari *et al.* (2025), where the use of ensemble and stochastic approaches (in particular, Monte Carlo dropout and bootstrap) improves the interpretability of models and enables the assessment of their generalisation ability under conditions of limited data.

From the perspective of classical metrology, the issue of uncertainty is addressed in the work by P. Harris *et al.* (2025), which demonstrates that, for complex distributed measurement systems, it is necessary to extend traditional approaches to uncertainty estimation to account for correlations, data heterogeneity and model complexity. Using the example of geometric measurements of parts manufactured using 3D printing technology, T. Kozior *et al.* (2024) demonstrated that the choice of measuring instrument and the number of samples under investigation significantly influence both the final result and the estimation of its uncertainty. Similar ideas can be found in the research by J. Hartmann & H. Alkhatib (2023), where the application of machine learning methods to model measurement errors enables greater accuracy in estimates by taking into account a multitude of influencing factors, as well as in the work by K. Zhou *et al.* (2025), which demonstrates the need to analyse error propagation in complex multi-component systems as a prerequisite for improving the accuracy of final results.

Ensuring the reproducibility and traceability of the lifecycle of machine learning models is also an important area of focus. In the work by D. Kreuzberger *et al.* (2023), the concept of MLOps is viewed as a set of principles and architectural solutions that ensure automation, version control and continuous integration of models, thereby creating the technical foundation for their reproducibility and metrological consistency. In turn, M. Micheli *et al.* (2023) emphasise the role of data and model documentation as a tool for ensuring transparency, traceability and compliance with regulatory requirements, which is a key element in building trust in AI systems in the context of their certification.

The literature review conducted has shown that current research covers specific aspects of AI application in software quality assessment, in particular defect prediction, automated code analysis, and ensuring reproducibility and traceability. At the same time, there is no integrated, metrology-oriented conceptual model that would combine requirements for uncertainty, repeatability and traceability of results in software certification procedures. This confirms the need to systematise existing approaches and to develop a generalised classification model for the application of AI methods in a metrological context.

MATERIALS AND METHODS

The research was carried out within the framework of a metrology-oriented approach to analysing the processes of AI application in software certification. The methodology is based on the principles of systems analysis, formalisation of measurement procedures, structural-logical modelling and multi-criteria classification. The general sequence of research stages is shown in Figure 1.

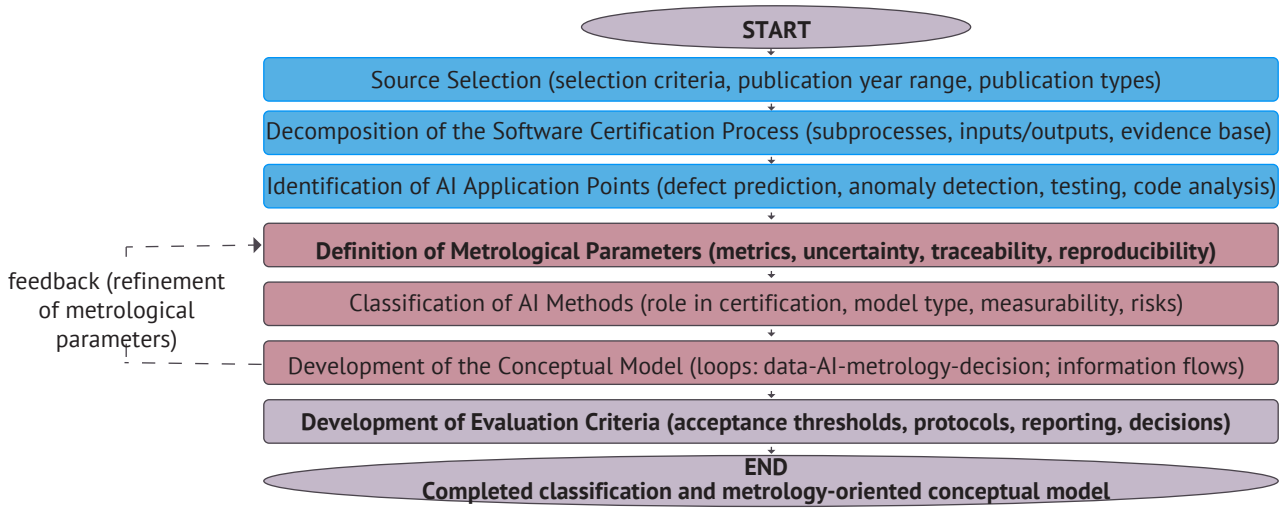


Figure 1. Stages of a metrology-oriented study into the use of AI in software certification

Source: compiled by the authors

In the first stage, the object and subject of the study were identified. The object was defined as the software certification process in the context of the use of AI methods, whilst the subject was the metrological characteristics of quality assessment procedures, including reproducibility, traceability and the assessment of uncertainty in results. The second stage involved compiling a sample of academic sources for analysis. The sample included peer-reviewed publications on software defect prediction, automated code analysis, ensuring the reproducibility of experiments, artefact tracing and procedures for assessing the conformity of intelligent systems. The criteria for selecting sources were the presence of formalised methods for assessing software quality, the use of machine learning or deep learning algorithms, the description of metrics for accuracy, stability or other quantitative indicators, and the possibility of interpreting the results in a metrological context.

In the third stage, a structural and functional decomposition of the software certification process was carried out with the aim of identifying specific points of integration for AI methods. The certification process was presented as a sequence of interrelated procedures: submission of an application, technical documentation, requirements specifications, test artefacts and datasets; an initial check of the completeness and consistency of the submitted materials; automated analysis of the quality of data, code, requirements and test scenarios using AI methods; validation or retraining of the model on a test or reference dataset; testing of software and intelligent components against defined metrics of accuracy, completeness, F-measure, stability and robustness; assessment of the uncertainty of the results obtained and establishment of an uncertainty budget for the certification decision; establishing traceability links between requirements, tests, models, datasets and evaluation results; compiling a certification report; making a decision on compliance, conditional compliance or non-compliance; post-certification monitoring of the stability of the software's performance characteristics. This decomposition made it possible to view AI methods not as an isolated automation tool, but as a component of the measurement and evidence-based certification procedure.

The fourth stage involved formalising the metrological characteristics of the assessment. These were defined using a four-step procedure. In the first step, a measurable outcome was established for each certification subprocess: for data preparation – completeness, balance, integrity and representativeness of the sample; for model training – accuracy, completeness, F-measure, stability and sensitivity to hyperparameters; for testing – coverage, defect detection capability, robustness and stability of test results; for decision-making – aggregated conformity assessment and the risk of a false conclusion. In the second step, relevant quantitative metrics were selected for each measured outcome; these had to be interpretable, reproducible and sensitive to changes in data quality or model parameters. In the third step, sources of uncertainty were identified, including data noise and incompleteness, sample bias, labelling errors, the stochastic nature of training algorithms, instability of the test environment, and subjectivity in expert aggregation. In the fourth step, mechanisms to ensure traceability and reproducibility were established for each characteristic: versioning of datasets and code, fixing of model parameters, control of random seeds, logging of experiments, preservation of test sets, and the use of a standardised runtime environment. Thus, the metrological characteristics were determined not declaratively, but by comparing the certification subprocess, the measured result, the metric, the source of uncertainty and the method of controlling it.

In the fifth stage, AI methods were classified according to their role in certification procedures. The classification criteria selected were functional purpose, covering defect prediction, anomaly detection, code structure analysis and decision support, as well as model type, specifically ensemble, Bayesian, deep learning and transformer approaches. Additionally, the level of metrological certainty of the results, the ability to ensure traceability, and sensitivity to data uncertainty were taken into account.

In the sixth stage, a structural-logical metrological model of the application of AI methods in software certification was constructed. The model was constructed by sequentially combining three levels: process, metrological and decision. At the process level, the stages of certification and the corresponding artefacts to be analysed were identified. At the metrological level, measurable characteristics, metrics, sources of uncertainty and means of traceability were established for each stage. At the decision-making level, rules were formulated for transitioning from individual quantitative indicators to an integrated conformity assessment.

To reach a conformity decision, threshold criteria were established for key evaluation metrics. The minimum acceptable values were set at accuracy and completeness of no less than 0.8, and an F-measure of no less than 0.75; the variation in results across a series of repeated experiments must not exceed 5%. An integrated conformity decision was made provided that all established criteria were satisfied simultaneously, while borderline values required additional expert analysis. To achieve the stated aim, systems analysis was used to decompose the certification process, measurement theory methods were applied to formalise the metrological characteristics, and structural and logical modelling was used to construct the conceptual model. In addition, classification analysis was employed to systematise AI methods, multi-criteria assessment methods were used to determine model quality indicators, and comparative analysis was applied to compare the characteristics of different types of algorithms. The reproducibility of the

study was ensured through the formalisation of all analytical procedures, the definition of source selection criteria, the standardisation of assessment metrics, and the description of the model construction algorithm. For each stage, a set of input data, processing procedures, and output parameters was established, allowing the study to be replicated under analogous conditions.

RESULTS AND DISCUSSION

The compiled sample of sources encompasses studies addressing reproducibility, traceability, uncertainty assessment, auditing, and the metrological assurance of AI systems used in software engineering. The generalised characteristics of the sample are presented in Table 1.

Table 1. Characteristics of the source sample

No.	Main focus of the study	Type/class of approach	Formalisation of metrological characteristics	Uncertainty quantification (UQ)	Traceability	Reproducibility	Suitability as an evidence base for certification
1	Reproducibility of Deep Learning in Software Engineering	An empirical study of DL (Deep Learning)	Reproducibility and replicability metrics	Not considered as a separate UQ loop	Partly due to the description of experiments	Analysis of reproducibility factors	Limited without UQ integration
2	Conformity assessment procedure (EU AI Act)	A procedural framework	Conformity and evidence criteria	UQ is not formalised quantitatively	High, due to documentation requirements	Depends on implementation by the evaluated entity	High for regulatory purposes
3	AI metrology in medicine	A metrological framework for AI	Formalised accuracy and robustness metrics	Quantitative estimation of uncertainty	Yes, due to calibration and reference procedures	Described to a limited extent	High
4	Expressing measurement uncertainty through machine learning	ML regression/ classification	Interval estimates of parameters	Yes, via intervals and calibration	Not the main focus	Not detailed	Medium (requires integration with traceability)
5	The impact of measurement uncertainty on machine learning	ML for sensor systems	RMSE (Root Mean Square Error) and error analysis	Yes, uncertainty propagation in accordance with GUM (Guide to the Expression of Uncertainty in Measurement)	Due to the description of the measurement chain	Partly due to reproducible experiments	High in measurement domains
6	Reproducibility of Deep Learning	An instrumental approach (dtoolAI)	Provenance metrics and experiment tracking	Not the main focus	Yes, due to the provenance chain	Yes, via a reproduction package	Medium (without UQ)
7	An overview of reproducibility in software engineering	Systematic review	Classification of reproducibility factors	Not considered	Partly due to the classification of artefacts	Description of practices	Medium
8	Traceability approaches	Feature model	Formalisation of traceability characteristics	Not considered	Yes (structural model)	Not the main focus	Medium
9	Traceability via BERT (Bidirectional Encoder Representations from Transformers)	NLP/ML model	Precision/ Recall/F1	UQ is not formalised	Yes (automatic linking)	Partly due to the experimental design	Medium
10	AI conformity framework	Design framework audit	Multi-criteria metrics (fairness, explainability)	UQ as a component of reliability	Due to the audit trail	Not described in technical detail	High regulatory significance

Continued Table 1.

No.	Main focus of the study	Type/class of approach	Formalisation of metrological characteristics	Uncertainty quantification (UQ)	Traceability	Reproducibility	Suitability as an evidence base for certification
11	Survey of uncertainty quantification (UQ) in ML	Review of UQ methods	Calibration, epistemic/ aleatoric uncertainty	Yes, methods are systematised	Not the main focus	Not considered	Medium (theoretical basis)
12	ML for uncertainty assessment	Applied ML UQ	Prediction intervals	Yes, the main result	Not the main focus	Not detailed	Medium
13	Impact of measurement uncertainty	ML fault diagnosis	Accuracy + error analysis	Yes	Partially	Not detailed	Medium
14	Sensor network metrology	Metrology of sensor networks	Traceability indicators	Yes, full uncertainty budget	Yes	Limited	High
15	Uncertainty of measurements	ML-assisted measurement	Uncertainty assessment for different methods	Yes	Partially	Not detailed	Medium
16	Uncertainty-aware ML	ML with integrated UQ	Prediction + uncertainty	Yes	Not the main focus	Not detailed	Medium
17	Uncertainty modelling point clouds	ML uncertainty modelling	Point cloud error models	Yes	Not the main focus	Not detailed	Medium
18	Vision measurement system UQ	End-to-end ML	Uncertainty analysis	Yes	Partially	Not detailed	Medium
19	NeurIPS Reproducibility Program	Software report	Checklist reproducibility	Not considered	Through standardised criteria	Yes (replay programme)	Medium
20	MLOps (Machine Learning Operations)	Architectural review	KPI monitoring and version control	Not a central element	Yes (versioning, pipeline)	Yes (via CI/CD – Continuous Integration / Continuous Delivery)	High infrastructure significance
21	Audit y EU AI Regulation	Regulatory guide	Post-market monitoring criteria	UQ indirectly	Yes (audit trail)	Depends on procedures	High
22	EU AI audit ecosystem	Policy and regulatory analysis	Governance indicators	UQ not formalised	Yes (institutional level)	Not technically applicable	Medium-high
23	AI documentation approaches	Analysis of model cards/ datasheets	Documentation standards	Not quantitative	Yes (through documentation artefacts)	Partially	High in terms of the evidence base
24	Transparent model cards	Model cards framework	Transparency/ usability indicators	Indirectly	Yes	Partially	High in medical domains
25	Reproducibility ML research	Review of barriers and drivers	Reproducibility factors	Not considered	Partially	Description of practices	Medium

Source: compiled by the authors

The analysis showed that the majority of studies focus on the quantitative assessment of uncertainty in model results, whilst the formalisation of reproducibility procedures and the availability of complete reproduction packages are covered only to a limited extent. This indicates an imbalance between the development of algorithmic approaches and the metrological assurance for their application in certification procedures. It was found that only a few studies provide a complete chain of traceability from data to the results obtained,

which makes it impossible to use them directly as an evidence base during software conformity assessment.

A structural and functional analysis of the certification process has made it possible to identify key points for the integration of AI methods, covering the stages of evidence base formation, quality assessment, decision support and post-certification monitoring. A map of these points is shown in Figure 2.

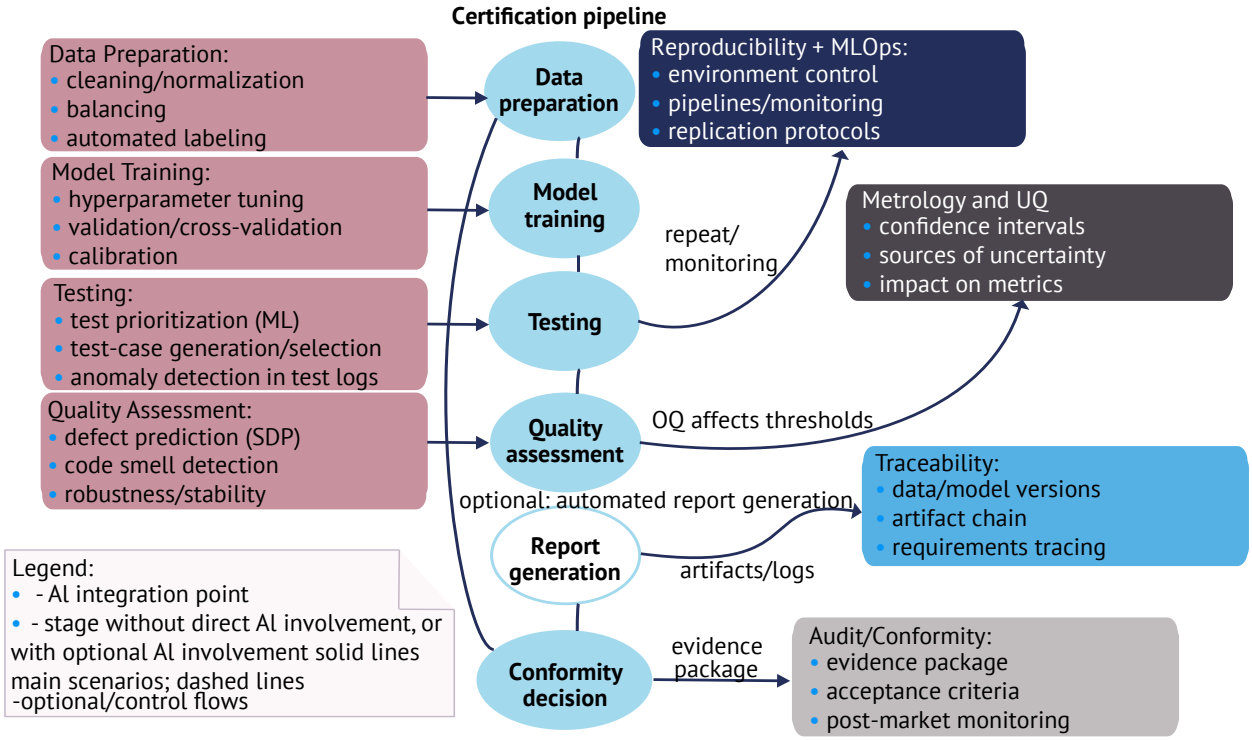


Figure 2. Map of points where AI methods are integrated into software certification procedures
Source: compiled by the authors

The results obtained show that the application of AI models is not limited solely to defect prediction or automated testing, but extends to procedures for tracing artefacts, analysing documentation, generating composite quality indicators and justifying certification conclusions. At the same time, each point of integration is characterised by a different level of metrological certainty: the processes for generating quality metrics are the most formalised, whilst the decision-making stages remain dependent on expert interpretation and the aggregation of indicators. The metrological characteristics for assessing AI models were formalised for the main certification subprocesses, and the results are summarised in Table 2.

Table 2. Metrological characteristics for the assessment of AI models by subprocess

Subprocess	Quality metrics	Sources of uncertainty	Uncertainty assessment method	Traceability	Reproducibility	Environment control
Data	data quality/ coverage; balance; integrity; data leakage metrics	measurement error/noise; sampling bias; omissions; labelling errors	Bootstrapping/ permutation methods; calibration; interval estimates for quality parameters	data versioning; description of origin (datasheets/ model cards); change log	sample fixation; randomness control during splits; repeated fold generation	containers; library/ operating system versioning; hardware specifications
Training	accuracy/F1/ Area Under the Curve/ RMSE; loss; training stability	stochasticity; Stochastic Gradient Descent; initialisation; hyperparameters; data drift	Ensemble/Monte Carlo Dropout/ Bayesian; predictive intervals; sensitivity analysis	code/model versioning; experiment register; training parameters	seed control; repeated runs; reproduction protocols; artefact repository	MLOps pipeline; Continuous Integration / Continuous Delivery; GPU/ CPU monitoring; dependency versioning

Continued Table 2.

Subprocess	Quality metrics	Sources of uncertainty	Uncertainty assessment method	Traceability	Reproducibility	Environment control
Testing	coverage; mutation score; fault revelation; robust metrics	insufficient testing; noise in oracles; instability of test environments	Uncertainty quantification (UQ) on outputs; bias testing/data outside the training sample; confidence calibration	trace links: requirements-tests-results; run log	repeatable tests; environment reproduction; preservation of test suites	managed environments; isolation; tool version control
Assessment/ solutions	acceptance thresholds; composite score; fairness/ explainability checks	uncertainty of metrics; aggregation; the human factor in interpretation	Combined UQ (metrics + model); uncertainty budget; decision risk	audit trail; reporting; decision log; version identification	verification of result reproducibility; independent verification	standardised report templates; access control; log immutability

Source: compiled by the authors

It has been established that, for the data preparation stage, the key factors are the quality and completeness of the sample, which are determined taking into account measurement errors, noise and distribution shifts. It was found that, for the model training stage, the dominant sources of uncertainty are the stochasticity of optimisation algorithms, the variability of hyperparameters and data drift, which necessitates the use of ensemble and Bayesian approaches to estimate interval characteristics of the results. During the testing phase, the need to combine classical coverage metrics with robustness assessment and the calibration of confidence intervals was identified. For the decision-making phase, the formation of an uncertainty budget is proposed, which aggregates the contributions of all previous subprocesses and allows for a quantitative assessment of the risk of a false compliance conclusion. An important result is the establishment of correspondence between metrological procedures and traceability mechanisms, in particular data versioning, experiment logging and the use of standardised execution environments. A classification of AI methods according to their role in certification procedures is given in Table 3.

Table 3. Classification of AI methods by role in certification

Task class	Model type	Certification stage	Level of metrological certainty	Risk	Need for UQ
Traceability	NLP/BERT (Bidirectional Encoder Representations from Transformers), IR models (Information Retrieval), graph models	building the evidence base; linking requirements, tests and artefacts	medium-high (link quality metrics are known)	medium (false correlations affect the audit)	limited (UQ is desirable for confidence in the relationships)
Reproducibility	experiment-tracking tools; pipelines; reproducibility protocols	the entire process: from data to decision	high (formalised protocols/artifacts)	high (risk of non-reproducible certification conclusions)	optional, but useful (for assessing stability)
Uncertainty Quantification (UQ)	Bayesian/ensembles/ MC-dropout, calibration, intervals	model evaluation; establishing acceptance criteria	high (subject to an uncertainty budget)	high (risk of false compliance/non-compliance)	high (main focus)
Audit/ conformity	procedural frameworks; checklists; post-market monitoring	decision-making; post-market monitoring	medium (often qualitative criteria + metrics)	high (regulatory/ ethical risks)	medium (UQ is required to justify reliability)
Code quality	Transformers/NLP, static analysis, classification	verification of implementation quality; evidence base	medium (code smell/pattern metrics)	medium	low-medium
Defect prediction	ensembles, Bayes, DL (Deep Learning)	assessment of defect risk prior to release	medium	medium-high (high-risk domains)	medium

Source: compiled by the authors

The results obtained have made it possible to identify functional classes corresponding to different levels of metrological uncertainty and risk. Artifact tracing methods are characterised by relatively high metric stability, but require additional assessment of the reliability of the relationships. Approaches to ensuring reproducibility form the basis of the metrological infrastructure, as they ensure the repeatability of results regardless of the execution environment. Methods for assessing uncertainty are crucial for establishing acceptance thresholds and determining confidence intervals for quality indicators. Audit and procedural frameworks integrate quantitative and qualitative criteria, enabling their use to justify certification decisions. Methods for defect prediction and code quality analysis are characterised by a moderate level of metrological certainty and require additional calibration for use in critical domains

The resulting model for the application of AI methods in certification procedures comprises three interrelated loops: the data generation and model training loop, the metrological evaluation of results loop, and the conformity decision-making loop. The interaction between these loops ensures the transfer of information regarding uncertainty, model parameters and measurement results between the stages of certification. The structural-logical model for the application of AI methods in certification procedures is shown in Figure 3.

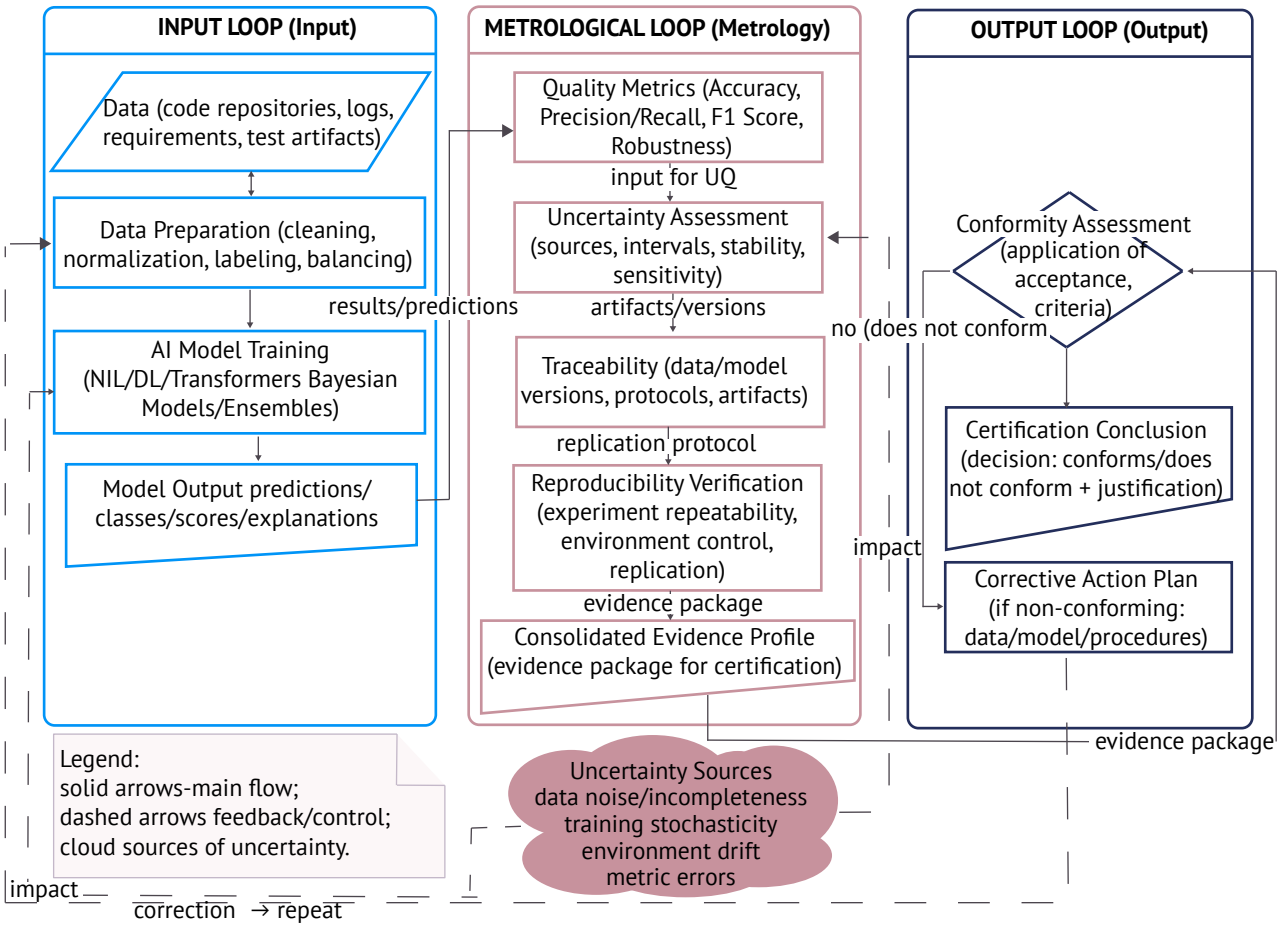


Figure 3. RA metrological model of AI integration into the software certification process
Source: compiled by the authors

The results obtained confirm that current research into AI in software engineering predominantly describes the functional performance of algorithms (accuracy/errors, test quality, performance), whilst metrological requirements (traceability, reproducibility, formalised uncertainty assessment and control of measurement conditions) remain scattered across different approaches and disciplines. This is clearly evident in the structure of the sample from this study: even when works focus on quality assurance or testing, they often fail to identify uncertainty as a controllable parameter in certification decisions, which is consistent with the criticism of the stochasticity and opacity of large models and datasets, which complicates the formation of an evidence base for compliance.

In this context, the findings of the study by J. Fehr *et al.* (2024), who conducted an empirical assessment of the public transparency of certified medical AI products

and found that even for solutions authorised for use in high-risk domains, the level of transparency regarding training data, ethical considerations, application limitations and risks remains insufficient; the significant gaps identified by the authors primarily concerned the documentation of training data, deployment caveats, safety monitoring and fairness-related characteristics. The results obtained in the present study align with these findings in that the mere declaration of AI use or even the existence of formal compliance procedures does not in itself provide sufficient evidence for a certification decision: just as the authors found, this study established that without a complete chain of traceability from data to outcome, without an explicit description of sources of uncertainty, and without a regulated presentation of the model's application limitations, the use of AI as an element of the evidence base remains limited. At the same time, there is a fundamental difference between these findings. Whilst the authors focus primarily on the issue of public transparency, the availability of product information and compliance with the ethical requirements of trustworthy AI in the healthcare sector, this study shifts the emphasis to a metrological interpretation of this issue, where transparency is viewed not as a self-sufficient requirement, but as one of the structural components of a broader system that includes reproducibility, uncertainty assessment, control of the conditions under which results are obtained, and the formalisation of acceptability criteria. In other words, J. Fehr *et al*. convincingly demonstrate a lack of information transparency even in a regulated environment, whereas the results obtained in the present study explain why a lack of transparency is not merely an ethical or regulatory deficiency but also a metrological problem, since it prevents verification of the stability, repeatability, and validity of the certification conclusion. It should also be noted that the authors of this source analyse primarily medical AI products and their public documentation, whereas the present study examines the broader scope of software certification, including points of AI integration in data preparation, testing, artefact tracing and compliance decision support, which allows the discussion to be shifted from the level of a transparency audit to that of constructing a comprehensive, metrology-oriented certification model.

The study by D. Manheim *et al.* (2025) highlights the limitations of current approaches to auditing AI systems, in particular the inadequacy of internal audits, the fragmented nature of standards, and the risk of formal "safety-washing". The results obtained in this study are consistent with these findings regarding the need for a systematic approach to assessing model quality, covering not only final metrics but also the entire life cycle – from data preparation to the certification decision. In particular, the proposed integration of metrological characteristics and the establishment of an uncertainty budget effectively corresponds to the idea of a shift from static checks to process-oriented auditing. At the same time, unlike D. Manheim *et al.*, where the main focus is on institutional mechanisms (the establishment of independent audit standards boards and the fostering of a culture of safety), this study shifts the emphasis to the formalisation of quantitative metrological assessment parameters, in particular metrics of quality, uncertainty and traceability. It should be noted separately that D. Manheim *et al.* devote considerable attention to aspects of process auditing, rather than merely product auditing, as well as to the need for continuous monitoring following the implementation of systems, which complements the findings regarding the inclusion of post-certification control as a separate evaluation loop. This indicates a convergence of contemporary research towards a transition to continuous, adaptive and multi-level AI evaluation models.

In their study, E. Bender *et al.* (2021) critically examine the nature of modern language models, emphasising their tendency to reproduce statistical dependencies without a deep understanding of meaning, which calls into question the reliability and interpretability of the results obtained. The results obtained in this study are consistent with these conclusions regarding the need to formalise trust in models through metrological characteristics, in particular the assessment of uncertainty and the assurance of traceability. At the same time, unlike the authors, who focus primarily on the ethical, social and semantic risks of using models, this study emphasises the quantitative interpretation of these risks through metrological indicators.

In their work, J. Pineau *et al*. (2020) highlight the problem of reproducibility of results in machine learning, linking it to insufficient transparency of experimental procedures, limited access to code and data, and incorrect statistical interpretation of results. The conclusions drawn from the present study are consistent with this approach regarding the need to improve the traceability of models and standardise the assessment of their quality. At the same time, unlike the cited work, where the main emphasis is placed on organisational mechanisms for ensuring reproducibility (checklists, code submission policies, reproduction challenges), this work proposes a more formalised approach focused on metrological assurance of the reliability of results.

The study by K. Shivashankar *et al*. (2025) examines the issues of maintainability and scalability of machine learning systems, which are linked to the increasing complexity of the model lifecycle, data dependency and the need to integrate engineering practices at the MLOps level. The results obtained are consistent with the findings of this study regarding the critical role of traceability, process controllability and the standardisation of model evaluation procedures. At the same time, unlike this work, which emphasises the metrological formalisation of quality and uncertainty metrics, K. Shivashankar *et al*. (2025) focus

primarily on the engineering aspects of the model lifecycle, such as dependency management, technical debt and infrastructure scaling.

In the work of A. Celepija *et al.* (2025), the concept of AI Product Cards is proposed as a formalised approach to documentation that combines descriptive characteristics of a system with its code and is intended to support the reuse and adaptation of AI solutions. The results obtained are consistent with the findings of this study regarding the need to improve transparency, standardisation and traceability throughout the model lifecycle. At the same time, unlike this study, where the main focus is on assessing the quality, reliability and compliance of models, the study cited shifts the emphasis to documentation as a tool for ensuring the reusability and scalable deployment of AI in the public sector. Furthermore, the authors consider another aspect of the issue – the integration of documentation with executable code and support for adapting models to new domains, which broadens the classical understanding of transparency in AI systems.

In a systematic review by R. Pan *et al.* (2021), approaches to the selection and prioritisation of test cases using machine learning methods are summarised, enabling testing efficiency to be improved by optimising the order in which tests are executed and reducing resource consumption. The results obtained correlate with the conclusions of this study regarding the advisability of integrating intelligent methods into software quality assessment processes and decision-making support in certification procedures. At the same time, unlike the metrology-oriented approach proposed in the present study, the cited work focuses primarily on algorithmic efficiency and testing performance, without formalising uncertainty, traceability, and reproducibility as measurable characteristics. The authors also consider the optimisation of the testing process within the software development life cycle, whereas the present study focuses on using these results as part of the evidence base for certification procedures and conformity assessment.

The study by S. Gilbert *et al.* (2025) highlights the concept of transparent "model cards" with multi-level information presentation as a tool for enhancing trust and safety in medical AI systems. The authors emphasise that the mere existence of documentation does not guarantee genuine transparency, as without verified data and links to deeper levels of information, it can become a tool for declarative rather than factual disclosure of model characteristics. These findings are consistent with the conclusions of the study regarding the need to ensure the traceability, reproducibility and measurability of AI system characteristics as the foundation of trust. At the same time, unlike the proposed approach, the authors' main focus is on user interpretability and the standardisation of information presentation for various stakeholders, whilst the issues of formalising metrological characteristics and quantitatively assessing uncertainty remain outside the scope of consideration.

The study by H. Semmelrock *et al.* (2024) systematically analyses the problem of reproducibility in machine learning through the lens of barriers and drivers, covering technical, methodological and cultural aspects. The authors emphasise that the low level of reproducibility is due not only to a lack of access to data and code, but also to the stochastic nature of algorithms, the ambiguity of definitions of reproducibility, and distortions in research practices (in particular, selective reporting and the "publish-or-perish" culture). The results obtained are consistent with the findings of the present study regarding the critical role of traceability, standardisation of procedures and formalisation of experimental conditions as the basis for the reliability of evaluation. At the same time, unlike the proposed metrological approach, the authors of the present study place the main emphasis on systematising barriers and practical drivers (in particular, containerisation, documentation standards and checklists), without a thorough formalisation of the quantitative assessment of uncertainty in results.

The results obtained confirm that the integration of AI methods into certification procedures requires a shift from the piecemeal use of models to their systematic metrological assurance, ensuring consistency between accuracy, reproducibility and the assessment of uncertainty. At the same time, a comparison with current research reveals a gap between algorithmic efficiency and formalised requirements for evidence, which opens up scope for further research into the standardisation of metrological approaches to the certification of AI systems.

CONCLUSIONS

This article addresses the research objective of establishing the metrological foundations for the application of AI methods in software certification procedures by systematising contemporary approaches, formalising measurable characteristics, and constructing a classification-based, metrology-oriented conceptual model. The results obtained are consistent with the stated aim and confirm the need to move from the purely functional use of intelligent algorithms to their metrologically controlled application in conformity assessment processes. Based on the compiled selection of scientific sources, it has been established that current research focuses predominantly on improving the accuracy of models and automating testing, whilst issues of traceability, reproducibility and the quantitative assessment of uncertainty are addressed only in a fragmentary manner. This results in the limited suitability of most approaches for use as an evidence base in certification procedures requiring formalised and repeatable measurements of quality indicators.

A structural and functional analysis of the certification process has enabled the identification of key points for integrating AI methods into the software evaluation lifecycle – from the formation of the evidence base and the tracing of artefacts to the support of certification decision-making and post-marketing monitoring. It has been shown that the level of metrological certainty of the results is uneven: the stages of generating quality metrics are the most formalised, whilst the aggregation of indicators and decision-making remain dependent on expert interpretation. A system of metrological characteristics for the evaluation of AI models by certification subprocesses is proposed, which includes quality metrics, sources of uncertainty, methods for its assessment, traceability mechanisms, reproducibility criteria and control of execution conditions. The concept of the uncertainty budget of a certification decision is introduced as an aggregated characteristic that takes into account the contribution of all stages of data processing and model training, enabling a quantitative assessment of the risk of a false conclusion regarding conformity.

AI methods have been classified according to their role in certification procedures, taking into account the level of metrological certainty, the risk of application and the need for uncertainty assessment. It is demonstrated that reproducibility assurance tools form the basis of the metrological infrastructure, uncertainty assessment methods are critical for establishing acceptability thresholds, and approaches to artefact traceability ensure the formation of a chain of evidence from data to the certification conclusion. The developed structural-logical model for integrating AI into software certification combines three interrelated loops – data and model training, metrological evaluation, and conformity decision-making – and ensures the transfer of information on uncertainty, model parameters and measurement results between stages. This provides a theoretical basis for incorporating intelligent methods into the digital quality infrastructure and for developing regulatory and methodological documents governing their application.

Prospects for further research lie in the development of formalised methods for calibrating AI models in certification environments, the creation of reference datasets for metrological validation, the definition of standardised procedures for calculating the uncertainty budget, as well as the experimental validation of the proposed model in domains with heightened reliability and safety requirements. The integration of the proposed approach with MLOps practices and compliance audit systems to ensure continuous metrological traceability throughout the entire life cycle of software systems requires separate consideration.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Albattah, W., & Alzahrani, M. (2024). Software defect prediction based on machine learning and deep learning techniques: An empirical approach. *AI*, 5(4), 1743-1758. doi: 10.3390/ai5040086.

[2] Ali, I., Rizvi, S.S.H., & Adil, S.H. (2025). Enhancing software quality with AI: A transformer-based approach for code smell detection. *Applied Sciences*, 15(8), article number 4559. doi: 10.3390/app15084559.

[3] Ali, M., Mazhar, T., Al-Rasheed, A., Shahzad, T., Yasin Ghadi, Y., & Amir Khan, M. (2024). Enhancing software defect prediction: A framework with improved feature selection and ensemble machine learning. *PeerJ Computer Science*, 10, article number e1860. doi: 10.7717/peerj-cs.1860.

[4] Batot, E.R., Gérard, S., & Cabot, J. (2022). A survey-driven feature model for software traceability approaches. In E.B. Johnsen & M. Wimmer (Eds.) *Fundamental approaches to software engineering. FASE 2022. Lecture notes in computer science* (pp. 23-48). Cham: Springer. doi: 10.1007/978-3-030-99429-7_2.

[5] Bender, E.M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623). Canada (Virtual Event): Association for Computing Machinery. doi: 10.1145/3442188.3445922.

[6] Celepija, A., Palmero Aprosio, A., Lepri, B., & Kazhamiakin, R. (2025). AI product cards: A framework for code-bound formal documentation cards in the public administration. *Data & Policy*, 7, article number e1. doi: 10.1017/dap.2024.55h.

[7] Cordeiro, A., & Oliveira Jr, E. (2025). Reviewing reproducibility in software engineering research. In *Proceedings of the 27th international conference on enterprise information systems* (pp. 364-371). Porto: INSTICC. doi: 10.5220/0013456000003929.

[8] Dorst, T., Schneider, T., Eichstädt, S., & Schütze, A. (2023). Influence of measurement uncertainty on machine learning results demonstrated for a smart gas sensor. *Journal of Sensors and Sensor Systems*, 12(1), 45-60. doi: 10.5194/jsss-12-45-2023.

[9] Escalante-Viteri, A., & Mauricio, D. (2025). Artificial intelligence in software testing: A systematic review of a decade of evolution and taxonomy. *Algorithms*, 18(11), article number 717. doi: 10.3390/a18110717.

[10] Fehr, J., Citro, B., Malpani, R., Lippert, C., & Madai, V.I. (2024). A trustworthy AI reality-check: the lack of transparency of artificial intelligence products in healthcare. *Frontiers in Digital Health*, 6, article number 1267290. doi: 10.3389/fdgth.2024.1267290.
[11] Floridi, L., Holweg, M., Taddeo, M., Amaya, J., Mökander, J., & Wen, Y. (2022). CapAI – a procedure for conducting conformity assessment of AI systems in line with the EU artificial intelligence act. *SSRN Electronic Journal*. doi: 10.2139/ssrn.4064091.
[12] Gilbert, S., Adler, R., Holoyad, T., & Weicken, E. (2025). Could transparent model cards with layered accessible information drive trust and safety in health AI? *NPJ Digital Medicine*, 8(1), article number 124. doi: 10.1038/s41746-025-01482-9.
[13] Harris, P., Østergaard, P.F., Tabandeh, S., Söderblom, H., Kok, G., van Dijk, M., Luo, Y., Pearce, J., Tucker, D., Vedurmudi, A.P., & Iturrate-Garcia, M. (2025). Measurement uncertainty evaluation for sensor network metrology. *Metrology*, 5(1), article number 3. doi: 10.3390/metrology5010003.
[14] Hartley, M., & Olsson, T.S.G. (2020). dtoolAI: Reproducibility for deep learning. *Patterns*, 1(5), article number 100073. doi: 10.1016/j.patter.2020.100073.
[15] Hartmann, D., de Pereira, J.R.L., Streitbörger, C., & Berendt, B. (2024). Addressing the regulatory gap: Moving towards an EU AI audit ecosystem beyond the AI Act by including civil society. *AI and Ethics*, 5(4), 3617-3638. doi: 10.1007/s43681-024-00595-3.
[16] Hartmann, J., & Alkhatib, H. (2023). Uncertainty modelling of laser scanning point clouds using machine-learning methods. *Remote Sensing*, 15(9), article number 2349. doi: 10.3390/rs15092349.
[17] Hernández-Molinos, M.J., Sánchez-García, A.J., Barrientos-Martínez, R.E., Pérez-Arriaga, J.C., & Ocharán-Hernández, J.O. (2023). Software defect prediction with bayesian approaches. *Mathematics*, 11(11), article number 2524. doi: 10.3390/math11112524.
[18] Kang, H., & Do, S. (2024). ML-based software defect prediction in embedded software for telecommunication systems (Focusing on the case of Samsung Electronics). *Electronics*, 13(9), article number 1690. doi: 10.3390/electronics13091690.
[19] Khalid, A., Badshah, G., Ayub, N., Shiraz, M., & Ghouse, M. (2023). Software defect prediction analysis using machine learning techniques. *Sustainability*, 15(6), article number 5517. doi: 10.3390/su15065517.
[20] Kokol, P. (2022). Software quality: How much does it matter? *Electronics*, 11(16), article number 2485. doi: 10.3390/electronics11162485.
[21] Kozior, T., Bochnia, J., Bochenek, A., Malara, D., Nawotka, M., Jansa, J., Hajnys, J., Wojtowicz, A., & Mesicek, J. (2024). Estimating the uncertainty of measurements for various methods and 3D printed parts. *Applied Sciences*, 14(8), article number 3506. doi: 10.3390/app14083506.
[22] Kreuzberger, D., Kühl, N., & Hirschl, S. (2023). Machine learning operations (MLOps): Overview, definition, and architecture. *IEEE Access*, 11, 31866-31879. doi: 10.1109/access.2023.3262138.
[23] Laracy, J.R., Meng, Z., Kirova, V.D., Ku, C.S., & Marlowe, T.J. (2025). Software quality assurance and AI: A systems-theoretic approach to reliability, safety, and security. *Software*, 4(4), article number 30. doi: 10.3390/software4040030.
[24] Lin, J., Liu, Y., Zeng, Q., Jiang, M., & Cleland-Huang, J. (2021). Traceability transformed: Generating more accurate links with pre-trained BERT models. In *2021 IEEE/ACM 43rd international conference on software engineering (ICSE)* (pp. 324-335). Madrid: IEEE. doi: 10.1109/icse43902.2021.00040.
[25] Liu, C., Gao, C., Xia, X., Lo, D., Grundy, J., & Yang, X. (2021). On the reproducibility and replicability of deep learning in software engineering. *ACM Transactions on Software Engineering and Methodology*, 31(1), 1-46. doi: 10.1145/3477535.
[26] Manheim, D., Martin, S., Bailey, M., Samin, M., & Greutzmacher, R. (2025). The necessity of AI audit standards boards. *AI & SOCIETY*, 40(8), 6609-6624. doi: 10.1007/s00146-025-02320-y.
[27] Micheli, M., Hupont, I., Delipetrev, B., & Soler-Garrido, J. (2023). The landscape of data and AI documentation approaches in the European policy context. *Ethics and Information Technology*, 25(4), article number 56. doi: 10.1007/s10676-023-09725-7.
[28] Mökander, J., Axente, M., Casolari, F., & Floridi, L. (2021). Conformity assessments and post-market monitoring: A guide to the role of auditing in the proposed european AI regulation. *Minds and Machines*, 32(2), 241-268. doi: 10.1007/s11023-021-09577-4.
[29] Montanari, C.E., Appleby, R.B., Di Croce, D., Giovannozzi, M., Pieloni, T., Redaelli, S., & Van der Veken, F.F. (2025). Machine learning techniques for uncertainty estimation in dynamic aperture prediction. *Computers*, 14(7), article number 287. doi: 10.3390/computers14070287.
[30] Nassif, A.B., Talib, M.A., Nasir, Q., & Dakalbab, F.M. (2021). Machine learning for anomaly detection: A systematic review. *IEEE Access*, 9, 78658-78700. doi: 10.1109/access.2021.3083060.
[31] Ogrizović, M., Drašković, D., & Bojić, D. (2024). Quality assurance strategies for machine learning applications in big data analytics: An overview. *Journal of Big Data*, 11(1), article number 156. doi: 10.1186/s40537-024-01028-y.
[32] Pan, R., Bagherzadeh, M., Ghaleb, T.A., & Briand, L. (2021). Test case selection and prioritization using machine learning: A systematic literature review. *arXiv*. doi: 10.48550/ARXIV.2106.13891.
[33] Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Larochelle, H. (2020). Improving reproducibility in machine learning research (A report from the NeurIPS 2019 reproducibility program) (Version 4). *arXiv*. doi: 10.48550/ARXIV.2003.12206.

[34] Polužanski, V., Kovacevic, U., Bacanin, N., Rashid, T.A., Stojanovic, S., & Nikolic, B. (2022). Application of machine learning to express measurement uncertainty. *Applied Sciences*, 12(17), article number 8581. doi: 10.3390/app12178581.
[35] Rabus, H., *et al*. (2025). Metrology for artificial intelligence in medicine. *Measurement: Sensors*, 38, article number 101789. doi: 10.1016/j.measen.2024.101789.
[36] Semmelrock, H., Ross-Hellauer, T., Kopeinik, S., Theiler, D., Haberl, A., Thalmann, S., & Kowald, D. (2024). Reproducibility in machine learning-based research: Overview, barriers and drivers. *arXiv*. doi: 10.48550/ARXIV.2406.14325.
[37] Shivashankar, K., Al Hajj, G., & Martini, A. (2025). Maintainability and scalability in machine learning: Challenges and solutions. *ACM Computing Surveys*, 57(12), 1-36. doi: 10.1145/3736751.
[38] Siddique, T., Mahmud, M., Keesee, A., Ngwira, C., & Connor, H. (2022). A survey of uncertainty quantification in machine learning for space weather prediction. *Geosciences*, 12(1), article number 27. doi: 10.3390/geosciences12010027.
[39] von Zahn, M., Zacharias, J., Lowin, M., Chen, J., & Hinz, O. (2025). Navigating AI conformity: A design framework to assess fairness, explainability, and performance. *Electronic Markets*, 35(1), article number 24. doi: 10.1007/s12525-025-00770-2.
[40] Zhou, K., Chu, W., & Zhao, P. (2025). Monocular visual measurement system uncertainty analysis and one-step end-end estimation upgrade. *Sensors*, 25(23), article number 7179. doi: 10.3390/s25237179.

Метрологічні засади застосування штучного інтелекту в сертифікації програмного забезпечення

Олег Любуський
Аспірант
Національний університет «Львівська політехніка»
79000, вул. Степана Бандери, 12, м. Львів, Україна
https://orcid.org/0009-0009-8796-4753
Тетяна Бубела
Доктор технічних наук, професор
Національний університет «Львівська політехніка»
79000, вул. Степана Бандери, 12, м. Львів, Україна
https://orcid.org/0000-0002-2525-9735

Анотація. Актуальність дослідження зумовлена необхідністю метрологічного забезпечення застосування методів штучного інтелекту у процедурах сертифікації програмного забезпечення, що потребують кількісно визначених показників якості, простежуваності результатів та оцінювання невизначеності сертифікаційних висновків. Метою роботи стало формування метрологічно орієнтованої концептуальної моделі використання інтелектуальних методів у процесах оцінювання відповідності програмного забезпечення. У дослідженні було застосовано системний аналіз, метод декомпозиції процесів сертифікації, класифікаційний підхід до узагальнення методів штучного інтелекту, а також методологію формування метрологічних характеристик вимірювання показників якості моделей. У результаті було сформовано репрезентативну вибірку сучасних наукових джерел, на основі якої визначено основні напрями використання інтелектуальних алгоритмів у сертифікаційних процедурах. Проведено структурно-функціональний аналіз життєвого циклу оцінювання програмного забезпечення та встановлено ключові точки інтеграції інтелектуальних методів у процеси формування доказової бази, трасування артефактів, оцінювання показників якості та підтримки прийняття рішень. Розроблено систему метрологічних характеристик для підпроцесів підготовки даних, навчання моделей, тестування та узагальнення результатів, що включає метрики якості, джерела невизначеності, методи її оцінювання, механізми простежуваності та критерії відтворюваності. Запропоновано підхід до формування бюджету невизначеності сертифікаційного рішення як агрегованої характеристики, що враховує внесок усіх етапів оброблення даних і функціонування моделей. Виконано класифікацію методів штучного інтелекту за їх роллю у сертифікаційних процедурах із урахуванням рівня метрологічної визначеності, ризику застосування та потреби в оцінюванні невизначеності. Сформовано структурно-логічну модель інтеграції інтелектуальних методів у сертифікацію програмного забезпечення, що об'єднує контури оброблення даних, метрологічного оцінювання та прийняття рішення про відповідність. Практична цінність отриманих результатів полягає у можливості використання запропонованої моделі для розроблення стандартизованих процедур оцінювання моделей штучного інтелекту, впровадження протоколів відтворюваності, забезпечення простежуваності сертифікаційних артефактів та кількісного обґрунтування висновків про відповідність програмного забезпечення

Ключові слова: метрологічне забезпечення; оцінювання відповідності; невизначеність вимірювань; простежуваність артефактів; відтворюваність результатів