UDC 004.8:006.91

Assessing the reproducibility of numerical outputs of machine-learning algorithms for measurement tasks

V. Ashchepkov[1], D. Byallovich[1,2], V. Skliarov[1,2]

[1]*National Scientific Centre "Institute of Metrology", Myronosytska Str., 42, 61002, Kharkiv, Ukraine*
ashhepkovvalera@gmail.com; biallovych@gmail.com; vladimir.skliarov@gmail.com

[2]*Kharkiv National University of Radio Electronics, Nauky Ave., 14, 61166, Kharkiv, Ukraine*

Abstract

The paper studies the reproducibility of results generated by stochastic machine-learning models for measurement tasks, an aspect that becomes critical when such algorithms are integrated into procedures governed by ISO 5725 and ISO/IEC 17025. Using the Isolation Forest algorithm, it is shown that even with identical input data the output may vary between runs, creating an additional source of variability relevant for metrological interpretation.

Two implementations are compared: the standard Isolation Forest algorithm and an improved version proposed earlier. The enhanced model does not eliminate stochasticity but reduces its effect by averaging the isolation-path characteristics, normalizing results across different contamination values, and determining the threshold from the structure of the anomaly-score distribution, which contributes to the stability of outputs.

Both models were repeatedly run under identical conditions and assessed through type-A standard uncertainty according to ISO 5725. The standard implementation exhibited significantly higher variability, whereas the improved version showed better reproducibility.

The results indicate that the internal variability of machine-learning algorithms functions as a metrological characteristic. Its standard uncertainty can be evaluated and incorporated into the overall measurement-uncertainty budget, supporting the harmonisation of algorithmic methods with principles of accuracy and repeatability.

Keywords: metrology; standard uncertainty; reproducibility; machine learning; measurement; processing of measurement data.

Received: 25.11.2025 Edited: 09.12.2025 Approved for publication: 12.12.2025

1. Introduction

In modern metrological practice, the use of machine-learning methods for processing of measurement results, detecting deviations, and improving data quality is being widely considered at an increasing rate. Such methods enable automation of analysis, enhance the efficiency of handling large data sets, and allow prompt identification of potentially incorrect values [1–3]. At the same time, it becomes necessary to assess whether these algorithms meet the requirements traditionally imposed on measurement procedures. If the numerical output of a model directly affects the final value of a measured quantity, the algorithm effectively becomes a component of the measurement process and shall comply with the stability and reliability requirements defined in the international standards ISO 5725 and ISO/IEC 17025 [4–5].

One of the key properties required for the acceptability of any method within a measurement context is its reproducibility [6]. In its conventional interpretation of ISO 5725-1, reproducibility involves changes in laboratories, operators, or measuring equipment. For stochastic machine-learning algorithms, however, the source of variability is not the change in external conditions but the internal randomness of computational procedures. This internal variability corresponds to the concept of *computational reproducibility* in Data Science. Therefore, in this paper, the term "reproducibility" is used in an extended sense – to denote the stability of numerical model outputs across repeated runs under unchanged input data. Such an extension is necessary when assessing the algorithms intended to function as components of measurement procedures.

Despite their potential advantages, many machine-learning models cannot ensure stable numerical results. Even with identical input data and fixed parameter settings, repeated application of an algorithm may yield different outputs due to random initializa-

tion, the sampling nature of ensemble methods, or implementation-specific numerical factors. In ordinary computational tasks, such differences are often insignificant. However, in a measurement context, they violate the fundamental requirement of reproducibility and introduce an additional component of uncertainty into the results [7–8].

2. Reproducibility of machine-learning model outputs and their specific aspects when in measurement tasks

The reproducibility problem is characteristic of most machine-learning methods because such algorithms rely on internal stochastic processes. Even when input data remain fully constant and model parameters unchanged, the results may differ from run to run. In measurement tasks, these variations are critical, since numerical outputs shall remain stable when the algorithm is repeatedly applied to the same dataset. The reasons for instability depend on the type of model: in neural networks it arises from random weight initialization and stochastic optimization procedures; in ensemble methods it results from random sampling of data subsets; and in algorithms with randomly formed decision rules it appears as changes in model structure during each retraining cycle. As a consequence, even minor variations in internal processes may lead to changes in numerical outputs, which is unacceptable in metrological applications.

A separate class of such methods consists of models based on decision trees. For them, stochasticity is an intrinsic property because the structure of each tree is determined by a random selection of data subsets and features. Even slight differences in the choice of split or threshold create a new tree structure, and therefore a new result. The output quantity ceases to be a deterministic reflection of the data and becomes dependent on internal random decisions of the algorithm, which complicates its use within the measurement process.

In this study, the Isolation Forest model is used – one of the most widely used stochastic ensembles of decision trees for detecting potential deviations in measurement datasets [9]. The model is configured by a set of parameters, including the number of trees, maximum depth, strategy for selecting subsets of observations, and the initial state of the random-number generator. It is the generator that determines the sequence of random decisions affecting the formation of splits, the selection of data subsets, and the isolation order of points. Because of the stochastic nature, the tree structure changes with every run, and as a result, the same point may receive different anomaly-score values even under fully constant input data [10].

Technically, the initial state of the random-number generator (random seed) may be fixed, ensuring the repetition of one specific sequence of stochastic operations. However, such fixation does not eliminate the underlying randomness of the algorithm; it only "locks in" one of the many possible ensemble configurations, artificially masking the inherent variability of the model. Since the purpose of this study is to assess the internal variability of the algorithm and its impact on the stability of numerical outputs, the random seed was not fixed in the experiments. This allows the true stochastic behaviour of the model to be represented, as well as ensures completeness of the reproducibility analysis.

To demonstrate the characteristic behaviour of the standard Isolation Forest implementation, a dataset of liquid mass-flow measurement results was used, reflecting a typical structure of metrological data. The dataset includes 11 values representing the differences between the indications of a measuring instrument and corresponding reference values, that is, the actual deviations recorded during calibration. Fig. 1 shows the raw values and corresponding anomaly-score estimates obtained during a single run of the model with fixed parameters. Values classified as normal are marked in green, while those identified as potential deviations are shown in red. The dashed line indicates the threshold used by the model to separate normal and anomalous observations.

To assess the reproducibility of numerical results, the model was run 100 times under fully constant

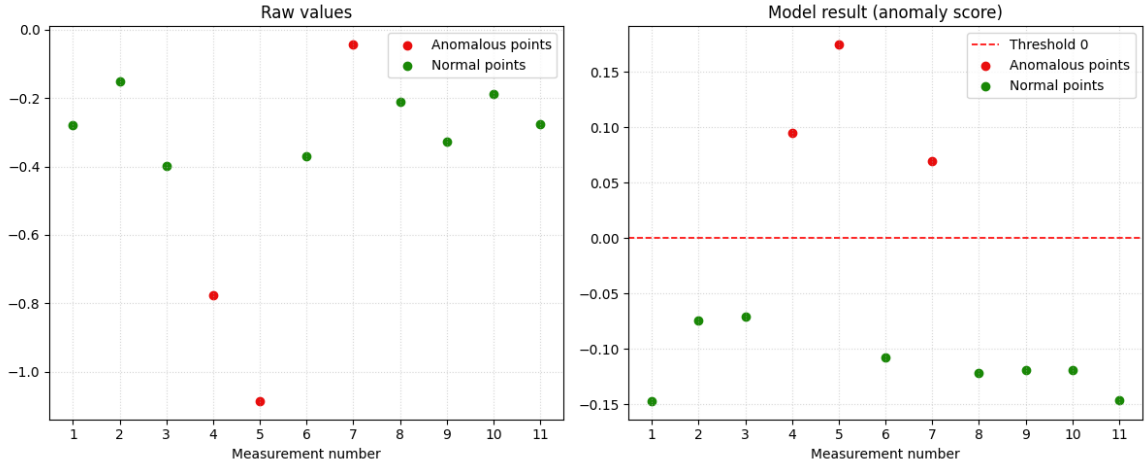

Fig. 1. Output of the Isolation Forest model

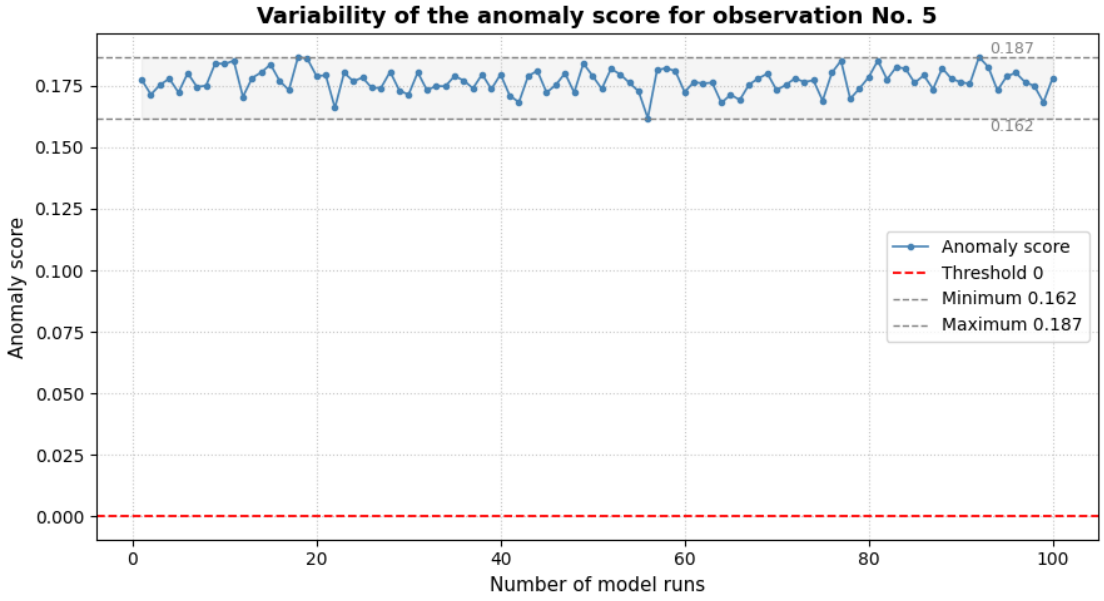

Fig. 2. Variability of the results of the standard Isolation Forest model

parameters and unchanged input data. Observation No. 5 – the point with the highest anomaly score – was selected for analysis. Fig. 2 illustrates the anomaly-score values obtained for this observation across all 100 runs. The findings demonstrate that, despite the data remaining constant, the score fluctuates within a specific range.

The standard model determines whether a value is normal or anomalous by comparing it to the fixed threshold of 0. When an observation lies very close to this threshold, even minor variations between runs can alter the classification outcome: the same point may be labelled as either normal or anomalous, depending on random internal factors. This sensitivity to stochastic variability constitutes the main reproducibility limitation of the standard approach.

3. Improved Isolation Forest model and its behaviour during repeated measurements

The Isolation Forest method was previously refined by the authors to address three fundamental limitations that restrict its applicability in a metrological context. The purpose of these modifications was not only to enhance reproducibility for the specific task of outlier detection but also to improve the stability of the output values in a broader class of measurement-related problems where the algorithm directly influences the numerical value of the measurand [11].

The first problem that needed to be addressed is the stochastic noise resulting from the random nature of tree construction. In the standard model, each tree is built from randomly selected subsets of data and random split thresholds, which introduces variability in the intermediate isolation paths. To reduce this effect, the model is run repeatedly under fixed parameters, and the intermediate characteristics of the isolation paths are averaged. This produces a value that approximates an "idealized" output in which random fluctuations are significantly reduced.

The second issue concerns the contamination parameter, which effectively scales the anomaly-score output in the standard implementation. As a result, different contamination settings can yield markedly different numerical outputs even for identical data, making scores incomparable across runs. In the improved version, the algorithm is run over a wide range of contamination values, and the resulting scores are normalized by averaging. This eliminates the dependence of the anomaly score on the chosen contamination level and provides a more stable, well-defined numerical interpretation.

The third limitation relates to the threshold used for separating normal and anomalous observations. In the standard approach, the threshold is imposed indirectly through an assumed proportion of expected anomalies, making the decision dependent on a user-defined parameter rather than on the data. In the refined model, the threshold is determined from the empirical distribution of sorted anomaly scores by identifying the transition region between the dense cluster of typical values and the sparse region of outliers. This yields a threshold that reflects the structure of the dataset rather than arbitrary settings.

To assess the effectiveness of the improved approach, the model was run 100 times under identical conditions, similar to the experiment with the standard algorithm. Fig. 3 presents the anomaly-score values obtained for observation No. 5 – the same point that exhibited substantial variability in the standard model.

It is evident that the range of output variation has narrowed substantially, and the numerical estimates have become noticeably more stable. Although full reproducibility is still not achieved and some residual variations remain, their influence on the final result

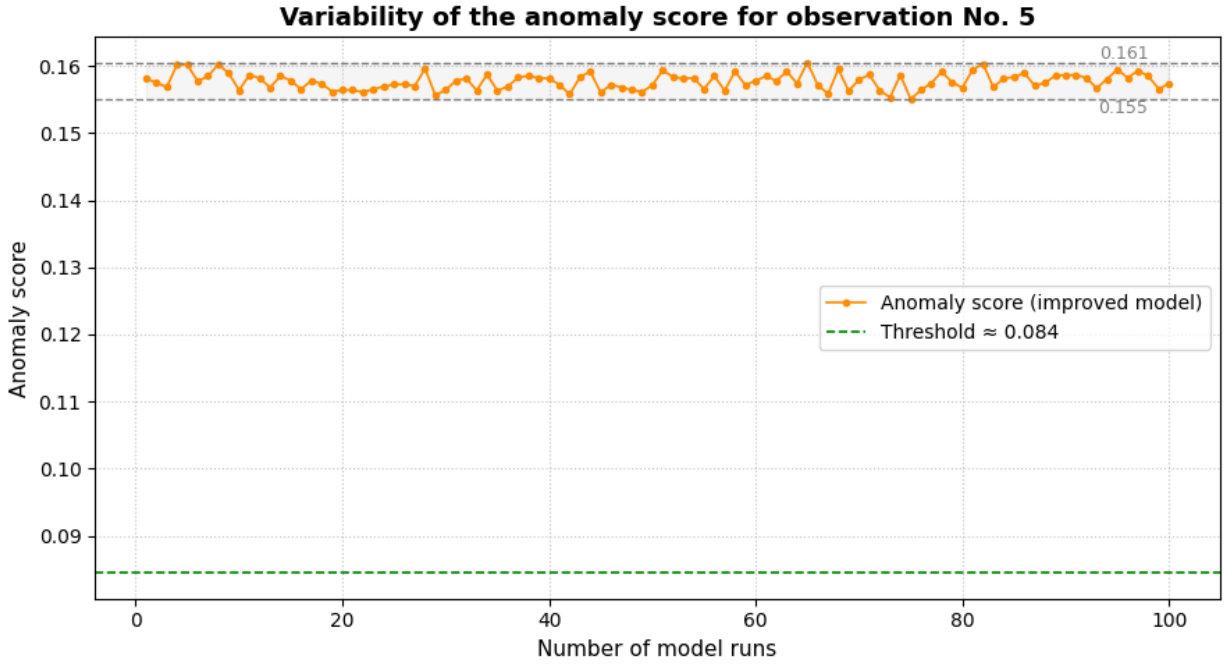

Fig. 3. Variability of the results of the improved Isolation Forest model

Table 1

Comparison of variability for the two model implementations

Model	Number of runs (n)	Maximum value (x_{max})	Minimum value (x_{min})	Range ($x_{max} - x_{min}$)	Type-A standard uncertainty (u_A)
Standard model	100	0.187	0.162	0.025	4.77×10^{-3}
Improved model	100	0.161	0.155	0.006	1.21×10^{-3}

is significantly lower compared to the original version of the algorithm.

4. Quantitative assessment of model reproducibility

For stochastic machine-learning models, complete elimination of internal variability is impossible: even after algorithmic improvements, a certain level of variations persists due to random processes involved in model formation. Therefore, in metrological applications, the key objective is not to remove randomness completely, but to quantify the residual variability that remains after refinement of the model. This quantification determines whether the algorithm can be integrated into measurement procedures without violating the requirements of relevant standards.

Since the Isolation Forest model produces a numerical output in the form of an anomaly score, its behaviour can be analysed using traditional metrological concepts. In ISO 5725, it is stated that for repeated measurements of the same quantity under identical conditions, the variability of results can be characterized through the type-A standard uncertainty derived from the statistical distribution of repeated observations. In this study, the repeated runs of a model – under fully constant input data and fixed parameters – play the role of repeated measurements.

This approach is not only consistent with ISO 5725 and widely accepted principles of variability assessment, but also harmonised with the justification procedures commonly applied in national metrology institutes when integrating a new algorithmic component into a measurement process. The evaluation of uncertainty becomes a necessary argument confirming the stability of the numerical output and enabling comparison between different implementations of the model.

For both models – the standard and the improved one – 100 runs were performed under identical conditions. Based on the obtained anomaly-score values, the type-A standard uncertainty was evaluated. The results are summarized in Table 1.

The results obtained demonstrate that the variability of a standard model is significantly higher. In particular, the standard uncertainty for the basic version was 0.0047, whereas for the improved model it was 0.0012, that is, four times lower. Although the algorithm residual randomness is not completely eliminated, this suggests a substantial improvement in reproducibility. The findings emphasize the importance of quantitative assessment of reproducibility when applying machine-learning algorithms in a measurement context. Even improved models require a metrological characterization, as it defines the limits

of their applicability, enables correct interpretation of numerical values, and ensures the possibility of further integration of such methods into real measurement procedures.

5. Conclusion

The study demonstrated that the stochastic nature of the Isolation Forest model significantly affects the reproducibility of its numerical results in measurement applications. Even with fully unchanged input data, the model produces different anomaly-score values due to random elements of the decision-tree construction. In metrological tasks, such behaviour leads to methodological uncertainty, since an algorithm that influences the measurement result shall ensure stability and independence of its outputs from internal fluctuations of its computational procedure. A quantitative assessment of reproducibility makes it possible to treat the algorithm internal variability as a component of the standard uncertainty accompanying the machine-learning output. Under this approach, the result of a model is interpreted as a computable quantity with its own uncertainty, which shall be incorporated into the overall measurement-uncertainty budget. This ensures consistency with metrological requirements and allows algorithmic methods to be integrated into measurement practice without violating the principles of accuracy and stability.

Оцінювання відтворюваності числових результатів алгоритмів машинного навчання у вимірювальних задачах

В.О. Ащепков[1], Д.Ю. Бяллович[1,2], В.В. Скляров[1,2]

[1]*Національний науковий центр "Інститут метрології", вул. Мироносицька, 42, 61002, Харків, Україна*
ashhepkovvalera@gmail.com; biallovych@gmail.com; vladimir.skliarov@gmail.com

[2]*Харківський національний університет радіоелектроніки, просп. Науки, 14, 61166, Харків, Україна*

Анотація

У роботі розглянуто проблему відтворюваності результатів стохастичних моделей машинного навчання у вимірювальних задачах, що набуває особливого значення під час інтеграції алгоритмів у вимірювальні процедури відповідно до підходів, визначених в ISO 5725 та ISO/IEC 17025. На прикладі алгоритму ізольованого лісу показано, що навіть за повністю сталих вхідних даних числовий результат моделі може змінюватися від запуску до запуску, утворюючи додаткову складову варіативності, яку необхідно враховувати під час аналізу результатів. Така мінливість обмежує можливість безпосереднього використання моделі як частини вимірювального процесу та потребує окремої метрологічної оцінки.

У дослідженні порівнюються дві реалізації алгоритму: стандартний варіант ізольованого лісу та удосконалена версія, запропонована в попередніх роботах авторів. Модифікована модель не усуває стохастичної природи алгоритму, однак зменшує її вплив завдяки усередненню проміжних характеристик ізоляційних шляхів, нормалізації результатів за різних значень параметра *contamination* та застосуванню більш обґрунтованого підходу до визначення порога на основі структури розподілу ступеня аномальності. Це забезпечує вищу стабільність вихідних значень порівняно зі стандартною моделлю.

Для обох реалізацій проведено серію повторних запусків за незмінних умов, після чого числові результати проаналізовано у метрологічних термінах через стандартну невизначеність за типом А відповідно до положень ISO 5725. Показано, що варіативність стандартної моделі істотно більша, тоді як удосконалена версія демонструє помітно вищу відтворюваність. Це свідчить про можливість більш коректного інтерпретування її вихідних значень у вимірювальному процесі.

Отримані результати підтверджують, що внутрішня варіативність алгоритмів машинного навчання може розглядатися як окрема метрологічна характеристика. Відповідна стандартна невизначеність може бути кількісно оцінена та включена до загального бюджету невизначеності вимірювання, забезпечуючи узгодженість алгоритмічних методів із принципами точності та повторюваності у вимірювальній практиці.

Ключові слова: метрологія; стандартна невизначеність; відтворюваність; машинне навчання; вимірювання; обробка вимірювальних даних.

References

1. Chandola V., Banerjee A., Kumar V. Anomaly detection: A survey. *ACM Computing Surveys*, 2009, vol. 41, issue 3, pp. 1–58. doi: https://doi.org/10.1145/1541880.1541882
2. Aggarwal C.C. Outlier Analysis. 2nd ed. Springer, 2017. 446 p. doi: https://doi.org/10.1007/978-3-319-47578-3
3. Aschepkov V. Methods of machine learning in modern metrology. *Measuring Equipment and Metrology*, 2024, vol. 85, no. 1, pp. 57–60. doi: https://doi.org/10.23939/istcmtm2024.01
4. DSTU EN ISO/IEC 17025:2019. General requirements for the competence of testing and calibration laboratories (EN ISO/IEC 17025:2017, IDT; ISO/IEC 17025:2017, IDT) (in Ukrainian).
5. DSTU GOST ISO 5725-1:2005. Accuracy (trueness and precision) of measurement methods and results. Part 1. General principles and definitions (GOST ISO 5725-1-2003, IDT) (in Ukrainian).
6. JCGM 100:2008. Evaluation of measurement data – Guide to the expression of uncertainty in measurement (GUM 1995 with minor corrections).
7. Nordling T., Melo Peralta T. A literature review of methods for assessment of reproducibility in science. *Research Square*, 2022. doi: https://doi.org/10.21203/rs.3.rs-2267847/v5
8. Semmelrock H., et al. Reproducibility in machine-learning-based research: Overview, barriers, and drivers. *AI Magazine*, 2025, vol. 46, issue 2. doi: https://doi.org/10.1002/aaai.70002
9. Liu F.T., Ting K.M., Zhou Z.-H. Isolation Forest. *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM)*, 2008, pp. 413–422. doi: https://doi.org/10.1109/ICDM.2008.17
10. Ashchepkov V.O. Vykorystannya modeli Isolation Forest dlya vyyavlennya anomaliy u danykh vymiryuvan [The use of the Isolation Forest model for anomaly detection in measurement data]. *Innovative technologies and scientific solutions for industries*, 2024, no. 1(27), pp. 236–245 (in Ukrainian). doi: https://doi.org/10.30837/ITSSI.2024.27.236
11. Ashchepkov V.O. Obrobka rezultativ vymiryuvan vytraty ridyny z vykorystannyam mashynnoho navchannya: dys. d-ra philosophii [Processing of liquid-flow measurement data using machine learning: PhD diss.]. Kharkiv, Kharkiv National University of Radioelectronics Publ., 2024. 198 p. (in Ukrainian). Available at: https://openarchive.nure.ua/handle/document/31412