UDC 004.8:006.91

Methodological challenges of outlier detection in metrology using machine learning models

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; vladimir.skliarov@gmail.com

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

Abstract

The paper addresses the challenges associated with applying machine learning models to detect outliers in metrological datasets. While such models ensure the possibility to identify complex deviations in the structure of a sample without relying on prior statistical assumptions, they do not provide normatively justified criteria for assessing the reliability of their decisions. Specifically, such models lack interpretable confidence indicators, metrological traceability, and formalised thresholds to determine whether an outlier is genuine. One proposed solution involves assessing the impact of eliminated anomalous values detected by the Isolation Forest model on the standard measurement uncertainty of Type A when the initial sample size is preserved through repeated measurements. This approach was validated using real-life measurements of liquid flow performed with Coriolis flowmeters of various diameters. The results empirically proved the effectiveness of the method in cases where the elimination of distortion-inducing values led to a significant reduction in measurement variability. However, several limitations were also identified, including the sensitivity of models to small sample sizes, the impracticality of performing repeated measurements in many real-life scenarios, and the lack of an objective threshold to determine the "significance" of uncertainty reduction. These findings highlight the need for further study of the formalization of confidence criteria in anomaly detection within the metrological domain, particularly in the context of compliance with international standards such as ISO/IEC 17025.

Despite these limitations, the application of machine learning models opens new opportunities for automating the analysis of metrological data and highlights the need to develop harmonized approaches for integrating such solutions into the regulatory framework.

Keywords: metrology; standard uncertainty; machine learning; outliers; anomalies; measurement data processing.

Received: 25.06.2025 Edited: 25.07.2025 Approved for publication: 30.07.2025

1. Introduction

"An outlier is a member of a dataset that is inconsistent with the other members of that set" [1]. In metrology, this means that an outlier is considered an individual value that contradicts the structure of the measured sample, even if it does not formally exceed the limits predefined by technical or physical conditions. The main difficulty lies in the fact that such values cannot always be formally identified without accounting for the conditions of reproducibility, repeatability, and the context of the measurement process.

To detect outliers, traditional approaches rely on statistical criteria and tests, including the **Grubbs test**, the **Shapiro–Wilk test**, **Student's t-test**, and others. These methods are based on assumptions regarding the data distribution, typically assuming normality. However, in practice, the presence of outliers can significantly distort the estimates of distribution parameters (e.g., mean or standard deviation), leading to a paradoxical situation in which the outlier influences the criterion that is supposed to detect it. For this reason, in metrological practice robust methods are preferable, such as the **interquartile range** (**IQR**) or the **median absolute deviation** (**MAD**), which are more resistant to the influence of outliers. Nevertheless, these approaches also require the definition of a threshold value beyond which an observation is considered an outlier. This threshold remains subjective and lacks a standardized regulatory basis [2].

These limitations have spurred interest in machine learning methods that do not require prior distribution assumptions and can capture complex internal relation-

ships among variables. The potential of such methods lies in the possibility to detect local structural anomalies in data while incorporating both statistical and metrological characteristics of the signal – without the need to formally establish a tolerance limit.

In machine learning terminology, the concept of an "outlier" corresponds to the notion of an "anomaly". Although these are not normative synonyms, in the context of automated data processing they both refer to observations that are inconsistent with the main structure of the dataset. An anomaly is defined as a sample that significantly deviates from others based on one or more criteria – density, distance, reconstruction error, and so on [3–4].

Anomaly detection methods can be roughly grouped into several categories:

• distance- and density-based methods (e.g., LOF, KNN, DBSCAN) [5];

• one-class classification models (e.g., One-Class SVM, Deep SVDD) [6–7];

• ensemble isolation methods (e.g., Isolation Forest, Random Cut Forest) [8];

• neural reconstruction models (e.g., Autoencoder, Variational AE) [9];

• statistical generative models (e.g., Gaussian Mixture Models, KDE) [10].

Unlike classical statistical tests that provide binary decisions (a value is either an outlier or not), machine learning models allow for:

• a continuous numerical estimate of the degree of anomaly (e.g., isolation depth in **Isolation Forest**, local density in **LOF**, reconstruction error in **Autoencoder**);

• ranking samples by the degree of their deviation from the normal structure;

• a spatial measurement of the distance from the boundary of the normal class (e.g., hyperplane in **One-Class SVM**);

• accounting for multidimensional correlations between parameters, which is fundamentally inaccessible to univariate approaches.

These properties provide a foundation for developing adaptive metrological criteria for outlier detection, which do not rely on rigid thresholds, but are instead flexible regarding the internal structure of the sample and the specific measurement conditions. For instance, rather than mechanically eliminating a value exceeding 3σ, one can apply a model that evaluates its distance from the centre of the normal distribution while accounting for measurement uncertainty, historical profile, and the stability of the measurement object.

2. Problems of applying machine learning methods in metrology

In conventional approaches to the processing of measurement data, the acceptability of any mathematical procedure is based on reproducible statistical principles: each stage of analysis – from hypothesis testing to uncertainty evaluation – has clear mathematical and regulatory justification. Machine learning methods, on the contrary, originate from a different paradigm – they are aimed at generalizing the behaviour of complex systems based on pattern recognition in data, without prior hypothesis formulations. This fundamental difference gives rise to several critical challenges in the metrological context.

First and foremost, the results of machine learning algorithms are typically the product of an internal loss function optimization process, which lacks direct physical or statistical interpretation in metrological terms. For example, the "anomaly index" assigned to a point by a model is not a function of standard error or any statistical significance criterion, but rather a heuristic value derived from the structure of the internal feature space representation. As a result, unlike classical statistical tests, there is no method that allows verification of a null hypothesis stating whether a given value is an outlier at a defined significance level.

A second major issue lies in the absence of a formalized system of confidence in the result: in conventional methods, one can calculate a confidence interval, evaluate standard or expanded measurement uncertainty, whereas machine learning lacks an equivalent tool. This prompts the question: with what probability can a given point be considered an outlier based on the decision of a model? How can such a deviation be justified in the context of metrological characteristics such as repeatability, stability of measurement, or known measurement error? Standard machine learning models fail to provide answers to these questions.

Furthermore, in cases of small datasets, which are typical for metrological experiments, machine learning methods face significant methodological limitations. Specifically, classification and anomaly detection algorithms require a sufficient amount of learnt data to build a model with acceptable generalization capability. With small sample sizes ($n \approx 5–10$), common in metrology, most machine learning models exhibit unstable behaviour: their internal structures, such as decision trees or latent spaces, adapt to random fluctuations that lack physical significance. In contrast, statistical methods may retain partial robustness when robust or specialized approaches adapted to small samples are applied. Therefore, even if an anomaly is correctly identified, it cannot be interpreted as a metrologically justified outlier.

Of particular importance is the issue of traceability. In classical data processing – especially with statistical methods – each step from data acquisition to decision-making can be documented, reproduced, and verified. In contrast, most machine learning models, especially deep neural networks, have opaque decision-making structures due to multilevel nonlinear transformations of input data. This makes it impossible to trace the reasoning behind each individual result or elimination. Such opacity contradicts key requirements

of metrological assurance, as defined, for example, in ISO/IEC 17025 [11], where traceability, interpretability, and justification of results are mandatory criteria for the credibility and suitability of a method.

3. Evaluation of the reliability of outlier detection results

From the standpoint of metrological analysis, it is fundamentally important not only to detect potential outliers in measurement data, but also to justify the correctness of the decision to eliminate them. In other words, it is necessary to assess whether the observed value is truly incompatible with the characteristic structure of the dataset, or whether it merely appears anomalous due to the peculiarities of the operation of a model. In traditional statistics, such justification is based on hypothetical distributions of measurement results and confidence probabilities, which allow for quantitative estimation of the risks of Type I and Type II errors. In contrast, most machine learning methods do not directly apply such criteria, as these models typically do not rely on parametric assumptions and lack established interpretations of statistical significance. As a result, even objectively detected deviations cannot be automatically interpreted as metrologically justified outliers without additional validation procedures.

To address this issue, it is advisable to focus on the variability of measurement results as an objective metrological criterion. One possible approach involves analysing the change in the standard uncertainty of Type A after the elimination of questionable outliers, with subsequently repeated measurements to restore the initial sample size. This approach is based on the following principles:

1. **Restoring the sample size (*n*):** Elimination of even objectively erroneous values changes the number of observations, which directly affects the uncertainty evaluation. To ensure metrologically valid comparison, the number of measurements shall be restored to the original level by performing repeated observations under identical conditions.

2. **Comparison of variability:** For each sample – before and after the elimination of anomalous values – the standard uncertainty (or another dispersion metric) is calculated. If, while maintaining the same number of measurements, a significant reduction in variability is observed, this indicates the incompatibility of the eliminated value with the main set.

3. **Uncertainty reduction criterion:** If the uncertainty after outlier elimination decreases only insignificantly, the value in question cannot be confidently regarded as a distortion, since it does not significantly affect the metrological homogeneity of the sample. Conversely, a substantial reduction – by a factor of several times – may serve as indirect evidence of its anomaly.

4. **Quality control of repeated measurements:** It is necessary to consider the possibility that new outliers may appear among the additional measurements. Therefore, it is advisable to reapply the chosen anomaly detection method to the updated dataset and verify its internal stability.

5. **Metrological interpretation of deviations:** In addition to quantitative analysis, it is important to account for the physical meaning and technical context of the result. A questionable value may arise not from random noise, but from a violation of measurement conditions, malfunctioning of the measuring instrument, external influences, or operational errors. Such interpretation increases confidence in the decision to eliminate the value.

The proposed approach was tested during the analysis of outlier detection results obtained using the Isolation Forest method, based on real-life measurements performed at the State Primary Measurement Standard of the unit of volumetric and mass flow rate of liquid, volume, and mass of liquid flowing through a pipeline (DETU 03-04-04) [12]. The objective of the study was to evaluate the effect of eliminating potential outliers on the standard measurement uncertainty of Type A.

As input data, the values of measurement errors obtained during the calibration of Coriolis flow meters with three different diameters were used. For each flow meter, measurements were performed at three fixed flow rate points; in each sample, 11 results were collected. Based on the initial data, the standard uncertainty of type A was calculated.

The Isolation Forest method was adapted to the specifics of the metrological task: it was applied separately to each sample to identify potentially anomalous values. Values classified by the algorithm as outliers were eliminated from the sample, after which additional measurements were performed in sufficient quantity to restore the initial number of observations *n*. Once the sample was updated, the method was applied again to check for remaining outliers. The cycle of "**detection – elimination – additional measurements**" was repeated until the model ceased to detect new anomalies.

At each stage, the current estimate of the standard uncertainty of type A was calculated. A comparison of the initial and final values made it possible to assess the degree to which the elimination of identified points affected the variability of the results. The summarized results of the study are presented in Table 1.

The values from the table are illustrated in Fig. 1.

Within the framework of the experimental study on the effectiveness of the Isolation Forest model for detecting outliers in metrological measurements, a comparison was made between the Type A standard uncertainty before and after the elimination of anomalous values identified by the algorithm. One of the key approaches to evaluating the reliability of detected outliers was the analysis of the dynamics of the dispersion of results, which quantitatively reflects the influence of individual points on the variability of the sample.

Table 1

Standard measurement uncertainty of Type A (%)

Designation	Flow rate point	Before outlier elimination	After outlier elimination	Number of outliers
Flow Meter №1 (DN50)	45 t/h	0.0330	0.0225	1
	25 t/h	0.0911	0.0308	3
	5 t/h	0.0202	0.0202	0
Flow Meter №2 (DN15)	5 t/h	0.0202	0.0177	1
	2.5 t/h	0.0312	0.0312	0
	1 t/h	0.0239	0.0202	1
Flow Meter №3 (DN6)	1 t/h	0.0325	0.0131	2
	0.5 t/h	0.0236	0.0212	1
	0.1 t/h	0.0607	0.0433	2

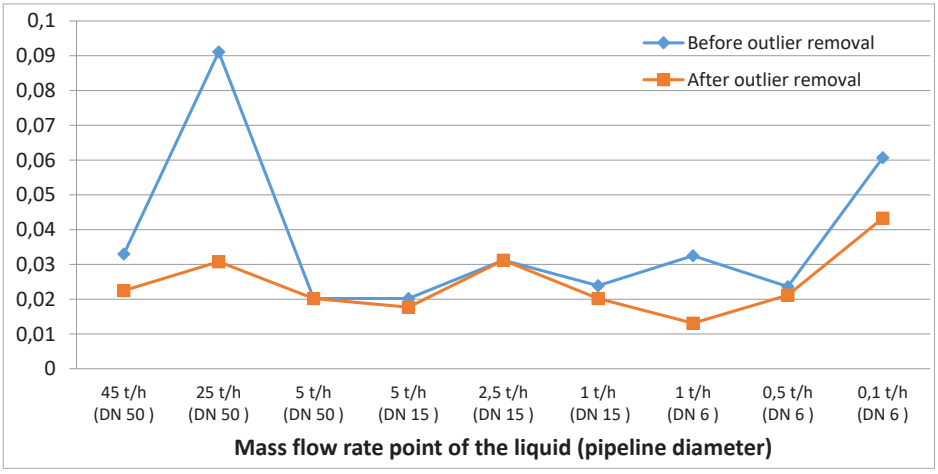

Fig. 1. Calculation of standard uncertainty of type A

Specifically, at the flow rate point of 25 t/h, the standard uncertainty decreased from 0.0911% to 0.0308%, that is, by more than three times. A similar effect was observed at the 1 t/h point, where the uncertainty decreased from 0.0325% to 0.0131%. Such a significant change, despite the relatively small number of eliminated measurements (three and two respectively), indirectly confirms that the values identified by the algorithm were indeed statistically and metrologically separated from the rest of the data – that is, inconsistent with the general structure of the sample.

This observation is consistent with an intuitive criterion: the elimination of a true outlier should result in a **noticeable decrease in internal dispersion**, provided the value in question had indeed distorted the estimate of variability.

However, the situation becomes significantly more complicated at those points where the number of detected outliers is small and the changes in standard uncertainty are marginal. For instance, at the 0.5 t/h flow rate, the elimination of a single measurement resulted in a decrease in standard uncertainty from 0.0236% to 0.0212%, i.e., which is about by 10%. Given the original number of observations (n = 11), such a reduction is relatively weak and does not provide a sufficient basis for confidently interpreting this value as an outlier. In such cases, the observed decrease in variability may be caused not due to the elimination of a true anomaly, but by random fluctuations or the sensitivity of the model to local deviations.

This leads to several key limitations of the proposed approach:

• First, there is **no normatively defined threshold** for how much the uncertainty must decrease for the eliminated value to be justifiably classified as an outlier. For example, a reduction of 20% or even 50% does not automatically carry metrological status if not supported by a formal procedure or confidence probability.

• Second, there remains **a risk of misinterpretation if the repeated measurements**, conducted after outlier exclusion, themselves contain implicit anomalies. In such a case, the observed decrease in variability may be falsely interpreted as evidence of validity, while it is merely a result of random sample stabilization.

• Third, **with small sample sizes**, even one or two values can disproportionately affect the estimate of uncertainty, increasing the sensitivity of the approach to random fluctuations that lack physical or metrological justification.

• Fourth, **repeated measurements are not always feasible under real-life conditions** – particularly in the case of unstable measurement objects, high experiment cost, or when verifying unique specimens using reference installations. Therefore, the proposed approach has limited applicability outside controlled laboratory environments.

Taken together, these limitations indicate that while the proposed method allows empirical verification of model output stability in individual cases, it does not eliminate the fundamental issue — the absence of an objective, formalized criterion of truth for an outlier. This, in turn, complicates the validation of machine learning models in metrology and highlights the need for further development of formalized approaches to assessing the reliability of results obtained from such methods.

4. Conclusions

The conducted study revealed a key methodological issue at the intersection of modern data processing algorithms and metrological practice: the absence of harmonized criteria for assessing the reliability of outlier detection decisions based on machine learning methods. Despite the high potential of such models – particularly their ability to identify structurally separated observations without prior assumptions about distribution – they do not provide a normatively justified rationale for eliminating specific values from measurement results.

The approach proposed in this paper, based on the analysis of changes in Type A standard uncertainty before and after the elimination of outliers under the condition of restoring the original sample size, demonstrated limited but potentially useful effectiveness as an empirical indicator of the separateness of anomalous values. At the same time, it revealed several fundamental limitations: the need for repeated measurements, the lack of guaranteed quality of new observations, the absence of a clearly defined threshold for the "significance" of uncertainty reduction, as well as sensitivity of the results to the number of eliminated points.

Thus, even when the correctness of actions is experimentally confirmed, there remains a need for the development of more formalized and normatively aligned mechanisms for reliability assessment. The proposed direction – comparing variability while preserving the number of measurements – should be considered a preliminary hypothesis outlining the construction of a comprehensive system for metrological validation of outlier detection algorithms.

Further study should aim to develop methodologically reasoned criteria for the reliability of analytical results that account for regulatory requirements, the nature of measurement errors, repeatability conditions, and the specifics of the measured quantities. This opens a new direction in applied metrology – the development of verifiable approaches to trust in artificial intelligence models, which is a critical prerequisite for their safe and regulated integration into the measurement domain.

Методологічні проблеми виявлення викидів у метрології з використанням моделей машинного навчання

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; vladimir.skliarov@gmail.com

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

Анотація

У статті розглядаються проблеми, пов'язані із застосуванням моделей машинного навчання для виявлення викидів у метрологічних вибірках. Попри здатність таких моделей ідентифікувати складні відхилення у структурі вибірки без необхідності попередніх статистичних припущень, вони не забезпечують нормативно обґрунтованих критеріїв оцінки достовірності прийнятих рішень. Зокрема, відсутні інтерпретовані показники довіри, метрологічна простежуваність і формалізовані порогові значення, які б дозволяли однозначно визначити, чи є те чи інше значення справжнім викидом. Одним із запропонованих рішень є оцінка впливу виключення аномальних значень, виявлених за допомогою моделі *Isolation Forest*, на стандартну невизначеність типу А за умови збереження початкового обсягу вибірки шляхом повторних вимірювань. Такий підхід було апробовано на основі реальних результатів вимірювань витрати рідини з використанням коріолісових витратомірів різних діаметрів. Отримані результати підтвердили ефективність підходу в тих випадках, коли видалення значень, що спотворювали результат, призводило до суттєвого зменшення варіативності вимірювань. Водночас було виявлено низку обмежень, зокрема чутливість моделі до малих вибірок, практичну неможливість проведення повторних вимірювань у багатьох реальних ситуаціях і відсутність об'єктивного критерію, який би визначав "суттєвість" зменшення невизначеності. Зазначені результа-

ти підкреслюють необхідність подальших досліджень щодо формалізації критеріїв довіри при виявленні аномалій у метрології, особливо в контексті відповідності міжнародним стандартам, таким як ISO/IEC 17025.

Попри зазначені обмеження, застосування моделей машинного навчання відкриває нові можливості для автоматизації аналізу метрологічних даних і вказує на потребу у створенні узгоджених підходів до інтеграції таких рішень у нормативне середовище.

Ключові слова: метрологія; стандартна невизначеність; машинне навчання; викиди; аномалії; обробка результатів вимірювань.

References

1. 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).
2. Ashchepkov V., Byallovych D., Skliarov V. Vplyv porogovykh znachen na standartnu nevyznachenist typu A pry vymiryuvannyakh masovoyi vytraty ridyny [The influence of threshold values on Type A standard uncertainty in mass flow rate measurements for liquids]. *Ukrainian Metrological Journal*, 2024, no. 3 (in Ukrainian). doi: https://doi.org/10.24027/2306-7039.3.2024.312469
3. 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
4. Aggarwal C.C. *Outlier Analysis*. 2nd ed. Springer, 2017. 446 p. doi: https://doi.org/10.1007/978-3-319-47578-3
5. Breunig M.M., Kriegel H.-P., Ng R.T., Sander J. LOF: Identifying density-based local outliers. *ACM SIGMOD Record*, 2000, vol. 29, issue 2, pp. 93–104. doi: https://doi.org/10.1145/335191.335388
6. Schölkopf B., Platt J., Shawe-Taylor J., Smola A.J., Williamson R.C. Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, 2001, no 13 (7), pp. 1443–1471. doi: https://doi.org/10.1162/089976601750264965
7. Ruff L., Vandermeulen R.A., Görnitz N., Deecke L. et al. Deep One-Class Classification. *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018, vol. 80, pp. 4393–4402.
8. Liu F.T., Ting K.M., Zhou Z.-H. Isolation Forest. *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2008, pp. 413–422. doi: https://doi.org/10.1109/ICDM.2008.17
9. Hinton G.E., Salakhutdinov R.R. Reducing the Dimensionality of Data with Neural Networks. *Science*, 2006, vol. 313, no. 5786, pp. 504–507. doi: https://doi.org/10.1126/science.1127647
10. Ahmed M., Mahmood A.N., Hu J. A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, 2016, vol. 60, pp. 19–31. doi: https://doi.org/10.1016/j.jnca.2015.11.016
11. 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).
12. 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. 98–113 (in Ukrainian). doi: https://doi.org/10.30837/ITSSI.2024.27.236