DRAFT International Standard

ISO/IEC DIS 4213

Information technology — Artificial Intelligence — Performance measurement for AI classification, regression, clustering and recommendation tasks

ICS:

ISO/IEC JTC **1**/SC **42**

Secretariat: **ANSI**

Voting begins on:
2026-07-28

Voting terminates on:
2026-10-20

THIS DOCUMENT IS A DRAFT CIRCULATED FOR COMMENTS AND APPROVAL. IT IS THEREFORE SUBJECT TO CHANGE AND MAY NOT BE REFERRED TO AS AN INTERNATIONAL STANDARD UNTIL PUBLISHED AS SUCH.

IN ADDITION TO THEIR EVALUATION AS BEING ACCEPTABLE FOR INDUSTRIAL, TECHNOLOGICAL, COMMERCIAL AND USER PURPOSES, DRAFT INTERNATIONAL STANDARDS MAY ON OCCASION HAVE TO BE CONSIDERED IN THE LIGHT OF THEIR POTENTIAL TO BECOME STANDARDS TO WHICH REFERENCE MAY BE MADE IN NATIONAL REGULATIONS.

RECIPIENTS OF THIS DRAFT ARE INVITED TO SUBMIT, WITH THEIR COMMENTS, NOTIFICATION OF ANY RELEVANT PATENT RIGHTS OF WHICH THEY ARE AWARE AND TO PROVIDE SUPPORTING DOCUMENTATION.

This document has not been edited by the ISO Central Secretariat.

Reference number
ISO/IEC DIS 4213:2026(en)

© ISO/IEC 2026



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2026

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword ..... **vi**

Introduction ..... **vii**

1 Scope **1**

2 Normative references **1**

3 Terms and definitions **1**
3.1 General 1
3.2 Classification metrics and related terms 2

4 Abbreviations **4**

5 General principles **5**
5.1 Generalized process for functional correctness assessment 5
5.2 Purpose of functional correctness assessment 6
5.3 Control criteria for functional correctness assessment 6
5.3.1 General 6
5.3.2 Preprocessing 6
5.3.3 Limiting information leakage 7
5.3.4 Limiting channel effects 7
5.3.5 Evaluation environment 8
5.3.6 Appropriate baselines 8
5.3.7 Functional correctness context 8

6 Assessing functional correctness for AI classification tasks **8**
6.1 General 8
6.2 Base elements for metric computation 9
6.2.1 General 9
6.2.2 Confusion matrix 9
6.2.3 Accuracy 9
6.2.4 Precision, recall and specificity 9
6.2.5 F1 score 10
6.2.6 F_β 10
6.2.7 Kullback-Leibler divergence 11
6.3 Binary classification 11
6.3.1 General 11
6.3.2 Confusion matrix for binary classification 11
6.3.3 Accuracy for binary classification 11
6.3.4 Precision, recall, specificity, F1 score and F_β for binary classification 12
6.3.5 Balanced accuracy for binary classification 12
6.3.6 Matthews correlation coefficient 12
6.3.7 Kullback-Leibler divergence for binary classification 13
6.3.8 Receiver operating characteristic curve and area under the receiver operating characteristic curve 13
6.3.9 Precision recall curve and area under the precision recall curve 13
6.3.10 Cumulative response curve 13
6.3.11 Lift Curve 14
6.4 Ternary classification with undecided 14
6.5 Multi-class classification 15
6.5.1 General 15
6.5.2 Accuracy for multi-class classification 15
6.5.3 Macro-average, weighted-average and micro-average 15
6.5.4 Balanced accuracy for multi-class classification 16
6.5.5 Distribution difference or distance metrics 17
6.6 Multi-label classification 17
6.6.1 General 17
6.6.2 Hamming loss and Total Hamming Loss 18
6.6.3 Exact match ratio 18

© ISO/IEC 2026 – All rights reserved

6.6.4 Jaccard index ... 19
6.6.5 Distribution difference or distance metrics ... 19

7 Assessing functional correctness for AI regression tasks ... 20
7.1 Overview ... 20
7.2 Residual Sum of Squares (RSS) ... 21
7.3 Mean Squared Error (MSE) or Mean Squared Deviation (MSD) ... 21
7.4 Root Mean Square Error (RMSE) or Root Mean Square Deviation (RMSD) ... 21
7.5 Mean Absolute Error (MAE) ... 21
7.6 Mean Absolute Percentage Error (MAPE) ... 22
7.7 R Squared (R^2) / Coefficient of Determination ... 22
7.8 Adjusted R Squared (R^2) ... 23
7.9 Max Error or Maximum Error Score ... 23
7.10 Actual / Predicted correlation ... 23

8 Assessing functional correctness for AI clustering tasks ... 24
8.1 General ... 24
8.2 Partitional clustering ... 24
8.2.1 Internal metrics ... 24
8.2.2 External metrics based on matching sets ... 28
8.2.3 External metrics based on peer-to-peer correlation ... 30
8.2.4 External metrics based on information theory ... 32
8.3 Hierarchical clustering ... 35
8.3.1 Internal metrics ... 35
8.3.2 External metrics ... 37

9 Assessing functional correctness for AI recommendation tasks ... 37
9.1 General ... 37
9.2 Predictive quality metrics ... 38
9.2.1 Precision @ K ... 38
9.2.2 Recall @ K ... 38
9.2.3 F1 @ K ... 39
9.3 Ranking quality metrics ... 39
9.3.1 Reciprocal rank ... 39
9.3.2 Mean reciprocal rank (MRR) ... 39
9.3.3 Average precision @ K ... 39
9.3.4 Average recall @ K ... 40
9.3.5 Mean average precision (mAP) ... 40
9.3.6 Hit rate @ K ... 41
9.3.7 Average reciprocal hit-rank @ K (ARHR@K) ... 41
9.3.8 Discounted Cumulative Gain (DCG) ... 42
9.3.9 Normalized Discounted Cumulative Gain (NDCG) ... 42
9.3.10 RBP (Rank-Biased Precision) ... 43
9.3.11 ERR (Expected Reciprocal Rank) ... 43
9.3.12 R-Score ... 43
9.3.13 Pearson Coefficient ... 44
9.3.14 Spearman Coefficient ... 44
9.3.15 Kendall Tau ... 45
9.4 Behavioural metrics ... 45
9.4.1 Diversity ... 45
9.4.2 Novelty ... 45
9.4.3 Serendipity ... 46
9.4.4 Popularity bias ... 46
9.4.5 Catalog coverage ... 47
9.4.6 Distributional coverage ... 47

10 Statistical tests of significance for classification ... 48
10.1 General ... 48
10.2 Paired Student's t-test ... 48
10.3 Analysis of variance ... 48
10.4 Kruskal-Wallis test ... 49

© ISO/IEC 2026 – All rights reserved

10.5 Chi-squared test 49
10.6 Wilcoxon signed-ranks test 49
10.7 Fisher's exact test 49
10.8 Central limit theorem 49
10.9 McNemar test 49
10.10 Accommodating multiple comparisons 50
10.10.1 General 50
10.10.2 Holm's method 50
10.10.3 False discovery rate 50
Annex A (informative) **Multi-class classification functional correctness illustration** **51**
Annex B (informative) **Illustration of ROC curve derived from classification results** **53**
Annex C (informative) **Summary information from benchmark tests** **57**
Annex D (Informative) **Chance-corrected cause-specific mortality fraction** **59**
Bibliography **60**

© ISO/IEC 2026 – All rights reserved

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 42, *Artificial intelligence.*

Any feedback or questions on this document should be directed to the user’s national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

© ISO/IEC 2026 – All rights reserved

Introduction

As academic, commercial and governmental researchers continue to improve AI models, a common framework of measures has evolved to quantify this improvement. Some types of advances in artificial intelligence are reported in terms of improved functional correctness (also referred to as performance by AI practitioners) relative to the state of the art or a reasonable baseline.

Functional correctness can be measured for AI classification, regression, clustering and recommendation tasks. For these tasks, the choice of an appropriate metric depends on the use case and domain constraints. Further, for machine learning (ML) systems, the chosen metric can differ from the metric used during training.

For example, the functional correctness of an AI model for classification can be represented through the following examples:

— A new AI model achieves 97,8 % classification accuracy on a dataset, where the state-of-the-art AI model achieves 96,2 % accuracy.

— A new AI model achieves classification accuracy equivalent to the state of the art, but it operates with much less training data than state-of-the-art approaches.

— A new AI model generates inferences 100x faster than state-of-the-art AI models while maintaining equivalent accuracy.

This document describes approaches and methods for AI classification, regression, clustering and recommendation tasks, mapping tasks to relevant metrics. Various AI stakeholder roles as defined in ISO/IEC 22989:2022 can apply the approaches and methods to measure functional correctness for AI classification, regression, clustering and recommendation tasks. For example, AI developers can use the approaches and methods when evaluating AI models.

The approaches and methods specified are not necessarily comprehensive. Additional approaches and methods beyond those presented in this document can be applied.

The term performance is used more widely than functional correctness when expressing the degree to which an AI system outputs correct results. However, performance can also be associated with AI system dimensions such as processing speed, resource utilisation, energy efficiency, latency, and throughput. Functional correctness more clearly and precisely expresses the concept of correct results or outputs than the term performance. Functional correctness is therefore used in this document in any context where the meaning is clear.

NOTE ISO/IEC 25059:2023 Annex C discusses the meaning of performance in AI versus software and systems engineering.

General principles can be applied when assessing functional correctness to ensure that results are fair and representative, such as limiting information leakage that potentially leads to misleading results. Clause 5 addresses this topic.

Clause 6 discusses assessment of functional correctness for AI classification tasks.

Clause 7 discusses assessment of functional correctness for AI regression tasks.

Clause 8 discusses assessment of functional correctness for AI clustering tasks.

Clause 9 discusses assessment of functional correctness for AI recommendation tasks. Many approaches discussed in this clause are also applicable to ranking tasks.

Statistical tests of significance are undertaken to establish the conditions under which functional correctness differs meaningfully for classification, regression, clustering and recommendation tasks. Clause 10 addresses these topics.

Annex A illustrates calculation of multi-class classification functional correctness, using examples of positive and negative classifications.

© ISO/IEC 2026 – All rights reserved

Annex B illustrates a receiver operating characteristic (ROC) curve derived from example data in Annex A.

Annex C summarizes results from AI benchmark tests.

Annex D discusses a chance-corrected cause-specific mortality fraction, an AI classification use case.

Apart from these, this document does not address any issues related to benchmarking, applications or use cases.

© ISO/IEC 2026 – All rights reserved

Information technology — Artificial Intelligence — Performance measurement for AI classification, regression, clustering and recommendation tasks

1 Scope

This document specifies methodologies for measuring the performance of AI models for classification, regression, clustering and recommendation tasks.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 22989:2022, *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*

ISO/IEC 23053:2022, *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*

ISO/IEC/TS 25058:2024, *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guidance for quality evaluation of artificial intelligence (AI) systems*

3 Terms and definitions

For the purposes of this document, the terms and definitions in ISO/IEC 22989:2022, ISO/IEC 23053:2022 and the following apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1 General

3.1.1
classification
method of structuring a defined data representation into classes and subclasses in accordance with their characteristics

[SOURCE: ISO 7200:2004, 3.1]

3.1.2
classifier
trained model and its associated mechanism used to perform *classification* (3.1.1)

3.1.3
functional correctness
degree to which a product or system provides the correct results with the needed degree of precision

Note 1 to entry: AI systems, and particularly those using machine learning models, do not usually provide functional correctness in all observed circumstances.

© ISO/IEC 2026 – All rights reserved

[SOURCE: ISO/IEC 25059:2023, 3.2.3]

3.1.4
variance
measure of the spread of a statistical distribution

[SOURCE: ISO/IEC 21836:2020]

3.1.5
information leakage
condition in which data from outside of an AI model's intended input is inadvertently used during AI model development, leading to over-optimistic performance estimates

3.2 Classification metrics and related terms

3.2.1
evaluation
<artificial intelligence> process of running an AI system or a model on a test dataset or validation dataset and assessing its outputs to quantify its functional correctness

Note 1 to entry: The result of the evaluation can be numerical or categorical.

Note 2 to entry: Outputs are typically assessed by comparing them with labels accompanying the dataset.

Note 3 to entry: The assessment can be fully automated or involve human subjects.

3.2.2
false negative
F_N
sample wrongly classified as negative

Note 1 to entry: Other terms in use for false negative include "miss" and "type II error".

3.2.3
false positive
F_P
sample wrongly classified as positive

Note 1 to entry: Other terms in use for false positive include "false alarm" and "type I error".

3.2.4
true positive
T_P
sample correctly classified as positive

3.2.5
true negative
T_N
sample correctly classified as negative

3.2.6
accuracy
number of correctly classified samples divided by all classified samples, calculated as $accuracy = (T_P + T_N) / (T_P + F_P + T_N + F_N)$

3.2.7
confusion matrix
matrix used to record the number of correct and incorrect *classifications* (3.1.1) of samples

3.2.8

© ISO/IEC 2026 – All rights reserved

F1 score
harmonic mean of *precision* (3.2.9) and *recall* (3.2.10), calculated as $F1 = 2T_P / \left(2T_P + F_P + F_N\right)$

3.2.9
precision
positive predictive value
number of samples correctly classified as positive divided by all samples classified as positive, calculated as $precision = T_P / \left(T_P + F_N\right)$

3.2.10
recall
sensitivity
hit rate
true positive rate
number of samples correctly classified as positive divided by all positive samples, calculated as $recall = T_P / \left(T_P + F_N\right)$

3.2.11
specificity
selectivity
true negative rate
number of samples correctly classified as negative divided by all negative samples, calculated as $specificity = T_N / \left(T_N + F_P\right)$

3.2.12
false negative rate
miss rate
number of samples incorrectly classified as negative divided by all positive samples, calculated as $false\ negative\ rate = F_N / \left(F_N + T_P\right)$

3.2.13
false positive rate
fall-out
number of samples incorrectly classified as positive divided by all negative samples, calculated as $false\ positive\ rate = F_P / \left(F_P + T_N\right)$

3.2.14
cumulative response curve
gain chart
graphical method of displaying *true positive rates* (3.2.10) and percentage of positive prediction in the total data across multiple thresholds

3.2.15
lift curve
graphical method of displaying on the y-axis the ratio of *true positive rate* (3.2.10) between the model and a random classifier, and on the x-axis the percentage of positive predictions in the total data across multiple thresholds

3.2.16
precision recall curve
PRC
graphical method for displaying *recall* (3.2.10) and *precision* (3.2.9) across multiple thresholds

Note 1 to entry: A PRC is more suitable than a ROC curve for showing functional correctness with imbalanced data.

3.2.17
receiver operating characteristic curve
ROC curve
graphical method for displaying *true positive rate* (3.2.10) and *false positive rate* (3.2.13) across multiple thresholds

© ISO/IEC 2026 – All rights reserved

3.2.18
cross-validation
method to estimate the functional correctness of an AI model using a single dataset

Note 1 to entry: Cross-validation is typically used for validating design choices before training the final model.

3.2.19
majority class
class with the most samples in a dataset

4 Abbreviations

AI	artificial intelligence
ANOVA	analysis of variance
ARHR	average reciprocal hit rate
ARP	average relative position
AUPRC	area under the precision recall curve
AUROC	area under the receiver operating characteristic curve
CLT	central limit theorem
CPCC	cophenetic correlation coefficient
CPU	central processing unit
CRC	cumulative response curve
CSMF	cause-specific mortality fraction
DBI	Davies-Bouldin Index
DCG	discounted cumulative gain
ERR	expected reciprocal rank
FNR	false negative rate
FPR	false positive rate
FC	fully connected
FDR	false discovery rate
IDCG	ideal discounted cumulative gain
IoU	intersection over union
GPU	graphics processing unit
MAE	mean absolute error
mAP	mean average precision
MAPE	mean absolute percentage error
MCC	Matthews correlation coefficient

© ISO/IEC 2026 – All rights reserved