Supplemental Table S1. Commonly Used Machine Learning Performance Metrics for Binary or Multiclass Classifications [8-11]

Metrics	Concepts and equations
Accuracy	• The proportion of correctly classified subjects among all subjects; best intuitive measure for a balanced dataset (approximately the same positive and negative values) • (TP+TN)/(TP+FP+FN+TN)
Precision (=positive predictive value)	• The proportion of correctly classified subjects (e.g., subjects with a disease in reality) among positively classified subjects by the machine learning model • TP/(TP+FP)
Recall (=sensitivity, true positive rate)	• The proportion of correctly classified subjects (classified as having a disease) among subjects with the disease in reality • TP/(TP+FN)
F1-score (=dice similarity coefficient; DSC)	• The harmonic mean of precision and recall. In the F1-score, precision and recall have the same weight (achieving the highest score in models with equally balanced precision and recall); useful for imbalanced datasets (with a relatively high difference between positive and negative values). • 2×(recall×precision)/(recall+precision)
Fβ-score	• F-score with different weights for recall and precision (β times the importance of recall relative to precision; F_2 [weights recall than precision; emphasis on impact of FNs] or $F_{0.5}$ [weights precision than recall; reducing the influence of FNs] are commonly used) • $(1+\beta^2)\times TP/[(1+\beta^2)\times TP+\beta^2\times FN+FP]$
Specificity	• The proportion of correctly classified subjects (not having a disease) among subjects without the disease in reality • TN/(TN+FP)
False positive rate (FPR)	• The proportion of incorrectly classified subjects (classified as having a disease) among subjects without the disease in reality • FP/(FP+TN)=1-specificity
False discovery rate (FDR)	• The rate of type 1 error in statistical testing (it is important to control the FDR efficiently with the preservation of statistical power in multiple testing problems with a large number of variables, such as genomics studies) • FP/(FP+TP)=1-PPV
AUROC (=c-statistic)	• The area under the receiver operating characteristic curve (plotting TPR against FPR) • A model with a curve close to the upper left corner has better classification performance, with more TPs and fewer FNs (higher AUROC, close to 1=better classifier).
AUPRC	• The area under the precision-recall curve (plotting precision against recall) • A model with a curve close to the upper right corner has better classification performance, with more TPs, fewer FPs, and fewer FNs (higher AUPRC, close to 1=better classifier). • May have an advantage over AUROC when comparing the performance of models in an imbalanced dataset [10].
Matthews correlation coefficient (MCC)	• Correlation coefficient between the true class and predicted class (MCC=1 for a perfect classifier [no FP and FN]; MCC=-1 for a classifier that always misclassifies [no TP and TN]) • (TP×TN–FP×FN)/sqrt[(TP+FP)×(TP+FN)×(TN+FP)×(TN+FN)] • Provides a relatively robust performance measurement for an imbalanced dataset when accuracy, the F1-score, precision, and recall show asymmetric performance for positive and negative classes [11].

TP, true positive; TN, true negative; FP, false positive; FN, false negative; PPV, positive predictive value; TPR, true positive rate; AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve.