

Exact and near-duplicate symptom profiles inflate the reported performance of machine learning classifiers for diabetes: a leakage-aware re-evaluation of the UCI Early-Stage Diabetes dataset

William Marin-Rodriguez

wmarin@unjfsc.edu.pe

Universidad Nacional José Faustino Sánchez Carrión

Daniel Andrade-Girón

Universidad Nacional José Faustino Sánchez Carrión

Viviana Vellon-Flores

Universidad Nacional José Faustino Sánchez Carrión

Enrique Diaz-Vega

Universidad Nacional José Faustino Sánchez Carrión

Timoteo Solano-Armas

Universidad Nacional José Faustino Sánchez Carrión

Research Article

Keywords: Data leakage, Cross-validation, Duplicate records, Diabetes, Ensemble learning, Model evaluation, Calibration, Reproducibility

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11235253/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Symptom-based machine learning classifiers for diabetes are commonly developed on the UCI Early-Stage Diabetes Risk Prediction dataset, for which accuracies of 97–99% are frequently reported. The dataset contains many repeated predictor profiles, and their effect on performance estimates has not been quantified. We measured how much row-level cross-validation overstates performance and whether it alters model rankings.

Methods

Ten ensemble classifiers, logistic regression and a 1-nearest-neighbour reference were evaluated on 520 records comprising 251 unique predictor profiles. Each model was assessed with stratified 5-fold cross-validation repeated 10 times under row-level random partitioning and under profile-grouped partitioning, in which identical profiles never occur in both training and test folds. A sensitivity analysis grouped records by symptom pattern, ignoring age. Hyperparameters were fixed a priori. We report classification, discrimination and calibration measures with profile-level cluster bootstrap confidence intervals and compared models with the Nadeau–Bengio corrected resampled t-test with Holm adjustment.

Results

Under random partitioning, 68.2% of test records had an identical profile in the training folds. Row-level cross-validation reproduced the accuracies reported in the literature (97.2–98.2% for tree-based ensembles) and overstated the Matthews correlation coefficient (MCC) by 0.105 to 0.157 and the area under the ROC curve by 0.023 to 0.037 relative to profile-grouped cross-validation. The stacking ensemble tied for first place under random partitioning but ranked eighth under grouped partitioning. With grouped partitioning, the highest MCC was 0.857 (95% confidence interval 0.786–0.915; extra trees), which differed significantly only from AdaBoost after adjustment. Grouping by symptom pattern lowered performance further for every model; the 1-nearest-neighbour reference, which matched the best ensembles under profile grouping, showed the largest additional loss.

Conclusions

Exact and near-duplicate predictor profiles are a large, correctable source of optimism on this benchmark. Studies using it should partition by profile at every level, report calibration and uncertainty, and avoid declaring a single best model without formal comparison. The estimates are internal and do not establish diagnostic or screening utility.

Background

Diabetes is a leading cause of morbidity and mortality worldwide, and its burden continues to grow across age groups [1, 2]. Because laboratory confirmation is not always accessible, many studies have proposed machine learning (ML) classifiers that use self-reported symptoms and basic demographic data to identify individuals who may need confirmatory testing. Ensemble methods such as random forests, gradient boosting and stacking dominate this literature [3].

The Early-Stage Diabetes Risk Prediction dataset in the UCI Machine Learning Repository is the most widely used benchmark for this task [4]. It contains 520 questionnaire records collected at Sylhet Diabetes Hospital, Bangladesh, with 16 predictors and a binary diabetes-status label. Reported accuracies on this dataset range from about 97% to 99% [5–7], and a recent study reported 98.1% accuracy with 10-fold cross-validation and 100% without it [8]. For a classifier that relies only on symptoms and age, such values are close to the ceiling and justify a closer look at how they were estimated.

Data leakage, in which information unavailable at prediction time reaches the model during training or evaluation, is a pervasive cause of over-optimistic results in ML-based science [9, 10]. In clinical neuroimaging, for example, repeated subjects across training and test data substantially inflated prediction performance [11]. In clinical ML more generally, record-wise cross-validation, which lets records from the same subject enter both training and test folds, can markedly overestimate accuracy relative to subject-wise cross-validation [12], and the same principle underlies grouped and blocked cross-validation for structured data [13]. A simple form arises when identical observations appear in both training and test data. In the UCI Early-Stage dataset, 269 of the 520 records repeat the complete predictor vector of another record. Under row-level random partitioning, copies of the same profile are routinely split between training and test folds, so part of the measured accuracy reflects recognition of training records rather than generalization to new profiles. To our knowledge, the size of this effect on this benchmark has not been quantified.

Two further practices limit the interpretation of published comparisons. Differences between models are often tested with procedures that treat cross-validation folds as independent, although overlapping training sets make such tests too liberal [14, 15]. Rank-based procedures such as the Friedman test with Nemenyi post hoc comparisons were designed for comparisons across independent datasets [16] and are sometimes applied instead to the folds of a single dataset. In addition, calibration is rarely assessed, even though poorly calibrated probabilities can mislead clinical decisions [17]. Current reporting guidance for prediction model studies calls for both uncertainty quantification and calibration assessment [18].

This study re-evaluated twelve classifiers on the UCI Early-Stage dataset with four aims: (i) to quantify how much row-level cross-validation overstates performance relative to profile-grouped cross-validation; (ii) to determine whether the partitioning scheme changes model rankings; (iii) to provide leakage-aware estimates with profile-level confidence intervals, calibration measures and statistically appropriate

model comparisons; and (iv) to assess the additional influence of near-duplicate profiles that differ only in age.

Methods

Data source and outcome

We used the Early-Stage Diabetes Risk Prediction dataset [4], which contains 520 records, 16 predictors and a binary outcome. The predictors are age (years) and 15 binary variables: sex, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia and obesity. Yes/No responses were coded 1/0 and sex was coded male = 1. The outcome was coded 1 for the repository label Positive ($n = 320$; 61.5%) and 0 for Negative ($n = 200$; 38.5%). The repository documentation does not report diagnostic criteria, diabetes type or the timing of symptom recording relative to diagnosis, so the outcome was treated as a dataset-provided status label rather than as confirmed or incident diabetes. There were no missing values.

Classifiers

Twelve classifiers were evaluated: ten ensembles (AdaBoost, gradient boosting, random forest, extra trees, bagging with an extra-tree base learner, XGBoost, LightGBM, CatBoost, a stacking ensemble combining random forest, extra trees and gradient boosting with a random forest meta-learner, and a soft-voting ensemble of the same three base learners), logistic regression as a conventional reference, and a 1-nearest-neighbour classifier as a reference that is maximally sensitive to exact and near-duplicate records. Each model was embedded in a pipeline of standardization, random oversampling of the minority class and the classifier, and all steps were fitted on training folds only. Hyperparameters were fixed before the analysis (Additional file 1: Table S1) rather than tuned. This isolates the effect of the partitioning scheme and avoids a second leakage route through hyperparameter search under row-level splits [19].

Validation schemes

Each pipeline was evaluated with stratified 5-fold cross-validation repeated 10 times (seeds 0–9) under two primary schemes. In the random scheme (StratifiedKFold), records were assigned to folds regardless of profile, so copies of a profile could appear in both training and test folds. In the profile-grouped scheme (StratifiedGroupKFold), all records sharing a profile were kept in the same fold. Both schemes used identical pipelines, so differences between them reflect partitioning alone. In each repetition every record received one out-of-fold (OOF) predicted probability. For each test fold we recorded the proportion of test records whose profile also occurred in the training folds (leakage exposure).

In a sensitivity analysis, the grouped scheme was repeated with symptom patterns as groups, so that records differing only in age were also kept in the same fold. This scheme is conservative: it removes near-duplicates but may also separate distinct individuals whose answers happen to coincide. We therefore treated profile grouping as the primary leakage-aware analysis and symptom grouping as a lower bound.

Performance measures

Classification at a 0.5 threshold was summarized with accuracy, balanced accuracy, weighted F1-score and the Matthews correlation coefficient (MCC). Discrimination was summarized with the area under the receiver operating characteristic curve (ROC-AUC) and the area under the precision–recall curve. Probabilistic accuracy and calibration were summarized with the Brier score, the expected calibration error (ECE; 10 equal-width bins), the calibration slope and calibration-in-the-large [17]. Calibration-in-the-large was estimated as the intercept of a logistic recalibration model with the logit of the predicted probability as an offset. Measures were computed on the OOF predictions of each repetition and averaged over the 10 repetitions. MCC was prespecified as the primary measure because it uses all four cells of the confusion matrix and is robust to the class imbalance in the data [20].

Statistical analysis

Optimism. For each model, measure and repetition, optimism was defined as the value under the random scheme minus the value under the profile-grouped scheme. We report the mean over repetitions together with the 2.5th and 97.5th percentiles of the 10 per-repetition differences.

Uncertainty. Confidence intervals for profile-grouped estimates were obtained with a profile-level cluster bootstrap. The 251 profiles were resampled with replacement (500 resamples), all records of a sampled profile were kept together, and in each resample the measure was computed for every repetition and averaged, matching the point estimate. Percentile 95% intervals are reported. Intervals are not reported for ECE because binned ECE is biased upward under resampling. These intervals describe variability in the evaluated records given the OOF predictions, not variability from refitting the models.

Model comparison. Fold-level MCC and ROC-AUC (50 folds per model) were compared between the model with the highest mean and each other ensemble and logistic regression. We used the corrected resampled t-test of Nadeau and Bengio, which multiplies the variance by $(1/k + n_{\text{test}}/n_{\text{train}})$ to account for overlapping training sets [15]. P-values were adjusted for each measure with the Holm procedure [21], across 10 comparisons under profile grouping and 11 under symptom grouping, where the 1-nearest-neighbour reference was also included. Agreement between the rankings obtained under the random and profile-grouped schemes was summarized with Kendall's τ , and ranking stability with the number of repetitions in which each model attained the highest MCC.

Software and reproducibility

Analyses were performed in Python 3.12.3 with scikit-learn 1.8.0 [22], imbalanced-learn 0.14.2, XGBoost 3.4.1, LightGBM 4.7.0, CatBoost 1.2.10, NumPy 2.4.4, pandas 3.0.2 and SciPy 1.17.1. Random seeds were fixed for all resampling and model fitting. The study re-evaluates existing modelling approaches and does not develop a model for clinical use; where applicable, reporting follows the TRIPOD + AI recommendations [18], and Additional file 2 summarizes how each section of the REFORMS checklist [23] is addressed.

Use of large language models

Large language models were used as assistive tools under the authors' supervision. This study builds on the authors' earlier version of this work, in which the authors audited the duplicate predictor profiles, designed the profile-grouped evaluation and selected the ensemble classifiers. Claude Opus 5.5 (Anthropic) was used between September and October 2026 to propose a re-analysis design quantifying leakage-induced optimism and the sensitivity analysis; to write and execute the analysis code; to perform the statistical analyses and produce the figures with plotting code; to draft the manuscript text; to identify and verify references; and to format the manuscript to the journal guidelines. ChatGPT-5.6 (OpenAI) was used to support the study design and language editing of an earlier version of this work. No generative image tools were used. The authors evaluated all proposals and outputs and made the final decisions on the study design, the interpretation of the results and the conclusions. One author (WMR) independently re-ran the complete analysis in a separate computing environment, using a copy of the dataset downloaded directly from the UCI repository; this re-run identified an error in the analysis code, which was corrected, and all reported values correspond to it. The code is publicly available for verification (see Availability of data and materials). All authors reviewed and approved the final text and take full responsibility for the content of this publication.

Results

Leakage exposure

Under the random scheme, an average of 68.2% of test records had an identical profile in the corresponding training folds. Under the profile-grouped scheme this proportion was zero by construction, but 31.0% of test records still had a training record with an identical symptom pattern that differed only in age (Table 1).

Table 1
Dataset structure and exposure of test records to identical training records

Characteristic	Value
Records	520
Predictors	16 (age and 15 binary variables)
Positive label, n (%)	320 (61.5)
Unique predictor profiles	251
Records duplicating another record's profile	269
Records in profiles occurring more than once, n (%)	376 (72.3)
Largest profile group	16 records
Profiles with discordant labels	0
Unique symptom patterns (age excluded)	214
Mean training / test fold size	416 / 104
Test records with identical profile in training, random CV	68.2%
Test records with identical profile in training, profile-grouped CV	0%
Test records with identical symptom pattern in training, profile-grouped CV	31.0%

Optimism attributable to duplicate profiles

Every model performed better under the random scheme on every measure and in every repetition (Table 2; Fig. 1). With random partitioning, the nine tree-based ensembles other than AdaBoost reached accuracies of 97.2–98.2% (weighted F1-score 0.972–0.982), consistent with the 96.8–99.0% reported in previous studies of this dataset [5–7]. With profile-grouped partitioning, their accuracies fell to 90.2–93.2%.

MCC was overstated by 0.105 to 0.157 across the ten ensembles, by 0.084 for logistic regression and by 0.064 for the 1-nearest-neighbour reference. ROC-AUC was overstated by 0.023 to 0.037 for the ensembles and by 0.016 for logistic regression. The Brier score was 0.041 to 0.057 higher under grouped partitioning for all ensembles except AdaBoost (0.009), whose probability estimates were poor under both schemes. The stacking ensemble showed the largest optimism in MCC (0.157).

Table 2
Performance under row-level random and profile-grouped cross-validation

Model	MCC R	MCC G	Δ MCC	AUC R	AUC G	Δ AUC	Δ Brier
Logistic regression	0.845	0.761	+ 0.084 (+ 0.040 to + 0.147)	0.971	0.955	+ 0.016 (+ 0.005 to + 0.028)	-0.022
1-nearest neighbour (reference)	0.935	0.870	+ 0.064 (+ 0.036 to + 0.110)	0.972	0.942	+ 0.030 (+ 0.016 to + 0.053)	-0.032
AdaBoost	0.831	0.708	+ 0.123 (+ 0.093 to + 0.178)	0.975	0.942	+ 0.033 (+ 0.026 to + 0.049)	-0.009
Gradient Boosting	0.940	0.794	+ 0.146 (+ 0.103 to + 0.193)	0.989	0.952	+ 0.037 (+ 0.021 to + 0.053)	-0.053
Random Forest	0.958	0.851	+ 0.106 (+ 0.079 to + 0.144)	0.999	0.973	+ 0.025 (+ 0.020 to + 0.030)	-0.044
Extra Trees	0.961	0.857	+ 0.105 (+ 0.081 to + 0.152)	0.999	0.976	+ 0.023 (+ 0.020 to + 0.029)	-0.044
Bagging Extra Tree	0.956	0.851	+ 0.105 (+ 0.068 to + 0.150)	0.998	0.974	+ 0.025 (+ 0.021 to + 0.029)	-0.044
XGBoost	0.942	0.810	+ 0.131 (+ 0.072 to + 0.201)	0.992	0.963	+ 0.029 (+ 0.018 to + 0.046)	-0.051
LightGBM	0.942	0.809	+ 0.133 (+ 0.085 to + 0.213)	0.996	0.962	+ 0.033 (+ 0.025 to + 0.042)	-0.055
CatBoost	0.944	0.836	+ 0.108 (+ 0.083 to + 0.135)	0.996	0.972	+ 0.024 (+ 0.017 to + 0.031)	-0.041
Stacking (RF meta)	0.961	0.804	+ 0.157 (+ 0.097 to + 0.220)	0.995	0.958	+ 0.037 (+ 0.028 to + 0.062)	-0.057
Soft voting	0.953	0.835	+ 0.119 (+ 0.082 to + 0.180)	0.998	0.973	+ 0.025 (+ 0.022 to + 0.030)	-0.044

R, row-level random cross-validation; G, profile-grouped cross-validation. Values are means over 10 repetitions of 5-fold cross-validation; optimism ($\Delta = R - G$) is shown with the 2.5th–97.5th percentiles of

the per-repetition differences. ΔBrier is negative because the Brier score is lower (better) under random partitioning.

Changes in model rankings

Rankings under the two schemes were positively but imperfectly associated (Kendall's $\tau = 0.71$ for MCC and 0.82 for ROC-AUC; Fig. 2). The stacking ensemble tied for the highest MCC under random partitioning (0.961) but ranked eighth under grouped partitioning (0.804). CatBoost moved from sixth to fourth and XGBoost from eighth to sixth. Under grouped partitioning the model with the highest MCC varied between repetitions: extra trees ranked first in 5 of 10 repetitions, bagging with extra trees in 2, and CatBoost, random forest and stacking in 1 each.

Leakage-aware performance and model comparison

Table 3 and Fig. 3 give the profile-grouped estimates. MCC ranged from 0.708 (AdaBoost) to 0.857 (extra trees; 95% CI 0.786–0.915), and ROC-AUC from 0.942 to 0.976 (extra trees; 95% CI 0.955–0.992). The confidence intervals of the six best ensembles overlapped widely. In corrected resampled t-tests against extra trees (Table 4), only AdaBoost had significantly lower MCC after Holm adjustment (mean fold-level difference 0.147; adjusted $p = 0.038$). Logistic regression had lower MCC (difference 0.096; unadjusted $p = 0.032$; adjusted $p = 0.291$) and lower ROC-AUC (difference 0.019; $p = 0.123$), but neither difference remained significant after adjustment. No ROC-AUC difference was significant after adjustment. The 1-nearest-neighbour reference reached an MCC similar to that of extra trees (0.870 vs. 0.857) but a lower ROC-AUC (0.942 vs. 0.976), and its predicted probabilities, being 0 or 1, provide no risk ranking or calibration.

Table 3
Leakage-aware performance under profile-grouped cross-validation

Model	MCC (95% CI)	F1w (95% CI)	ROC-AUC (95% CI)	Brier (95% CI)	ECE	Slope	CITL
Logistic regression	0.761 (0.680–0.835)	0.886 (0.850–0.922)	0.955 (0.930–0.975)	0.083 (0.060–0.108)	0.059	0.70	0.26
1-nearest neighbour (reference)	0.870 (0.808–0.913)	0.936 (0.908–0.958)	0.942 (0.914–0.963)	0.064 (0.043–0.093)	0.064	n.a.	n.a.
AdaBoost	0.708 (0.609–0.792)	0.862 (0.821–0.900)	0.942 (0.911–0.966)	0.167 (0.158–0.177)	0.250	6.62	0.26
Gradient Boosting	0.794 (0.708–0.862)	0.902 (0.866–0.933)	0.952 (0.924–0.975)	0.081 (0.054–0.109)	0.060	0.64	0.17
Random Forest	0.851 (0.780–0.909)	0.930 (0.897–0.958)	0.973 (0.951–0.991)	0.062 (0.043–0.083)	0.062	1.40	0.02
Extra Trees	0.857 (0.786–0.915)	0.932 (0.899–0.960)	0.976 (0.955–0.992)	0.059 (0.038–0.082)	0.056	1.20	-0.06
Bagging Extra Tree	0.851 (0.780–0.909)	0.929 (0.896–0.957)	0.974 (0.952–0.991)	0.062 (0.042–0.084)	0.068	1.35	0.02
XGBoost	0.810 (0.725–0.873)	0.909 (0.876–0.939)	0.963 (0.938–0.981)	0.075 (0.050–0.103)	0.066	0.56	0.44
LightGBM	0.809 (0.729–0.874)	0.909 (0.873–0.939)	0.962 (0.936–0.981)	0.079 (0.052–0.109)	0.078	0.40	0.75
CatBoost	0.836 (0.759–0.901)	0.922 (0.886–0.952)	0.972 (0.951–0.987)	0.064 (0.040–0.090)	0.043	0.90	0.23
Stacking (RF meta)	0.804 (0.719–0.868)	0.906 (0.869–0.938)	0.958 (0.927–0.979)	0.072 (0.048–0.100)	0.070	0.45	-1.18
Soft voting	0.835 (0.761–0.893)	0.922 (0.888–0.949)	0.973 (0.951–0.989)	0.062 (0.042–0.086)	0.042	1.17	0.02

Point estimates are means over 10 repetitions; 95% confidence intervals from a profile-level cluster bootstrap (500 resamples). F1w, weighted F1-score; ECE, expected calibration error; Slope, calibration slope; CITL, calibration-in-the-large; n.a., not applicable (predicted probabilities are 0 or 1).

Table 4
Comparison of each classifier with extra trees under profile-grouped cross-validation

Comparator	Δ MCC	p	p-Holm	Δ AUC	p	p-Holm
Logistic regression	+ 0.096	0.032	0.291	+ 0.019	0.123	0.986
AdaBoost	+ 0.147	0.004	0.038	+ 0.030	0.026	0.255
Gradient Boosting	+ 0.062	0.119	0.955	+ 0.022	0.074	0.669
Random Forest	+ 0.005	0.761	1.000	+ 0.002	0.638	1.000
Bagging Extra Tree	+ 0.006	0.730	1.000	+ 0.001	0.778	1.000
XGBoost	+ 0.046	0.191	1.000	+ 0.010	0.299	1.000
LightGBM	+ 0.047	0.256	1.000	+ 0.010	0.320	1.000
CatBoost	+ 0.019	0.484	1.000	+ 0.002	0.782	1.000
Stacking (RF meta)	+ 0.055	0.310	1.000	+ 0.020	0.187	1.000
Soft voting	+ 0.021	0.363	1.000	+ 0.002	0.767	1.000

Nadeau–Bengio corrected resampled t-test on 50 folds; extra trees had the highest mean MCC and ROC-AUC. Differences are extra trees minus the comparator; p-Holm, Holm-adjusted p-value across 10 comparisons.

Calibration

Calibration differed markedly between models (Table 3; Fig. 4). Extra trees had the lowest Brier score (0.059). CatBoost had the calibration slope closest to 1 (0.90) and, together with soft voting, the lowest ECE (0.043 and 0.042). Gradient boosting, XGBoost, LightGBM and stacking produced overconfident probabilities (slopes 0.40–0.64), whereas random forest, extra trees, bagging and soft voting produced probabilities that were too moderate (slopes 1.17–1.40). AdaBoost probabilities were strongly compressed toward 0.5 (slope 6.62; ECE 0.250).

Sensitivity analysis: near-duplicate profiles

Grouping by symptom pattern lowered MCC and ROC-AUC further for every model (Table 5; Fig. 1). Among the ensembles, MCC fell by 0.002 (AdaBoost) to 0.065 (stacking) and ROC-AUC by 0.002 to 0.020. The 1-nearest-neighbour reference showed the largest loss of all models (MCC 0.870 to 0.799; ROC-AUC 0.942 to 0.902), and its ROC-AUC became the lowest of the twelve classifiers. Extra trees retained the highest MCC (0.826) and ROC-AUC (0.972). Under symptom grouping, no difference between extra trees and any other model was significant after Holm adjustment, either for MCC or for ROC-AUC.

Table 5
Sensitivity analysis: profile-grouped versus symptom-pattern-grouped cross-validation

Model	MCC profile	MCC symptom	Δ MCC	AUC profile	AUC symptom	Δ AUC	Brier symptom
Logistic regression	0.761	0.734	-0.027	0.955	0.951	-0.005	0.090
1-nearest neighbour (reference)	0.870	0.799	-0.071	0.942	0.902	-0.040	0.096
AdaBoost	0.708	0.706	-0.002	0.942	0.939	-0.003	0.168
Gradient Boosting	0.794	0.771	-0.024	0.952	0.945	-0.007	0.090
Random Forest	0.851	0.804	-0.047	0.973	0.969	-0.004	0.073
Extra Trees	0.857	0.826	-0.031	0.976	0.972	-0.003	0.070
Bagging Extra Tree	0.851	0.810	-0.041	0.974	0.971	-0.002	0.070
XGBoost	0.810	0.776	-0.034	0.963	0.957	-0.006	0.084
LightGBM	0.809	0.767	-0.042	0.962	0.951	-0.012	0.097
CatBoost	0.836	0.794	-0.042	0.972	0.965	-0.006	0.072
Stacking (RF meta)	0.804	0.738	-0.065	0.958	0.938	-0.020	0.100
Soft voting	0.835	0.794	-0.041	0.973	0.966	-0.007	0.073

Values are means over 10 repetitions. Δ = symptom-grouped minus profile-grouped.

Discussion

This re-evaluation shows that duplicate predictor profiles are a major source of optimism on the most widely used symptom-based diabetes benchmark. Row-level cross-validation exposed about two thirds of test records to identical training records and reproduced the near-perfect accuracies reported in the literature. When profiles were kept together, MCC fell by 0.10–0.16 for every ensemble, the apparent leader changed, and most pairwise differences between ensembles could no longer be distinguished. Removing near-duplicates that differed only in age lowered performance further for every model. This pattern agrees with evidence from clinical neuroimaging, where repeated subjects across training and test data inflated prediction performance [11].

Interpretation of previously reported results

These findings do not show that any particular published study used row-level partitioning, because partitioning procedures are often not reported in enough detail to judge. They do show that the accuracies commonly reported for this dataset are what row-level evaluation produces, and that profile-disjoint evaluation yields values of about 90–93%. Claimed improvements of one or two percentage points for a new ensemble fall within the optimism attributable to duplicates, and within the variation between repetitions of a single model under grouped cross-validation (MCC 0.816–0.883 for extra trees). A recent large-scale study across 2,047 tabular benchmark datasets concluded that random cross-validation is usually adequate for datasets without temporal structure [24]. The present results show that this does not hold for a benchmark in which about half of the records duplicate another record, which supports auditing widely used datasets individually rather than relying on average behaviour across benchmarks.

Why optimism differed between models

Optimism was not uniform. The stacking ensemble, which uses internal cross-validation to generate the inputs of its meta-learner, lost the most under both the profile-grouped and the symptom-grouped schemes. A plausible explanation is that duplicates within the training data also leak between these internal folds, so that the meta-learner is trained on over-optimistic base-learner predictions. The same mechanism affects hyperparameter tuning and early stopping whenever internal splits are made at the record level [19]. Leakage control therefore needs to apply at every level at which the data are split, not only in the outer evaluation loop.

Near-duplicates and the nearest-neighbour reference

Under profile grouping, a 1-nearest-neighbour rule matched the best ensembles on MCC. The sensitivity analysis explains this result. Almost one third of test records still had a training record with the same symptoms and a different age, and a nearest-neighbour rule benefits most from such near-copies. Once they were kept in the same fold, its MCC fell by 0.071 and its ROC-AUC became the lowest of all models, whereas the leading ensembles lost much less. The apparent parity therefore reflected residual near-duplicate leakage rather than genuine competitiveness. Exact-match grouping is necessary but may not be sufficient when records differ only in a single variable, and analysts should examine near-duplicates as well as exact copies.

Model selection and calibration

Under both leakage-aware schemes, the choice of a single best model was not supported by corrected tests, except against AdaBoost under profile grouping, and the leading model varied between repetitions. With about 100 records per test fold, threshold-based measures such as MCC have limited power to separate models. ROC-AUC and calibration were more informative: extra trees and random forest discriminated best, whereas CatBoost produced the most reliable probabilities. Because the proportion of positive records (61.5%) is far higher than diabetes prevalence in screening populations, absolute probabilities would in any case require recalibration before use in another setting [17].

Recommendations

For studies using this or similar datasets, we recommend (1) auditing and reporting exact duplicates and near-duplicate records; (2) defining the unit of analysis and partitioning by it in every split, including the internal cross-validation used for stacking, tuning or early stopping; (3) reporting uncertainty at the level of the partitioning unit; (4) comparing models with tests that account for overlapping training sets and refraining from declaring a best model when differences are not supported; and (5) reporting calibration alongside discrimination, as recommended by TRIPOD + AI [18]. The REFORMS checklist offers complementary reporting items for ML-based science [23].

Limitations

This study has several limitations. First, it used a single dataset, and the size of the optimism depends on the proportion of duplicates, so it will differ in other data. Second, the origin of the duplicate profiles is unknown. If some represent different individuals with identical answers, similar profiles may also occur among future patients, and performance for new patients would lie between the grouped and random estimates. Whether grouping is required depends on whether records within a group are genuinely dependent, a question also debated for subject-wise cross-validation [25]. The symptom-grouped estimates are conservative for the same reason. Third, hyperparameters were fixed. Tuning could change absolute values but would not remove the exposure of test records to identical training records. Fourth, random oversampling and a 0.5 threshold were used throughout, and other choices could shift threshold-based measures. Fifth, the outcome definition, diabetes type and timing of symptom recording are undocumented, and the data come from a single hospital with a case mix unlike that of screening populations. The results are internal estimates for a dataset-provided label and do not establish diagnostic or screening utility, which would require external and prospective validation.

Conclusions

Row-level cross-validation on the UCI Early-Stage Diabetes dataset substantially overstates classifier performance because about two thirds of test records have identical copies in the training data. Profile-grouped evaluation lowers MCC by 0.10–0.16, changes model rankings and shows that most ensembles cannot be distinguished statistically, and near-duplicates that differ only in age add a further, model-dependent optimism. Studies using this benchmark should partition by predictor profile at every level, examine near-duplicates, report calibration and uncertainty, and interpret results as internal evidence only.

Abbreviations

CI
confidence interval
CV
cross-validation

ECE
expected calibration error
F1w
weighted F1-score
LLM
large language model
MCC
Matthews correlation coefficient
ML
machine learning
OOF
out-of-fold
ROC-AUC
area under the receiver operating characteristic curve
TRIPOD + AI
Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis, updated for artificial intelligence

Declarations

Ethics approval and consent to participate

Not applicable. The study used a publicly available, anonymized secondary dataset; no human participants were contacted and no identifiable information was accessed.

Clinical trial number

Clinical trial number: not applicable

Consent for publication

Not applicable.

Availability of data and materials

The dataset supporting the conclusions of this article is available in the UCI Machine Learning Repository, <https://doi.org/10.24432/C5VG8H> [4]. Analysis code: Project name: Leakage-aware re-evaluation of the UCI Early-Stage Diabetes dataset. Project home page and archived version: <https://doi.org/10.5281/zenodo.23027767> (version 1.2). Operating system(s): Platform independent.

Programming language: Python (3.12 or later). Other requirements: scikit-learn 1.8.0, imbalanced-learn 0.14.2, XGBoost 3.4.1, LightGBM 4.7.0, CatBoost 1.2.10. License: MIT. Any restrictions to use by non-academics: None.

Competing interests

The authors declare that they have no competing interests.

Funding

No funding was received for this study.

Authors' contributions

WMR and DAG conceived the study, designed the methodology, prepared the software and carried out the investigation. WMR independently re-ran the complete analysis and verified all reported values. WMR, DAG and EDV analysed the data, and EDV validated the results. WMR and DAG drafted and revised the manuscript. DAG curated the data and prepared the figures. EDV supervised the study. VVF and TSA administered the project. DAG, EDV, VVF and TSA provided resources. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. *Nat Rev Endocrinol*. 2018;14:88–98. <https://doi.org/10.1038/nrendo.2017.151>.
2. Xu ST, Sun M, Xiang Y. Global, regional, and national trends in type 2 diabetes mellitus burden among adolescents and young adults aged 10–24 years from 1990 to 2021: a trend analysis from the Global Burden of Disease Study 2021. *World J Pediatr*. 2025;21:73–89. <https://doi.org/10.1007/s12519-024-00861-8>.
3. Mahajan P, Uddin S, Hajati F, Moni MA. Ensemble learning for disease prediction: a review. *Healthc (Basel)*. 2023;11:1808. <https://doi.org/10.3390/healthcare11121808>.
4. Islam MMF, Ferdousi R, Rahman S, Bushra HY. Early stage diabetes risk prediction [dataset]. UCI Machine Learning Repository; 2020. <https://doi.org/10.24432/C5VG8H>. Accessed 28 Sep 2026.

5. Moniruzzaman M, Zaman AGM, Tasnia R, Biswas S, Khanam M. Early-stage diabetes prediction using data mining algorithms. In: Proceedings of the 2nd International Conference on Computing Advancements (ICCA 2022). New York: ACM; 2022. <https://doi.org/10.1145/3542954.3542990>
6. Laila UE, Mahboob K, Khan AW, Khan F, Taekeun W. An ensemble approach to predict early-stage diabetes risk using machine learning: an empirical study. *Sensors (Basel)*. 2022;22(14):5247. <https://doi.org/10.3390/s22145247>. PMID: 35890927.
7. Pradhan G, Thapa G, Pradhan R, Khandelwal B, Panigrahi R, Bhoi AK, et al. Optimized forest framework with a binary multineighborhood artificial bee colony for enhanced diabetes mellitus detection. *Int J Comput Intell Syst*. 2024;17:194. <https://doi.org/10.1007/s44196-024-00598-2>.
8. Sarker P, Nahid AA, Choi K, Samad MA. Feature reduction using swarm optimization and random forest classifiers for early diabetes risk prediction. *Sci Rep*. 2026;16:14355. <https://doi.org/10.1038/s41598-026-35984-7>.
9. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*. 2023;4:100804. <https://doi.org/10.1016/j.patter.2023.100804>.
10. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: formulation, detection, and avoidance. *ACM Trans Knowl Discov Data*. 2012;6:1–21. <https://doi.org/10.1145/2382577.2382579>.
11. Rosenblatt M, Tejavibulya L, Jiang R, Noble S, Scheinost D. Data leakage inflates prediction performance in connectome-based machine learning models. *Nat Commun*. 2024;15:1829. <https://doi.org/10.1038/s41467-024-46150-w>.
12. Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP. The need to approximate the use-case in clinical machine learning. *GigaScience*. 2017;6(5):1–9. <https://doi.org/10.1093/gigascience/gix019>.
13. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*. 2017;40:913–29. <https://doi.org/10.1111/ecog.02881>.
14. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput*. 1998;10:1895–923. <https://doi.org/10.1162/089976698300017197>.
15. Nadeau C, Bengio Y. Inference for the generalization error. *Mach Learn*. 2003;52:239–81. <https://doi.org/10.1023/A:1024068626366>.
16. Demšar J. Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res*. 2006;7:1–30. <https://jmlr.org/papers/v7/demsar06a.html>.
17. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17:230. <https://doi.org/10.1186/s12916-019-1466-7>.
18. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. <https://doi.org/10.1136/bmj-2023-078378>.
19. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 2006;7:91. <https://doi.org/10.1186/1471-2105-7-91>.

20. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21:6.
<https://doi.org/10.1186/s12864-019-6413-7>.
21. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6:65–70.
<https://doi.org/10.2307/4615733>.
22. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
<https://jmlr.org/papers/v12/pedregosa11a.html>.
23. Kapoor S, Cantrell EM, Peng K, Pham TH, Bail CA, Gundersen OE, et al. REFORMS: consensus-based recommendations for machine-learning-based science. Sci Adv. 2024;10:eadk3452.
<https://doi.org/10.1126/sciadv.adk3452>.
24. Roth S. Which leakage types matter? A quantitative landscape across 2,047 benchmark datasets [Preprint]. arXiv. 2026. <https://doi.org/10.48550/arXiv.2604.04199>
25. Little MA, Varoquaux G, Saeb S, Lonini L, Jayaraman A, Mohr DC et al. Using and understanding cross-validation strategies. Perspectives on Saeb GigaScience. 2017;6(5):1–6.
<https://doi.org/10.1093/gigascience/gix020>

Figures

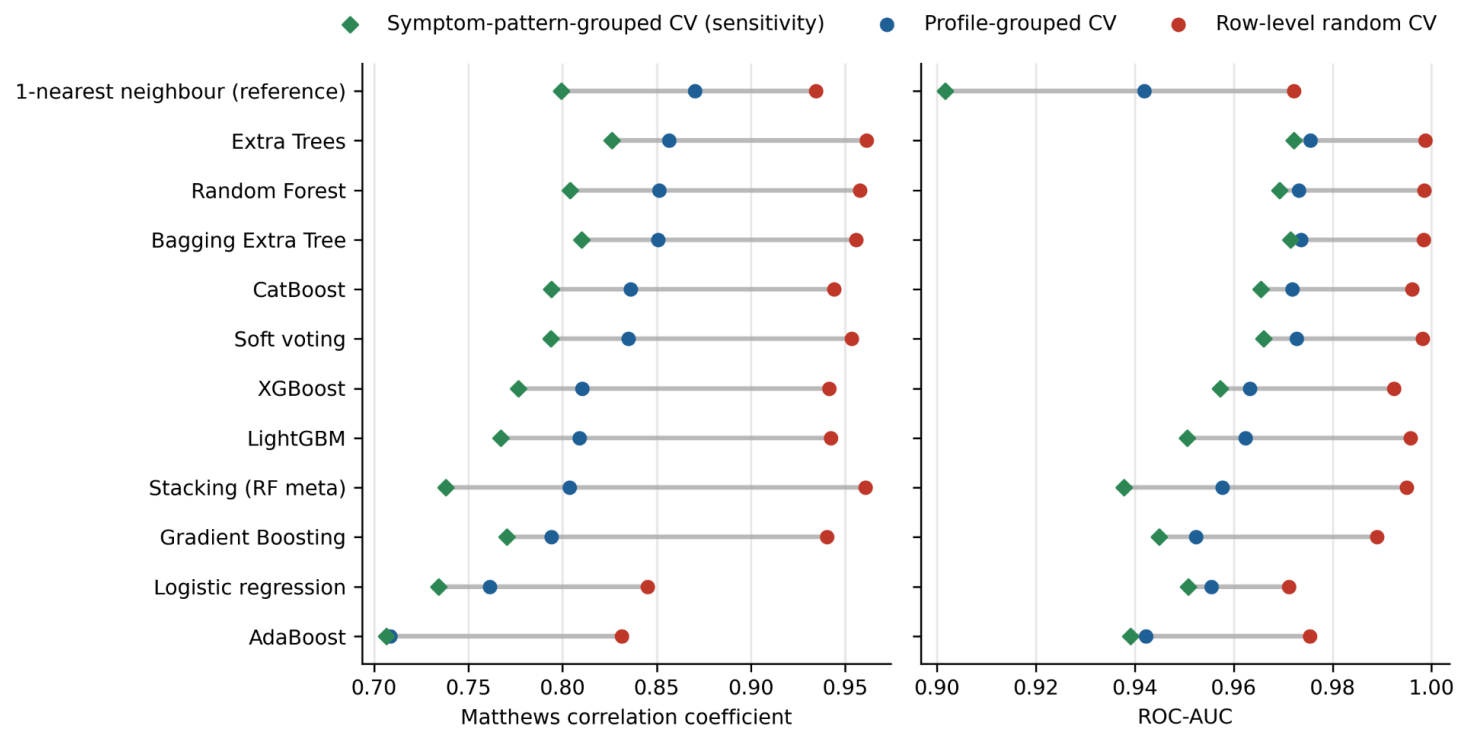


Figure 1

Optimism attributable to exact and near-duplicate profiles. Matthews correlation coefficient (left) and ROC-AUC (right) for each classifier under row-level random cross-validation (red), profile-grouped cross-validation (blue) and symptom-pattern-grouped cross-validation (green, sensitivity analysis).

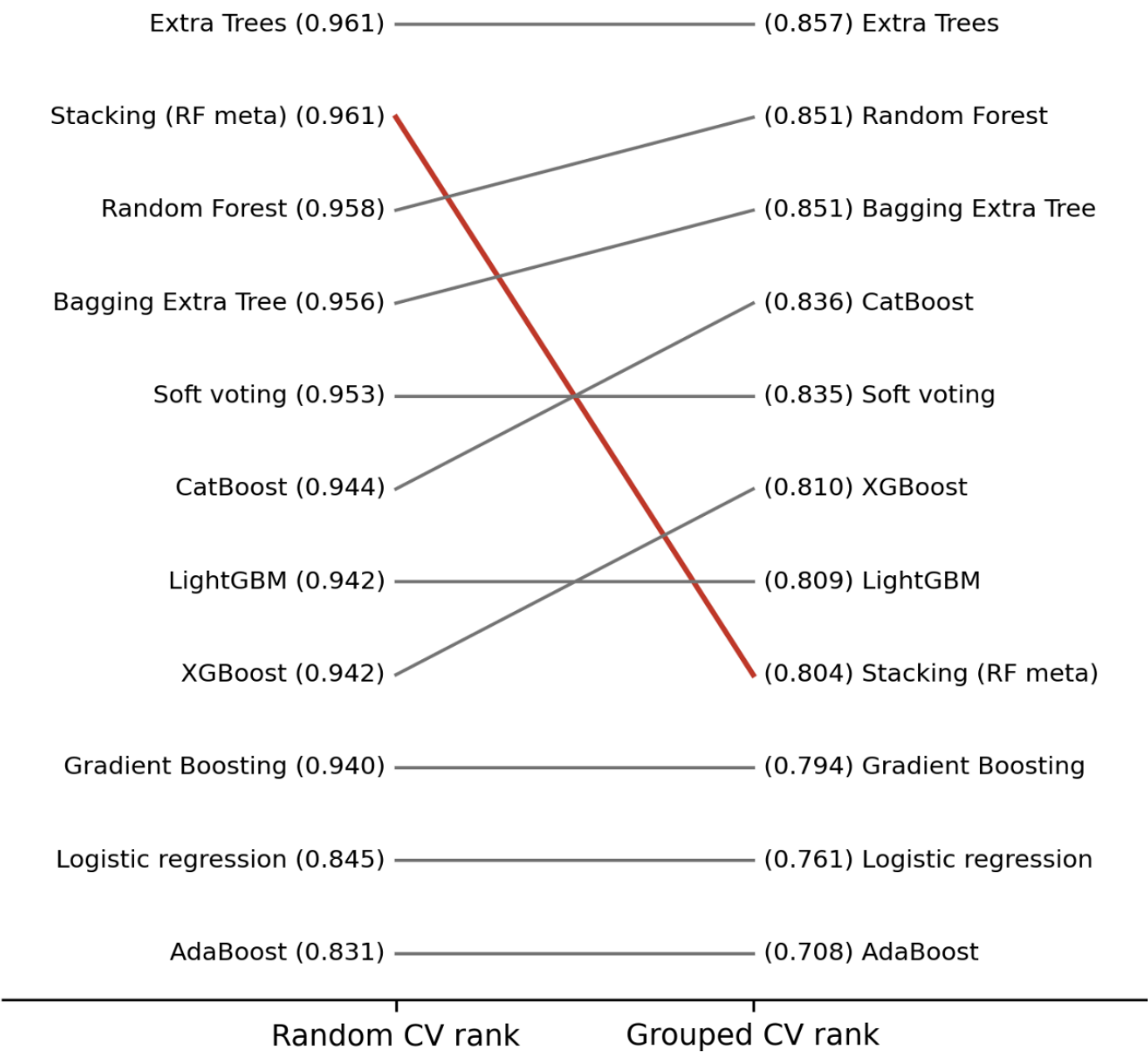


Figure 2

Ranking of classifiers by mean MCC under random and profile-grouped cross-validation (MCC in parentheses). The highlighted line marks the stacking ensemble, whose rank changed by six positions.

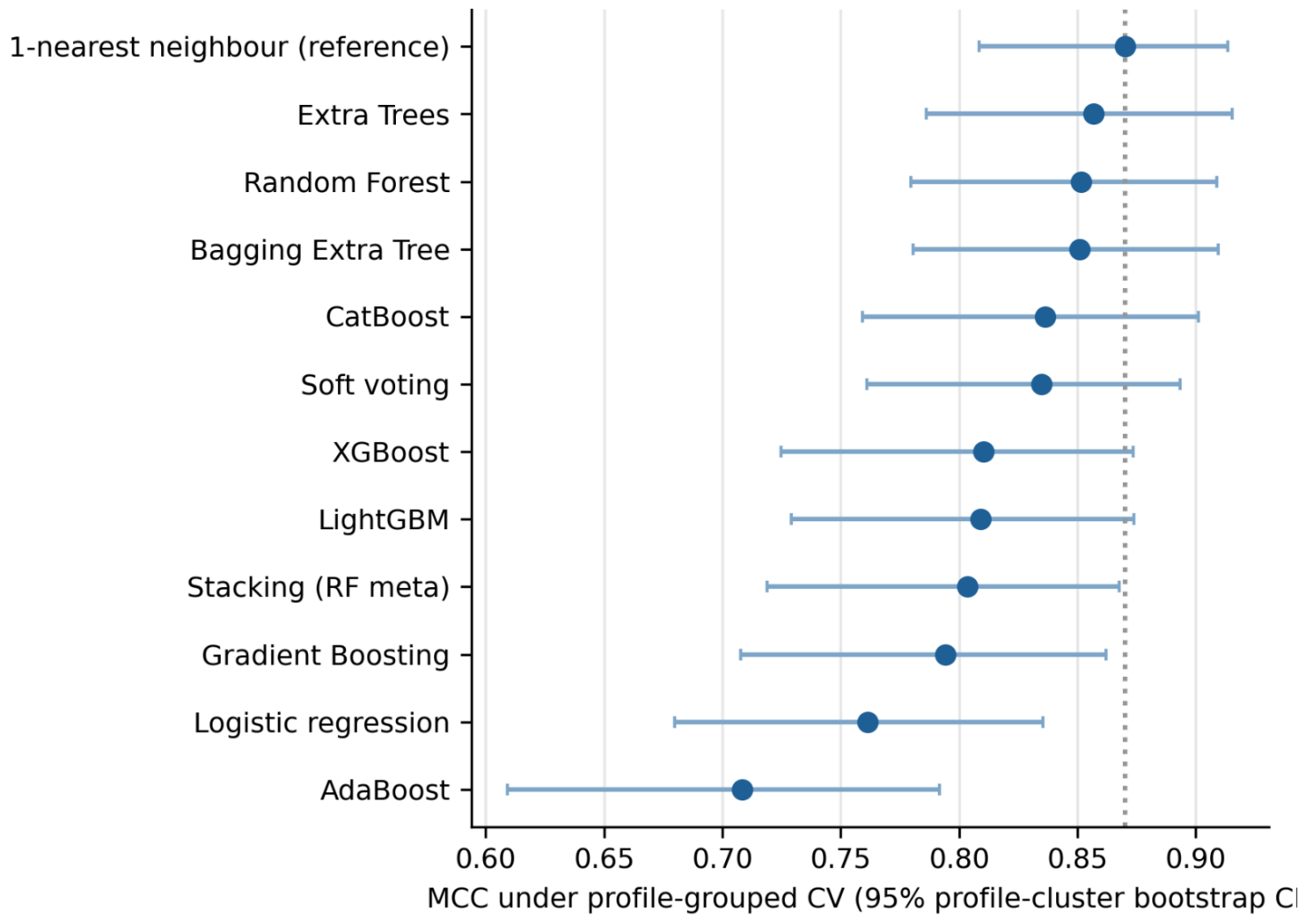


Figure 3

MCC under profile-grouped cross-validation with 95% profile-level cluster bootstrap confidence intervals. The dotted line marks the highest point estimate.

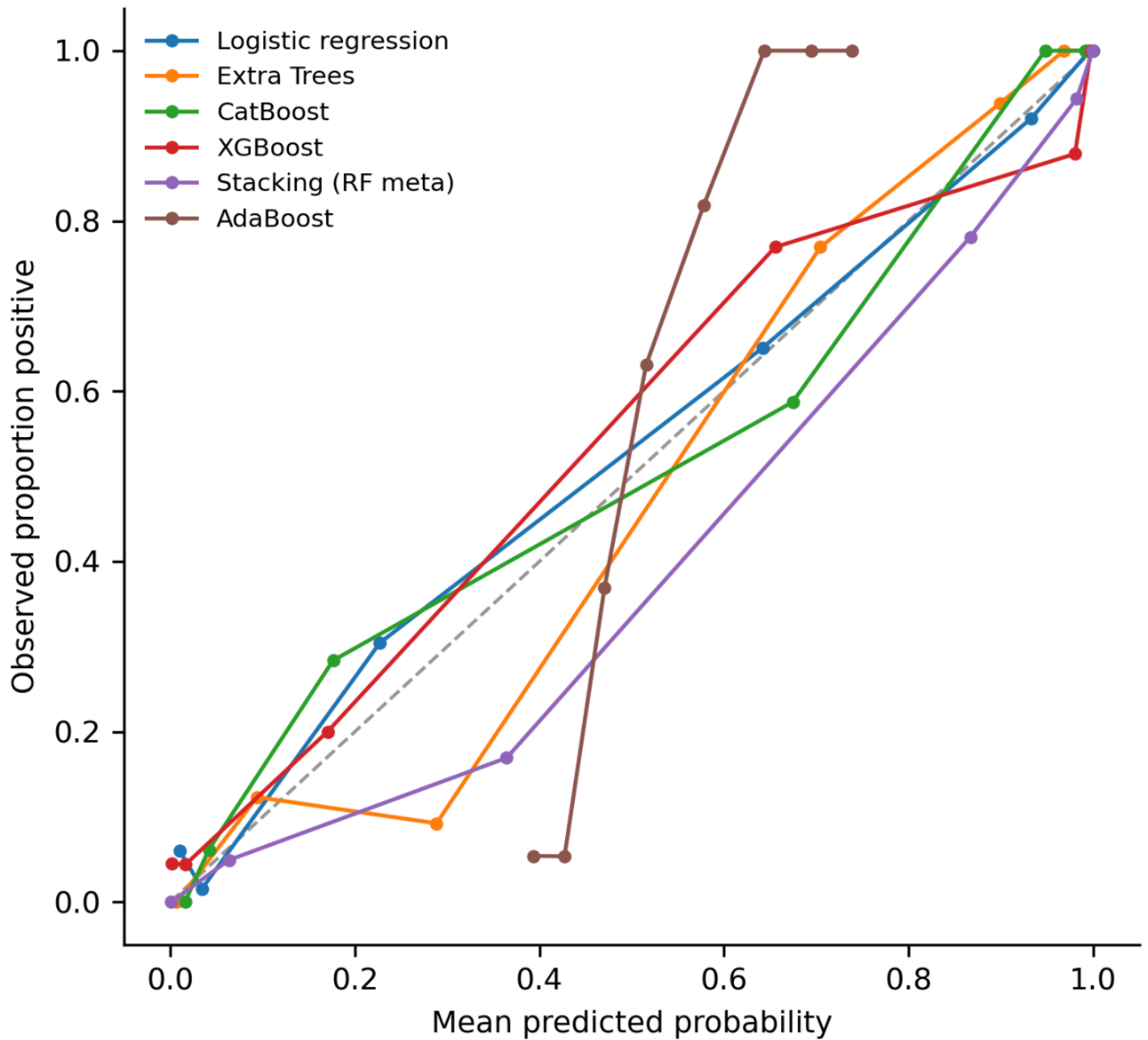


Figure 4

Calibration curves under profile-grouped cross-validation for selected classifiers (repetition-averaged out-of-fold probabilities, 8 quantile bins). The dashed line indicates perfect calibration.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1.docx](#)
- [Additionalfile2.docx](#)