Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

How Many Benchmark Datasets Contradict Themselves? Identical Inputs With Different Labels and the Error Floor They Impose

Maikel Yelandi Leyva-Vázquez

mleyvaz@gmail.com

Universidad Bernardo O'Higgins

Research Article

Keywords: Bayes error, data quality, duplicate records, inconsistency rate, OpenML, reproducibility

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11569047/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

How Many Benchmark Datasets Contradict Themselves? Identical Inputs With Different Labels and the Error Floor They Impose

Maikel Yelandi Leyva-Vázquez[1,2,3*]

ORCID: https://orcid.org/0000-0001-7911-5879.
[1*]Universidad Bernardo O'Higgins, Santiago, Chile.
[2]Universidad Bolivariana del Ecuador, Durán, Ecuador.
[3]Universidad de Guayaquil, Guayaquil, Ecuador.

Corresponding author(s). E-mail(s): mleyvaz@gmail.com;

Abstract

When two rows of a classification dataset have the same recorded inputs and different labels, no deterministic classifier that sees only those inputs can get both right. Following a protocol fixed in advance after a pilot, we count such contradictory duplicates in four OpenML benchmark suites (163 of 164 tasks, 134 distinct dataset versions) and report the smallest error attainable on the whole dataset, the same floor on the official cross-validation test folds and, under an independent-sampling model, lower confidence bounds on the Bayes error of the recorded inputs. In OpenML-CC18, 20 of 72 datasets contradict themselves and 9 have a floor of at least 1%; an exact-binomial 95% bound on the Bayes error is at least 1% for 5 of them. In analcatdata_dmft about half of the rows cannot be classified correctly by any function of the inputs. Ten-fold cross-validation scatters most contradictory groups, so the floors on the official folds are small. On the 101,766 rows of Diabetes130US, keeping only the seven numeric inputs of a derived benchmark version raises the floor from zero to 9.8%, with rows and labels unchanged. A contradictory group shows that the recorded inputs underdetermine the label; it does not identify why. The audit provides empirical floors for dataset and split-level reporting, together with population lower bounds whose interpretation requires an explicit sampling model.

Keywords: Bayes error, data quality, duplicate records, inconsistency rate, OpenML, reproducibility

1 Introduction

Benchmark suites of tabular classification datasets are the common ground on which learning algorithms are compared. OpenML-CC18 [1], the AutoML Benchmark [2] and the suites of Grinsztajn et al. [3] are run thousands of times, and differences of a fraction of a percentage point in accuracy are read as progress. How much of the remaining error can be removed at all is rarely asked. Several lines of work bear on the question. Audits of label errors combine a model with human validation [4]; studies of leakage examine how information crosses the split between training and test data [5]; and estimators of the Bayes error rely on soft labels [6] or on a fitted generative model [7], and have been compared on real data [8].

This paper looks at the simplest evidence a dataset gives about its own irreducible error: rows

whose recorded inputs are identical but whose labels differ. A deterministic classifier that sees only the inputs must give such rows the same prediction, so it errs on all but the most frequent label of each group. The statistic is standard. Feature-selection research calls it the inconsistency rate [9]; rough-set theory builds on the same partition of the rows into classes with identical attribute values [10]; a data-validation tool flags such rows as conflicting labels [11], and others flag exact and near duplicates [12]; and a contemporaneous preprint audits feature-identical label conflicts in anomaly-detection benchmarks and derives the evaluation ceiling they create [13]. A conflict does not, by itself, establish a labelling error: two correctly labelled records may share the recorded inputs while having different outcomes. A contradictory group is evidence that the recorded inputs underdetermine the label; whether the cause is a wrong label, a variable that was not recorded or another mechanism is not identified by the file. The contribution of this paper is applied: an audit of the supervised suites the field reports on, under definitions fixed in advance, and what it implies for the accuracies reported on them. We do not propose a new estimator of the Bayes error.

We ask four questions, written down before the full suites were downloaded: how many benchmark datasets contradict themselves (Q1); what error floor the contradictions impose on the full data, on the official folds and on the Bayes error (Q2); how much of the error of the best reported results they account for, and whether any reported result exceeds the ceiling they imply (Q3); and how the answers change when identity is relaxed to near-identity, and what kind of duplication produces the contradictions (Q4). The analysis is descriptive and makes no claim of error by dataset creators or by authors of reported runs.

In brief, 20 of the 72 OpenML-CC18 datasets, and 34 of the 134 dataset versions in the four suites, contain contradictory duplicates. Most floors are below 1%; in 14 versions they are larger, up to 50.1%. We contribute (i) a fixed-protocol census of exact label conflicts in 134 dataset versions from four supervised OpenML suites, with the grouping checked by a second parser and row-level witnesses for the main examples; (ii) their implications for multiclass 0–1 loss on full files and on the official cross-validation folds; (iii) a controlled comparison of input representations in Diabetes130US; and (iv) reproducible audit outputs and guidance for reporting and investigating conflicts (Section 6). An additional exact-binomial lower confidence bound on the population Bayes error is provided under an explicit i.i.d. sampling model. The underlying inconsistency statistic and the general identifiability principle are established ideas.

2 Related work

Label errors and noise. Northcutt et al. [4] estimated that on average at least 3.3% of the labels in the test sets of ten widely used image, text and audio benchmarks are wrong, using confident learning and human validation; Frénay and Verleysen [14] survey classification under label noise. Contradictory duplicates answer a narrower question. They need neither a model nor annotators, because they are visible in the data, but they do not reveal which label, if any, is wrong.

Duplicates and leakage. Duplicated records inflate test performance when copies fall on both sides of a split [5]. Near-duplicates were removed from CIFAR test sets by Barz and Denzler [15]; duplication was studied in code corpora by Allamanis [16] and in language-model training data by Lee et al. [17]. Single-dataset audits of tabular benchmarks report repeated records, for example a heart-disease dataset with 723 redundant rows out of 1,025 [18]. The NASA software-defect datasets, several of which are in CC18, have been audited for repeated and inconsistent records [19, 20]. Deng [13] audited 690 anomaly-detection datasets, found feature-identical label conflict in 147 of them, derived an attained AUROC ceiling and audited label conflicts in the released OddBench training and test splits; the idea that exact conflicts limit what a benchmark can identify is therefore not new. The present study quantifies multiclass 0–1 error floors across four supervised OpenML suites, pools them over official cross-validation folds, and reports population lower confidence bounds under an i.i.d. sampling model. Cardoso et al. [21] estimated instance difficulty in CC18 with item response theory and found that not all of its datasets are useful for comparing algorithms; difficulty for classifiers does not separate reducible from irreducible error.

Bayes error. The Bayes error L^* is the smallest error any classifier can reach [22]. The asymptotic error of the nearest-neighbour rule brackets it [23], but only in the large-sample limit and for a chosen metric; finite-sample estimators use soft labels [6] or compute the exact Bayes error of a fitted generative model [7], and Renggli et al. [8] compared several of them on real data. Data-complexity measures [24] and instance hardness [25] describe class overlap without giving a bound. For discrete inputs, $\hat{F}$ is the resubstitution error of the histogram (majority-vote) rule, and the optimistic bias of that error is well known [22, Chapter 23] and has been studied for discrete classifiers [26]; Section 3 uses it as an audit bound. The bound is specific to the recorded representation and is zero whenever no input repeats, so it does not replace those estimators. Hand [27] argued that much reported progress in classifier technology is small compared with the uncertainty in the problem itself; the floor makes one part of that uncertainty concrete.

3 Definitions and lower bounds on the Bayes error

Let a dataset consist of rows $i = 1, \dots, n$ with recorded input x_i and label $y_i \in \mathcal{Y}$. Group the rows by identical input; for a group g write n_g for its size and $n_{g,y}$ for the number of its rows with label y. A group is contradictory if it contains at least two labels. The full-data floor is

$$\hat{F} = \frac{K}{n}, \qquad K = \sum_g \left(n_g - \max_{y \in \mathcal{Y}} n_{g,y} \right). \tag{1}$$

Any deterministic function h of the input gives the same prediction to every row of a group, so it errs on at least $n_g - \max_y n_{g,y}$ of them, and predicting the majority label attains this. Hence $\hat{F}$ is the smallest 0–1 error that any deterministic classifier using only the inputs can reach on these n rows, even one fitted to them; the same holds on any subset of rows. This is a statement about a fixed file and needs no probability model. A predictor that draws fresh randomness for each row can do better on a given file by luck, but its expected number of errors in a group is still at least $n_g - \max_y n_{g,y}$. For an official OpenML task with test folds T_k, applying (1) within each test fold and pooling the forced errors over the folds gives the split floor $\hat{F}_{\text{split}}$, and $C = 1 - \hat{F}_{\text{split}}$ is a ceiling on the pooled cross-validated accuracy of any classifier whose prediction for a test row is a deterministic function of that row's input within each fold. Splitting a group can only reduce its forced errors, so $\hat{F}_{\text{split}} \le \hat{F}$; because a different classifier is fitted in each fold, $\hat{F}$ itself is not a floor for cross-validated error.

Under a sampling model the floor also says something about the population. Suppose that $(x_i, y_i)_{i=1}^n$ are independent and identically distributed draws from a common distribution of (X, Y), let $L^* = E\{1 - \max_y P(Y = y \mid X)\}$ be the Bayes error for predicting Y from the recorded X, and let $R_x = n^{-1} \sum_i \{1 - \max_y P(Y = y \mid X = x_i)\}$ be the Bayes risk at the observed inputs, so that $E(R_x) = L^*$. Write m for the number of rows that lie in groups of size at least two, and $\ell_\delta(k, n)$ for the one-sided exact-binomial lower confidence limit of Clopper and Pearson [28]: $\ell_\delta(0, n) = 0$ and, for $k \ge 1$, $\ell_\delta(k, n)$ is the δ quantile of the Beta$(k, n - k + 1)$ distribution.

Proposition 1 (known facts combined with standard tools). (a) $E(\hat{F} \mid x_1, \dots, x_n) \le R_x$ and hence $E(\hat{F}) \le L^*$. (b) For every $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$L^* \ge \ell_\delta(K, n). \tag{2}$$

(c) For every $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$L^* \ge \hat{F} - \sqrt{\frac{\log(1/\delta)}{2n}}. \tag{3}$$

(d) If the labels are independent given the inputs and $m \ge 1$, then for every $\delta \in (0, 1)$, with probability at least $1 - \delta$ given the inputs,

$$R_x \ge \hat{F} - \frac{1}{n} \sqrt{\frac{m \log(1/\delta)}{2}}; \tag{4}$$

if $m = 0$ then $\hat{F} = 0$ and the statement $R_x \ge 0$ is trivial.

Proof. (a) $\hat{F}$ is the training error of the rule that minimises training error over all functions of the input. A Bayes rule h^* is one such function, so $\hat{F}$ is at most the training error of h^*, whose conditional expectation given the inputs is R_x. (b) Let B be the number of rows on which h^* errs. Under independent sampling B is binomial with parameters n and L^*, and $K \le B$ on every sample,

because K is the smallest number of errors of any function of the input. The limit $\ell_\delta(k, n)$ is non-decreasing in k and satisfies $P\{\ell_\delta(B, n) > L^*\} \le \delta$ [28], so $P\{\ell_\delta(K, n) > L^*\} \le P\{\ell_\delta(B, n) > L^*\} \le \delta$. The argument does not require K to be binomial. (c) Replacing one row changes K by at most one, so $\hat{F}$ has bounded differences $1/n$ and McDiarmid's inequality [29] gives $P\{\hat{F} - E(\hat{F}) \ge t\} \le \exp(-2nt^2)$; combine with (a). (d) Given the inputs, $\hat{F}$ depends only on the m labels of rows in repeated groups (a singleton group contributes no forced error), each of which moves $\hat{F}$ by at most $1/n$; McDiarmid's inequality over these m labels gives $P\{\hat{F} - E(\hat{F} \mid x) \ge t \mid x\} \le \exp(-2n^2t^2/m)$; combine with (a). □

Five remarks. First, part (a) is the optimistic bias of the resubstitution error of the histogram rule [22, Chapter 23]; (b) to (d) add a confidence statement. Bound (3) is the one fixed in the preregistration. Bound (2) was added after the preregistration, following an AI-assisted adversarial review of a draft (see the statement on the use of AI tools); it uses the same assumption and the same count K, and in every dataset studied here it is at least as large as (3), often much larger. We therefore report (2) as the main bound and (3) for transparency. Second, (2) is positive whenever $K \ge 1$, at any confidence level, so whether a bound is positive carries no information beyond $\hat{F} > 0$; and a zero value of the looser bound (3) is not evidence that the Bayes error is zero. What is informative is the level that the bound certifies, and we count the datasets in which it reaches 1%; a small bound leaves the Bayes error unconstrained from above. Third, the targets differ: (2) and (3) bound the population quantity L^*, while (4) bounds R_x, the Bayes risk at the inputs that happened to be sampled. They are not two intervals for one quantity, and we report (4) only as a secondary result. Fourth, all of them refer to the recorded representation: adding a variable can lower L^* and R_x, so a large bound says that the columns provided do not determine the label, not that the labels are noisy. Fifth, the bounds are conditional on a sampling model that we cannot verify for any benchmark file. Repeated encounters of one patient, repeated measurements, records entered twice and independent units that share coarse inputs all produce repeated inputs. Exact input repetition alone neither establishes nor refutes the i.i.d. assumption; the sampling units and the data-generation process must also be considered. We therefore read the bounds as what the file would imply if its rows were independent draws, not as certified properties of the prediction problem.

A simulation with 2,000 replicates and $\delta = 0.05$ illustrates both the conservatism and the dependence on the model. In two designs with independent rows (a dense design with 30 possible input values, $n = 400$, and a sparse one in which 90% of the inputs are continuous, $n = 2000$) the three bounds exceeded their targets in at most 0.50% (2), 0.00% (3) and 0.00% (4) of the replicates. In the sparse design, where $L^* = 30.4\%$, the mean of (2) was 2.4% and that of (3) 0.3%: when few inputs repeat, every bound of this kind is far below the Bayes error. With dependent rows the guarantee is lost. In a design with three partly dependent copies of each unit ($n = 1200$) the miscoverage of (2) rose to 3.15%, and in a counterexample that arose in the same review (one constant input, $n = 400$; with probability 0.2 a uniformly selected half of the labels are 1 and the rest are 0, otherwise all labels are 0, so that each label is 1 with marginal probability 0.1 and $L^* = 0.1$) all three bounds failed in 20.2% of the replicates. Exact duplicates are also distance-zero nearest neighbours, so they could feed the bracket of Cover and Hart [23]; that route needs the nearest-neighbour risk of the whole distribution, which the repeated atoms alone do not give, and we do not pursue it.

Near-duplicates. Exact identity is sensitive to rounding. As a sensitivity analysis, each numeric column is min–max scaled and rounded to a grid of width $h \in \{0.001, 0.01, 0.05\}$, missing values staying missing, and (1) is recomputed on the coarsened inputs, giving $\hat{F}_h$. Rounding is a binning rule (two values on either side of a cell boundary stay apart) and min–max scaling is sensitive to outliers, so $\hat{F}_h$ is a rough indication of how close the data are to contradicting themselves, not a bound for models that see the raw inputs.

4 Data and protocol

What was fixed in advance, and what was already known. A protocol was committed to the project's git repository at 02:31 UTC on 12 October 2026, immediately before the download of the feature values, labels and results of the full suites began.

It fixed the definitions of $\hat{F}$, $\hat{F}_{\text{split}}$, bound (3), the near-duplicate grid, the populations and the outputs. It was written after a pilot, which it discloses: a smoke test on ten classic OpenML datasets (blood-transfusion among them, with a floor of 6.8%) and on roughly forty CC18 datasets, eleven of which had a positive floor (among them analcatdata_dmft, mfeat-morphological, cmc and several NASA defect sets), and a per-fold check on blood-transfusion. The census is therefore a fixed-protocol enumeration, not an independent confirmation of a hypothesis: the largest CC18 floors were known when the plan was frozen. New after the freeze were the CC18 datasets outside the pilot, the three secondary suites, the official splits of all tasks except the one piloted and the reported accuracies. Added after the preregistration, and exploratory in status: the Bonferroni counts, bounds (2) and (4), the simulation designs beyond the first, the controlled Diabetes130US comparison and the row-level witnesses. The time stamp is a local git commit, not a deposit in a public registry. Deviations are listed in the supplement.

Populations and files. The primary population is OpenML-CC18 [1] (OpenML study 99, 72 tasks). Secondary populations are the AutoML Benchmark classification suite [2] (study 271, 71 tasks) and the categorical and numerical classification suites of Grinsztajn et al. [3] (studies 334 and 337, 7 and 16 tasks). Together these are 164 tasks on 135 distinct dataset versions, because the suites share datasets; 163 tasks on 134 versions were analysed (excluded: KDDCup09-Upselling, a 1.8 GB file with 14,892 attributes whose analysis had not finished after 6.5 hours on the available laptop and was stopped (log: DEVIATIONS D6)). All files were fetched from OpenML [30] on 12 October 2026 (UTC): the canonical ARFF file of each dataset version (7.7 GB in total; the MD5 checksum matched the OpenML record for all 135 files), the official split file of each task (all are ten-fold cross-validation with one repeat), and the highest predictive-accuracy evaluations listed by the OpenML API for each task (at most 1,000). The supplement contains a frozen copy of the retrieved evaluation lists and checksums of the split files, since a later query of the live API returns different lists.

Grouping and its verification. The input vector of a row is every attribute except the task target, the dataset's ignore attributes and its row-identifier attribute, as in the usual OpenML loaders. Numeric values are compared as parsed double-precision numbers, nominal and string values as strings, and a missing value is a value of its own. Rows are grouped through a pandas row hash of the normalised input vector. Because a hash can collide and the reader normalises values (it strips surrounding whitespace and quotes), the grouping was checked: a separate ARFF parser (liac-arff), with values compared exactly as decoded and without the pandas hash, reproduced the floor count, the number of groups and the number of contradictory groups for all 134 files (112 files of at most 50 MB grouped as exact tuples; 22 larger files read as a stream and grouped on a 128-bit BLAKE2b digest of each row). For eight named datasets the supplement lists row-level witnesses: the row numbers, labels and shared input vector of the contradictory groups with most forced errors. On a laptop with little free memory the median computing time per task was 0.9 seconds and 91% of the tasks took less than a minute; the largest file (KDDCup99, 4,898,431 rows) took 6.6 hours. The preregistered typology labels each dataset with a positive floor as *coarse* (fewer distinct input vectors than half the rows), *concentrated* (called "degenerate record" in the preregistration: otherwise, and one input vector holds at least half of the contradictory rows) or *scattered*.

5 Results

5.1 How many datasets contradict themselves?

Table 1 gives the census. Repeated inputs are common: 37 of the 72 CC18 datasets contain at least one input vector that occurs twice. 20 CC18 datasets (27.8%) contain at least one contradictory group, at the upper end of the 10 to 20 expected from the pilot. The suite is enumerated in full, so this count has no sampling error; purely as an illustration of precision if CC18 were regarded as a sample of curated benchmarks, a 95% Wilson interval for the share would be 18.8% to 39.0%. In the AutoML Benchmark the count is 19 of 70. In the Grinsztajn suites it is 2 of 7 and 4 of 16, even though their construction removed rows with missing values and numerical features with fewer than ten distinct values

[3, Section 3.2]. Over the 134 distinct dataset versions, 34 contradict themselves.

5.2 How large are the floors?

Most floors are small (Table 2, Figure 1). The median full-data floor among the CC18 datasets that have one is 0.44%, and 9 of them reach 1%. The four largest CC18 floors are those of analcatdata_dmft (50.1%), blood-transfusion-service-center (6.8%), mfeat-morphological (5.3%) and cmc (4.5%). In analcatdata_dmft, 687 of 797 rows sit in contradictory groups: from the four recorded inputs (the decayed, missing and filled teeth index before and after the programme, gender and ethnicity) no function can recover the prevention strategy of half of the individuals. Its six strategies are almost equally frequent, so a rule that ignores the inputs errs on 80.6% of the rows; if the strategy was assigned independently of the inputs, as in a randomised design, the Bayes error is close to that value and the floor understates it. Outside CC18 the largest floors are in airlines (19.9%; 280,982 of 539,383 rows in contradictory groups), compas-two-years (19.0%), Diabetes130US (7.4%) and sf-police-incidents (3.4%). In airlines the inputs describe the scheduled flight (carrier, flight number, airports, day of week, departure time, length) but not its date, so the same scheduled flight appears many times, sometimes delayed and sometimes not; these rows are repeated observations of one schedule, a case in which the independent-sampling reading is doubtful.

A controlled comparison: Diabetes130US. Two versions of this dataset are in the suites. The AutoML Benchmark version has 101,766 rows (hospital encounters of 71,518 patients), 49 inputs and three classes, and no contradictions; the version in the Grinsztajn numerical suite has 71,090 rows, 7 numeric inputs and two classes, and a floor of 7.4%. The two files differ in rows, inputs and target, so the difference between their floors does not isolate the effect of any one step. We therefore recomputed the floor on the same 101,766 rows of the first version, changing one thing at a time (Table 3). Two of its 49 inputs are encounter and patient identifiers, which make every row unique by construction; without them the remaining 47 inputs still give 101,766 distinct input vectors and a floor of zero. Keeping only the seven numeric inputs of the reduced version, with the rows and the three-class target unchanged, leaves 79,546 distinct vectors and a floor of 9.8%. Restricting the target to its two largest classes, the rule described by Grinsztajn et al. [3, Section 3.2], removes 11,357 rows and gives 7.8%; the distributed reduced file has a floor of 7.4%. That file carries no row identifiers, so we cannot align it row by row; it is compatible with being a class-balanced subsample of that subset (its multiset of inputs and labels is contained in it), which is what the published preprocessing rule describes, but containment does not prove how the file was made. On fixed rows and a fixed target, then, the reduction of the inputs alone creates the contradictions; the later row filtering changes the floor by about two percentage points. The rows are encounters, not patients, so this is a statement about the file.

Two features of Table 2 matter for practice. First, the split floor is much smaller than the full-data floor: ten-fold cross-validation scatters each contradictory group over several test folds, so few forced errors remain inside any one fold (blood-transfusion: 6.8% on the full data, 3.3% pooled over the folds, between 1.3% and 8.0% in single folds). The fold ceiling is therefore a weak constraint on published accuracies. Second, the population bounds do not depend on the folds, but they depend on the sampling model and on which inequality is used. Under independent sampling the exact bound (2) is at least 1% for 5 CC18 datasets (analcatdata_dmft, blood-transfusion-service-center, mfeat-morphological, cmc, first-order-theorem-proving) and at least 5% for 2; over all 134 versions the counts are 10 and 5. These are 95% statements for each dataset separately; with a Bonferroni correction over the datasets of each row of Table 1 the counts at 1% are 5 and 10, and at 5% they are 1 and 4. The preregistered bound (3) is looser: for blood-transfusion it is 2.34% against 5.36%, and it is positive for only 5 CC18 datasets (2 with Bonferroni) and 10 versions (7), although (2) is positive for all 34 versions with a positive floor. The median of (2) over those versions is 0.11%: for most versions that contradict themselves the certified lower bound is small and does not materially constrain attainable accuracy. It says nothing about how large the Bayes error is; the true value may be far above the bound, as in the sparse simulation design of Section 3. The counts above are counts of versions for which the analysis certifies at least the stated error level

Table 1 Contradictory duplicates in four OpenML suites (number of datasets).

Suite	N	Rep.	$\hat{F} > 0$	$\hat{F} \geq 1\%$	(2) $\geq 1\%$	Bonf.	(3) > 0	$\hat{F}_{\text{split}} > 0$	$\hat{F}_{0.01} > 0$
OpenML-CC18	72	37	20	9	5	5	5	13	26
AutoML Benchmark	70	33	19	7	5	5	5	13	28
Grinsztajn categorical	7	3	2	1	1	1	1	2	4
Grinsztajn numerical	16	9	4	2	2	2	2	4	9
All versions (unique)	134	65	34	14	10	10	10	24	52

N: datasets analysed. Rep.: at least one input vector occurs twice. (2) $\geq 1\%$: the exact lower bound on L^* with $\delta = 0.05$, valid under independent sampling, is at least 1% (not preregistered). Bonf.: the same with $\delta = 0.05$ divided by the number of datasets in the row. (3) > 0: the preregistered bound with $\delta = 0.05$ is positive. $\hat{F}_{0.01}$: floor after rounding min–max scaled numeric inputs to a grid of width 0.01. The last row counts each dataset version once.

Table 2 CC18 datasets with contradictory duplicates, by full-data floor.

Dataset	n	Groups	Rows in groups	$\hat{F}$	$\hat{F}_{\text{split}}$	Bound (2)	Bound (3)	Type
analcatdata_dmft	797	151	687	50.06%	16.31%	47.09%	45.73%	coarse
blood-transfusion-service-center	748	31	200	6.82%	3.34%	5.36%	2.34%	scattered
mfeat-morphological	2,000	106	214	5.35%	0.55%	4.55%	2.61%	scattered
cmc	1,473	62	140	4.55%	0.75%	3.69%	1.36%	scattered
first-order-theorem-proving	6,118	118	334	2.39%	0.47%	2.07%	0.82%	scattered
kc2	522	6	118	1.53%	1.15%	0.76%	0	concentrated
kc1	2,109	20	253	1.14%	0.52%	0.79%	0	scattered
jm1	10,885	88	889	1.12%	0.44%	0.96%	0	scattered
PhishingWebsites	11,055	64	357	1.03%	0.39%	0.88%	0	scattered
pc1	1,109	6	13	0.54%	0	0.24%	0	scattered
segment	2,310	2	51	0.35%	0.26%	0.17%	0	concentrated
pc3	1,563	3	9	0.26%	0.13%	0.09%	0	scattered
dresses-sales	500	1	2	0.20%	0	0.01%	0	concentrated
Internet-Advertisements	3,279	3	54	0.09%	0.06%	0.02%	0	concentrated
pc4	1,458	1	3	0.07%	0	<0.01%	0	concentrated
spambase	4,601	3	6	0.07%	0	0.02%	0	scattered
dna	3,186	1	2	0.03%	0	<0.01%	0	concentrated
splice	3,190	1	2	0.03%	0	<0.01%	0	concentrated
adult	48,842	5	10	0.01%	<0.01%	<0.01%	0	scattered
nomao	34,465	2	8	0.01%	0	<0.01%	0	concentrated

Groups: number of contradictory groups. Rows in groups: rows in those groups. Bounds on L^* with $\delta = 0.05$ under independent sampling: (2) exact binomial, (3) preregistered, truncated at zero. A positive value that rounds to 0.00% is shown as <0.01%; 0 is an exact zero. Type: preregistered typology.

under the model, not counts of versions whose irreducible error is or is not important in practice. The conditional bound (4), whose target is R_x, is positive for 7 CC18 datasets and 15 versions. Of the 10 versions in which (2) reaches 1%, 1 owes it to a single contradictory group (heloc); such a bound describes how the file encodes one input vector rather than the prediction problem.

5.3 What do the floors mean for reported accuracies?

For the CC18 tasks we compared 69,654 reported accuracies (the best 1,000 per task as listed by OpenML; at least 43 for every task) with the fold ceilings (Table 4, Figure 2). No reported accuracy exceeds its fold ceiling. This is a sanity check rather than a finding: a prediction that is a deterministic function of the input on each fold cannot exceed the ceiling, and the absence of exceedances says nothing about whether the runs are free of leakage or otherwise correct. Among the 13 tasks with a positive split floor, the ratio of the split floor to the error of the best reported run has a median of 3.1%, is at least 10% in 4 tasks, and is at most 22.1% (analcatdata_dmft). The ratio uses the floor pooled over the folds, whereas a listed accuracy may be a mean over folds; the two ceilings differ by at most 0.00004, so the ratio is a close approximation and not an exact decomposition of a run's errors. Nor does it identify which errors of a run were forced.

The population bounds cannot be compared with a reported accuracy in the same way. A bound refers to the expected error of any classifier on new draws under the sampling model, whereas a reported accuracy is a cross-validated estimate for one run selected as the best of those listed, and

Table 3 Diabetes130US: floor on the rows of the AutoML Benchmark version when inputs, target and rows are changed one at a time (exploratory).

Rows	Inputs	Target	Distinct inputs	K	$\hat{F}$
101,766	49 (as distributed)	3 classes	101,766	0	0
101,766	47 (no identifiers)	3 classes	101,766	0	0
101,766	7 numeric	3 classes	79,546	9,934	9.76%
90,409	47 (no identifiers)	2 largest classes	90,409	0	0
90,409	7 numeric	2 largest classes	71,206	7,084	7.84%
71,090	7 numeric	2 classes, balanced (reduced file)	58,703	5,233	7.36%

Rows 1 to 5 use OpenML data id 4541; the last row is data id 45022 as distributed. K: number of forced errors.

the number of runs listed differs widely between tasks. We give the figures side by side, without a ratio, for three datasets. In analcatdata_dmft the best of 1,000 listed runs has an error rate of 73.7%, bound (2) is 47.1% and the majority-class error is 80.6%; in blood-transfusion (1,000 runs) the best error is 19.3% and the bound 5.4%; for airlines only 4 runs are listed, with a best error of 36.4% and a bound of 19.8%.

5.4 Near-duplicates, missing values and where the contradictions sit

Exact identity understates how close many datasets are to contradicting themselves (Figure 3). Rounding min–max scaled numeric inputs to a grid of width 0.01 raises the number of CC18 datasets with a positive floor from 20 to 26, and a grid of width 0.05 to 30; in the numerical Grinsztajn suite the counts go from 4 to 9 and 13 of 16. The most instructive case is mfeat-zernike, which has no exact contradiction but a floor of 5.2% on a grid of width 0.001. All of its near-identical groups, like all 106 contradictory groups of mfeat-morphological, mix the classes coded 7 and 10, that is, the digits 6 and 9 (the original files store the digits in blocks of 200 from 0 to 9, and the original morphological file shows the same 106 collisions). Features that do not change when a digit is rotated cannot tell a 6 from a 9. How rows with missing inputs are treated barely matters; dropping them changes the floor of 3 CC18 datasets.

The preregistered typology separates three situations among the 20 CC18 datasets with a positive floor: 1 coarse, 8 concentrated and 11 scattered (over all versions: 4, 12 and 18). It is a blunt description. Coarse datasets such as analcatdata_dmft, airlines and compas-two-years have few distinct input vectors, and their contradictions are consistent with omitted variables (in airlines, the date); KDDCup99 is also coarse, because most of its rows repeat, yet its floor is only 0.0004%. In 8 of the 12 concentrated versions there is only one contradictory group, which meets the rule trivially. We inspected the largest contradictory group of five selected versions, not all of them; in each the shared vector is constant and looks like a code for missing information rather than a measurement: all sixteen numeric inputs equal to zero in segment, zero height, width and aspect ratio in Internet-Advertisements, the code -9 in every input of 557 heloc rows, and -999 in all fifty inputs of the two MiniBooNE versions. Scattered datasets, including the NASA defect sets kc1, jm1, pc1 and pc3 [19, 20], have many small contradictory groups. The data cannot tell label noise from omitted variables, and we do not try.

6 Discussion

Contradictory duplicates occur in about a quarter of the datasets of each suite, and in most of them the floor is small. We see three uses. First, benchmark suites could publish $\hat{F}$ and $\hat{F}_{\text{split}}$ as dataset metadata, next to the number of instances and classes, so that users see when part of the error cannot be removed by a deterministic function of the columns provided; the population bounds can be added where a sampling model is defensible. Second, when a benchmark version is derived from another by dropping columns, the floor recomputed on the same rows before and after shows what the reduction removes, as in the Diabetes130US comparison. Third, a concentrated floor identifies an input pattern for source-level investigation. Curators should document its provenance and missing-value coding; records should be corrected or removed only when

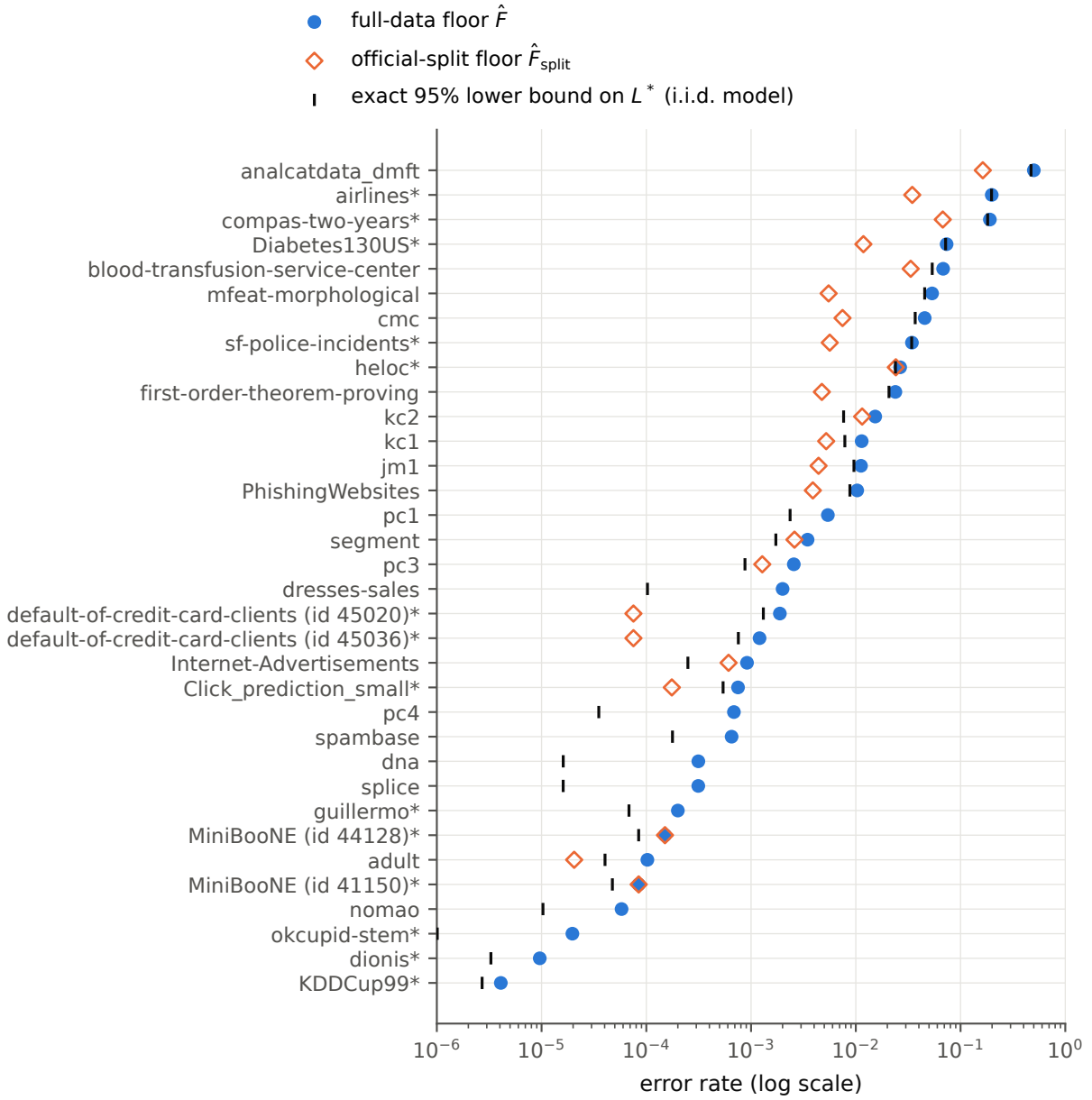


Fig. 1 Full-data floor, official-split floor and exact lower bound (2) on L^* under independent sampling, for every dataset version with a positive floor (log scale). Zeros are not drawn: 10 of the 34 versions have a split floor of zero. An asterisk marks versions that are not in OpenML-CC18

their invalidity is established, with the resulting change in the evaluation population reported. Table 5 summarises these uses and the limit of each.

A general lesson runs through these cases: a positive floor certifies that the columns provided do not determine the label, but it does not identify why. The Diabetes130US comparison establishes that removing columns is sufficient to create empirical label conflicts while holding the rows and labels fixed. It does not establish that every label is correct, nor does it identify which omitted variables would support accurate prediction on new observations. In airlines, omission of the flight date is a plausible source of repeated input vectors; in the NASA defect datasets, the file

Table 4 Best reported OpenML accuracy and the fold ceiling implied by contradictory duplicates (CC18 tasks with a positive full-data floor).

Dataset	Runs	Best acc.	Fold ceiling C	$\hat{F}_{\text{split}}$	Split ratio
analcatdata_dmft	1,000	0.263	0.837	16.31%	22.1%
blood-transfusion-service-center	1,000	0.807	0.967	3.34%	17.4%
mfeat-morphological	1,000	0.772	0.995	0.55%	2.4%
cmc	1,000	0.576	0.993	0.75%	1.8%
first-order-theorem-proving	1,000	0.639	0.995	0.47%	1.3%
kc2	1,000	0.870	0.989	1.15%	8.8%
kc1	1,000	0.873	0.995	0.52%	4.1%
jm1	1,000	0.825	0.996	0.44%	2.5%
PhishingWebsites	1,000	0.976	0.996	0.39%	16.1%
pc1	1,000	0.945	1.000	0	0
segment	1,000	0.981	0.997	0.26%	14.0%
pc3	1,000	0.909	0.999	0.13%	1.4%
dresses-sales	1,000	0.658	1.000	0	0
Internet-Advertisements	1,000	0.980	0.999	0.06%	3.1%
pc4	1,000	0.929	1.000	0	0
spambase	1,000	0.963	1.000	0	0
dna	1,000	0.969	1.000	0	0
splice	1,000	0.988	1.000	0	0
adult	1,000	0.876	1.000	<0.01%	<0.1%
nomao	1,000	0.976	1.000	0	0

Runs: evaluations listed (at most 1,000); the best accuracy is the maximum over that many runs. Fold ceiling: pooled over the ten test folds; the ceiling computed as a mean over folds differs from it by at most 0.00004, and the exceedance check used the larger of the two. Split ratio: $\hat{F}_{\text{split}}/(1 - \text{best acc.})$ (preregistered). A positive value that rounds to zero is shown as an inequality.

alone does not identify the source of the conflicts. Reporting the floor without this caution would invite the reading that the labels are noisy.

The analysis has limits. The floor is a property of a file; it does not say whether the conflicting labels are errors. The bounds only use exact repeats, so they are zero for most continuous data and say nothing about their Bayes error; where they are large they can still be far below it, as in analcatdata_dmft. They hold under i.i.d. sampling from a common distribution, an assumption we did not establish for any benchmark. Deduplication, class balancing, outcome-dependent selection and repeated measurements can alter the sampling structure, so the bounds should be interpreted conditionally on the stated model; the simulation shows that dependence can destroy the coverage. The fold ceiling holds for predictions that are deterministic functions of the input; a randomised predictor, or one that uses row order, could exceed it on a given fold. The census was specified after a pilot that had already shown the largest CC18 floors, and its time stamp is a local git commit rather than a public registration. We audit the files as OpenML distributes them, which is what benchmark users load; conversion from the original sources may have added or removed duplicates (for mfeat-morphological the original file has the same collisions). The reader treats the percent sign as a comment marker and single quotes as the quote character; a second parser reproduced the grouping for all 134 files. The reported accuracies are those listed by OpenML on the day of retrieval; they are dominated by a few automated pipelines, and outside CC18 they are scarce; results published elsewhere on other splits are not covered. Near-duplicate floors depend on the chosen grid and are descriptive only.

7 Conclusion

About one in four OpenML-CC18 datasets (20 of 72) contains rows that contradict each other. In 11 of them the resulting floor is below 1%; in 9 it is larger, up to approximately half of the rows. If the rows are read as independent draws, an exact 95% lower bound on the Bayes error of the recorded inputs reaches 1% in 10 of the 134 dataset versions studied. The empirical floor requires no sampling model. Full-data floors can accompany dataset

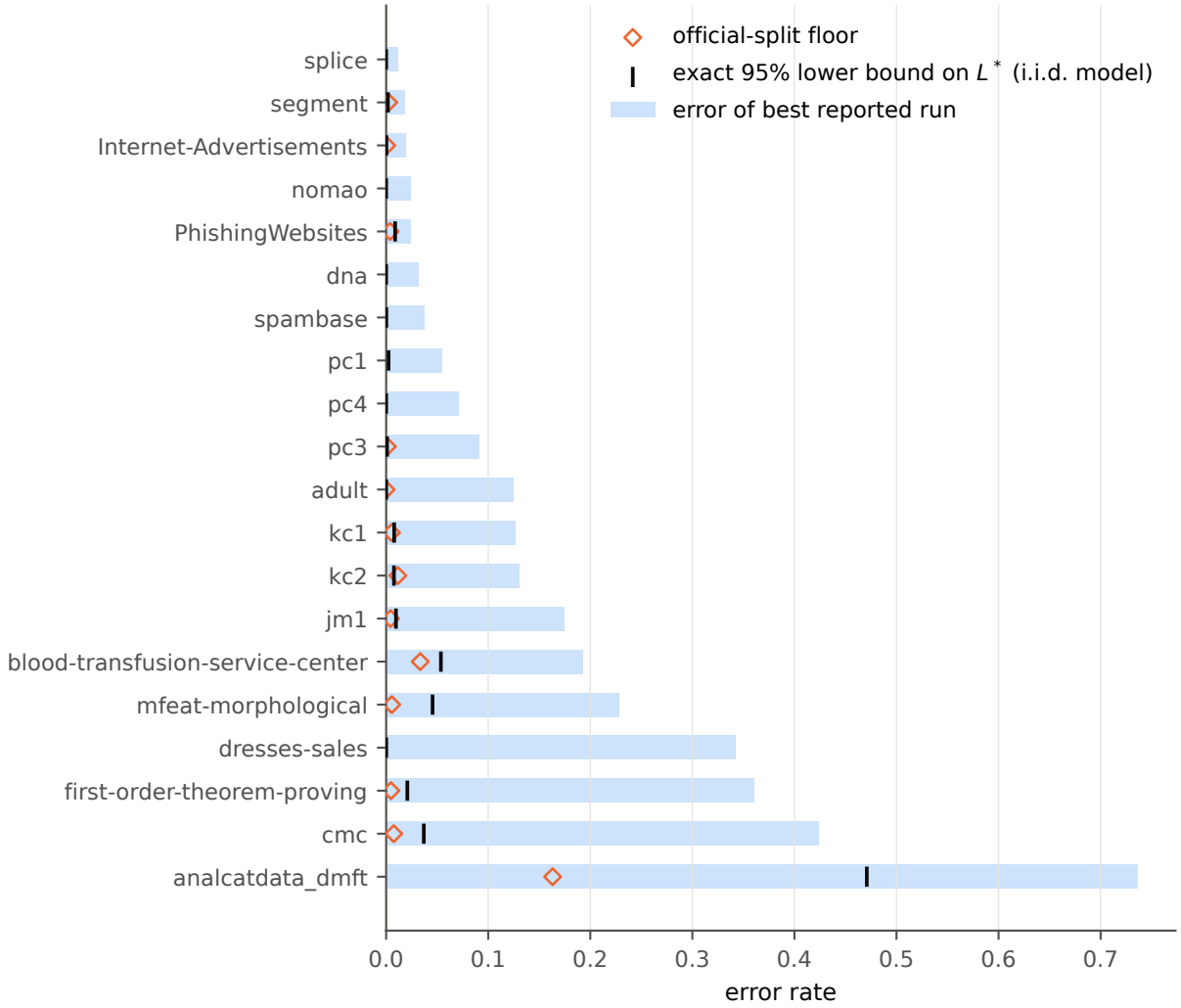


Fig. 2 Error of the best reported OpenML run (bars) against the official-split floor and the exact lower bound (2) on the Bayes error under independent sampling, for CC18 tasks with a positive full-data floor. The bound and the bars are different quantities (see text)

descriptions, while split-specific floors can accompany cross-validated accuracies. Population lower confidence bounds should be reported separately with their sampling assumptions. None of these quantities identifies which labels, if any, are wrong.

Statements and Declarations

- **Competing interests.** The author has no relevant financial or non-financial interests to disclose.
- **Funding.** This research received no specific grant from any funding agency.
- **Data availability.** All data are public OpenML datasets, tasks and evaluations (https://www.openml.org), identified in the supplementary material by dataset and task id with MD5 checksums.
- **Code availability.** Code (Python), the preregistration and its deviation log, per-task results, row-level witnesses, the frozen evaluation lists, checksums of all input files, the simulation of the bounds and the scripts that rebuild every number, table and figure from the OpenML files are provided as supplementary material (Online Resource 1). A public deposit with a DOI will be made before publication.

Table 5 Interpreting audit findings: recommended action and limit of interpretation.

Audit finding	Recommended action	Interpretation limit
Positive full-data floor	Report the input representation, the conflicting groups and the empirical floor	Does not establish label errors
Positive official-fold floor	Report the ceiling with the same fold weighting as the accuracy	Does not identify which errors of a fitted model were forced
Floor increases after feature removal	Document the transformation and compare on fixed rows and labels	Does not validate the labels or the discarded predictors
Conflicts concentrate in a constant vector	Investigate source records and missing-value coding	Does not justify automatic deletion
Positive population lower bound	State the sampling model and the confidence level	Does not establish that the benchmark satisfies that model

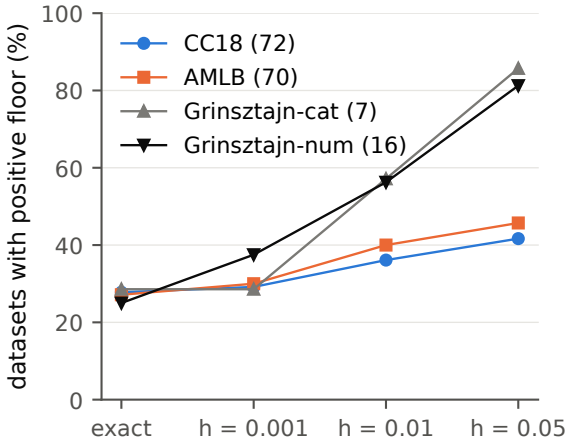


Fig. 3 Share of datasets with a positive floor when numeric inputs are compared exactly or after rounding to grids of width h (min–max scale). Numbers in the legend are dataset versions per suite

- **Use of AI tools.** The author used Claude (Anthropic), through the Claude Code command-line tool in October 2026, for coding assistance, literature search support, drafting and language editing, and ChatGPT (OpenAI, October 2026) as an adversarial reviewer of drafts; bound (2) and the dependence counterexample of Section 3 came out of that review. This was not editorial peer review. All AI-assisted material was checked by the author, who takes full responsibility for the content, accuracy, originality and integrity of the manuscript.
- **Ethics approval and consent.** Not applicable. The study uses public, de-identified benchmark files and involves no new data from human participants.
- **Author contributions.** M.Y.L.-V. is the sole author: conceptualization, methodology, software, analysis and writing.

References

[1] Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Pieter Gijsbers, Frank Hutter, Michel Lang, Rafael G. Mantovani, Jan N. van Rijn, and Joaquin Vanschoren. OpenML benchmarking suites. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. arXiv:1708.03731.

[2] Pieter Gijsbers, Marcos L. P. Bueno, Stefan Coors, Erin LeDell, Sébastien Poirier, Janek Thomas, Bernd Bischl, and Joaquin Vanschoren. AMLB: an AutoML benchmark. *Journal of Machine Learning Research*, 25 (101):1–65, 2024.

[3] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems 35 (Datasets and Benchmarks Track)*, pages 507–520, 2022. https://doi.org/10.52202/068431-0037.

[4] Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. arXiv:2103.14749.

[5] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4 (9):100804, 2023. https://doi.org/10.1016/j.patter.2023.100804.

[6] Takashi Ishida, Ikko Yamane, Nontawat Charoenphakdee, Gang Niu, and Masashi Sugiyama. Is the performance of my deep

network too good to be true? A direct approach to estimating the Bayes error in binary classification. In *International Conference on Learning Representations (ICLR)*, 2023. https://openreview.net/forum?id=FZdJQgy05rz.

[7] Ryan Theisen, Huan Wang, Lav R. Varshney, Caiming Xiong, and Richard Socher. Evaluating state-of-the-art classification models against Bayes optimality. In *Advances in Neural Information Processing Systems 34*, pages 9367–9377, 2021.

[8] Cedric Renggli, Luka Rimanic, Nora Hollenstein, and Ce Zhang. Evaluating Bayes error estimators on real-world datasets with FeeBee. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. arXiv:2108.13034.

[9] Manoranjan Dash and Huan Liu. Consistency-based search in feature selection. *Artificial Intelligence*, 151(1–2): 155–176, 2003. https://doi.org/10.1016/S0004-3702(03)00079-1.

[10] Zdzisław Pawlak. Rough sets. *International Journal of Computer and Information Sciences*, 11(5):341–356, 1982. https://doi.org/10.1007/BF01001956.

[11] Deepchecks. Conflicting labels check (tabular data integrity), documentation version 0.6. Software documentation, 2022. https://docs.deepchecks.com/0.6/checks_gallery/tabular/integrity/plot_conflicting_labels.html, accessed 12 October 2026.

[12] Cleanlab. Datalab issue types: near duplicate issue (documentation of cleanlab version 2.7.1). Software documentation, 2025. https://docs.cleanlab.ai/v2.7.1/cleanlab/datalab/guide/issue_type_description.html, accessed 13 October 2026.

[13] Jie Deng. When identical rows disagree: from benchmark identifiability to replication-robust anomaly detection. Preprint, arXiv:2609.29580, 2026. https://doi.org/10.48550/arXiv.2609.29580.

[14] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014. https://doi.org/10.1109/TNNLS.2013.2292894.

[15] Björn Barz and Joachim Denzler. Do we train on test data? Purging CIFAR of near-duplicates. *Journal of Imaging*, 6(6):41, 2020. https://doi.org/10.3390/jimaging6060041.

[16] Miltiadis Allamanis. The adverse effects of code duplication in machine learning models of code. In *Proceedings of the 2019 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software (Onward! 2019)*, pages 143–153, 2019. https://doi.org/10.1145/3359591.3359735.

[17] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, 2022. https://doi.org/10.18653/v1/2022.acl-long.577.

[18] Oğuzhan Kilim. Duplicate clinical profiles as a source of data leakage in heart disease classification: a benchmark audit of validation, calibration, and explainability. *Advances in Artificial Intelligence Research*, 6:52–66, 2026. https://doi.org/10.54569/aair.1998842.

[19] David Gray, David Bowes, Neil Davey, Yi Sun, and Bruce Christianson. The misuse of the NASA Metrics Data Program data sets for automated software defect prediction. In *15th Annual Conference on Evaluation and Assessment in Software Engineering (EASE 2011)*, pages 96–103, 2011. https://doi.org/10.1049/ic.2011.0012.

[20] Martin Shepperd, Qinbao Song, Zhongbin Sun, and Carolyn Mair. Data quality: some comments on the NASA software defect datasets. *IEEE Transactions on Software*

Engineering, 39(9):1208–1215, 2013. https://doi.org/10.1109/TSE.2013.11.

[21] Lucas F. F. Cardoso, Vitor C. A. Santos, Regiane S. Kawasaki Francês, Ricardo B. C. Prudêncio, and Ronnie C. O. Alves. Data vs classifiers, who wins? arXiv:2107.07451, 2021. https://arxiv.org/abs/2107.07451.

[22] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996. https://doi.org/10.1007/978-1-4612-0711-5.

[23] Thomas M. Cover and Peter E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1): 21–27, 1967. https://doi.org/10.1109/TIT.1967.1053964.

[24] Ana C. Lorena, Luís P. F. Garcia, Jens Lehmann, Marcilio C. P. Souto, and Tin Kam Ho. How complex is your classification problem? A survey on measuring classification complexity. *ACM Computing Surveys*, 52(5): 107:1–107:34, 2019. https://doi.org/10.1145/3347711.

[25] Michael R. Smith, Tony Martinez, and Christophe Giraud-Carrier. An instance level analysis of data complexity. *Machine Learning*, 95(2):225–256, 2014. https://doi.org/10.1007/s10994-013-5422-z.

[26] Ulisses Braga-Neto and Edward R. Dougherty. Exact performance of error estimators for discrete classifiers. *Pattern Recognition*, 38(11):1799–1814, 2005. https://doi.org/10.1016/j.patcog.2005.02.013.

[27] David J. Hand. Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–14, 2006. https://doi.org/10.1214/088342306000000060.

[28] C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. https://doi.org/10.1093/biomet/26.4.404.

[29] Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics, 1989*, pages 148–188. Cambridge University Press, 1989. https://doi.org/10.1017/CBO9781107359949.008.

[30] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014. https://doi.org/10.1145/2641190.2641198.

Supplementary Files

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

- supplementarymaterialJDSA.zip