(Appendix) Measuring the Validity of Clustering Validation Datasets

APPENDIX A
ACKERMAN & BEN-DAVID'S [1] WITHIN-DATASET AXIOMS

(Axiom W1) ***Scale Invariance***
A measure f satisfies scale invariance if for every clustering C of (X, d), and every positive α, $f(C, X, d) = f(C, X, \alpha d)$ (where αd is defined by setting, for every pair of domain points x, y, $\alpha d(x, y) = \alpha \cdot d(x, y)$)

C-consistent variant (Used in W2)
Given a clustering C over (X, d), a distance function d' is C-consistent variant of d, if $d'(x, y) \leq d(x, y)$ for all $x \sim_C y$, and $d'(x, y) \geq d(x, y)$ for all $x \nsim_C y$.

(Axiom W2) ***Consistency***
A measure f satisfies consistency if for every clustering C over (X, d), whenever d' is a C-consistent variant of d, then $f(C, X, d') \geq f(C, X, d)$

(Axiom W3) ***Richness***
A measure f satisfies richness if for each non-trivial clustering C of X, there exists a distance function d over X such that $X = \arg\max f(C_i, X, d)$.

Clustering Isomorphism (Used in W4)
Two clustering C and C' over the same domain, (X, d), are isomorphic, denoted $C \approx_d C'$, if there exists a distance-preserving isomorphism $\phi : X \to X$, such that for all $x, y \in X$, $x \sim_C y$ if and only if $\phi(x) \sim_{C'} \phi(y)$.

(Axiom W4) ***Isomorphism Invariance***
A measure f is isomorphism invariant if for all clusterings C, C' over (X, d) where $C \approx_d C'$, $f(C, X, d) = f(C', X, d)$

APPENDIX B
ALLEVIATING THE IMBALANCE OF HUMAN-DRIVEN SEPARABILITY SCORES

We use the human-driven separability score of Gaussian clusters [2], [3] to calibrate the logistic growth rate k of IVM_A (Section 4 T4-c) and compute the error in the ablation study (Section 6.1). We need to weigh each separability score to compensate for their non-uniform distribution. Here we provide a pseudo-code for computing these weights. Note that we assume that the scores range between 0 and 1 (Algorithm 1).

APPENDIX C
GENERALIZATION OF REMAINING IVMS

Here, we describe the adjustment process of the remaining IVMs, which we explain briefly in Section 5.2.

C.1 Adjusting the Dunn Index (DI)

Dunn Index [4] is defined as:

$$DI(C, X, d) = \frac{\min\limits_{1 \leq i \leq |C|} \min\limits_{1 \leq j \leq |C|} \min\limits_{x \in C_i, y \in C_j} d(x, y)}{\max\limits_{1 \leq k \leq |C|} \max\limits_{x, y \in C_k} d(x, y)}. \tag{1}$$

DI satisfies none of the across-dataset axioms; thus we applied every possible protocols we provided (Section 4).

Applying T1 (Data-cardinality invariance) Both the denominator (maximum intra-class distance) and numerator (minimum inter-class distance) of DI compute the minimum or maximum over point pairwise distances, which is very sensitive

Algorithm 1: Assigning weights to each dataset to compensate for the imbalance of human-driven separability scores

Input: list of scores $S = [s_1, s_2, \cdots, s_n]$ with length n
Input: number of bins b
Output: list of weights $W = [w_1, w_2, \cdots, w_n]$
$bins = []$, $w\_bins = []$; /* initialize empty list */
for $i = 1$ **to** b **do**
  $bins$.**append**(0) ; /* create bins */
end
$bin\_size = 1/b$;
for $k = 1$ **to** n **do**
  $bin\_idx = \lceil s_k / bin\_size \rceil$; /* quantize scores */
  $bins[bin\_idx] += 1$; /* count frequencies */
end
for $i = 1$ **to** b **do**
  $w\_bins[i] = 1/bins[i]$; /* compute bin weights */
end
$W = []$;
for $k = 1$ **to** n **do**
  W.**append**$(w\_bins[\lceil s_k / bin\_size \rceil])$
end
return W

to noise. We thus first change the denominator and numerator into the maximum average intra-class distance and minimum average inter-class distance, respectively, leading to:

$$DI_1(C, X, d) = \frac{\min\limits_{1\le i\le |C|} \min\limits_{1\le j\le |C|} \frac{\sum_{x\in C_i, y\in C_j} d(x,y)}{|C_i|\cdot|C_j|}}{\max\limits_{1\le k\le |C|} \frac{\sum_{x,y\in C_k} d(x,y)}{(k\cdot(k-1))/2}}. \tag{2}$$

Applying T2 (Shift invariance) As both denominator and numerator of DI_1 consist of type-1 distances, we can directly apply the exponential protocol, hence:

$$DI_2(C, X, d) = \frac{\exp\left(\frac{1}{\sigma_d} \min\limits_{1\le i\le |C|} \min\limits_{1\le j\le |C|} \frac{\sum\limits_{x\in C_i, y\in C_j} d(x,y)}{|C_i|\cdot|C_j|}\right)}{\exp\left(\frac{1}{\sigma_d} \max\limits_{1\le k\le |C|} \frac{\sum\limits_{x,y\in C_k} d(x,y)}{(k\cdot(k-1))/2}\right)}, \tag{3}$$

where $\sigma_d = \text{std}(\{d(x,c) | x \in X\})$.

Applying T4 (Range invariance) In the case of maximum score, as $\max(DI_2) \to +\infty$, we apply the logistic function (T4-b), resulting in $DI_3 = 1/(1 + e^{-k\cdot DI_2})$, to achieve $DI_{3\max} \to 1$. As we did for CH, we estimate the worst score as the expectation of DI_3 over random clustering partitions C^π (T4-a): $DI_{3\min} = E_\pi(DI_3(C^\pi, X, d))$. We then get $DI_4 = (DI_3 - DI_{3\min})/(DI_{3\max} - DI_{3\min})$. We set the growth rate k by calibrating DI_4 using human-judgement separability scores (T4-c).

Applying T3 (class-cardinality invariance) As in CH_A (Section 5), we make DI_A class-cardinality invariant by averaging class-pairwise scores:

$$DI_A = \frac{1}{\binom{|C|}{2}} \sum_{S\subseteq C, |S|=2} DI_4(C, X, d). \tag{4}$$

Approaching DI_3's minimum bound We found that $DI_{3\min}$ can get closer to the minimum bound of DI_3 by defining $DI_{3\min} = DI_3(C^\psi, X, d)$; $C^\psi = \{C_1^\psi, C_2^\psi\}$ satisfying $|C_1^\psi| = \{x\}$ and $|C_2^\psi| = X \setminus \{x\}$ where x is the geometric median of X (i.e., $x = \arg\min_{x'\in X} \sum_{y\in X} d(x', y)$). $DI_3(C^\psi, X, d)$ is always smaller than $DI_3(C^\pi, X, d)$, thus always closer to the worst case (Proof in Appendix D).

Computational Complexity The computational complexity of DI, DI_1, DI_2, DI_3 is $O(|X|^2\Delta_X)$. If we use the geometric median-based worst-case approximation, the performance of $DI_{3\,\min}$ also depends on the time complexity of the algorithm that computes the geometric median. We use the one proposed by Vardi and Zhang [5]; the complexity is $O(|X|\Delta_X\tau)$ where τ is the number of iterations needed to reach convergence. Thus, $DI_A = O(|X|^2\Delta_X(|C|+\tau))$.

C.2 Adjusting the I-Index (II)

I-Index [6] is defined as

$$II(C,X,d) = \left(\frac{1}{|C|} \cdot \frac{\sum_{x\in X} d(x,c)}{\sum_{i=1}^{|C|}\sum_{x\in C_i} d(x,c_i)} \cdot \max_{1\le i,j\le |C|} d(c_i,c_j) \right)^p, \tag{5}$$

where p is a constant that controls the discrimination between good and bad-quality clusterings, where we set $p=1$ for the rest of the adjustment. II satisfies data-cardinality invariance but does not satisfy remaining invariance axioms. The adjustment procedure is as follows:

Applying T2 (Shift invariance) As both the denominator and the numerator of the main fractional term are type-2 distances, we can easily make the term shift-invariant by (1) changing the distance function to the square of the Euclidean distance (T2-c), then by (2) converting the sum of distances into the average (T2-b), then applying the exponential function normalized by σ_{d^2} (T2-a). On the other hand, we do not need an exponential protocol for the third term (max) as the term only consists of type-3 distances; we only divide it by σ_{d^2} to ensure scale invariance. The resulting formula is as follows:

$$II_1(C,X,d^2) = \frac{e^{\left(\sum_{x\in X} \frac{d^2(x,c)}{|X|\sigma_{d^2}}\right)}}{|C|e^{\left(\sum_{i=1}^{|C|}\sum_{x\in C_i} \frac{d^2(x,c_i)}{|X|\sigma_{d^2}}\right)}} \max_{1\le i,j\le |C|} \frac{d^2(c_i,c_j)}{\sigma_{d^2}}. \tag{6}$$

Applying T4 (Range invariance) As $\max(II_1)\to+\infty$, we define $II_2 = 1/(1+e^{-k\cdot II_1})$ so that $II_{2\,\max}\to 1$ (T4-b). As for CH and DI, we get $II_3 = (II_2 - II_{2\,\min})/(II_{2\,\max} - II_{2\,\min})$, where $II_{2\,\min} = E_\pi(II_2(C^\pi,X,d^2))$ (T4-a). The k value are set by calibrating II_3 using human-judgment separability scores (T4-c).

Applying T3 (Class-cardinality invariance) We design II_A to be class-cardinality invariant by averaging class-pairwise scores:

$$II_A = \frac{1}{\binom{|C|}{2}} \sum_{S\subseteq C,|S|=2} II_3(C,X,d). \tag{7}$$

Removing Monte-Carlo simulations As we did for CH_A, we can accelerate the computation of II_A by removing Monte-Carlo simulations for estimating $II_{2\,\min}$. As randomly permuting class labels makes every $C_i\in C$ satisfy $C_i \overset{D}{=} X, c\simeq c_i \forall c_i$. Therefore $d^2(c_i,c_j)\simeq 0$, which leads to $II_1(C^\pi,X,d^2)\simeq 0$ and $II_{2\,\min} = E_\pi(II_2(C^\pi,X,d)) = E_\pi(1/2) = 1/2$.

Computational Complexity II, II_1, II_2 is $O(|X|\Delta_X)$, thus $II_{2\,\min} = O(|X|\Delta_X T)$, where T is the number of Monte Carlo simulations to compute $II_{2\,\min}$. II_3 is thus $O(|X|\Delta_X T)$ and $II_A = O(|X|\Delta_X T|C|)$. Removing the Monte-Carlo simulations reduces the complexity of II_A to $O(|X|\Delta_X|C|)$.

C.3 Adjusting Xie-Beni index (XB)

Xie-Beni Index [7] is defined as:

$$XB(C,X,d^2) = \frac{\sum_{i=1}^{|C|}\sum_{x\in C_i} d^2(x,c_i)}{|X|\cdot \min_{1\le i,j\le |C|, i\ne j} d^2(c_i,c_j)}. \tag{8}$$

XB misses all axioms except data-cardinality invariance. For the adjustment, we first take the inverse formula so that the higher score implies better clustering, resulting in:

$$XB_1(C,X,d^2) = \frac{|X|\cdot \min_{1\le i,j\le |C|, i\ne j} d^2(c_i,c_j)}{\sum_{i=1}^{|C|}\sum_{x\in C_i} d^2(x,c_i)} \tag{9}$$

Applying T2 (Shift invariance) As the denominator consists of type-2 distance while the numerator consists of type-3 distance, we add a factor term with type-2 distances to the numerator to equalize shifting. As we did for CH, we add the

sum of squared distances of the data points to their centroid. We also remove the $|X|$ term from the numerator to ensure the data-cardinality invariance. This leads to the following:

$$XB_2(C,X,d^2) = \frac{\sum_{x\in X} d^2(x,c)}{\frac{|C|}{\sum_{i=1}\sum_{x\in C_i} d^2(x,c_i)}} \cdot \min_{1\le i,j\le |C|, i\neq j} d^2(c_i,c_j). \tag{10}$$

We then apply the exponential protocol (T2-a, b) to the first term, while maintaining the second term as type-3 distances do not shift (we only normalized it by σ_{d^2} to ensure scale-invariance).

$$XB_3(C,X,d^2) = \frac{\exp\left(\sum_{x\in X} \frac{d^2(x,c)}{|X|\sigma_{d^2}}\right)}{\exp\left(\sum_{i=1}^{|C|}\sum_{x\in C_i} \frac{d^2(x,c_i)}{|X|\sigma_{d^2}}\right)} \min_{\substack{1\le i,j\le |C| \\ i\neq j}} \frac{d^2(c_i,c_j)}{\sigma_{d^2}}. \tag{11}$$

Here, when we limit $|C|=2$, $XB_3(C,X,d^2) = 2\cdot II_1(C,X,d^2)$, as $\min_{1\le i,j\le 2, i\neq j} d^2(c_i,c_j) = \max_{1\le i,j\le 2} d^2(c_i,c_j) = d^2(c_1,c_2)$. Therefore, by applying to XB_3 the adjustment procedure of II_1, we obtain XB_A which is identical to II_A:

$$\forall C,X, \quad XB_A(C,X,d^2) = II_A(C,X,d^2)$$

C.4 Adjusting Davies-Bouldin Index (DB)

Davies-Bouldin index [8] is defined as:

$$DB(C,X,d) = \frac{1}{|C|}\sum_{i=1}^{|C|} \max_{\substack{1\le j\le |C| \\ i\neq j}} \frac{\sum_{k\in\{i,j\}} \frac{1}{|C_k|}\sum_{x\in C_k} d(x,c_k)}{d(c_i,c_j)}. \tag{12}$$

DB satisfies data-cardinality invariance but misses all other across-dataset axioms. We first inverse the formula, making higher scores imply better clustering (i.e., $DB_1(C,X,d) = DB(C,X,d)^{-1}$).

Applying T2 (Shift invariance) As the denominator use type-3 distances while the numerator consists of type-2 distances, we add an additional term consisting of type-2 distances to the denominator and change the distance function to d^2 to equalize shifting. We add the average of squared distances of data points to their centroid, resulting in:

$$DB_2(C,X,d^2) = \left(\frac{1}{|C|}\sum_{i=1}^{|C|} \max_{\substack{1\le j\le |C| \\ i\neq j}} \frac{\sum_{k\in\{i,j\}} \frac{1}{|C_k|}\sum_{x\in C_k} d^2(x,c_k)}{\frac{d^2(c_i,c_j)}{|X|}\sum_{x\in X} d^2(x,c)} \right)^{-1} \tag{13}$$

We then apply the exponential protocol (T2-a, b) to the first fractional term, while only normalizing the second term with σ_{d^2} to ensure scale invariance:

$$DB_3(C,X,d^2) = \left(\frac{1}{|C|}\sum_{i=1}^{|C|} \max_{\substack{1\le j\le |C| \\ i\neq j}} \frac{\sum_{k\in\{i,j\}} e^{\frac{1}{|C_k|}\sum_{x\in C_k}\frac{d^2(x,c_k)}{\sigma_{d^2}}}}{\frac{d^2(c_i,c_j)}{\sigma_{d^2}} e^{\frac{1}{|X|}\sum_{x\in X}\frac{d^2(x,c)}{\sigma_{d^2}}}} \right)^{-1} \tag{14}$$

Applying T4 (Range invariance) We apply min-max scaling to make DB_A range invariant. As $\max(DB_3)\to+\infty$, we transform the function by applying the logistic function (T4-b), resulting in $DB_4 = 1/(1+e^{-k\cdot DB_3})$, so that $DB_{4\max}\to 1$. We then get $DB_5 = (DB_4 - DB_{4\min})/(DB_{4\max} - DB_{4\min})$, where $DB_{4\min} = E^\pi(DB_4(C^\pi,X,d^2)$. k value is calibrated based on human-judgment separability scores (T4-c).

Removing Monte-Carlo simulations We remove the Monte-Carlo simulation estimating $DB_{4\min}$ to further accelerate the computation. As randomly permuting class labels makes $C_i \overset{D}{=} X\ \forall C_i\in C$, thus $c_i\simeq c_j$ and $d^2(c_i,c_j)\simeq 0\ \forall c_i,c_j$, $DB_3(C^\pi,X,d^2)\simeq 0$, which leads to $DB_{4\min} = E^\pi(DB_4(C^\pi,X,d^2) = E^\pi(1/2) = 1/2$.

Computational Complexity The computational complexity of DB, DB_1, DB_2, DB_3, and DB_4 is $O(|X|\Delta_X)$. Therefore $DB_{4\min} = O(|X|\Delta_X T)$ and $DB_A = O(|X|\Delta_X T|C|)$ where T is the number of Monte-Carlo simulations to compute $DB_{4\min}$. By removing the simulations, we can reduce the complexity of DB_A to $O(|X|\Delta_X|C|)$.

C.5 Adjusting the Silhouette Coefficient (SC)

Silhouette Coefficient [9] is defined as:

$$SC(C, X, d) = \frac{1}{|C|} \cdot \sum_{i=1}^{|C|} \frac{1}{|C_i|} \sum_{x \in C_i} \frac{b(x) - a(x)}{\max(b(x), a(x))}. \tag{15}$$

For a data point $x \in C_I$,

$$a(x) = \frac{1}{|C_I| - 1} \sum_{y \in C_I, x \neq y} d(x, y), \tag{16}$$

and

$$b(x) = \min_{1 \leq J \leq |C|, I \neq J} \frac{1}{|C_J|} \sum_{y \in C_J} d(x, y). \tag{17}$$

As SC satisfies data-cardinality invariance (A1) and range invariance (A4) (Appendix D), we only need to apply the shift (T2) and class-cardinality (T4) protocols.

Applying T2 (Shift invariance) As $a(x)$ and $b(x)$ are already robust estimators of population statistics (T1) and also consist of type-1 distances, we simply apply the exponential protocol (T2-a,b) to the terms, resulting in:

$$SC_1(C, X, d) = \frac{1}{|C|} \cdot \sum_{i=1}^{|C|} \frac{1}{|C_i|} \sum_{x \in C_i} \frac{b'(x) - a'(x)}{\max(b'(x), a'(x))}, \tag{18}$$

where

$$a'(x) = e^{a(x)/\sigma^d} \text{ and } b'(x) = e^{b(x)/\sigma^d}. \tag{19}$$

Applying T3 (Class-cardinality invariance) We make SC_A satisfy class-cardinality invariance by averaging class-pairwise scores:

$$SC_A(C, X, d) = \frac{1}{\binom{|C|}{2}} \sum_{S \subseteq C, |S|=2} SC_1(C, X, d). \tag{20}$$

While SC misses two across-dataset axioms, SC_A satisfies them all.

Computational Complexity The complexity of SC and SC_1 is $O(|X|^2 \Delta_X)$, thus $SC_A = O(|X|^2 \Delta_X |C|)$.

APPENDIX D
THEOREMS AND PROOFS

D.1 About the Shift of Distances

Here, we prove the theorems about the shift of type-1 and type-2 distances, which form the bases of the equalizing shifting step of the shift protocol (Section 4 T2-c).

Theorem 1. *$\forall X' \subset X$, $\forall \beta > 0$, and for any Euclidean distance functions d_L and d_H satisfying $d_H^2 = d_L^2 + \beta$, we have $\sum_{x \in X'} d_H^2(x, c) = \sum_{x \in X'} d_L^2(x, c) + \beta/2$, where $c = \overline{X'}$.*

Proof. According to Hopcroft & Kannon [10], for any set of data points $X' \subset X$, the following equality holds:

$$2n \cdot \sum_{x \in X'} d^2(x, c) = \sum_{x \in X'} \sum_{y \in X'} d^2(x, y). \tag{21}$$

Here,

$$\begin{aligned} 2n \cdot \sum_{x \in X'} d_H^2(x, c) &= \sum_{x \in X'} \sum_{y \in X'} d_H^2(x, y). \\ &= \sum_{x \in X'} \sum_{y \in X'} (d_L^2 + \beta)(x, y) \\ &= n^2 \beta + \sum_{x \in X'} \sum_{y \in X'} d_L^2(x, y) \\ &= n^2 \beta + 2n \cdot \sum_{x \in X'} d_L^2(x, c) \\ &= 2n \cdot \left(\frac{n\beta}{2} + \sum_{x \in X'} d_L^2(x, c) \right) \\ &= 2n \cdot \sum_{x \in X'} \left(d_L^2(x, c) + \frac{\beta}{2} \right) \end{aligned} \tag{22}$$

Therefore, $\sum_{x\in X'} d_H^2(x,c) = \sum_{x\in X'} d_L^2(x,c) + \beta/2$. □

Theorem 2. *$\forall X', X'' \subset X$, $\forall \beta > 0$, and for any d_L and d_H satisfying $d_H^2 = d_L^2 + \beta$, we have $d_H^2(c', c'') = d_L^2(c', c'')$, where $c' = \overline{X'}$ and $c'' = \overline{X''}$.*

Proof. According to Apostol & Mnatsakanian [11], the following equality holds for any set of points X', X'':

$$d^2(c_1,c_2) = \frac{1}{n_{X'}n_{X''}} \sum_{x\in X'} \sum_{y\in X''} d^2(x,y) - \frac{1}{2n_{X'}^2} \sum_{x\in X'} \sum_{y\in X'} d^2(x,y) - \frac{1}{2n_{X''}^2} \sum_{x\in X''} \sum_{y\in X''} d^2(x,y). \tag{23}$$

Here, using Theorem 1,

$$\begin{aligned}
d_H^2(c_1,c_2) &= \frac{1}{n_{X'}n_{X''}} \sum_{x\in X'} \sum_{y\in X''} d_H^2(x,y) - \frac{1}{2n_{X'}^2} \sum_{x\in X'} \sum_{y\in X'} d_H^2(x,y) - \frac{1}{2n_{X''}^2} \sum_{x\in X''} \sum_{y\in X''} d_H^2(x,y). \\
&= \frac{1}{n_{X'}n_{X''}} \sum_{x\in X'} \sum_{y\in X''} (d_L^2+\beta)(x,y) - \frac{1}{2n_{X'}^2} \sum_{x\in X'} \sum_{y\in X'} (d_L^2+\beta)(x,y) - \frac{1}{2n_{X''}^2} \sum_{x\in X''} \sum_{y\in X''} (d_L^2+\beta)(x,y). \\
&= \frac{1}{n_{X'}n_{X''}} \sum_{x\in X'} \sum_{y\in X''} d_L^2(x,y) - \frac{1}{2n_{X'}^2} \sum_{x\in X'} \sum_{y\in X'} d_L^2(x,y) + \beta - \beta/2 - \beta/2 - \frac{1}{2n_{X''}^2} \sum_{x\in X''} \sum_{y\in X''} d_L^2(x,y) \\
&= \frac{1}{n_{X'}n_{X''}} \sum_{x\in X'} \sum_{y\in X''} d_L^2(x,y) - \frac{1}{2n_{X'}^2} \sum_{x\in X'} \sum_{y\in X'} d_L^2(x,y) - \frac{1}{2n_{X''}^2} \sum_{x\in X''} \sum_{y\in X''} d_L^2(x,y) \\
&= d_L^2(c_1,c_2)
\end{aligned} \tag{24}$$

D.2 About Class-Cardinality Invariance

Ackerman and Ben-David [1] showed that if a function f' satisfies any of the within-dataset axioms, then its aggregation $f(C,X,\delta) = \text{agg}_{S\subseteq C,|S|=2} f'(S,X,\delta)$ with agg $\in \{\min, \text{avg}, \max\}$, also satisfies that axiom, which is represented by the following theorem from [1]:

Theorem 3. *If f' satisfies a within-dataset axiom, then $\forall agg \in \{\min, avg, \max\}$, $f(C,X,\delta) = agg_{S\subseteq C,|S|=2} f'(S,X,\delta)$ also satisfies the axiom.* [1]

Here, we use Theorem 3 to prove the same property for across-dataset axioms.

Theorem 4. *If f' satisfies data-cardinality invariance (A1), then $f(C,X,\delta) = agg_{S\subseteq C,|S|=2} f'(S,X,\delta)$ is also data-cardinality invariant.*

Proof. At first, $f(C,X,\delta) = \text{agg}_{S\subseteq C,|S|=2} f'(S,X,\delta)$ and $f(\underline{C}_\alpha, X_\alpha, \delta) = \text{agg}_{S\subseteq \underline{C}_\alpha,|S|=2} f'(S,X_\alpha,\delta)$ by definition (A3). Then, as $\text{agg}_{S\subseteq C,|S|=2} f'(S,X,\delta) = \text{agg}_{S\subseteq \underline{C}_\alpha,|S|=2} f'(S,X_\alpha,\delta)$ (A1), $f(C,X,\delta) = f'(\underline{C}_\alpha, X_\alpha, \delta)$. Thus, f satisfies data-cardinality invariance. □

Theorem 5. *If f' satisfies shift invariance, then $f(C,X,\delta) = agg_{S\subseteq C,|S|=2} f'(S,X,\delta)$ is also shift-invariant (A2).*

Proof. By definition (A3), $f(C,X,\delta) = \text{agg}_{S\subseteq C,|S|=2} f'(S,X,\delta)$ and $f(C,X,\delta+\beta) = \text{agg}_{S\subseteq C,|S|=2} f'(S,X,\delta+\beta)$. Then, as $\text{agg}_{S\subseteq C,|S|=2} f'(S,X,\delta) = \text{agg}_{S\subseteq C,|S|=2} f'(S,X,\delta+\beta)$ (A1), $f(C,X,\delta) = f(C,X,\delta+\beta)$. Thus, f satisfies shift invariance. □

Theorem 6. *If f' is range invariant over any pair of classes, then $f(C,X,\delta) = agg_{S\subseteq C,|S|=2} f'(S,X,\delta)$ is also Range Invariant (A4).*

Proof. Assume f' satisfies Range Invariance; i.e.,

$$\forall S \subseteq C, |S| = 2, f'_{\min}(S,X,\delta) = 0 \text{ and } f'_{\max}(S,X,\delta) = 1. \tag{25}$$

From the assumption, we get the following:

(1) $\min_S(f'_{\min}) \le \min_S(f') \le \min_S(f'_{\max})$ so $\min_S(f') \in [0,1]$

(2) $\max_S(f'_{\min}) \le \max_S(f') \le \max_S(f'_{\max})$ so $\max_S(f') \in [0,1]$

(3) $\text{avg}_S(f'_{\min}) \le \text{avg}_S(f') \le \text{avg}_S(f'_{\max})$ so $\text{avg}_S(f') \in [0,1]$

From (1), (2), and (3), we get $\forall C, \forall \text{agg} \in \{\min, \text{avg}, \max\}$,
$\text{agg}_{S\subseteq C,|S|=2}(f'(S,X,d)) \in [0,1]$ and so $f_{\min} = 0$ and $f_{\max} = 1$. Thus, f satisfies Range Invariance if f' does so. □

D.3 About Baseline IVMs and Across-Dataset Axioms

Our main claim is that standard IVM cannot properly evaluate and compare CLM across datasets as they do not satisfy the across-dataset axioms. We provide the proof that verifies the claim.

D.3.1 About Calinski-Harabasz Index (CH)

CH [12] is defined as:

$$CH(C, X, d^2) = \frac{|X| - |C|}{|C| - 1} \cdot \frac{\sum_{i=1}^{|C|} |C_i| d^2(c_i, c)}{\sum_{i=1}^{|C|} \sum_{x \in C_i} d^2(x, c_i)}, \tag{26}$$

Theorem 7. *CH does not satisfy data-cardinality invariance (W1).*

Proof. In the definition of CH, terms $d^2(c_i, c)/(|C| - 1)$ and $\sum_{i=1}^{|C|} \sum_{x \in C_i} d^2(x, c_i)/(|X| - |C|)$ are robust estimators of population statistics (average of distances; $|C|$ is the number of classes), thus robust to changes in data cardinality. However, the $|C_i|$ term in the numerator (amount of data in each class) makes CH proportional to the number of points, thus making the function dependent on the data cardinality. □

Theorem 8. *CH does not satisfy shift invariance (W2).*

Proof. For any clustering C over (X, d),

$$CH(C, X, d^2 + \beta) = \frac{|X| - |C|}{|C| - 1} \cdot \frac{|X|\beta + \sum_{i=1}^{|C|} |C_i| d^2(c_i, c)}{|X|\beta + \sum_{i=1}^{|C|} \sum_{x \in C_i} d^2(x, c_i)} \neq CH(C, X, d) \tag{27}$$

CH does not satisfy Shift Invariance. □

Theorem 9. *CH does not satisfy class-cardinality invariance (W3).*

Proof. CH does not aggregate scores over pairs of clusters. Its formula cannot be algebraically transformed to an aggregation $(\min, \max, \text{avg})$ of cluster-pairwise formula. Therefore, CH does not satisfy class-cardinality invariance by design. □

Theorem 10. *CH does not satisfy Range Invariance (W4).*

Proof. The range of CH is $[0, +\infty[$. Therefore, CH is not range invariant as its maximum value is not bounded. □

D.3.2 About Dunn Index (DI)

Dunn Index is defined as

$$DI(C, X, d) = \frac{\min\limits_{C_i \in C} \min\limits_{C_j \in C, C_i \neq C_j} \min\limits_{x \in C_i, y \in C_j} d(x, y)}{\max\limits_{C_k \in C} \max\limits_{x, y \in C_k, x \neq y} d(x, y)}. \tag{28}$$

Theorem 11. *DI does not satisfy data-cardinality invariance (A1).*

Proof. DI is a fraction of minimum inter-cluster distances and maximum intra-cluster distances. The minimum and maximum functions are not robust to subsampling, there value can change drastically if the current min or max is not part of the subsample. Thus, DI is not data-cardinality invariant. □

Theorem 12. *DI does not satisfy Shift Invariance (A2).*

Proof.

$$DI(C, X, d + \beta) = \frac{\beta + \min\limits_{C_i \in C} \min\limits_{\substack{C_j \in C \\ i \neq j}} \min\limits_{\substack{x \in C_i \\ y \in C_j}} d(x, y)}{\beta + \max\limits_{C_k \in C} \max\limits_{\substack{x, y \in C_k \\ x \neq y}} d(x, y)} \neq DI(C, X, d). \tag{29}$$

□.

Theorem 13. *DI does not satisfy class-cardinality invariance (A3).*

Proof. DI does not aggregate scores over pairs of clusters. Its formula cannot be algebraically transformed to an aggregation $(\min, \max, \text{avg})$ of cluster-pairwise formula. Therefore, DI does not satisfy class-cardinality invariance by design. □.

Theorem 14. *DI does not satisfy Range Invariance (A4).*

Proof. The range of DI is $[0, +\infty[$. Therefore, DI is not range invariant as its maximum value is not bounded. □.

D.3.3 About I Index (II)

I-Index is defined as

$$II(C, X, d) = \left(\frac{1}{|C|} \cdot \frac{\sum\limits_{x \in X} d(x, c)}{\sum\limits_{i=1}^{|C|} \sum\limits_{x \in C_i} d(x, c_i)} \cdot \max_{1 \leq i, j \leq |C|} d(c_i, c_j) \right)^p, \tag{30}$$

Theorem 15 *II satisfies data-cardinality invariance (A1).*

Proof. For any C over (X, d), II uses averages of distances towards the average (center of gravity) of each class (c_i) and of the whole set of points (c), and the maximum distance between such centers. Average is a robust estimator of population statistics. Thus, II is data-cardinality invariant. □

Theorem 16. *II does not satisfy Shift Invariance (A2).*

Proof. For any C over (X, d),

$$II(C, X, d + \beta) = \left(\frac{1}{|C|} \cdot \frac{|X|\beta + \sum_{x \in X} d(x, c)}{|X|\beta + \sum_{i=1}^{|C_i|} \sum_{x \in C_i} d(x, c_i)} \times \left(\beta + \max_{1 \leq i,j \leq |C|} d(c_i, c_j) \right) \right)^p \neq II(C, X, d). \tag{31}$$

□.

Theorem 17. *II does not satisfy class-cardinality invariance (A3).*

Proof. II does not aggregate scores over pairs of clusters. Its formula cannot be algebraically transformed to an aggregation $(\min, \max, \text{avg})$ of cluster-pairwise formula. Therefore, II does not satisfy class-cardinality invariance by design □.

Theorem 18. *II does not satisfy Range Invariance (A4).*

Proof. The range of II is $[0, +\infty[$. Therefore, II is not range invariant as its maximum value is not bounded. □.

D.3.4 About Xie-Beni Index (XB)

Xie-Beni Index is defined as:

$$XB(C, X, d^2) = \frac{\sum_{i=1}^{|C|} \sum_{x \in C_i} d^2(x, c_i)}{|X| \min_{1 \leq i,j \leq |C|, i \neq j} d^2(c_i, c_j)}. \tag{32}$$

Theorem 19. *XB satisfies data-cardinality invariance (A1).*

Proof. XB uses averages of distances towards the average (center of gravity) of each class (c_i) and the minimum distance between such centers. Averages are robust estimators of population statistics. Thus, XB is data-cardinality invariant. □

Theorem 20. *XB does not satisfy shift invariance (A2).*

Proof. For any C over (X, d),

$$XB(C, X, d^2 + \beta) = \frac{|X|\beta + \sum_{i=1}^{|C|} \sum_{x \in C_i} d^2(x, c_i)}{|X| \left(\beta + \min_{1 \leq i,j \leq |C|, i \neq j} d^2(c_i, c_j) \right)} \neq XB(C, X, d^2). \tag{33}$$

Thus, XB does not satisfy Shift Invariance. □

Theorem 21. *XB does not satisfy class-cardinality invariance (A3).*

Proof. XB does not aggregate scores over pairs of clusters. Its formula cannot be algebraically transformed to an aggregation $(\min, \max, \text{avg})$ of cluster-pairwise formula. Therefore, XB does not satisfy class-cardinality invariance by design. □.

Theorem 22. *XB does not satisfy Range Invariance (A4).*

Proof. The range of XB is $[0, +\infty[$. Therefore, XB is not range invariant as its maximum value is not bounded. □.

D.3.5 About Davies-Bouldin Index (DB)

Davies-Bouldin index is defined as:

$$DB(C, X, d) = \frac{1}{|C|} \sum_{i=1}^{|C|} \max_{1 \leq j \leq |C|, i \neq j} \frac{\sum_{k \in \{i,j\}} \frac{1}{|C_k|} \sum_{x \in C_k} d(x, c_k)}{d(c_i, c_j)}. \tag{34}$$

Theorem 23. *DB satisfies data-cardinality invariance (A1).*

Proof. DB uses averages of distances towards the average (center of gravity) of each class (c_i) and the minimum distance between such centers. Averages are robust estimators of population statistics. Thus, DB satisfies data-cardinality invariance. □

Theorem 24. *DB does not satisfy shift invariance (A2).* For any C over (X, d),

$$DB(C, X, d + \beta) = \frac{1}{|C|} \sum_{i=1}^{|C|} \max_{\substack{j \in \{1, \cdots, |C|\} \\ i \neq j}} \frac{\frac{1}{|C_i|} \sum_{x \in C_i} d(x, c_i) + \frac{1}{|C_j|} \sum_{x \in C_j} d(x, c_j) + 2\beta}{d(c_i, c_j) + \beta} \neq DB(C, X, d). \tag{35}$$

Thus, DB does not satisfy shift invariance.

Theorem 25. *DB does not satisfy class-cardinality invariance (A3).*

Proof. DB does not aggregate scores over pairs of clusters. Its formula cannot be algebraically transformed to an aggregation ($\min, \max,$ avg) of cluster-pairwise formula. Therefore, DB does not satisfy class-cardinality invariance by design. □

Theorem 26. *DB does not satisfy Range Invariance (A4).*

Proof. The range of DB is $[0, +\infty[$. Therefore, DB is not range invariant as its maximum value is not bounded. □.

D.3.6 About Silhouette Coefficient (SC)

Silhouette Coefficient is defined as:

$$SC(C, X, d) = \frac{1}{|C|} \cdot \sum_{i=1}^{|C|} \frac{1}{|C_i|} \sum_{x \in C_i} \frac{b(x) - a(x)}{\max(b(x), a(x))}. \tag{36}$$

For a data point $x \in C_I$,

$$a(x) = \frac{1}{|C_I| - 1} \sum_{y \in C_I, x \neq y} d(x, y), \tag{37}$$

and

$$b(x) = \min_{1 \leq J \leq |C|, I \neq J} \frac{1}{|C_J|} \sum_{y \in C_J} d(x, y). \tag{38}$$

Theorem 27. *SC satisfies data-cardinality invariance (A1).*

Proof. $a(x)$ and $b(x)$, two building blocks of SC, are both robust estimators of population statistics (average distances), and thus are data-cardinality invariance. Therefore, SC is data-cardinality invariance.

Theorem 28. *SC does not satisfy shift invariance (A2).*

Proof. For any C over (X, d),

$$SC(C, X, d + \beta) = \frac{1}{|C|} \cdot \sum_{i=1}^{|C|} \frac{1}{|C_i|} \sum_{x \in C_i} \frac{b(x) - a(x)}{\beta + \max(b(x), a(x))} \neq SC(C, X, d). \tag{39}$$

Thus, SC is not shift invariant. □

Theorem 29. *SC does not satisfy class-cardinality Invariance (A3).*

Proof. SC does not aggregate scores over pairs of clusters. Its formula cannot be algebraically transformed to an aggregation ($\min, \max,$ avg) of cluster-pairwise formula. Therefore, SC does not satisfy class-cardinality invariance by design. □

Theorem 30. *SC satisfies Range Invariance (A4).*

Proof. The theoretical range of SC is $[-1, 1]$, and the expected minimum of the measure (based on randomly permuted class labels) is 0. Therefore, SC is range invariant.

D.4 About CH_A and Across-Dataset Axioms

Theorem 31. CH_5 satisfies data-cardinality invariance (A1).

Proof. CH_A uses averages of distances towards the center of gravity of each class (c_i) or the entire dataset (c), or the average of distances between the centers. Averages are robust estimators of population statistics. Thus, CH_5 satisfies data-cardinality invariance. Due to Theorem 4, CH_A is also data-cardinality invariant as it is the class-pairwise aggregation of CH_5. □

Theorem 32. CH_A satisfies shift invariance (A2).

Proof. For any C over (X, d),

$$\begin{aligned} CH_3(C, X, d^2 + \beta) &= \frac{e^{\frac{\sum_{x \in X} (d^2(x,c) + \beta)}{\sigma_{d^2} \cdot |X|}}}{e^{\frac{\sum_{i=1}^{|C|} \sum_{x \in C_i} (d^2(x,c_i) + \beta)}{\sigma_{d^2} \cdot |X|}}} \cdot \frac{\sum_{i=1}^{|C|} |C_i| (d^2(c_i, c) + \beta)}{\sigma_{d^2} \cdot |X| \cdot (|C| - 1)} \\ &= \frac{e^{\frac{\sum_{x \in X} d^2(x,c)}{\sigma_{d^2} \cdot |X|}} e^{\frac{\beta}{\sigma_{d^2}}}}{e^{\frac{\sum_{i=1}^{|C|} \sum_{x \in C_i} d^2(x,c_i)}{\sigma_{d^2} \cdot |X|}} e^{\frac{\beta}{\sigma_{d^2}}}} \cdot \frac{\sum_{i=1}^{|C|} |C_i| d^2(c_i, c)}{\sigma_{d^2} \cdot |X| \cdot (|C| - 1)} \\ &= CH_3(C, X, d^2). \end{aligned} \tag{40}$$

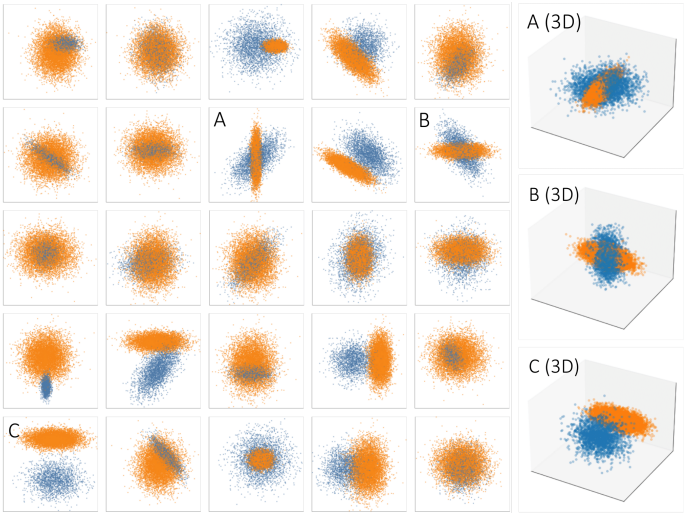

Fig. 1. Subset of 1,000 base datasets used in the ablation study (Section 6.1) and logistic growth rate calibration (Section 4 T4-c); we displayed the first two (left) and three (right) axes of each dataset. Note that only the first two axes are used to calibrate the logistic growth rate of IVM_A.

Therefore, CH_3 is shift invariant, thus CH_4 and CH_5 are also shift invariant. Due to Theorem 5, CH_A is also shift invariant as it is the class-pairwise aggregation of CH_5. □

Theorem 33. CH_A satisfies class-cardinality invariance (A3).

Proof. CH_A is defined as the class-pairwise aggregation of CH_5, thus is class-cardinality invariant □.

Theorem 34. CH_A satisfies range invariance (A4)

Proof. The range of CH_5 is $[0, 1]$. thus CH_5 is range invariant. Therefore, due to Theorem 6, CH_A is also range invariant as it is the class-pairwise aggregation of CH_5. □

APPENDIX E
BASE GAUSSIAN DATASETS

For the ablation study (Section 6.1) and logistic growth rate calibration (Section 4 T4-c), we used datasets where each consisting of two Gaussian clusters with diverse hyperparameters. Figure 1 shows a subset of 28 scatterplot representations of these datasets among 1000 available.

APPENDIX F
EXPERIMENTAL SETTING

Classifiers settings. As aforementioned in Section 5.2., we exploited Bayesian Optimization [13] to find the best hyper-parameter setting of the classifiers for each fold of cross validation. We used the implementation of Nogueira [14]. For each classifiers, we performed five random exploration and 25 steps of optimization, following the default setting of the implementation. For the bounds of the hyperparameters, we exploited the settings provided by `auto-sklearn` [15]. For the hyperparameters that are not mentioned below, we used the default provided by `sklearn`.

- **Support Vector Machine** (SVM)
  - `C`: $[2^{-5}, 2^5]$
  - `gamma`: $[2^{-15}, 2^3]$
- **k-Nearest Neighbor** (kNN)
  - `n_neighbors`: $[1, 2 \cdot \sqrt{|X|}]$, where $|X|$ denotes the number of data points
- **Multilayer Perceptron** (MLP)
  - `hidden_layer_depth`: [1, 4]
  - `num_nodes_per_layer`: [16, 256]
  - `activation`: {`identity`, `logistic`, `tanh`, `relu`}
  - `alpha`: $[10^{-7}, 10^{-1}]$
  - `learning_rate_init`: $[10^{-4}, 0.5]$
- **Naive Bayesian networks** (NB)
  - `var_smoothing`: $[10^{-12}, 1]$
- **Random forest** (RF)

  - `n_estimators`: [1, 200]
  - `min_samples_split`: [2, 20]
  - `min_samples_leaf`: [1, 20],
  - `criterion`: {`gini`, `entropy`, `log_loss`}
- **Logistic regression** (LR)
  - `penalty`: {`L1`, `L2`, `Elasticnet`}
  - `C`: $[2^{-5}, 2^{5}]$
  - `tol`: $[10^{-5}, 10^{-1}]$
  - `l1_ratio`: [0, 1]
- **Linear discriminant analysis** (LDA)
  - `tol`: $[10^{-5}, 10^{-1}]$
  - `solver`: {`svd`}
- **XGBoost** (XBG)
  - `n_estimators`: [1, 200]
  - `learning_rate`: [0.01, 1]
  - `gamma`: [0, 1]
  - `min_child_weight`: [1, 20]
  - `subsample`: [0.5, 1]
  - `colsample_bytree`: [0.5, 1]

For the ensemble of classifiers, we used the max combination of the classifiers (i.e., chose the highest score among the classifiers). Finally, we estimated the CLM ranking of the labeled datasets based on the classification scores.

Settings for computing ground-truth CLM. We used `sklearn` implementation for DBSCAN, K-Means, Birch, Agglomerative clustering, and used `sklearn-extra` implementation for K-Medoids. We used `pyclustering` implementation for X-Means. For HDBSCAN, we used the implementation of McInnes et al. [16]. To find the best hyperparameter setting for each clustering technique, we used Nogueira's Bayesian optimization implementation [14] with default setting, as we did for classification techniques. The bounds of each technique's hyperparameter is set as follows:

- **DBSCAN**
  - `epsilon`: [0.01, 1.0]
  - `min_samples`: [1, 10]
- **HDBSCAN**
  - `epsilon`: [0.01, 1.0]
  - `min_samples`: [1, 10]
  - `min_cluster_size`: [2, 50]
- ***K*-Means**
  - K (number of clusters): [2, $|C|\cdot$ 3], where $|C|$ denotes the number of classes in data
- ***K*-Medoids**
  - K: [2, $|C|\cdot$ 3]
- ***X*-Means**
  - `kmax`: [2, 50]
  - `tolerance`: [0.01, 1.0]
- **Birch**
  - `threshold`: [0.1, 1.0]
  - `branching_factor`: [10, 100]
- **Agglomerative Clustering**
  - number of clusters: [2, $|C|\cdot$ 3]

We used Single, Average, and Complete variants of Agglomerative Clustering. After finding the best hyperparameter setting of a clustering technique for a given datasets, we ran the technique 20 times with the same setting and report the average value as the final score.

Apparatus. We run all the experiments in a Linux server machine equipped with 40-core Intel Xeon Silver 4210 CPUs, TITAN RTX, and 224GB RAM.

APPENDIX G
LABELED DATASETS

Please refer to the end of this document for the tables containing the list of labeled datasets used in our across-dataset rank correlation analysis (Section 5.2; Table 1 and 2). Note that before applying the clustering techniques, classifiers, and measures, we normalized the datasets and removed the rows containing missing data.

REFERENCES

[1] M. Ackerman and S. Ben-David, "Measures of clustering quality: A working set of axioms for clustering," in *Advances in Neural Information Processing Systems*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., vol. 21. Curran Associates, Inc., 2008.

[2] M. M. Abbas, M. Aupetit, M. Sedlmair, and H. Bensmail, "Clustme: A visual quality measure for ranking monochrome scatterplots based on cluster patterns," *Computer Graphics Forum*, vol. 38, no. 3, pp. 225–236, 2019.

[3] M. Aupetit, M. Sedlmair, M. M. Abbas, A. Baggag, and H. Bensmail, "Toward perception-based evaluation of clustering techniques for visual analytics," in *30th IEEE Visualization Conference, IEEE VIS 2019 - Short Papers, Vancouver, BC, Canada, October 20-25, 2019*. IEEE, 2019, pp. 141–145.

[4] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of cybernetics*, vol. 4, no. 1, pp. 95–104, 1974.

[5] Y. Vardi and C.-H. Zhang, "The multivariate l 1-median and associated data depth," *Proceedings of the National Academy of Sciences*, vol. 97, no. 4, pp. 1423–1426, 2000.

[6] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1650–1654, 2002.

[7] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 13, no. 8, pp. 841–847, 1991.

[8] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[9] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0377042787901257

[10] J. Hopcroft and R. Kannan, "Foundations of data science," 2014.

[11] T. M. Apostol and M. A. Mnatsakanian, "Sums of squares of distances in m-space," *The American Mathematical Monthly*, vol. 110, no. 6, pp. 516–526, 2003. [Online]. Available: https://doi.org/10.1080/00029890.2003.11919989

[12] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/03610927408827101

[13] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems*, vol. 25. Curran Associates, Inc., 2012.

[14] F. Nogueira, "Bayesian Optimization: Open source constrained global optimization tool for Python," 2014–. [Online]. Available: https://github.com/fmfn/BayesianOptimization

[15] M. Feurer, A. Klein, J. Eggensperger, Katharina Springenberg, M. Blum, and F. Hutter, "Efficient and robust automated machine learning," in *Advances in Neural Information Processing Systems 28 (2015)*, 2015, pp. 2962–2970.

[16] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering." *J. Open Source Softw.*, vol. 2, no. 11, p. 205, 2017.

[17] "Kaggle," https://www.kaggle.com, 2010.

[18] A. Asuncion and D. Newman, "Uci machine learning repository," 2007.

[19] I.-C. Yeh, K.-J. Yang, and T.-M. Ting, "Knowledge discovery on rfm model using bernoulli sequence," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5866–5871, 2009.

[20] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory, "A taxonomy of visual cluster separation factors," in *Computer Graphics Forum*, vol. 31, no. 3pt4. Wiley Online Library, 2012, pp. 1335–1344.

[21] M. Patrício, J. Pereira, J. Crisóstomo, P. Matafome, M. Gomes, R. Seiça, and F. Caramelo, "Using resistin, glucose, age and bmi to predict the presence of breast cancer," *BMC cancer*, vol. 18, no. 1, pp. 1–8, 2018.

[22] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology." *Proceedings of the national academy of sciences*, vol. 87, no. 23, pp. 9193–9196, 1990.

[23] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[24] M. Hon, D. Stello, and J. Yu, "Deep learning classification in asteroseismology," *Monthly Notices of the Royal Astronomical Society*, vol. 469, no. 4, pp. 4578–4583, 2017.

[25] S. Nene, S. Nayar, H. Murase *et al.*, "Columbia object image library (coil-20)," *Data available at http://www. cs. columbia. edu/CAVE/software/softlib/coil-20. php*, 1996.

[26] B. A. Johnson and K. Iizuka, "Integrating openstreetmap crowdsourced data and landsat time-series imagery for rapid land use/land cover (lulc) mapping: Case study of the laguna de bay area of the philippines," *Applied Geography*, vol. 67, pp. 140–149, 2016.

[27] M. Koklu, R. Kursun, Y. S. Taspinar, and I. Cinar, "Classification of date fruits into genetic varieties using image analysis," *Mathematical Problems in Engineering*, vol. 2021, 2021.

[28] B. Antal and A. Hajdu, "An ensemble-based system for automatic screening of diabetic retinopathy," *Knowledge-based systems*, vol. 60, pp. 20–27, 2014.

[29] M. Koklu and I. A. Ozkan, "Multiclass classification of dry beans using computer vision and machine learning techniques," *Computers and Electronics in Agriculture*, vol. 174, p. 105507, 2020.

[30] E. Kaya and I. Saritas, "Towards a real-time sorting system: identification of vitreous durum wheat kernels using ann based on their morphological, colour, wavelet and gaborlet features," *Computers and Electronics in Agriculture*, vol. 166, p. 105016, 2019.

[31] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger, "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state," *Physical Review E*, vol. 64, no. 6, p. 061907, 2001.

[32] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[33] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[34] D. Ayres-de Campos, J. Bernardes, A. Garrido, J. Marques-de Sa, and L. Pereira-Leite, "Sisporto 2.0: a program for automated analysis of cardiotocograms," *Journal of Maternal-Fetal Medicine*, vol. 9, no. 5, pp. 311–318, 2000.

[35] L. Sharan, R. Rosenholtz, and E. Adelson, "Material perception: What can you see in a brief glance?" *Journal of Vision*, vol. 9, no. 8, pp. 784–784, 2009.

[36] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *International workshop on ambient assisted living*. Springer, 2012, pp. 216–223.

[37] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, 2017, pp. 512–515.

[38] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *The American journal of cardiology*, vol. 64, no. 5, pp. 304–310, 1989.

[39] I. Guyon, A. Saffari, G. Dror, and G. Cawley, "Agnostic learning vs. prior knowledge challenge," in *2007 International Joint Conference on Neural Networks*. IEEE, 2007, pp. 829–834.

[40] R. J. Lyon, B. Stappers, S. Cooper, J. M. Brooke, and J. D. Knowles, "Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach," *Monthly Notices of the Royal Astronomical Society*, vol. 459, no. 1, pp. 1104–1123, 2016.

[41] L. Rachakonda, S. P. Mohanty, E. Kougianos, and P. Sundaravadivel, "Stress-lysis: A dnn-integrated edge device for stress level detection in the iomt," *IEEE Transactions on Consumer Electronics*, vol. 65, no. 4, pp. 474–483, 2019.

[42] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 2011, pp. 142–150.

[43] P. Van Der Putten and M. Van Someren, "A bias-variance analysis of a real world learning problem: The coil challenge 2000," *Machine learning*, vol. 57, no. 1, pp. 177–195, 2004.

[44] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[45] M. Elter, R. Schulz-Wendtland, and T. Wittenberg, "The prediction of breast cancer biopsy outcomes using two cad approaches that both emphasize an intelligible decision process," *Medical physics*, vol. 34, no. 11, pp. 4164–4172, 2007.

[46] J. Xia, Y. Zhang, J. Song, Y. Chen, Y. Wang, and S. Liu, "Revisiting dimensionality reduction techniques for visual cluster analysis: An empirical study," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 529–539, 2022.

[47] A. L. Cambridge, "The database of faces," 1994. [Online]. Available: https://cam-orl.co.uk/facedatabase.html

[48] M. Little, P. Mcsharry, S. Roberts, D. Costello, and I. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *Nature Precedings*, pp. 1–1, 2007.

[49] M. Sadikin, "EHR dataset for patient treatment classification," 2020. [Online]. Available: https://data.mendeley.com/datasets/7kv3rctx7m/1

[50] J. W. Smith, J. E. Everhart, W. Dickson, W. C. Knowler, and R. S. Johannes, "Using the adap learning algorithm to forecast the onset of diabetes mellitus," in *Proceedings of the annual symposium on computer application in medical care*. American Medical Informatics Association, 1988, p. 261.

[51] İ. A. Özkan, M. Köklü, and R. Saraçoğlu, "Classification of pistachio species using improved k-nn classifier," *health*, vol. 2, p. 3, 2021.

[52] M. Koklu, S. Sarigil, and O. Ozbek, "The use of machine learning methods in classification of pumpkin seeds (cucurbita pepo l.)," *Genetic Resources and Crop Evolution*, vol. 68, no. 7, pp. 2713–2726, 2021.

[53] İ. ÇINAR, M. KOKLU, and Ş. TAŞDEMİR, "Classification of raisin grains using machine vision and artificial intelligence methods," *Gazi Mühendislik Bilimleri Dergisi (GMBD)*, vol. 6, no. 3, pp. 200–209, 2020.

[54] I. Cinar and M. Koklu, "Classification of rice varieties using artificial intelligence methods," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 7, no. 3, pp. 188–194, 2019.

[55] M. Charytanowicz, J. Niewczas, P. Kulczycki, P. A. Kowalski, S. Łukasik, and S. Żak, "Complete gradient clustering algorithm for features analysis of x-ray images," in *Information technologies in biomedicine*. Springer, 2010, pp. 15–24.

[56] M. Sikora *et al.*, "Application of rule induction algorithms for analysis of data collected by seismic hazard monitoring systems in coal mines," *Archives of Mining Sciences*, vol. 55, no. 1, pp. 91–114, 2010.

[57] D. Kotzias, M. Denil, N. De Freitas, and P. Smyth, "From group to individual labels using deep features," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 597–606.

[58] J. J. Thompson, M. R. Blair, L. Chen, and A. J. Henrey, "Video game telemetry as a critical tool in the study of complex skill learning," *PloS one*, vol. 8, no. 9, p. e75129, 2013.

[59] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, "Contributions to the study of sms spam filtering: new collection and results," in *Proceedings of the 11th ACM symposium on Document engineering*, 2011, pp. 259–262.

[60] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," 2011.

[61] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," 2008.

[62] M. B. Er and I. B. Aydilek, "Music emotion recognition by using chroma spectrogram and deep visual features," *International Journal of Computational Intelligence Systems*, vol. 12, no. 2, pp. 1622–1634, 2019.

[63] H. T. Kahraman, S. Sagiroglu, and I. Colak, "The development of intuitive knowledge classifier and the modeling of domain dependent data," *Knowledge-Based Systems*, vol. 37, pp. 283–295, 2013.

[64] E. Ventocilla and M. Riveiro, "A comparative user study of visualization techniques for cluster analysis of multidimensional data sets," *Information visualization*, vol. 19, no. 4, pp. 318–338, 2020.

[65] N. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing detection based associative classification data mining," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5948–5959, 2014.

[66] B. A. Johnson, R. Tateishi, and N. T. Hoan, "A hybrid pansharpening approach and multiscale object-based image analysis for mapping diseased pine and oak trees," *International journal of remote sensing*, vol. 34, no. 20, pp. 6969–6982, 2013.

[67] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision support systems*, vol. 47, no. 4, pp. 547–553, 2009.

[68] J. G. Rohra, B. Perumal, S. J. Narayanan, P. Thakur, and R. B. Bhatt, "User localization in an indoor environment using fuzzy hybrid of particle swarm optimization & gravitational search algorithm with neural networks," in *Proceedings of Sixth International Conference on Soft Computing for Problem Solving*. Springer, 2017, pp. 286–295.

TABLE 1
96 Labeled datasets used in across-dataset rank correlation analysis (Ordered by source frequency then alphabetically)

Dataset	Objects	Features	Classes	Source
Birds Bones and Living Habits	413	10	6	Kaggle [17]
Cardiovascular Study	2,927	15	2	Kaggle
Credit Risk Classification	976	11	2	Kaggle
Customer Classification	1,000	11	4	Kaggle
Fraud Detection Bank	20,468	112	2	Kaggle
Heart Attack Analysis & Prediction	303	13	2	Kaggle
Microbes	30,527	24	10	Kaggle
Mobile Price Classification	2,000	20	4	Kaggle
Music Genre Classification	1,000	26	10	Kaggle
Orbit Classification For Prediction / NASA	1,748	11	6	Kaggle
Paris Housing Classification	10,000	17	2	Kaggle
Predicting Pulsar Star	9273	8	2	Kaggle
pH-recognition	653	3	15	Kaggle
Rice Seed (Gonen&Jasmine)	18,185	10	2	Kaggle
Siberian Weather Stats	1,439	11	9	Kaggle
Smoker Condition	1,012	7	2	Kaggle
Water Quality	2,011	9	2	Kaggle
Wine Customer Segmentation	178	13	3	Kaggle
Banknote Authentication	1,372	4	2	UCI ML Repo [18]
Breast Cancer Wisconsin (Prognostic)	569	30	2	UCI ML Repo
Breast Tissue	106	9	6	UCI ML Repo
CNAE-9	1,080	856	9	UCI ML Repo
Dermatology	358	34	6	UCI ML Repo
Echocardiogram	61	10	2	UCI ML Repo
Ecoli	336	7	8	UCI ML Repo
Glass Identification	214	9	6	UCI ML Repo
Harberman's Survival	306	3	2	UCI ML Repo
Hepatitis	80	19	2	UCI ML Repo
Image Segmentation	210	19	7	UCI ML Repo
Ionosphere	351	34	2	UCI ML Repo
Iris	150	4	3	UCI ML Repo
Letter Recognition	20,000	16	26	UCI ML Repo
MAGIC Gamma Telescope	19,020	10	2	UCI ML Repo
Optical Recognition of Handwritten Digits	3,823	64	10	UCI ML Repo
Pen-Based Recognition of Handwritten Digits	7,494	16	10	UCI ML Repo
Planning Relax	182	12	2	UCI ML Repo
SECOM	1,567	590	2	UCI ML Repo
Spambase	4,601	57	2	UCI ML Repo
SPECTF Heart	80	44	2	UCI ML Repo
Statlog (German Credit)	1,000	24	2	UCI ML Repo
Statlog (Image Segmentation)	2,310	19	7	UCI ML Repo
Taiwanese Bankruptcy Prediction	6,819	95	2	UCI ML Repo
Wine	178	13	3	UCI ML Repo
Yeast	1,484	8	10	UCI ML Repo
Zoo	101	16	7	UCI ML Repo
Blood Transfusion Service Center	748	4	2	Yeh et al. [19]
Boston	154	13	3	Sedlmair et al. [20]
Breast Cancer Coimbra	116	9	2	Patrício et al. [21]
Breast Cancer Wisconsin (Original)	683	9	2	Wolberg et al. [22]
CIFAR10	3,250	1,024	10	Krizhevsky et al. [23]
Classification in Asteroseismology	1,001	3	2	Hon et al. [24]
COIL20	1,440	400	20	Nene et al. [25]
Crowdsourced Mapping	10,545	28	6	Johnson et al. [26]
Date Fruit	898	34	7	Koklu et al. [27]
Diabetic Retinopathy Debrecen	1,151	19	2	Antal et al. [28]
Dry Bean	13,611	16	7	Koklu et al. [29]

Dataset	Objects	Features	Classes	Source
Durum Wheat Features	9,000	236	3	Kaya et al. [30]
Epileptic Seizure Recognition	5,750	178	5	Andrzejak et al. [31]
ExtyaleB	319	30	5	Georghiades et al. [32]
Fashion-MNIST	3,000	784	10	Xiao et al. [33]
Fetal Health Classification	2,126	21	3	Ayres-de-Campos et al. [34]
Flickr Material Database	997	1,536	10	Sharan et al. [35]
HAR	735	561	6	Anguita et al. [36]
Hate Speech	3,221	100	3	Davidson et al. [37]
Heart Disease	297	13	5	Detrano et al. [38]
HIVA	3,076	1,617	2	Guyon et al. [39]
HTRU2	17,898	8	2	Lyon et al. [40]
Human Stress Detection	2,001	3	3	Rachakonda et al. [41]
IMDB	3,250	700	2	Maas et al. [42]
Insurance Company Benchmark	5,822	85	2	Putten et al. [43]
Labeled Faces in the Wild	2,200	5,828	2	Huang et al. [44]
Mammographic Mass	830	5	2	Elter et al. [45]
MNIST64	1,082	64	6	Xia et al. [46]
Olivetti Faces	400	4,096	40	AT&T Lab. Cambridge [47]
Parkinsons	195	22	2	Little et al. [48]
Patient Treatment Classification	4,412	10	2	Mujiono Sadikin [49]
Pima Indians Diabetes Database	768	8	2	Smith et al. [50]
Pistachio	2,148	28	2	Ozkan et al. [51]
Pumpkin Seeds	2,500	12	2	Koklu et al. [52]
Raisin	900	7	2	Cinar et al. [53]
Rice Dataset Cammeo and Osmancik	3,810	7	2	Cinar et al. [54]
Seeds	210	7	3	Charytanowicz et al. [55]
Seismic Bumps	646	24	2	Sikora et al. [56]
Sentiment Labeled Sentences	2,748	200	2	Kotzias et al. [57]
SkillCraft1 Master Table Dataset	3,338	18	7	Thompson et al. [58]
SMS Spam Collection	835	500	2	Almeida et al. [59]
Street View House Numbers	732	1,024	10	Netzer et al. [60]
Student Grade	395	29	2	Cortez et al. [61]
Turkish Music Emotion	400	50	4	Er et al. [62]
User Knowledge Modeling	258	5	4	Kahraman et al. [63]
Weather	365	192	7	Ventocilla et al. [64]
Website Phishing	1,353	9	3	Abdelhamid et al. [65]
Wilt	4,339	5	2	Johnson et al. [66]
Wine Quality	4,898	11	7	Cortez et al. [67]
Wireless Indoor Localization	2,000	7	4	Rohra et al. [68]
World12d	150	12	5	Sedlmair et al. [20]

TABLE 2
96 Labeled datasets ordered by CH_A score. The top and bottom 32 datasets (`top-1/3` $\mathcal{P}^+$ and `bottom-1/3` $\mathcal{P}^-$, respectively, in Section 7.1) are separated by double horizontal lines.

Dataset	Objects	Features	Classes	CH_A
Weather	365	192	7	1.000
Olivetti Faces	400	4,096	40	0.956
MNIST64	1,082	64	6	0.956
Optical Recognition of Handwritten Digits	3,823	64	10	0.941
Seeds	210	7	3	0.925
Wireless Indoor Localization	2,000	7	4	0.915
COIL20	1,440	400	20	0.887
Iris	150	4	3	0.875
Pen-Based Recognition of Handwritten Digits	7,494	16	10	0.862
Rice Seed (Gonen&Jasmine)	18,185	10	2	0.861
Breast Cancer Wisconsin (Original)	683	9	2	0.800
pH-recognition	653	3	15	0.780
Echocardiogram	61	10	2	0.770
Fashion-MNIST	3,000	784	10	0.739
Mobile Price Classification	2,000	20	4	0.732
Human Stress Detection	2,001	3	3	0.722
Dry Bean	13,611	16	7	0.694
HAR	735	561	6	0.690
Rice Dataset Cammeo and Osmancik	3,810	7	2	0.654
Wine Customer Segmentation	178	13	3	0.619
Wine	178	13	3	0.619
Zoo	101	16	7	0.605
Image Segmentation	210	19	7	0.576
Boston	154	13	3	0.574
Statlog (Image Segmentation)	2,310	19	7	0.514
Letter Recognition	20,000	16	26	0.511
User Knowledge Modeling	258	5	4	0.485
Ecoli	336	7	8	0.472
Website Phishing	1,353	9	3	0.407
Date Fruit	898	34	7	0.372
Music Genre Classification	1,000	26	10	0.340
Pistachio	2,148	28	2	0.308
Crowdsourced Mapping	10,545	28	6	0.297
Raisin	900	7	2	0.294
Breast Cancer Wisconsin (Prognostic)	569	30	2	0.294
Yeast	1,484	8	10	0.269
Dermatology	358	34	6	0.254
Glass Identification	214	9	6	0.253
Classification in Asteroseismology	1,001	3	2	0.249
Breast Tissue	106	9	6	0.216
Mammographic Mass	830	5	2	0.202
Banknote Authentication	1,372	4	2	0.197
Birds Bones and Living Habits	413	10	6	0.186
ExtyaleB	319	30	5	0.148
Flickr Material Database	997	1,536	10	0.137
CNAE-9	1,080	856	9	0.130
Fetal Health Classification	2,126	21	3	0.126
Durum Wheat Features	9,000	236	3	0.121
Smoker Condition	1,012	7	2	0.104
Student Grade	395	29	2	0.098
Turkish Music Emotion	400	50	4	0.097
CIFAR10	3,250	1,024	10	0.070
Ionosphere	351	34	2	0.069
SPECTF Heart	80	44	2	0.066
Microbes	30,527	24	10	0.050
Hate Speech	3,221	100	3	0.045
Predicting Pulsar Star	9273	8	2	0.033

Dataset	Objects	Features	Classes	CH_A
Parkinsons	195	22	2	0.033
HTRU2	17,898	8	2	0.033
Siberian Weather Stats	1,439	11	9	0.030
Patient Treatment Classification	4,412	10	2	0.029
SMS Spam Collection	835	500	2	0.027
MAGIC Gamma Telescope	19,020	10	2	0.027
Orbit Classification For Prediction / NASA	1,748	11	6	0.024
Harberman's Survival	306	3	2	0.024
IMDB	3,250	700	2	0.022
Pumpkin Seeds	2,500	12	2	0.022
World12d	150	12	5	0.021
Heart Attack Analysis & Prediction	303	13	2	0.021
Diabetic Retinopathy Debrecen	1,151	19	2	0.020
Seismic Bumps	646	24	2	0.017
Hepatitis	80	19	2	0.017
Statlog (German Credit)	1,000	24	2	0.014
Wine Quality	4,898	11	7	0.013
Sentiment Labeled Sentences	2,748	200	2	0.013
Pima Indians Diabetes Database	768	8	2	0.013
Blood Transfusion Service Center	748	4	2	0.013
Heart Disease	297	13	5	0.012
Cardiovascular Study	2,927	15	2	0.011
Insurance Company Benchmark	5,822	85	2	0.010
Street View House Numbers	732	1,024	10	0.008
SkillCraft1 Master Table Dataset	3,338	18	7	0.006
HIVA	3,076	1,617	2	0.006
Spambase	4,601	57	2	0.005
Wilt	4,339	5	2	0.005
Breast Cancer Coimbra	116	9	2	0.005
SECOM	1,567	590	2	0.005
Customer Classification	1,000	11	4	0.003
Credit Risk Classification	976	11	2	0.003
Planning Relax	182	12	2	0.002
Taiwanese Bankruptcy Prediction	6,819	95	2	0.002
Labeled Faces in the Wild	2,200	5,828	2	0.002
Water Quality	2,011	9	2	0.001
Epileptic Seizure Recognition	5,750	178	5	0.001
Paris Housing Classification	10,000	17	2	0.000
Fraud Detection Bank	20,468	112	2	0.000