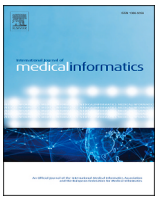
International Journal of Medical Informatics 222 (2026) 106738

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Instance level analysis of equitable learning via dissimilar variable grouping on healthcare datasets

Hyeonggeun Yun [a,*], Amanda S. Barnard [a], Hanna Suominen [a,b,c]

[a] School of Computing, The Australian National University, Canberra, 2601, ACT, Australia
[b] School of Medicine and Psychology, The Australian National University, Canberra, 2601, ACT, Australia
[c] Department of Computing, University of Turku, Turku, FI-20014, Finland

HIGHLIGHTS

- SHIELD uses CMI-driven dissimilarity across multiple grouping strategies.
- Decoder-mapped SHAP preserves attribution in the original variable space.
- Grouping reduced explanation disparity across three healthcare datasets.
- Fairness gains were accompanied by context-dependent predictive trade-offs.

ARTICLE INFO

Keywords:
Machine learning
Conditional mutual information
Explainable artificial intelligence
Equitable learning
Health informatics

ABSTRACT

Machine learning models are increasingly deployed in healthcare, but their complexity can propagate biases and obscure decision-making that disproportionately affect vulnerable populations. While existing explainability methods focus on variable-level attribution, they often fail to capture how predictions and explanations behave at the instance level. To address this gap, we introduce SHIELD: a SHapley and Information-theory based framework for Equitable Learning via Dissimilar variable grouping. In this framework, variables are grouped using conditional mutual information to weaken correlations that may encode sensitive attributes. Group-specific autoencoders learn latent representations that preserve a mapping to the original variables, allowing SHapley Additive exPlanations (SHAP) to be computed in the original variable space for faithful instance-level attribution. Experiments on three clinical datasets demonstrated that dissimilarity-based grouping produced a more even distribution of variable importance and increased total attribution magnitude, suggesting broader usage of variables in model decision-making. On average, grouping improved the six evaluated fairness measures, covering both outcome and explanation disparities by 17.09%. Nevertheless, predictive performance decreased by 6.18% across six metrics. These changes represent a context-dependent trade-off rather than directly comparable gains and losses. Overall, SHIELD provides a principled and reproducible framework for equitable and explainable machine learning in health informatics, with code available to support future research.

1. Introduction

Artificial Intelligence (AI) and Machine Learning (ML) systems are increasingly used in healthcare for diagnosis, triage, and decision support, where model outputs directly affect patient outcomes. In such high-stakes settings, explanations must reliably justify individual predictions. When the basis of a prediction is opaque, clinicians cannot assess whether a model is reasoning appropriately for a specific patient, and failures in transparency have already contributed to disparities in care [1,2]. Recent literature emphasises that explainability must combine interpretability, fidelity, and clinical value [3,4].

Most existing explainability approaches target variable-level attribution, identifying which variables contribute to a model's predictions [5–7]. While valuable, this perspective does not capture how predictions and explanations behave for individual patients. Instance-level explanations are therefore essential for clinically trustworthy predictions across diverse patient groups.

* Corresponding author.
Email address: Geun.Yun@anu.edu.au (H. Yun).

https://doi.org/10.1016/j.ijmedinf.2026.106738
Received 17 June 2026; Received in revised form 11 August 2026; Accepted 22 September 2026
Available online 23 September 2026

These challenges are further amplified in clinical data. Clinically relevant outcomes, such as diagnosis or risk prediction, are often rare, and minority groups are underrepresented [8–10]. In addition, imbalance, sparsity, and heterogeneity in clinical datasets increase error rates and can yield unequal explanation quality across patients [9,11]. These properties make it particularly difficult to ensure both accurate and equitable model behaviour.

Unfairness in ML models often arises when variables are correlated with sensitive attributes, allowing models to rely on proxy signals that indirectly encode protected information and lead to systematic disparities across patient groups [12]. Standard training further exacerbates this issue by prioritising a small number of dominant predictors [13], potentially amplifying hidden dependencies and producing misleadingly concentrated variable contributions [11,14]. Popular explainers, such as SHapley Additive exPlanations (SHAP) [5], provide a principled way to attribute model predictions to individual variables and instances. However, transparency alone does not guarantee fairness, as models can produce plausible explanations while still relying on biased or proxy-driven decision pathways.

Many fairness interventions attempt to reduce dependence on sensitive attributes by modifying latent representations. While effective in improving statistical fairness metrics, such approaches can obscure clinically meaningful variables and reduce interpretability [15,16]. In healthcare, where decisions must be transparent and justifiable, this trade-off poses a significant limitation. Clinicians rely on explanations grounded in the original variable space to assess whether model reasoning is appropriate for individual patients. Achieving fairness without sacrificing such interpretability would enable more trustworthy and equitable decision support in clinical practice.

This study addresses these challenges by proposing and validating an instance-level analysis of SHIELD: a Shapley and Information-theory based framework for Equitable Learning via Dissimilar variable grouping. By redistributing reliance across variable groups and analysing how dissimilarity-based grouping affects both predictions and local SHAP explanations, this study investigates the relationship between predictive performance, explanation behaviour, and group fairness. SHIELD is used to characterise these trade-offs across datasets, models, and grouping configurations, rather than assuming that improvements in fairness compensate for reductions in predictive performance.

2. Methodology

2.1. Data collection

All data used in this study were existing, de-identified clinical datasets obtained from the UCI Machine Learning Repository [17]. Ethical approval for the secondary analysis of these publicly available datasets was granted by The Australian National University Human Research Ethics Committee (Protocol H/2025/0248).

The tasks of classifying Diabetes, Heart Disease, and Obesity were selected for their clinical relevance and prior use in the literature [18–20] (Table 1). The datasets collectively covered various multiclass tasks, continuous/categorical predictors, varying ratios of variables to instances, and different sensitive attributes. In particular, the group fairness analysis compared Caucasians and non-Caucasians in the Diabetes dataset, and males and females in the Heart Disease and Obesity datasets (Section 2.5.3).

Table 1
Datasets used in the study [21–23].

Attribute	Diabetes	Heart Disease	Obesity
D (# of variables)	47	13	16
N (# of instances)	101,766	303	2111
C (# of classes)	3	5	7
Class balanced?	No	No	Yes
Sensitive attribute	Race	Sex	Sex

2.2. Data preprocessing

Preprocessing involved imputation, encoding, and standardisation. Missing values were imputed using an XGBoost-based classifier [24], which captures nonlinear relationships in heterogeneous clinical data better than mean or mode imputation [25]. Categorical variables were encoded using one-hot, ordinal, or label encoding as appropriate [26]. Z-score standardisation was applied to reduce scale effects [27].

2.3. Variable grouping by dissimilarity

Variable grouping encouraged equitable learning of models by reducing the influence of correlated and proxy variables that can implicitly encode sensitive information [2,14]. Grouping by dissimilarity (not similarity) distributed correlated/proxy variables across different groups, weakening their joint effect and reducing the risk of biased latent representations [28]. If similar variables are grouped together, their combined influence may reinforce dependencies with sensitive attributes [14,29], so spreading them apart instead could attenuate this effect and act as a natural regulariser.

All grouping strategies relied on a dissimilarity matrix derived from conditional mutual information (CMI) [30]. For variables X_i and X_j and target variable Y, CMI measures how much information the two variables share given the outcome:

$$I(X_i; X_j \mid Y) = \sum_{x_i, x_j, y} p(x_i, x_j, y) \log \frac{p(x_i, x_j \mid y)}{p(x_i \mid y) p(x_j \mid y)}.$$

Dissimilarity was defined as the complement of CMI:

$$D_{i,j} = 1 - \mathrm{CMI}_{i,j},$$

where higher values denote more distinct pairs.

Two metrics assessed grouping quality. For each group G_k containing variable indices $i, j \in G_k$, diversity measured average intra-group dissimilarity [31]:

$$\text{Diversity} = \sum_{\forall k \in \{1,\ldots,K\}} \sum_{(i<j) \in G_k} \frac{D_{i,j}}{K|G_k|},$$

where K is the number of groups. Dispersion enforced a conservative minimum dissimilarity threshold [31]:

$$\text{Dispersion} = \min_{\forall k \in \{1,\ldots,K\}} \{ \min_{(i<j) \in G_k} \{ D_{i,j} \} \}.$$

To examine whether the grouping strategies produced structurally distinct partitions, grouping behaviour was evaluated across $K = 2, \ldots, 10$ using diversity and dispersion. Random grouping was also included as an unstructured control.

2.3.1. Bicriterion approach

Bicriterion anticlustering [31] jointly maximised diversity and dispersion via a weighted objective:

$$\text{obj} = \alpha \cdot \text{Diversity} + \beta \cdot \text{Dispersion},$$

where α, β quantified the priorities of each criterion. It used local search heuristics to swap variables between groups whenever the objective improved (Appendix A). This prevented configurations where groups appeared diverse overall but contained a tightly related subset of proxy variables (as the search also considered dispersion).

2.3.2. K-plus anticlustering approach

K-plus anticlustering [32] generalised K-means anticlustering by matching higher-order distributional statistics, including variance, skewness, and kurtosis, across groups (Appendix B). Polynomial variable expansions $X^{(2)}, X^{(3)}, \ldots, X^{(r)}$ were constructed, and the composite K-plus objective combined standard Sum of Squares Error (SSE) with additional moment-based SSE terms weighted by $\lambda_1, \ldots, \lambda_r$. Similar to the bicriterion method, variable swaps between groups were accepted

when they reduced the overall objective. This produced groups that were internally heterogeneous while statistically aligned across multiple higher-order moments.

2.3.3. Greedy approach

This study introduces a deterministic Greedy dissimilarity-grouping procedure (Appendix C). Unlike Bicriterion and K-plus anticlustering, which rely on iterative optimisation or higher-order distributional objectives, the proposed method selects globally dissimilar seed variables and assigns remaining variables according to their average dissimilarity to each current group. It therefore provides a computationally simple, directly CMI-driven alternative that preserves the study's core objective of separating conditionally dependent variables.

2.3.4. Latent representations of groups

Each variable group was compressed using an autoencoder with a nonlinear encoder and a linear decoder. For group k, the encoder consisted of an affine transformation followed by a Rectified Linear Unit (ReLU), while the decoder consisted of an affine output layer with identity activation. The encoder was therefore defined as

$$f_k(x) = \mathrm{ReLU}\left(W_{\mathrm{enc}}^{(k)} x + b_{\mathrm{enc}}^{(k)}\right),$$

and the decoder as

$$g_k(z) = W_{\mathrm{dec}}^{(k)} z + b_{\mathrm{dec}}^{(k)}.$$

Although the decoder was linear, the complete reconstruction mapping $g_k \circ f_k$ remained nonlinear because of the ReLU activation in the encoder. This allowed the autoencoder to represent potentially nonlinear relationships without being restricted to a single linear projection, as in commonly used dimensionality reduction techniques such as Principal Component Analysis (PCA). Meanwhile, the decoder weights preserved an explicit mapping from the latent coordinates to the original variables.

Formally, let $X^{(k)} \in \mathbb{R}^{n_k \times d_k}$ denote the matrix of d_k variables in group k across n_k instances. The autoencoder had a shallow architecture $d_k \to m_k \to d_k$, where m_k denoted the bottleneck dimension. The encoder $f_k : \mathbb{R}^{d_k} \to \mathbb{R}^{m_k}$ and decoder $g_k : \mathbb{R}^{m_k} \to \mathbb{R}^{d_k}$ were trained by minimising the mean squared reconstruction loss:

$$\min_{f_k, g_k} \left(\frac{1}{n_k} \sum_{i=1}^{n_k} \left\| X_i^{(k)} - g_k\left(f_k\left(X_i^{(k)}\right)\right) \right\|^2 \right).$$

Candidate bottleneck dimensions $m_k \in \{1, \ldots, 5\}$ were evaluated sequentially. The smallest dimension achieving a validation mean squared reconstruction error of at most 10^{-2} was selected. The resulting latent vector $z = f_k(X_i^{(k)})$ served as the representation of group k for instance i. To preserve interpretability, SHIELD used the linear decoder weights to redistribute latent space SHAP attributions across the variables within each group. Under the column-vector convention,

$$W_{\mathrm{dec}}^{(k)} \in \mathbb{R}^{d_k \times m_k},$$

where each column of $W_{\mathrm{dec}}^{(k)}$ described the contribution of one latent coordinate to the original variables d_k. Let the latent-space SHAP attribution vector for group k be

$$\phi^{(k)} = \left[\phi_1^{(k)}, \ldots, \phi_{m_k}^{(k)}\right]^T.$$

The absolute decoder weights were first obtained as

$$\left|W_{\mathrm{dec}}^{(k)}\right|_{ij} = \left|W_{\mathrm{dec}}^{(k)}[i, j]\right|.$$

For each latent coordinate j, weights were then normalised across original variables to sum to one:

$$\tilde{W}_{ij}^{(k)} = \frac{\left|W_{\mathrm{dec}}^{(k)}\right|_{ij}}{\sum_{i'=1}^{d_k} \left|W_{\mathrm{dec}}^{(k)}\right|_{i'j}} \quad \text{for} \quad (i = 1, \ldots, d_k; j = 1, \ldots, m_k).$$

The final variable-level attribution vector was then

$$\phi_{\mathrm{original}}^{(k)} = \tilde{W}^{(k)} \phi^{(k)} \in \mathbb{R}^{d_k},$$

with each variable receiving

$$\left[\phi_{\mathrm{original}}^{(k)}\right]_i = \sum_{j=1}^{m_k} \tilde{W}_{ij}^{(k)} \phi_j^{(k)}.$$

Concatenating these decomposed vectors across all K groups yielded full-variable attributions, enabling both instance-level explanation analysis and fairness auditing after dimensionality reduction. Importantly, autoencoders were trained only once on the training split and then fixed to ensure application of identical encoding functions f_k to the test data. This prevented information leakage, avoiding drift that could have biased downstream evaluation.

2.4. Training

This study evaluated the effectiveness of fairness-aware variable grouping strategies via Logistic Regression [33], Support Vector Machine (SVM) [34], Multi-Layer Perceptron (MLP) [35], Random Forest [7], and XGBoost [24]. This heterogeneity allowed for a robust assessment of the generalisability and fairness implications of the proposed methodology [13]. Hyperparameters (Appendix D) were tuned using 30 iterations of Bayesian optimisation with five-fold stratified cross-validation on the training set [36–38].

2.5. Evaluation

2.5.1. Performance metrics

Predictive performance was evaluated using accuracy and macro-averaged versions of precision, recall, F1-score, Receiver Operating Characteristic–Area Under the Curve (ROC-AUC), and Average Precision (AP). Macro averaging weighted each outcome class equally regardless of prevalence, thus reducing the influence of majority classes in imbalanced datasets. For each class, ROC-AUC and AP were calculated by treating that class as positive and all remaining classes as negative. Class-specific scores were then averaged with equal class weights. Each metric was reported for the ungrouped baseline and all dissimilarity-grouped configurations to highlight predictive trade-offs introduced by grouping strategies.

2.5.2. SHAP

SHAP provides instance-level additive attributions grounded in cooperative game theory, ensuring the axiomatic properties of local accuracy, missingness, and consistency [5]. For an input x and model f, the SHAP value for variable i is:

$$\phi_i = \sum_{S \subseteq F \setminus i} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup i}(x) - f_S(x)],$$

where F is the full variable set and S is a coalition excluding i.

For ungrouped models, SHAP values were computed directly for each variable. For grouped models, explanations were produced via the latent representation z_k for group k. SHAP values were first obtained in latent space and then mapped back to original variables, as explained in Section 2.3.4.

2.5.3. Fairness metrics

Group fairness was evaluated by comparing model behaviour across demographic groups defined by binary sensitive attribute A (`race` and `sex` in this study). Equal Opportunity, Equalised Odds, and Predictive Parity [39] were measured as group-level fairness metrics. Equal Opportunity required equality of true positive rates, ensuring that individuals who genuinely should qualify for a beneficial outcome are treated similarly across demographic groups. Equalised Odds extended this by also enforcing equal false positive rates. Predictive Parity

evaluated the equality of positive predictive values across demographic groups. Group-level differences alone could be misleading (e.g., with small demographic groups). Using the mean ϵ_i and variance σ_i^2 of group-specific error rates, the N-Sigma index accounted for statistical uncertainty [40]:

$$N\text{-Sigma} = \frac{|\epsilon_1 - \epsilon_0|}{\sqrt{\frac{\sigma_1^2+\sigma_0^2}{2}}}.$$

This normalised disparity reduced the risk of over-interpreting noisy fairness gaps.

Explanation bias was evaluated through differences in SHAP attributions because outcome-centric metrics alone would not guarantee fair internal reasoning (seemingly fair models at the prediction level could still produce biased explanations, undermining trust in (healthcare) contexts where interpretability is critical) [11]:

$$B_j^{\text{exp}} = \left|\mathbb{E}[\phi_j \mid A = 1] - \mathbb{E}[\phi_j \mid A = 0]\right|.$$

This explanation parity was plotted against prediction parity on a Cartesian plane called the 'bias quadrant'. Two complementary metrics were then used to quantify fairness behaviour in this space. First, the average distance to the origin measured the magnitude of combined prediction-explanation bias at the instance level. With points $p_i = (\phi_i^{(A)},\ \hat{p}_i - \mu_{A_i})$ after base-rate centring, this metric was computed as

$$\text{dist} = \frac{1}{n}\sum_i \|p_i\|_2,$$

where smaller values indicate closer alignment with ideal fairness at the origin.

Separability quantified how strongly privileged and unprivileged demographic groups diverged. This was measured using the Bhattacharyya distance between empirical Gaussian approximations of the distributions of the two demographic groups:

$$D_B = \tfrac{1}{8}(\mu_1 - \mu_2)^\top \bar{\Sigma}^{-1}(\mu_1 - \mu_2) + \tfrac{1}{2}\ln\left(\frac{\det \bar{\Sigma}}{\sqrt{\det \Sigma_1 \det \Sigma_2}}\right),$$

where $\bar{\Sigma} = \frac{1}{2}(\Sigma_1 + \Sigma_2)$; larger values indicated greater divergence between group-specific prediction-explanation behaviour in the bias quadrant [41].

3. Results

3.1. Grouping behaviour

Fig. 1 shows the CMI-based pairwise dissimilarity matrix for the Heart Disease dataset. The matrix exhibited a non-uniform structure, indicating that the variables shared different degrees of outcome-conditioned information. Several relationships were consistent with plausible clinical dependencies. For example, `age` showed comparatively lower dissimilarity with resting blood pressure, `trestbps` (0.704), and serum cholesterol, `chol` (0.654), reflecting the known association of ageing with blood pressure and serum cholesterol levels [42,43]. In contrast, pairs such as `age` and `sex` had higher dissimilarity (0.938), indicating that they retained relatively little shared information after conditioning on the disease outcome. These patterns suggest that the CMI-based measure captured meaningful differences in variable dependence rather than treating the feature space as homogeneous.

Fig. 2 shows that the grouping behaviour was method-dependent and generally most stable for values up to $K = 6$, for which singleton groups were not structurally required in the 13-variable Heart Disease dataset. Bicriterion provided the most balanced performance over this range, maintaining a relatively high diversity of 0.911-0.940, while simultaneously increasing dispersion from 0.65 to 0.91. This reflects its explicit optimisation of both the average and minimum pairwise

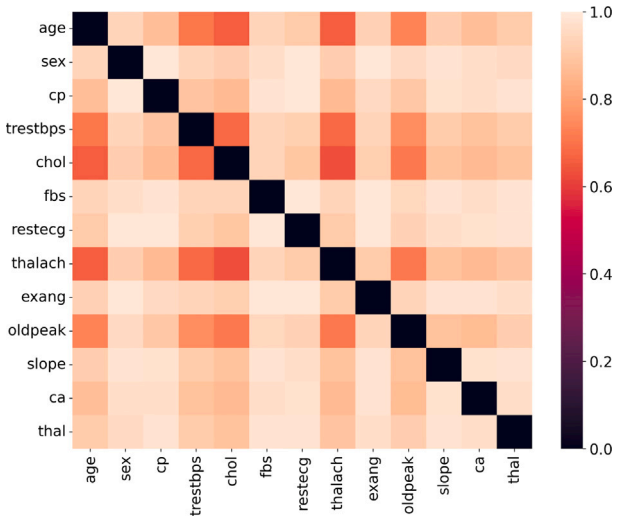


Fig. 1. CMI-based dissimilarity matrix for the Heart Disease dataset. Lower values indicate stronger similarity.

dissimilarity, allowing it to form groups that were broadly heterogeneous without retaining a highly similar variable pair. K-plus achieved the highest diversity increasing from 0.936 to 0.951, but with substantially lower dispersion of 0.63–0.76. This is consistent with its objective of balancing higher-order statistical properties across groups rather than directly maximising the minimum CMI-based dissimilarity, meaning that groups could be diverse overall while still containing at least one relatively similar pair. The proposed Greedy method showed a comparable upward trend in both measures up to K = 6–7, indicating that sequentially assigning variables according to their average dissimilarity produced structured and increasingly heterogeneous groups, although it did not optimise the worst-case pair as directly as Bicriterion. In contrast, Random grouping fluctuated across K without a clear trend, demonstrating that favourable diversity or dispersion at an isolated value of K could occur by chance rather than through systematic use of the dissimilarity matrix.

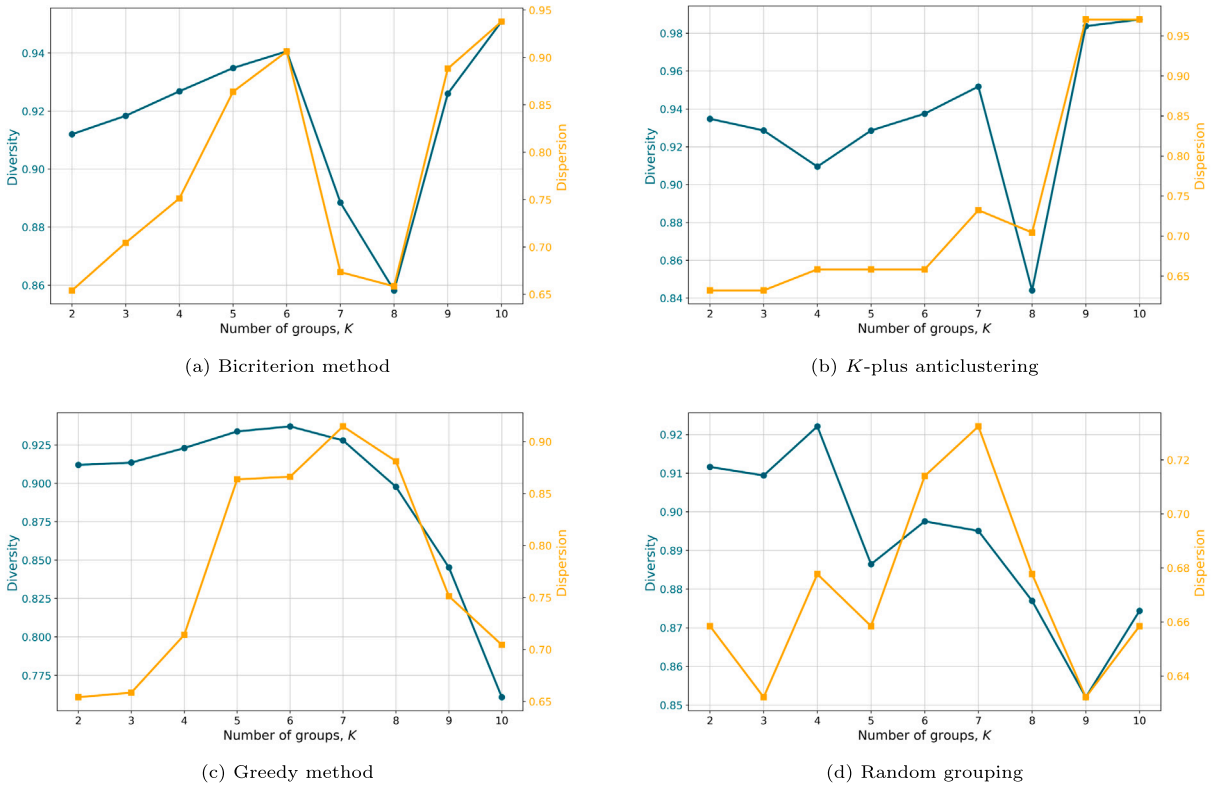
Beyond K = 6–7, the increasing number of singleton and two-variable groups made the metrics dependent on fewer multi-variable groups. For Greedy, highly dissimilar seed variables could remain as singletons and therefore be excluded from the pairwise metrics. Bicriterion operated over increasingly constrained swap configurations, while the sharp K-plus increases at K = 9–10 reflected only a few remaining dissimilar pairs. Thus, the high-K values indicated increasing fragmentation rather than a general improvement or degradation across all groups.

Overall, Bicriterion produced the strongest balance between average and worst-case dissimilarity, K-plus prioritised broad diversity over minimum separation, Greedy provided a computationally simple structured alternative, and Random grouping lacked consistent behaviour across K.

3.2. Performance metrics

Accuracy consistently appeared as the highest-performing metric across all datasets and grouping strategies (Table 2) which is a common pattern in medical classification settings where class imbalance may inflate overall correctness without improving sensitivity. This was most evident in Diabetes and Heart Disease, where recall and F1-scores were 44.4% and 56.6% lower than accuracy, respectively. Dataset-specific variation exceeded grouping effects, with average standard deviations of 26.36% and 4.20% across datasets and grouping methods, respectively.

On average, grouped models performed worse than the ungrouped baseline across all metrics, with reductions of 5.31% in accuracy, 7.13% in F1-score, 9.14% in precision, 5.26% in recall, 2.16% in ROC-AUC and 8.09% in AP. The reduction was smaller for ROC-AUC than



(a) Bicriterion method (b) K-plus anticlustering

(c) Greedy method (d) Random grouping

Fig. 2. Diversity and dispersion versus the number of groups K on the Heart Disease dissimilarity matrix (Fig. 1) for the grouping strategies.

Table 2
Mean predictive performance across datasets and grouping methods. The best value within each dataset and metric is shown in bold.

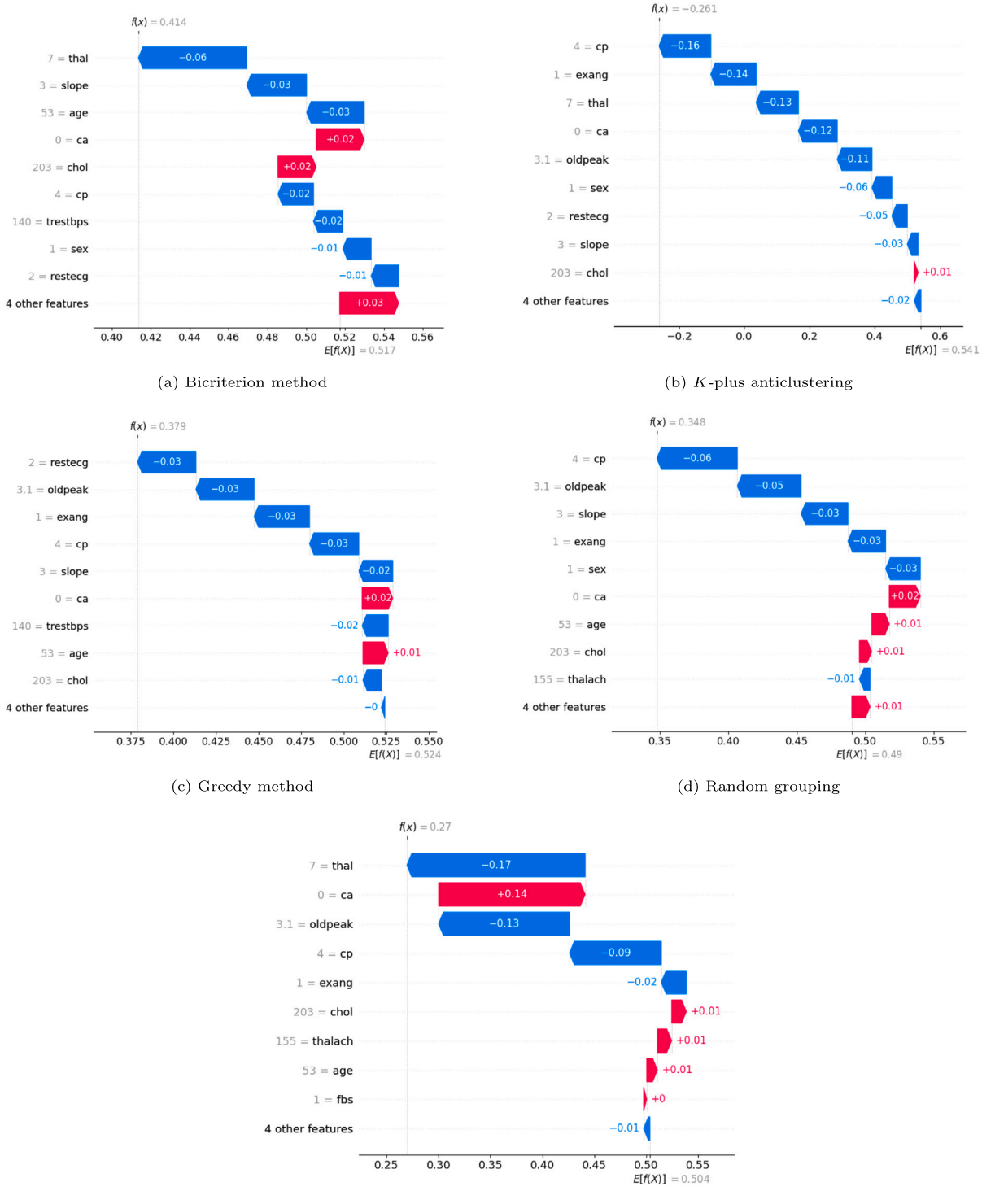
Dataset	Grouping method	Accuracy	F1-score	Precision	Recall	ROC-AUC	AP
Diabetes	Bicriterion	0.558	0.341	0.496	0.262	0.625	0.430
	K-plus	0.559	0.337	0.368	**0.315**	0.624	0.427
	Greedy	0.555	0.338	**0.506**	0.255	0.630	0.434
	Random	0.565	0.354	0.471	0.284	0.637	0.438
	Ungrouped	**0.570**	**0.359**	0.505	0.279	**0.654**	**0.456**
Heart Disease	Bicriterion	0.624	0.256	0.255	0.259	0.801	0.392
	K-plus	**0.656**	**0.302**	**0.324**	0.284	0.798	**0.430**
	Greedy	0.634	0.271	0.256	**0.287**	**0.801**	0.403
	Random	0.619	0.264	0.274	0.257	0.799	0.403
	Ungrouped	0.633	0.285	0.295	0.280	0.798	0.418
Obesity	Bicriterion	0.886	0.883	0.883	0.882	0.978	0.905
	K-plus	0.741	0.733	0.735	0.731	0.955	0.830
	Greedy	0.860	0.855	0.856	0.854	0.979	0.905
	Random	0.791	0.785	0.787	0.783	0.954	0.822
	Ungrouped	**0.933**	**0.932**	**0.932**	**0.932**	**0.997**	**0.981**
Average	Bicriterion	0.689	0.493	0.545	0.468	0.802	0.576
	K-plus	0.652	0.457	0.476	0.443	0.792	0.562
	Greedy	0.683	0.488	0.539	0.465	0.803	0.581
	Random	0.658	0.468	0.511	0.441	0.797	0.555
	Ungrouped	**0.712**	**0.525**	**0.577**	**0.497**	**0.816**	**0.618**

for AP, indicating that grouping had a more pronounced effect on precision-recall performance than on overall class discrimination. K-plus presented a notable exception. On the Heart Disease dataset, it outperformed the baseline on every metric, with gains up to 9.83% in precision and 5.96% in F1-score. In contrast, it underperformed other grouping methods across all metrics by an average of 9.85% on the Obesity dataset. These findings indicate that the effectiveness of a particular grouping strategy depends on dataset complexity and class imbalance, although grouping as a whole systematically introduced reductions in predictive performance.

3.3. SHAP and fairness metrics

3.3.1. Instance-level explanations

At the instance level, grouped waterfalls consistently displayed fewer extreme bars and smoother transitions from the base value to the final output than the ungrouped baseline (Fig. 3). This indicated that no single variable dominated the prediction, and that contributions were instead distributed across a broader set of variables. Bicriterion and K-plus showed the most pronounced smoothing, consistent with their higher intra-group heterogeneity. These visual patterns aligned with



(a) Bicriterion method

(b) K-plus anticlustering

(c) Greedy method

(d) Random grouping

(e) Ungrouped

Fig. 3. Comparison of SHAP plots for Heart disease classification by Random Forest across grouping methods, including the ungrouped case.

the global summary trends and illustrated the core mechanism through which grouping reduces explanation concentration.

Table 3 quantified these instance-level effects by aggregating absolute SHAP magnitudes into the sums contributed by the top nine variables versus the rest. Grouping consistently increased the 'rest' component across all models, ranging from the smallest increase for Logistic Regression (0.15 to 0.23) to the largest increase for Random Forest (0.06 to 0.48). These expansions were accompanied by increases in total absolute attribution averaging 52.9% and approximately doubling for SVM and MLP. This implied that grouping simultaneously redistributed attributions equitably and amplified total explanatory signal, reducing reliance on narrow, potentially proxy-driven pathways.

3.3.2. *Bias quadrant analysis at the instance level*

The bias-quadrant analysis (Fig. 4) evaluated whether these changes in the explanation structure translated into reduced fairness disparity

Table 3
Decomposition of total absolute SHAP magnitude into the top 9 variables and the remaining 21 variables for ungrouped and grouped (K-plus) configurations on the Diabetes dataset.

Model	Configuration	Top 9 variables	Rest (21 variables)	Total |SHAP|
Logistic Regression	Ungrouped	0.78	0.15	0.93
	Grouped	0.79	0.23	1.02
SVM	Ungrouped	0.31	0.15	0.46
	Grouped	0.53	0.35	0.88
MLP	Ungrouped	0.34	0.13	0.47
	Grouped	0.46	0.50	0.96
Random Forest	Ungrouped	0.66	0.06	0.72
	Grouped	0.38	0.48	0.86
XGBoost	Ungrouped	0.80	0.08	0.88
	Grouped	0.66	0.57	1.23

at the instance level. Grouping consistently moved points toward the vertical axis, indicating lower disparities in local SHAP attributions for protected attributes. K-plus produced the tightest clustering in the low-bias region and avoided quadrants I and IV, where explanation biases were high. Bicriterion also contracted points toward the origin but showed a distinctive pattern. The degree of contraction was stronger for instances with negative prediction deviation values (quadrants III and IV) than for those with positive attributions, which was consistent with its conservative emphasis on minimum intra-group similarity (dispersion). Greedy and Random were more variable, often landing in quadrants where biases at prediction and explanation levels diverged. In the ungrouped case, privileged and unprivileged instances occupied largely distinct regions of the bias quadrant, with privileged instances concentrated in quadrants I and IV and unprivileged instances in quadrants II and III. Variable grouping substantially reduced this separation, producing greater overlap between demographic groups, as shown in the remaining panels of Fig. 4. These quadrant patterns provided complementary evidence that grouping reduced explanation bias, even when outcome disparity was only partially improved.

3.3.3. Fairness metrics

The effect of grouping on outcome-based parities depended strongly on the dataset (Table 4). The exception was Heart Disease, where grouping reduced Equal Opportunity and Equalised Odds by 46.2% on average relative to the ungrouped case, but the effect was negligible or inconsistent in other datasets. N-Sigma also did not systematically improve under grouping, indicating that some apparent fairness gains were not statistically robust after accounting for variance and size of each demographic group.

In contrast, explanation-parity metrics showed a clearer pattern. Grouping reduced average bias-quadrant distance by 8.04%, with the largest reductions under Bicriterion (9.47%) and K-plus (16.60%). Similarly, the separability decreased significantly by 89.65% on average, with the largest reductions under K-plus (94.35%) and random (97.64%).

4. Discussion

4.1. Clinical safety and ethical interpretation of fairness-performance trade-off

The reductions in predictive performance require careful interpretation in healthcare settings. Improvements in statistical fairness or explanation parity do not, by themselves, justify lower predictive performance, because fairness and predictive metrics quantify different properties and do not share a common utility scale. They should therefore be considered jointly rather than combined into a single measure of net benefit. In particular, reductions in recall or AP may increase the risk of missed positive cases, whereas reduced precision may expose patients to unnecessary treatment. The consequences of these changes depend on the clinical task, outcome prevalence, decision threshold, and relative harms of false-negative and false-positive predictions. Accordingly, SHIELD should not be used as an automatic rule for replacing an ungrouped model with a grouped configuration. A candidate configuration should first satisfy task-specific predictive requirements, after which its fairness and explanation behaviour can be considered as additional selection criteria. For example in an acute care triage setting, a small reduction in sensitivity may be unacceptable even if explanation parity improves. In a lower-risk retrospective auditing setting, the same configuration may still be informative for mitigating dependence on proxy variables and guiding further model development.

4.2. Principal findings

Grouping variables by CMI and decoding latent attributions into the native variable space improved instance-level explainability and several fairness measures, but these changes were accompanied by dataset and method dependent reductions in predictive performance. Across datasets, grouped models redistributed explanatory credit more evenly across instances and variables, reducing over-reliance on a narrow set of correlated variables. The accompanying MIT-licensed code release [44] supports reproducibility through an auditable implementation of SHIELD.

4.3. Significance of the study

These findings have practical implications for the design and evaluation of clinical prediction models. First, by distributing explanatory attribution across a broader set of variables, SHIELD can reduce the extent to which model explanations are dominated by a small number of correlated or proxy variables. Second, decoder-mapped SHAP preserves traceability to the original variables, ensuring that any retained variables remain clinically interpretable and auditable. Finally, the improved instance-level fairness, reflected by reduced SHAP concentration and lower separability and average distance in the bias quadrant, suggests that SHIELD produces more consistent explanations across privileged and unprivileged groups. This consistency may support fairness auditing when developing and validating clinical prediction models.

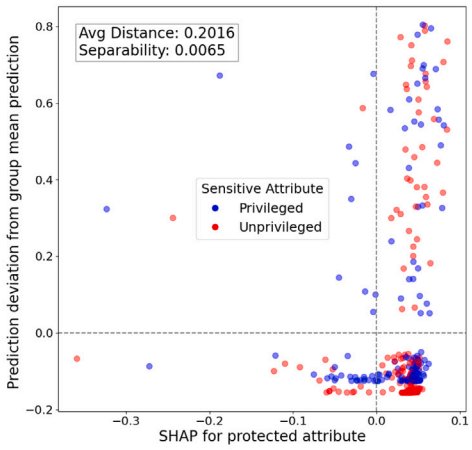
4.4. Limitations

The search over grouping and model hyperparameters was restricted, especially for the number of groups K. Although Bayesian optimisation was used in principle, practical runtime constraints limited the breadth of configurations explored, meaning that some fairness or accuracy improvements may depend on non-optimal selections.

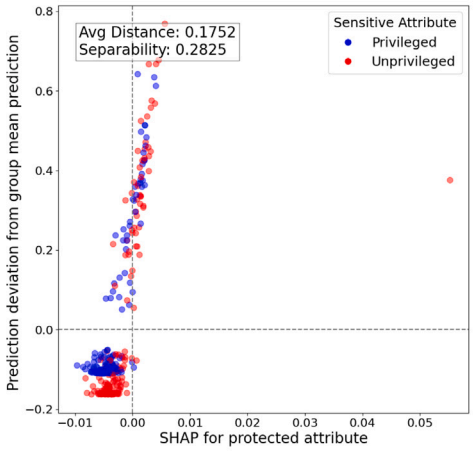
The evaluation relied on conventional stratified splits. Without adversarial splits or splitting that is aware of distribution-shift [45], robustness might be overestimated, particularly in edge cases where populations differ from the training distribution. This may have limited the sensitivity of the evaluation to changes in the N-Sigma error rate.

Fairness assessments covered outcome metrics and explanation-level audits, but these captured only selected definitions of fairness and did not consider individual fairness and causal/counterfactual criteria [46]. The group fairness analysis considered one sensitive attribute at a time and did not examine intersectional groups. Consequently, the reported fairness results may not reveal disparities affecting intersectional groups defined jointly by sensitive attributes, such as race, sex, and age.

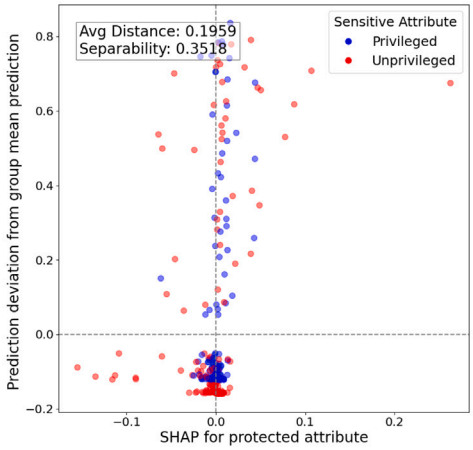
The evaluation used static tabular benchmark datasets and did not include an external site, temporal holdout, or geographically distinct cohort. Therefore, the results demonstrate methodological feasibility within the evaluated datasets but do not establish generalisability across institutions, time periods, or patient populations [47]. Contemporary



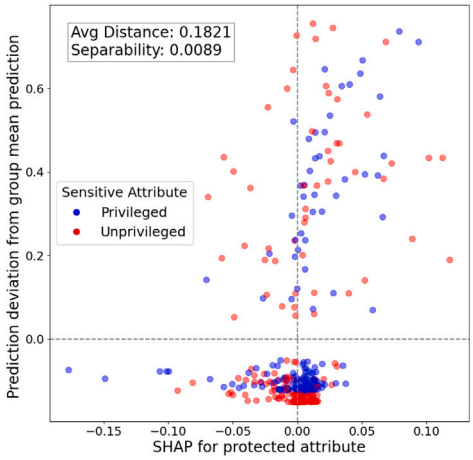
(a) Bicriterion method



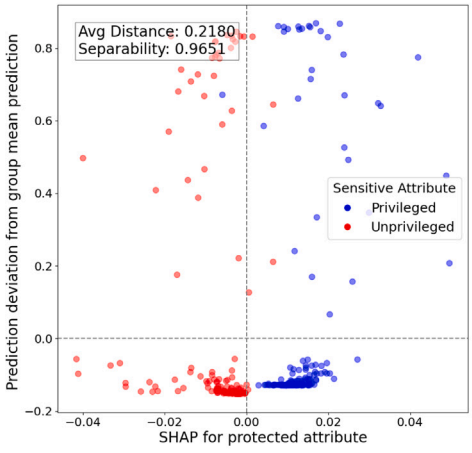
(b) K-plus anticlustering



(c) Greedy method



(d) Random grouping



(e) Ungrouped

Fig. 4. Comparison between bias quadrant plots of the Obesity dataset via SVM across different grouping methods.

electronic health record data may also differ in coding practices, missingness, temporal structure, and variable availability. Applying SHIELD in such environments would require separate data harmonisation, system integration, and prospective validation work beyond the scope of this study. Hence, SHIELD should be interpreted as a framework for research design and model development or validation rather than as a deployment-ready clinical system [47].

4.5. Future work

Future evaluations should extend the group fairness analysis to multiple sensitive attributes and their intersections, including race–sex and sex–age groups, where adequate sample sizes permit reliable estimation. Future work should develop theoretical bounds linking grouping-induced excess risk to encoder–decoder distortion. Stability analyses for

Table 4
Mean fairness metrics across datasets and grouping methods. Lower values indicate lower disparity for all metrics. The best value within each dataset and metric is shown in bold.

Dataset	Grouping method	Equal Opportunity	Equalised Odds	Predictive Parity	N-Sigma	Average Distance	Separability
Diabetes	Bicriterion	0.004	0.016	**0.021**	**0.042**	0.108	0.034
	K-plus	**0.002**	**0.004**	0.036	0.073	0.106	0.022
	Greedy	0.004	0.025	0.034	0.067	**0.105**	0.005
	Random	0.003	0.010	**0.021**	0.046	0.113	**0.003**
	Ungrouped	0.005	0.013	0.024	0.052	0.111	2.306
Heart Disease	Bicriterion	0.232	**0.214**	0.660	0.233	0.113	0.695
	K-plus	0.373	0.373	0.750	0.282	**0.104**	0.193
	Greedy	0.312	0.281	0.850	0.359	0.131	0.152
	Random	**0.214**	0.241	**0.511**	0.315	0.148	**0.077**
	Ungrouped	0.520	0.520	0.647	**0.175**	0.136	2.095
Obesity	Bicriterion	0.091	0.099	0.074	0.133	0.197	0.044
	K-plus	0.219	0.221	0.158	0.200	**0.169**	0.056
	Greedy	0.128	0.122	0.063	0.146	0.206	0.511
	Random	0.103	0.116	0.138	**0.123**	0.183	**0.027**
	Ungrouped	**0.030**	**0.063**	**0.050**	0.148	0.216	0.826
Average	Bicriterion	0.109	**0.110**	0.252	0.136	0.139	0.258
	K-plus	0.198	0.199	0.315	0.185	**0.126**	0.090
	Greedy	0.148	0.143	0.316	0.191	0.147	0.223
	Random	**0.107**	0.122	**0.223**	0.161	0.148	**0.036**
	Ungrouped	0.185	0.199	0.240	**0.125**	0.154	1.742

decoded SHAP attributions could clarify when reconstructed explanations remain faithful, while information-theoretic analyses of CMI could formalise how grouping restricts fairness leakage. Extending SHIELD to regression tasks would further test whether dissimilarity-based grouping yields similar fairness and explainability benefits for continuous prediction outcomes.

4.6. Broader impact statement

This work contributes to methodological research on equitable and explainable AI-based decision support by providing a model-agnostic, auditable pipeline for analysing instance-level explanations under variable dependence. SHIELD does not introduce opaque adversarial components or modify clinical ground-truth labels. Instead, it enhances interpretability directly in the native variable space, emphasising reproducible and clinically meaningful evaluation rather than model novelty alone. Although evaluated on static tabular datasets, the framework promotes important principles for model development, including fairness auditing, methodological transparency, and reproducibility. By outlining the need for external, temporal, and prospective validation, the study provides a foundation for further methodological and translational research rather than evidence of immediate clinical deployment readiness.

5. Conclusion

This study introduced SHIELD, a dissimilarity-based variable grouping framework for equitable and explainable clinical prediction. By combining CMI-driven grouping with decoder-mapped SHAP, SHIELD redistributed attributional mass away from dominant proxy variables while preserving traceability to the original variable space. Across healthcare datasets, grouping broadened attribution use, reduced explanation bias, and improved several fairness metrics, with variable predictive trade-offs. These findings suggest that structured redistribution of variable dependence offers a methodological direction toward more accountable AI systems in healthcare.

CRediT authorship contribution statement

Hyeonggeun Yun: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Amanda S. Barnard:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition, Conceptualization. **Hanna Suominen:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition, Conceptualization.

Funding

This study was supported by the Medical Research Future Fund, Australia under award number GA187319 (AB and HS). Computational resources for this project were supplied by the National Computational Infrastructure (NCI) [grant numbers p00 (AB) and ny83 & va80 (HS)].

Declaration of competing interest

The authors declare the following financial interests/personal relationships that may be considered potential competing interests:

Amanda S. Barnard reports that financial support was provided by the Australian Government Medical Research Future Fund. Amanda S. Barnard reports that equipment, drugs, or supplies were provided by the National Computational Infrastructure. Hanna Suominen reports that financial support was provided by the Australian Government Medical Research Future Fund. Hanna Suominen reports that equipment, drugs, or supplies were provided by the National Computational Infrastructure. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Pseudocode for bicriterion anticlustering

Algorithm 1 Bicriterion Anticlustering.

Input: Dissimilarity matrix D, number of groups K, weights α, β
Randomly initialise groups $G_1, \ldots, G_K$
repeat
  Compute Diversity and Dispersion for current partition P
  for each pair of variables (x, y) in different groups **do**
    Swap x and y if it improves obj(P)
  end for
until no further improvement in objective
Output: Optimised groups maximising bicriterion objective

Appendix B. Pseudocode for K-plus anticlustering

Algorithm 2 K-plus Anticlustering.

Input: Variable matrix X, number of groups K, maximum order r, weights $\lambda_1, \ldots, \lambda_r$
Construct polynomial variables $X^{(2)}, \ldots, X^{(r)}$
Initialise groups using K-means++ or random assignment
repeat
  Compute $\mathrm{SSE}_{\mathrm{K+}}$ for current partition
  for each pair of variables (x, y) in different groups **do**
    Swap x and y if it reduces $\mathrm{SSE}_{\mathrm{K+}}$
  end for
until convergence or no improvement
Output: Balanced variable groups with matched statistical properties

Appendix C. Pseudocode for greedy variable grouping

Algorithm 3 Greedy Variable Grouping.

Input: Dissimilarity matrix $D \in \mathbb{R}^{d \times d}$, number of groups K
for each variable $i \in \{1, \ldots, d\}$ **do**
  Compute its row-wise dissimilarity score:

$$r_i = \sum_{j=1}^{d} D_{ij}$$

end for
Select the K variables with the largest values of r_i as seed variables
Initialise each group G_k with one distinct seed variable
Let U be the set of all remaining unassigned variables
while $U \neq \emptyset$ **do**
  for $k = 1, \ldots, K$ **do**
    if $U = \emptyset$ **then**
      break
    end if
    for each variable $f \in U$ **do**
      Compute its average dissimilarity to the current group:

$$a(f, G_k) = \frac{1}{|G_k|} \sum_{j \in G_k} D_{fj}$$

    end for
    Select the remaining variable

$$f^* = \arg\max_{f \in U} a(f, G_k)$$

    Assign f^* to G_k
    Remove f^* from U
  end for
end while
Output: K dissimilar groups

Appendix D. Description of hyperparameters

See Table D.5.

Table D.5
Hyperparameters used in variable grouping and model training.

Symbol and/or Name	Range
Variable Grouping Stage	
K: Number of groups	[2, 10]
α, β: Bicriterion weights	[0, 1]
w_2, w_3, w_4: Moment weights	[0, 1]
ϵ: Smoothing constant	$[10^{-10}, 10^{-6}]$
Model Training Stage	
C: Regularisation (LR, SVM)	[0.01, 100]
γ: Kernel coefficient (SVM)	[0.001, 1]
Number of trees (RF, XGB)	[50, 500]
Maximum tree depth (RF, XGB)	[3, 15]
Learning rate (XGB, MLP)	[0.001, 0.3]
Hidden layer sizes (MLP)	Varies
L_2 penalty (MLP)	[0.0001, 0.1]

Data availability

All data used in this study are publicly available at the UCI Machine Learning Repository [17].

References

[1] A. Rajkomar, J. Dean, I. Kohane, Machine learning in medicine, N. Engl. J. Med. 380 (14) (2019) 1347–1358, https://doi.org/10.1056/NEJMra1814259
[2] S. Barocas, A.D. Selbst, Big data's disparate impact, Calif. Law Rev. 104 (3) (2016) 671–732, https://doi.org/10.15779/Z38BG31
[3] J.M. Bauer, M. Michalowski, Human-centered explainability evaluation in clinical decision-making: a critical review of the literature, J. Am. Med. Inform. Assoc. 32 (9) (2025) 1477–1484, https://doi.org/10.1093/jamia/ocaf110
[4] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V.X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed, P.N. Ossorio, S. Thadaney-Israni, A. Goldenberg, Do no harm: a roadmap for responsible machine learning for health care, Nat. Med. 25 (9) (2019) 1337–1340, https://doi.org/10.1038/s41591-019-0548-6
[5] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 4768–4777, https://doi.org/10.48550/arXiv.1705.07874.
[6] M.T. Ribeiro, S. Singh, C. Guestrin, Why should I trust you? Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144, https://doi.org/10.1145/2939672.2939778
[7] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324
[8] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, PLoS One 10 (3) (2015), https://doi.org/10.1371/journal.pone.0118432
[9] S.A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M.A. Riegler, P. Halvorsen, S. Parasa, On evaluation metrics for medical applications of artificial intelligence, Sci. Rep. 12 (2022), https://doi.org/10.1038/s41598-022-09954-8
[10] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, J. Dean, A guide to deep learning in healthcare, Nat. Med. 25 (2019) 24–29, https://doi.org/10.1038/s41591-018-0316-z
[11] A. Jain, M. Ravula, J. Ghosh, Biased models have biased explanations, 2020, https://doi.org/10.48550/arXiv.2012.10986
[12] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to Manage the health of populations, Science 366 (6464) (2019) 447–453, https://doi.org/10.1126/science.aax2342

[13] S.A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E.P. Hamilton, D. Roth, A comparative study of fairness-enhancing interventions in machine learning, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, ACM, 2019, pp. 329–338, https://doi.org/10.1145/3287560.3287589
[14] A. Datta, M. Fredrikson, G. Ko, P. Mardziel, S. Sen, Proxy non-discrimination in data-driven systems, 2017, https://doi.org/10.48550/arXiv.1707.08120
[15] B.H. Zhang, B. Lemoine, M. Mitchell, Mitigating unwanted biases with adversarial learning, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, pp. 335–340, https://doi.org/10.1145/3278721.3278779
[16] F. Kamiran, T. Calders, Data preprocessing techniques for classification without discrimination, Knowl. Inf. Syst. 33 (1) (2012) 1–33, https://doi.org/10.1007/s10115-011-0463-8
[17] University of California, Irvine, UC Irvine machine learning repository, 2025, https://archive.ics.uci.edu/.
[18] J.A.M. Sidey-Gibbons, C.J. Sidey-Gibbons, Machine learning in medicine: a practical introduction, BMC Med. Res. Methodol. 19 (2019), https://doi.org/10.1186/s12874-019-0681-4
[19] A. Dhurandhar, K. Shanmugam, R. Luss, Leveraging simple model predictions for enhancing its performance, in: ICML 2020, 2020, pp. 1–20, https://doi.org/10.48550/arXiv.1905.13565
[20] Z. Zhou, Z.-J. Zhou, H. Hao, S. Li, X. Chen, Y. Zhang, M. Folkert, J. Wang, Constructing multi-modality and multi-classifier radiomics predictive models through reliable classifier fusion, 2017, https://doi.org/10.48550/arXiv.1710.01614
[21] F.M. Palechor, A. de la Hoz Manotas, Estimation of obesity levels based on eating habits and physical condition, 2019, https://doi.org/10.24432/C5H31Z.
[22] J. Clore, K. Cios, J. DeShazo, B. Strack, Diabetes 130-US hospitals for years 1999-2008, 2014, https://doi.org/10.24432/C5230J.
[23] A. Janosi, W. Steinbrunn, M. Pfisterer, R. Detrano, Heart disease, 1989, https://doi.org/10.24432/C52P4X.
[24] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785
[25] Y. Deng, T. Lumley, Multiple imputation through XGBoost, J. Comput. Graph. Stat. 33 (2) (2024) 352–363, https://doi.org/10.1080/10618600.2023.2252501
[26] K. Potdar, T. Pardawala, C. Pai, A comparative study of categorical variable encoding techniques for neural network classifiers, Int. J. Comput. Appl. 175 (4) (2017) 7–9, https://doi.org/10.5120/ijca2017915495
[27] S.G.K. Patro, K.K. Sahu, Normalization: a preprocessing stage, Int. Adv. Res. J. Sci. Eng. Technol. 2 (3) (2015), https://doi.org/10.17148/IARJSET.2015.2305
[28] G. Brown, A. Pocock, M.-J. Zhao, M. Luján, Conditional likelihood maximisation: a unifying framework for information theoretic feature selection, J. Mach. Learn. Res. 13 (1) (2012) 27–66, https://dl.acm.org/doi/10.5555/2503308.2188387.
[29] A.E.R. Prince, D. Schwarcz, Proxy discrimination in the age of artificial intelligence and big data, Iowa Law Rev. 105 (3) (2020) 1257–1318, https://scholarship.law.umn.edu/faculty_articles/682.
[30] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423, https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
[31] M.J. Brusco, D.J. Cradit, D. Steinley, Combining diversity and dispersion criteria for anticlustering: a bicriterion approach, Br. J. Math. Stat. Psychol. 73 (3) (2020) 375–396, https://doi.org/10.1111/bmsp.12186
[32] M. Papenberg, K-Plus anticlustering: an improved K-means criterion for maximizing between-group similarity, Br. J. Math. Stat. Psychol. 77 (1) (2024) 80–102, https://doi.org/10.1111/bmsp.12315
[33] D.W. Hosmer, S. Lemeshow, R.X. Sturdivant, Applied Logistic Regression, John Wiley & Sons, 2013, https://doi.org/10.1002/9781118548387
[34] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297, https://doi.org/10.1023/A:1022627411411
[35] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (6088) (1986) 533–536, https://doi.org/10.1038/323533a0
[36] J. Snoek, H. Larochelle, R.P. Adams, Practical Bayesian optimization of machine learning algorithms, in: NIPS'12: Proceedings of the 26th International Conference on Neural Information Processing Systems, vol. 2, 2012, pp. 2951–2959, https://dl.acm.org/doi/10.5555/2999325.2999464.
[37] M.W. Browne, Cross-validation methods, J. Math. Psychol. 44 (1) (2000) 108–132, https://doi.org/10.1006/jmps.1999.1279
[38] J. Bergstra, D. Yamins, D. Cox, Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures, in: S. Dasgupta, D. McAllester (Eds.), Proceedings of the 30th International Conference on Machine Learning, Vol. 28 of Proceedings of Machine Learning Research, PMLR, Atlanta, Georgia, USA, 2013, pp. 115–123, https://proceedings.mlr.press/v28/bergstra13.html.
[39] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, Curran Associates Inc., Red Hook, NY, USA, 2016, pp. 3323–3331. https://dl.acm.org/doi/10.5555/3157382.3157469.
[40] D. DeAlcala, I. Serna, A. Morales, J. Fierrez, J. Ortega-Garcia, Measuring bias in AI models: an statistical approach introducing N-Sigma, in: 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC), 2023, pp. 1167–1172, https://doi.org/10.1109/COMPSAC57700.2023.00176
[41] T. Kailath, The divergence and bhattacharyya distance measures in signal selection, IEEE Trans. Commun. Technol. 15 (1) (1967) 52–60, https://doi.org/10.1109/TCOM.1967.1089532
[42] S.S. Franklin, W. Gustin, N.D. Wong, M.G. Larson, M.A. Weber, W.B. Kannel, D. Levy, Hemodynamic patterns of age-related changes in blood pressure, Circulation 96 (1) (1997) 308–315, https://doi.org/10.1161/01.CIR.96.1.308
[43] M.A. Berns, J.H. de Vries, M.B. Katan, Determinants of the increase of serum cholesterol with age: a longitudinal study, Int. J. Epidemiol. 17 (4) (1988) 789–796, https://doi.org/10.1093/ije/17.4.789
[44] H. Yun, SHIELD, 2025, https://github.com/geun-yun/SHIELD.
[45] A.S. Barnard, BenchMake: turn any scientific data set into a reproducible benchmark, Mach. Learn. Sci. Technol. 6 (3) (2025) 030502, https://doi.org/10.1088/2632-2153/adf810
[46] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (6) (2021), https://doi.org/10.1145/3457607
[47] M. Ghassemi, L. Oakden-Rayner, A.L. Beam, The false hope of current approaches to explainable artificial intelligence in health care, Lancet Digit. Health 3 (11) (2021), https://doi.org/10.1016/S2589-7500(21)00208-9